# Genetic variability of *Dasheen mosaic virus* and consequences for detection

Subuhi Khan

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

2012

School of Applied Science

Supervisor: Dr.Colleen Higgins

# Table of Contents

# List of Figures

## Chapter 1



## Chapter 2

## Chapter 3

# List of Tables

## Chapter 2



## Chapter 3

# Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

**Subuhi Khan**

# Acknowledgments

This work was undertaken during the year 2011-2012 at the Faculty of Health and Environmental Sciences of Auckland University of Technology (AUT). I would like to acknowledge my work to people who, with their utmost support and inspiration made this project a success.

I must give my deep and respectful gratitude to my supervisor, Dr. Colleen Higgins for her guidance, supervision and help throughout this project. I have learned a lot throughout this year, with many challenging yet valuable experiences in order to complete this project. My endless thanks to Dr. Colleen Higgins for giving me a chance to explore a new dimension of mine, as well as for giving me precious advises in order to improve myself to become a better person. Her understanding and guidance have provided a good base for this thesis.

I owe my sincere appreciation to the technicians present in the laboratories of AUT who have also done a great deed by aiding me in this project.

I would like to express my gratitude towards my family and friends who have always been there for me no matter where I am, and for all the unconditional love and patience. This thesis would not have been possible without their encouragement, support and understanding.

I would like to conclude with a vote of thanks to all the other people for their help and support whose names cannot be given in a short space.

# Abstract

The genus *Potyvirus* is one of the largest groups of plant viruses having RNA genomes. RNA viruses show high genetic diversity which can influence virus infectivity and host selection. In this study, deep sequencing profiles of a potyvirus, namely *Dasheen mosaic virus* (DsMV) were analysed to determine the level of genetic variation occurring in two different DsMV strains (DsMV-NZ1 and DsMV-B) and isolates (DsMV-NZ1.1, NZ1.2, B1.1 and B1.2).

Potyviruses are one of the most important groups of plant viruses that cause severe damage to agricultural, horticultural and ornamental crops. Within this genus, DsMV is an important viral pathogen of cultivated aroids worldwide, causing a 40-60% reduction in total crop yield. Therefore, the precise and accurate identification of DsMV infection is important for developing appropriate control mechanisms. The first objectives of this study was to test the effectiveness of universal potyvirus primers in detection and identification of DsMV infection in taro. Three pairs of universal primers, namely HPFor-HPRev, CIFor-CIRev and NIb2F-NIb3R were tested in this study using reverse-transcriptase polymerase reaction (RT-PCR). After testing each of these primer pairs, it was concluded that these primers can be used for detection and identification of the DsMV infection of taro, but only under the experimental conditions used in this study.

The second objective was to determine the level of variation within each strain/isolate of the DsMV. For that Illumina HiSeq 2000 platform was used for creating the deep sequencing profiles of each sample. The bioinformatics analysis revealed that the level of variation across all the samples was consistent and an average 20% of genetic variation was observed within each strain/isolate from the DsMV reference genome. It was also found that the level of variation within the isolates of a same strain was lower than variation between the strains. A phylogenetic tree constructed based on the whole genome nucleotide sequences of the DsMV showed that DsMV belongs to the *Bean common mosaic virus* (BCMV) group. The phylogenetic analysis also revealed that DsMV is most closely related to *Vanilla mosaic virus* (VaMV) than to any other virus of the BCMV group.

The third objective of this study was to determine the level of variation within each region of the DsMV genome. The bioinformatics analysis of the Illumina sequencing data showed that the variation within the 5’ and 3’ untranslated regions was

~28% and 30%, respectively. With the protein coding genes, the P1 gene showed the highest level of genetic variation (~30%) while NIa-VPg gene showed the lowest (~16%). The low level of variation in the NIa-VPg gene suggested that this gene may be functionally very important for the DsMV while the high variation of the P1 may provide a mechanism for altering the host range. Each region needs to be investigated in more detail, which may help in identification of the regions of conservation within the potyviral genome.

The results showed that genetic variability can interfere with the efficient detection and identification of potyviral infection in plants. The findings of this study, helped in understanding the pattern of genetic variability among potyviruses. and in developing a better understanding of the functionality of the potyviral genome. The findings of this study can also be used for identifying the conserved regions within the potyviral genome, which would be expected to assist designing better universal primers for precise and accurate detection of potyvirus infection in plants, as well as deepening or understanding of viral processes such as translation replication.

# Chapter 1

# General Introduction

# 1. Introduction

Viruses were first discovered in plants and the first virus was named *Tobacco mosaic virus* (TMV) by Beijerinck in 1898 (Lewsey & Carr, 2009). RNA viruses cause several diseases to humans, plants and animals (Malpica et al., 2002). TMV was the first positive-sense ssRNA virus to be identified and was used extensively for many years as a model for investigating many important concepts in plant biology. Around 80% of the viruses infecting plants have RNA genomes with the majority (~65%) having single-stranded (ss) positive RNA genomes (Lewsey & Carr, 2009).

## 1.1 Potyvirus

The *Potyviridae* family is one of the largest groups of known plant viruses, consisting of seven different genera: *Potyvirus*; *Rymovirus*; *Tritimovirus*; *Bymovirus*; *Maclurovirus*; *Ipomovirus*; *Brambyvirus* and *Poacevirus* (McDonald et al., 2010; Tatineni, Ziems, Wegulo, & French, 2009). Potyviruses are the largest and economically most important group of plant-infecting viruses. The genus *Potyvirus* contains at least 180 definitive and possible members that represent more than 30% of all known plant viruses. These viruses are transmitted by aphids in a non-persistent manner and cause significant loss to commercially important crops and ornamental plants (Shukla, Frenkel, & Ward, 1991).

### 1.1.1 Potyvirus genome

The potyviral virion is a flexuous, filamentous 680-900 nm long rod-shaped particle (figure 1.1) (Shukla et al., 1991). The particles contains a single-stranded (ss) positive sense RNA molecule of around 10,000 bases (Shukla et al., 1991) encapsidated by approximately 2000 copies of the capsid protein (CP) (figure 1.2) (Matthews, 1991). The N-terminal end of the CP has been found to be highly variable and is associated with virus specific antigenic determinants, whereas the remaining core protein shows high level of conservation between different potyviral species (Marie-Jeanne, Ioos, Peyre, Alliot, & Signoret, 2000).

Figure 1. 1: Electron micrograph showing the filamentous rod-shaped structure of *Potyvirus* (Brunt et al., 2012b)


Figure 1.2: Structure showing the Potyviral genome encapsidated in capsid protein (*ViralZone*, 2012)

The potyviral RNA genome shows similarity with genomes of como-,nepo-, and picornaviruses in having a VPg protein at its 5' end, a single open reading frame (ORF) and a 3'UTR (Riechmann, Laín, & García, 1992; Shukla et al., 1991). The RNA is polyadenylated at the 3' end (Nicolas & Laliberté, 1992). The ORF encodes a single large polyprotein that is co- and/or post translationally cleaved by three viral encoded proteinases (P1, HC-Pro and NIa) (Shukla et al., 1991). The Proteinase 1 (P1) and Helper Component-protease (HC-Pro) facilitate their own cleavage from the polyprotein while the Nuclear Inclusion protein a (NIa) causes the C-terminal cleavage of the remaining two-thirds of the polyprotein. As shown in figure 1.3, these proteolytic activities give rise to ten functional proteins, namely, the P1, HC-Pro, third protein (P3), 6K1, cylindrical inclusion protein (CI), 6K2, small nuclear inclusion protein (NIa;

including both VPg and protease (NIa-Pro) domains), large nuclear inclusion protein (NIb) and CP (Ha et al., 2008; Marie-Jeanne et al., 2000). In addition, a small ORF encoding a recently identified 25 kDa protein, P3N-PIPO, is embedded in the P3 coding region as a plus 2 frameshift sequence (Chung, Miller, Atkins, & Firth, 2008).

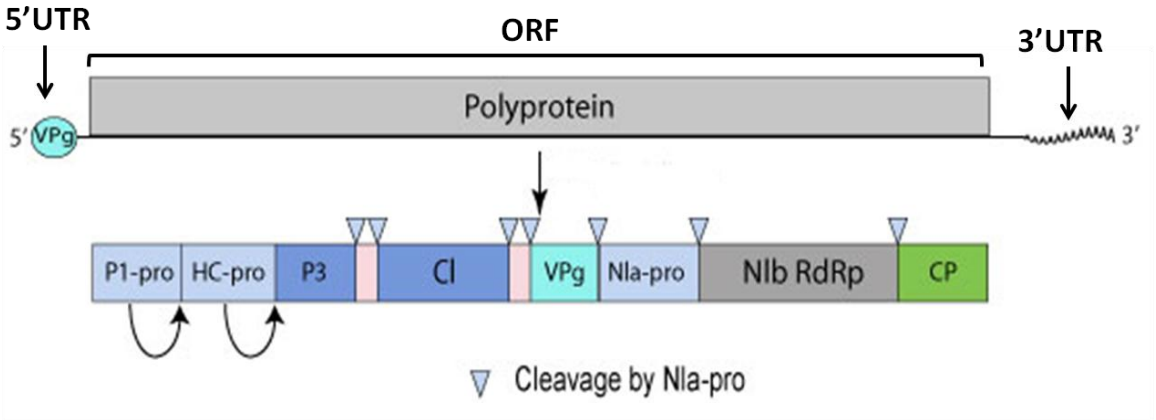


Figure 1.3: The potyviral polyprotein gives rise to ten functional subunits. The arrows represent the protein cleavage sites (modified from (*ViralZone*, 2012))

## 1.2 Dasheen mosaic virus

*Dasheen mosaic virus* (DsMV) belongs to the genus *Potyvirus.* DsMV is known as a common and widespread pathogen of ornamental and edible aroids throughout the Pacific and worldwide (Elliot, Zettler, & Brown, 1997). Similar to other members of the genus *Potyvirus*, DsMV is also composed of flexuous, filamentous particles of around 750 nm in length containing a positive sense single-stranded (ss) RNA genome (Chen, Chen, Chen, & Adams, 2001). The DsMV genome sequence (GenBank accession number: NC_003537) was found to be 9991 nucleotides in length (Chen et al., 2001) and to have the same structure as other potyviruses (Brunt et al., 2012a).

DsMV is transmissible by three methods, that is either through vectors such as aphids, or by vegetative and/or mechanical methods (Nelson, 2008). DsMV is commonly transmitted by aphids in a non-persistant manner (Zettler & Hartman, 1987) and infects species in at least 16 genera of *Araceae* (Elliot et al., 1997). For instance, among edible aroids most of the species belongs to the genera *Colocasia* and *Xanthosoma*, while ornamental genera include *Alocasia*, *Caladium*, *Zantedeschia* and several others (Abo El-Nil, Zettler, & Hiebert, 1977; Zettler & Hartman, 1987). Many

studies have reported that DsMV severely affects the yield of ornamental and edible aroids. For instance, DsMV has been found to cause severe reduction in vine length (around 66%) in *Philodendron scandens* and 30%-40% reduction in corm weight of edible aroids (Elliot et al., 1997).

### 1.2.1 Symptoms

Symptoms induced by DsMV are different from those caused by other pathogens infecting aroids. The symptoms include leaf mosaic (i.e. normal leaf colouration is altered by uneven light and dark patterns), ring spots on the leaves, some degree of leaf distortion, decline in plant vigour, and plant stunting (figure 1.4). The occurrence of symptoms depends on the host and severity of symptoms varies between hosts (Elliot et al., 1997).

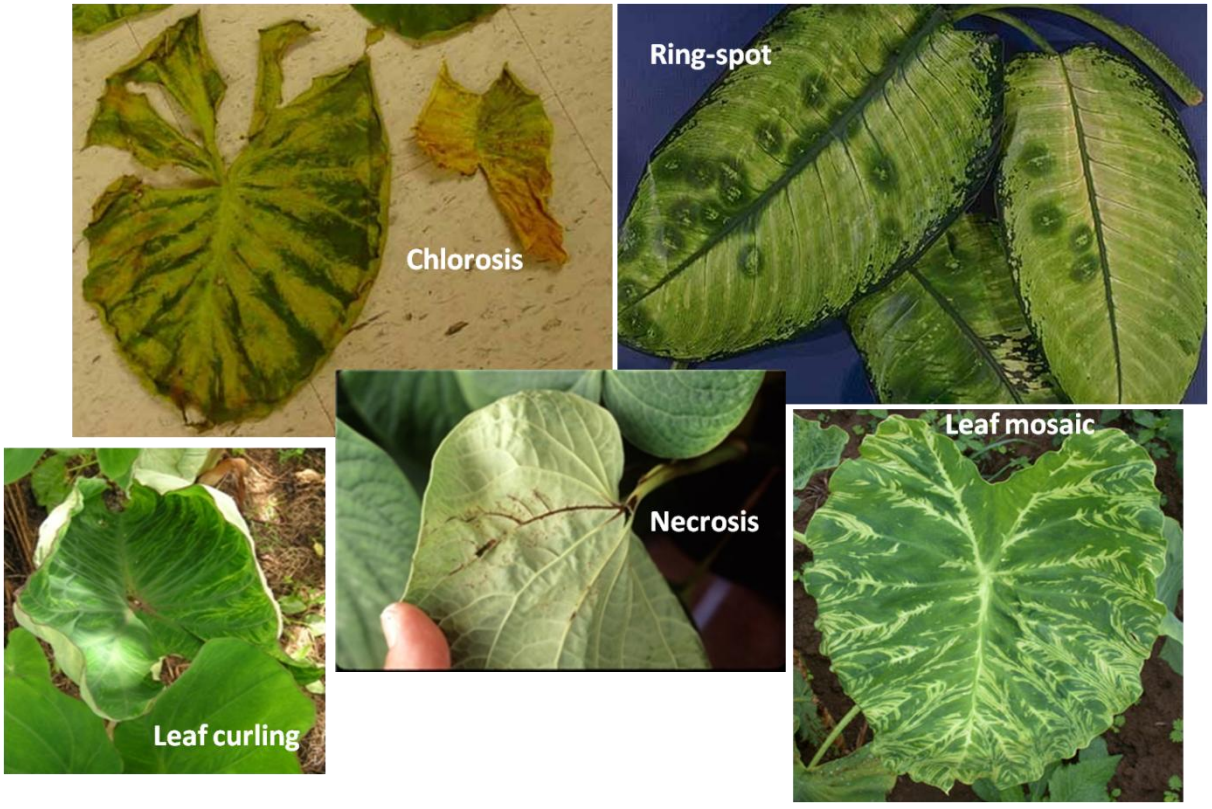


Figure 1.4: Some of the common symptoms shown by plants infected with DsMV (Harding, 2012; Nelson, 2012; Pestnet, 2012; PNW-Handbook, 2012)

### 1.2.2 Disease spread

The disease cycle of DsMV is shown in figure 1.5. Similar to other potyviruses, DsMV is also dispersed by several species of aphids or by the movement of infected plants or as infected plant sap on pruning tools (Nelson, 2008). DsMV arrives at the site of infection (a wounded plant cell) in an aphid's mouthparts or on an infested cutting tool. DsMV enters the plant cell either by aphid injection or on an infested tool. DsMV particles multiply within host cells and move between adjacent cells through plasmodesmata and long-distance within the plants via vascular tissues. Once inside a host cell, the virus unencapsidates and undergoes translation and replication. DsMV particles survive within a living host cell or it can survive in the mouthparts of aphids for short periods (less than an hour, (Karam & Heinze, 2003)) or as infested sap on tools. DsMV cannot survive in dead plant tissues or dead aphids (Nelson, 2008).

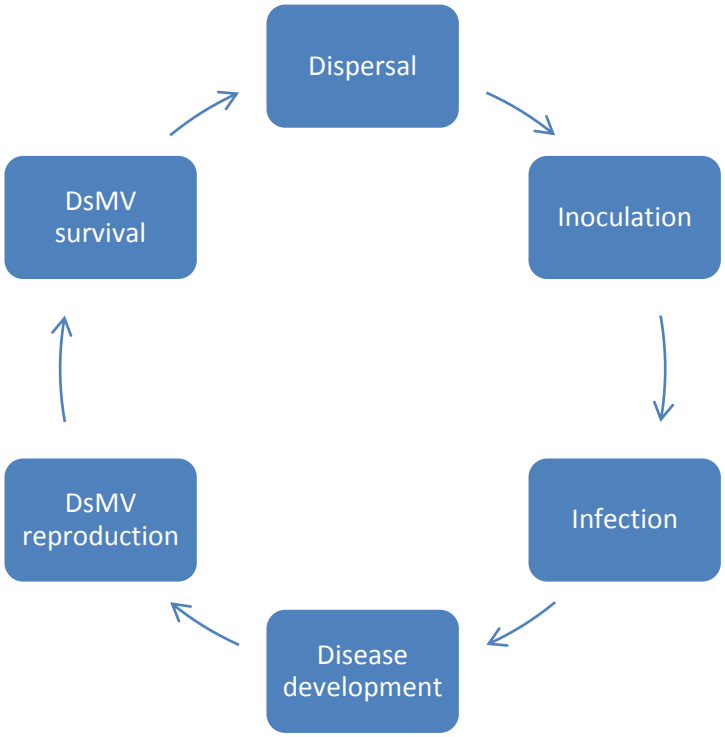


Figure 1.5: Disease cycle for DsMV (Nelson, 2008)

## 1.3 Potyviral Translation and Replication

### 1.3.1 Translation initiation

The infection cycle of *Potyvirus* starts when viral particles enter the host cell through a wound or aphids. Once the virus enters the host cell, viral particles disassemble and release the viral RNA into the host cytoplasm. Potyviral RNA is then translated by the host ribosomes. For most cellular mRNAs present in eukaryotes, translation is initiated by the recognition of the 5' cap structure (m7GppN) by eukaryotic initiation factor eIF4E (eIF4E and eIF(iso)4E in plants) (Bailey-Serres, 1999; Browning, 1996). In contrast to the cap structure present in cellular mRNA, potyviruses possess a viral-encoding protein (VPg) covalently linked to the 5'UTR. However, it has been shown that an intact VPg at the 5'end of the potyviral genome is not required for achieving infection (Riechmann, Laín, & García, 1989). Thus, it has been proposed that initiation of translation within potyviruses might be similar to that of the *Picornaviridae* family, where the ribosomal recruitment is mediated by formation of stable secondary structures (internal ribosome entry sites, IRES) within the 5'UTR (Pelletier & Sonenberg, 1988). While no IRES structures have been identified in potyviruses, cap-independent regulatory elements (CIREs) have been found in the 5'UTR of *Tobacco etch virus* (TEV), suggesting that potyviruses may use a cap independent mechanism for translation (Niepel & Gallie, 1999). Further, this region may be associated with the regulation of translation efficiency (Niepel & Gallie, 1999). The 5'UTR (143 nt) of TEV mediates cap-independent translation (Carrington & Freed, 1990), and promotes translation efficiency in conjugation with the poly (A) tail (Gallie, Tanguay, & Leathers, 1995; Niepel & Gallie, 1999). This is facilitated by two CIREs that span nucleotides (nts) 28-118 in the 5' UTR (Niepel & Gallie, 1999). Smaller segments of the TEV 5'UTR, 1-20, 37-65, 67-113 and 110-114 nt were also reported to enhance translation (Kawarasaki et al., 2000). Zeenko and Gallie (2005) recently reported that the 5' proximal domain (38-75 nt) can fold into a 45 nt RNA pseudoknot (PK1), that is required for full cap-independent translation (Zeenko & Gallie, 2005).

Some studies have also pointed towards a role for the VPg in the initiation of viral translation. The VPg of *Turnip mosaic virus* (TuMV) and TEV have been shown to interact with the eIF(iso)4E factor of their host, *Arabidopsis thaliana* (Lellis, Kasschau, Whitham, & Carrington, 2002; Wittmann, Chatel, Fortin, & Laliberté, 1997).

It was found that mutants of Arabidopsis in which the expression of the eIF(iso)4E was compromised lost susceptibility to TuMV and TEV (Lellis et al., 2002). Loss of interaction between eIF(iso)4E and the VPg meant that viral translation could not take place, thus no virus infection could occur. Consequently, it has been suggested that the role of the VPg in initiating translation could be similar to that of a 5'cap structure, that is serving as an assembly site for the translation initiation factor complex and subsequent recruitment of the 40S ribosomal subunits (Lellis et al., 2002).

Thus, it is unclear if potyviruses use a cap-dependent or independent mechanism for translation. They may be able to use either as circumstances demand. Once translation occurs and the individual proteins cleaved from the polyprotein, replication can then occur.

### 1.3.2 Replication

Similar to other positive sense ssRNA viruses, potyviruses replicate through a complementary negative strand RNA molecule generated by the viral-encoded RNA-dependent RNA polymerase (RdRp); the nuclear inclusion protein b (NIb). The generation of the negative strand RNA molecule is mediated by the recognition of the 3' end of the positive strand RNA molecule by the RdRp, which probably requires the presence of specific secondary structures in the 3'UTR (Haldeman-Cahill, Daros, & Carrington, 1998; Teycheney et al., 2000). Once the negative strand RNA molecule is synthesised, it is used as a template for the generation of subsequent positive-stranded RNA molecules. Potyviral replication takes place in association with cellular membranes (Bienz, Egger, Pfister, & Troxler, 1992; Schaad, Jensen, & Carrington, 1997)

Several proteins encoded by the viral genome are involved in viral replication. The proteins putatively involved in the potyvirus replication complex (CI, VPg, NIaPro and NIb) bind RNA (Merits, Guo, & Saarma, 1998) and interact with the non-structural proteins P1 and P3 (Merits, Guo, Järvekülg, & Saarma, 1999). It has been found that the RNA polymerase activity of the NIb protein and helicase activity of the CI protein play a significant role in the replication of the viral genome (Hong & Hunt, 1996; Lain, Riechmann, & Garcia, 1990). It is speculated that the NIa recruits the NIb to the site of

replication through protein-protein interactions (Daròs, Schaad, & Carrington, 1999). In addition, the potyviral CP can interact with the NIa and NIb (Hong et al., 1995). The P3 protein has no known enzymatic activity but contains a putative trans-membrane domain and may be involved in anchoring the replication complex to membranes (Klein, Klein, Rodriguez-Cerezo, Hunt, & Shaw, 1994; M. T. Martin, Cervera, & Garcia, 1995; Rodríguez-Cerezo & Shaw, 1991). The coding regions for the small 6K1 and 6K2 peptides in the polyprotein flank the CI domain. Mutations that prevent processing of the 6K1 peptide from the P3 protein, or the 6K2 peptide from CI, are not lethal to *Tobacco etch virus* (TEV) (Restrepo-Hartwig & Carrington, 1994) and *Plum pox virus* (PPV) (Riechmann, Cervera, & García, 1995). It has been hypothesized that the 6K2 peptide may function as a membrane anchor for the CI and VPg during TEV replication (Restrepo-Hartwig & Carrington, 1994; Schaad, Haldeman-Cahill, Cronin, & Carrington, 1996; Schaad et al., 1997). Subsequently, the replication complex may be released from the membrane by catalysis of cleavage at the polyprotein proteolytic site between 6K2 and VPg (Schaad et al., 1996; Schaad et al., 1997). Once the RNA has been replicated, it is then packaged in CP and moved throughout the plant. The HC-Pro is also attached to the particle. This protein, together with a DAG amino acid motif within the CP are essential for aphid transmission to other plants (Maia, Haenni, & Bernardi, 1996).

## 1.4 Quasi-species

For a virus to infect a given host, it must be able to spread from initial infection foci to the rest of the plant. This means that several viral processes such as viral replication, translation, cell-to-cell movement of the virus, and long-distance movement of viral particles must be accomplished within the host. These processes are regulated by proteins encoded by the potyviral genome that interact with the host and variations within these proteins can result in an incompatible interaction with the host, which are deleterious for these viral processes (Klein et al., 1994). Therefore, the genetic composition of a virus is constrained by its host (Simmonds, 2004). However, a *Potyvirus* species is not represented by a unique genomic sequence, but rather by a pool of variants or mutants known as a quasi-species (Eigen, 1996), which are centred on a reference sequence (Roossinck, 1997). The existence of genetic variants within a

potyviral population increases the probability of survival and helps in broadening the host range of potyviruses (Domingo, 2002).

Nucleotide substitutions are common within Potyviral genomes. As RNA viruses, they are dependent on their own RdRp rather than a host polymerase for replication. The high mutation rates of RNA viruses is due to the low efficiency or absence of proofreading activity of the viral RdRp (Steinhauer, Domingo, & Holland, 1992). It has been estimated that the error rates of viral RdRp are approximately $10^{-4}$ (Domingo et al., 1996). This means that for every 10,000 nt replicate, 1 nucleotide error can be expected and as the copies are made, the number of errors can increase rapidly. In contrast, the mutation rates for cellular RNA polymerases are much lower, from $10^{-8}$-$10^{-11}$ (Domingo et al., 1996; Roossinck, 1997).

Studies focussing on genetic variability between different isolates of the same viral species have found that certain regions within the potyviral genome are more susceptible to mutations and show high levels of genetic variability. For example, it has been suggested that within potyviruses, the 5'UTR, P1 and N-terminus of the CP show high level of genetic variation while the CI, NIa and NIb appear to be more conserved. This suggests that the pattern of genetic variability is dependent upon the region analysed (Aleman-Verdaguer, Goudou-Urbino, Dubern, Beachy, & Fauquet, 1997; Fuji & Nakamae, 2000; Kekarainen, Merits, Oruetxebarria, Rajamäki, & Valkonen, 1999).

As discussed before, a pool of genetic variants generated around a reference sequence forms a viral population. However, the size of viral population is limited and is usually constrained by various factors. For example, the mutants produced in *Nicotiana benthamiana* after a single passage was higher for *Cucumber mosaic virus* (CMV) (43%) than for TMV (29%), the mutation frequencies remaining constant for TMV during serial vegetative propagation of infected plants. CMV accumulated to lower levels than TMV in the same host, indicating that the difference in the genetic diversity was not due to difference in the degree of viral replication (Schneider & Roossinck, 2000). Thus, the pattern of genetic diversity can vary among different virus species infecting a common host. In addition, the level of genetic diversity found within the same virus can differ among different host species. For instance, the percentage of CMV mutants found in *N. benthamiana* was lower than that observed in *N. tabacum* (43% versus 70%, respectively). Furthermore, TMV showed a higher percentage of mutations in *N. tabacum* (46%) than in *N. benthamiana* (29%) (Schneider & Roossinck,

2001). These findings suggest that the host plays a crucial role in both generation and limitation of genetic diversity.

### 1.4.1 Impact of variation on evolution

The population of potyviruses are genetically diverse. As in the case with all living organisms, reproduction leads to the generation of individuals that differ genetically from their parents; these are called mutants or variants. The distribution frequency of these variants within a given population may differ with time and may lead to the appearance of new species. This process is called evolution (García-Arenal, Fraile, & Malpica, 2003). In potyviruses, studying the pattern of evolution is important for the development of efficient control and stable control strategies, as these often fail because of high virus mutation rates. This high rate of mutation helps them in overcoming the host defence mechanisms and broadening their host range (García-Arenal et al., 2003). The level of genetic variation observed within a *Potyvirus* species is the result of errors that occurs during the viral replication process. It has been found that the high mutation rates observed in potyviruses are not because of their adaptive nature, but rather than being an evolutionary strategy it appears to be due to the need for rapid replication of their chemically unstable RNA genomes (Drake & Holland, 1999). The population diversity depends on three parameters: the number of variants present within a population, the frequency of occurrence of each variant within a population and genetic distances between the existing variants (García-Arenal et al., 2003). As the virus continues to replicate the population of one variant may also change with time within an infected plant. Therefore, it may be possible that one infected plant may have two different sets of dominant sequences at two separate time points (Higgins, Cassidy, Teycheney, Wongkaew, & Dietzgen, 1998). If the sequences change over time, this may influence the direction of evolution. Identification of conserved regions within a viral genome also allows these areas to be targeted for disease diagnosis (Eigen, 1996; Higgins et al., 1998).

The high mutation rate of viral RNA genomes may be influenced by several factors such as ionic concentration or concentration of nucleotide substrates within the host cells, or by environmental stress affecting the host plant. The increase in mutation rate can facilitate the increase in the population of some variants and therefore those

variants will become dominant in the population. Different dominant sequences may show different infectivity or host selection. This may increase viral pathogenicity resulting in more severe infections to the hosts (Domingo et al., 1996). Therefore, it becomes important to understand the pattern of diversity of plant RNA viruses. While studying the genetic variability of viruses it becomes interesting to identify regions that do not vary during viral replication. The conserved regions within the viral genome may hold structural or functional importance. These highly conserved regions may be necessary for the virus to replicate and continue infecting and may be crucial for pathogenic activity of the viruses. The identification of conserved sequences may also give better insight into the life cycle of pathogenic viruses. This implies that better understanding of the viral genome and functionality can help in developing novel anti-viral mechanisms which can help protect agricultural plants from viral infections (Ashkenazy, Erez, Martz, Pupko, & Ben-Tal, 2010).

### 1.4.2 Impact of variation on detection and diagnosis

Precise and accurate identification of potyviruses that infect commercially important crops is required to monitor disease incidence, to identify virus reservoirs and to recommend appropriate control mechanisms (Grisoni et al., 2006). However, precise detection and identification of a *Potyvirus* species has always been problematic (Van Regenmortel, International, & Fauquet, 2000). The traditional methods used for discriminating between a potyviral species and its isolates were based on serology and biological criteria such as host-range, cross-protection, and symptomatology (Shukla et al., 1991). Recently, some researchers have used the genomic composition of the virus, mainly the coat protein (CP) and the 3'UTR sequences (Van Regenmortel et al., 2000). Members of the *Potyvirus* genus tend to have a narrow host range, confined to closely related genera (Shukla, McKern, Gough, Tracy, & Letho, 1988). However, *Potato virus Y* (PVY), the type species of the genus can infect species within Chenopodiaceae, Comelinaceae and Solanaceae families. Members with even broader host range exist, such as *Lettuce mosaic virus* (LMV) and *Bean yellow mosaic virus* (BYMV) that can infect species of eight and thirteen different plant families, respectively (Bos, Kowalska, & Maat, 1974; German-Retana, Walter, & Le Gall, 2008). Even though host range assays and symptomatology are helpful tools for the initial screening of potyviruses,

conflicting results have been reported apparently due to differences in the environmental conditions affecting test plant growth or the use of different genotypes of plant species (Brunt, 2000).

The use of antibodies raised against the viral CP (serology) is widely used for detection and identification of plant viruses. This approach is also applied for identification of new viruses and to determine their phylogenetic relationships with existing viruses (Shukla et al., 1991). When dealing with potyviruses, the use of serological relationships between virus species gets hampered either due to cross-reactions between closely related species or due to the lack of its ability for detecting all the variants present within a viral population. Cross-reactions are probably due to the recognition by antibodies of the highly conserved core region of the CP. Lack of identification of some strains is mainly due to the variability in the N-terminal domain of the CP, which also may be lost due to proteolysis during virus extraction from plant tissues (Shukla et al., 1991). The present studies of viral genetic diversity rely on techniques such as reverse transcriptase- polymerase chain reaction (RT-PCR) and sequencing of PCR products using traditional dideoxy sequencing methods (Jan, Azam, Warsi, Ali, & Haq, 2011). These techniques identify a sub-region of the genomes, most commonly the CP and can only detect one or two variants within this region among the viral population. For example, success of detection of potyviruses using RT-PCR is dependent on annealing of primers, which may be a problem if sequences vary. Therefore, to overcome this limitation of RT-PCR based detection methods, degenerate primers have been designed (Langeveld et al., 1991). The usefulness of degenerate primers in detection and identification of potyviruses will be discussed in chapter 2.

However, the development of new and powerful sequencing techniques, so called next generation sequencing such as 454 pyrosequencing and Illumina/Solexa technology, have made it possible to carry out precise and accurate detection of potyviruses. Next generation sequencing can also be used to identify and sequence all the variants present in a viral population because it generates thousands of sequences, thereby providing sufficient data to study the full extent of viral genomic diversity (Coetzee et al., 2010). This means that if the infected host transcriptome is sequenced then the whole viral genome can be determined. Otherwise, a region of the genome can be amplified by RT-PCR and sequenced to great depth (discussed in chapter 3). Development of bioinformatics tools such as GeneiousPro has also made it feasible to handle information from large amounts of sequence data (Drummond et al., 2009).

Recently, this kind of approach has been used for identifying rare variants of HIV-1 to develop effective therapeutic treatment against HIV (García-Arenal et al., 2003; Hedskog et al., 2010).

## 1.5 Aims of the study

The overall aim of this study was to understand the pattern of genetic variability within the genus *Potyvirus*. This research focussed on genetic variability occurring within two different strains of DsMV i.e. NZ 1 and B 1 and their isolates i.e. NZ 1.1, NZ 1.2, B 1.1, and B 1.2 produced by serial vegetative propagation.

The specific aims of this study were:

- To test the effectiveness of published universal *Potyvirus* primers in detection and identification of DsMV.
- To determine the level of genetic variation within the whole genome of each strain of DsMV and their isolates
- To determine the level of variation of each gene within each DsMV strain and their isolates
- To compare the level of genetic variation found between different strains of DsMV and to determine if there are any regions of conservation within the DsMV genome.

# Chapter 2

# Testing potyvirus universal primers

## 2.1 Introduction

The genus *Potyvirus* forms one of the largest groups of plant viruses and causes significant damage to agricultural, horticultural, pasture and ornamental crops (Shukla et al., 1991). Within this genus, *Dasheen mosaic virus* (DsMV) is one of the most important viral pathogens of cultivated aroid plants worldwide (Chen et al., 2001). DsMV was first identified by Zettler *et al.* in 1970 from dasheen, also known as taro (*Colocasia esculenta*) (Elliot et al., 1997). DsMV infects both cultivated plants and aroids across the world causing severe damage to many ornamental plants and agricultural crops such as *Caladium bicolour*, *Dieffenbachia maculata*, and *Colocasia esculenta* to name a few (Zettler & Hartman, 1987). DsMV infection is naturally restricted to aroids (Babu, Hegde, Makeshkumar, & Jeeva, 2011) and is transmitted in a non-persistent manner by several widely distributed aphid species such as *Myzus persicae* and *Aphis gossypii* (Babu et al., 2011; Ha et al., 2008). Plants infected by DsMV show mosaic, leaf puckering and filliformy or shoestring like symptoms (Zettler & Hartman, 1987) (figure 2.1). In some cases, DsMV infection may lead to 40%-60% reduction in total crop yield causing around 40% reduction in the average corm weight of edible aroids. (Babu et al., 2011; Elliot et al., 1997).


Figure 2.1: Pictures showing a healthy taro plant, edible corm and a taro plant infected with DsMV (Davis, 2012; Finegardening, 2012; Harding, 2012)

### 2.1.1 Virus disease diagnosis

Precise and accurate identification of Potyviruses that infect commercially important crops is required to monitor disease incidence, to identify virus reservoirs and

to recommend appropriate control mechanisms (Grisoni et al., 2006). Detection of virus by using sensitive diagnostic tools can assist controlling the spread of viral diseases among plants. Development of an efficient diagnostic technique will allow the detection of viral infection during various stages of plant development. Many different approaches are used for analysing plant viruses (Alvarez, 2004; R. R. Martin, James, & Lévesque, 2000). Initially, viral disease can be quickly diagnosed by visual examination of symptoms that they cause in a host. However, some viruses infect broad host ranges but may be symptomless or produce symptoms similar to other viruses, thus requiring molecular tests for diagnosis that are more sensitive (García-Arenal, Fraile, & Malpica, 2001; Sharma, Singh, Nagpal, Virk, & Zaidi, 2009). In the last few years, significant advancements have occurred in the available diagnostic techniques in order to achieve rapid and reliable detection of plant viruses. These techniques include wood indexing, electron microscopy (EM), serological and molecular methods and arrays, each with particular advantages and disadvantages (Jan et al., 2011).

#### 2.1.1.1 Electron microscopy

The electron microscope was developed during the 1930's in Germany and viruses were among the first objects to be seen with it (Madeley, 1997). The electron microscope uses a beam of electrons to create an image of a specimen and has greater resolving power than a light microscope (Erni, Rossell, Kisielowski, & Dahmen, 2009). There are two types of electron microscopes: transmission electron microscope (TEM) and scanning electron microscope (SEM) (Egerton, 2005). TEM involves production of high voltage electron beam emitted by an electrode under the influence of magnetic field and when the electron beam passes through the specimen, it generates an image (JIC, 2012). In contrast, SEM builds an image by detecting the secondary electrons emitted from the surface of a specimen due to excitation by the primary electron beam (JIC, 2012). TEM has been more helpful to virologists in the morphological characterisation of viral particles in crude and purified samples (Kitajima, 2004). Electron microscopy can help in rapid diagnosis of serious infections but it requires expensive equipment and experienced staff. EM is able to show the shape and dimension of an infecting virus, potentially providing information about the virus family or genus; however, it does not provide information about the species. EM is unable to detect viruses with high sensitivity and specificity (Jan et al., 2011).

### 2.1.1.2 Serology

Serology is based on the recognition of virus antigens with antibodies raised against them (Babu et al., 2011; Jan et al., 2011). Serological methods have been successfully used for many years to identify the virion of different viruses; enzyme linked immunosorbent assay (ELISA) is commonly used to detect viruses such as Cucumber mosaic virus (Sun, Jiao, Zhang, Zhang, & Zhang, 2001). However, when dealing with potyviruses, the use of serological relationships between virus species gets hampered either due to cross-reactions between closely related species (Brunt, 1992) or due to the lack of its ability for detecting all the variants present within a viral population. Cross-reactions are probably due to the recognition by antibodies of the highly conserved core of the coat protein (CP). Lack of determination of some strains is mainly due to the aforementioned variability in the N-terminal domain of the CP, which also may be lost due to proteolysis during virus extraction from plant tissues (Gibbs & Mackenzie, 1997; Shukla et al., 1991). Serological techniques such as ELISA can be used for identifying even small amount of viruses. It can be used for testing multiple plants at a time for a single virus or a single plant can be tested for many viruses. Serological diagnosis of Potyvirus is often imprecise, because of the frequent serological cross-reactions between species, low titre values, and inability to detect virus of unknown origin (Babu et al., 2011; Jan et al., 2011).

### 2.1.1.3 Molecular hybridisation

Molecular hybridisation as a diagnostic tool in plant virology was first used to detect viroids in purified RNA preparations (Owens & Diener, 1981) and later applied to plant viral detection (Garger et al., 1983; Maule, Hull, & Donson, 1983). Owens and Diener (1981) used hybridisation to detect Potato spindle tuber viroid in potato tuber extracts by using cDNA as a hybridisation probe. Later, Garger *et al.* (1983) used hybridisation method to detect cucumber mosaic virus (CMV) and tobacco mosaic virus (TMV). The sensitivity and efficiency of hybridisation depends on several factors such as the concentration and distribution of viruses within an infected plant, virus recovery during sample preparation and the quality of probes that were used for virus detection. The detection of viral pathogens is based on the specific interaction between the nucleotide sequence of the virus and the probe. The specificity of the hybridisation

reaction is dependent upon the stringency of hybridisation and stability of the hybrid complex (Jan et al., 2011). The most commonly used hybridisation approach for virus detection is non-isotopic dot blot hybridisation using digoxigenin-labelled probes. This technique has been used for detection of many viruses, such as apple mosaic virus (ApMV), Prune dwarf virus (PDV), and Apple chlorotic leaf spot virus (ACLSV) (Pallas, Mas, & Sanchez-Navarro, 1998).

### 2.1.1.4 Polymerase chain reaction

Polymerase chain reaction (PCR) was developed by Kary Mullis in the 1980's. It allows selective amplification of a selected region of a DNA molecule (Brown, 2010). The principle behind PCR is based on the ability of DNA polymerase to synthesise a new complementary strand against a template strand (Brown, 2010). As DNA polymerase can add nucleotides only at the free 3'-OH group, it needs a primer (i.e. short stretch of nucleotides) to which it can attach the first nucleotide. This limitation of DNA polymerase makes it possible to amplify a specific region of the template sequence (NCBI, 2012). The most commonly used DNA polymerase for PCR is *Taq* polymerase (source: Thermus aquaticus) (Brown, 2010).

PCR consists of three steps; namely, denaturation, annealing, and extension (Brown, 2010; NCBI, 2012) (figure 2.2). In denaturation, the reaction mixture is heated at 94˚C for around 30 seconds that causes the melting of a double-stranded template by disrupting the hydrogen bonds to give rise to single stranded DNA molecules (Brown, 2010). In annealing, the reaction mixture is cooled down to 50˚C-60˚C, which facilitates the primers in annealing to the template (Brown, 2010). In extension, the temperature is increased to around 72˚C to allow the DNA polymerase to synthesise a new complementary DNA strand against the template by adding dNTPs that are complementary to the template in 5'-3' direction (Brown, 2010; NCBI, 2012). Reverse-transciptase PCR is similar to PCR, but it involves the conversion of RNA to single-stranded cDNA that then acts as the template in a PCR (NCBI, 2012).

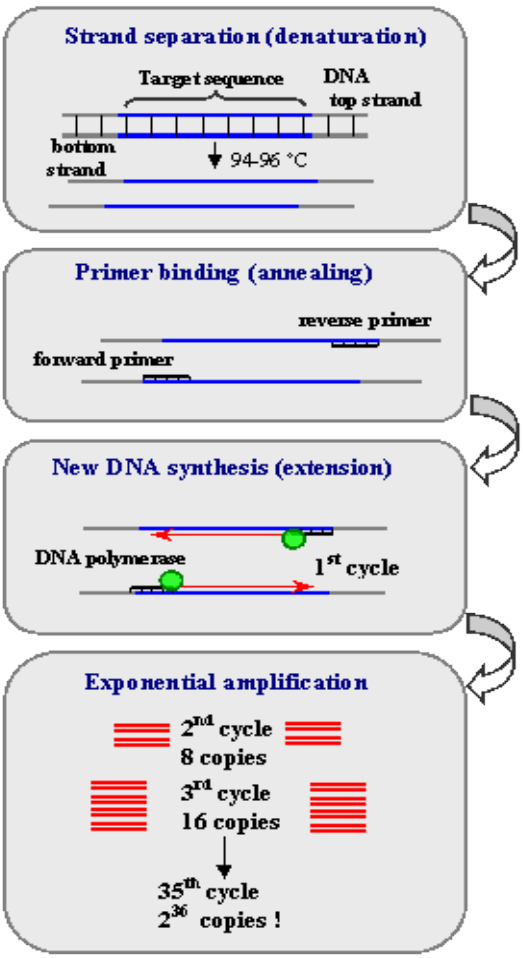


Figure 2.2: The basic steps involved in a PCR i.e. denaturation, annealing and extension (NCBI, 2012)

PCR and RT-PCR have emerged as the most sensitive techniques for detecting and amplifying particular plant-infecting viruses, such as potyviruses (Antoniw, 1995). PCR based techniques are extremely sensitive, inexpensive and require minimal skills. The high sensitivity of this technique is its major advantage; RT-PCR has been successfully used in detection of potyviral infection even in the absence of visible symptoms (Babu et al., 2011; Ha et al., 2008; Jan et al., 2011; Khan, Jan, Aquil, & Haq, 2011). With the increase in potyvirus sequence information, it has become possible to identify potyviruses using PCR based methods (Langeveld et al., 1991). PCR based techniques for the detection and identification of potyviruses is based on the use of primers which target a specific region within the viral genome. Primer design may allow strain specific, species specific, genus specific or family specific amplification and thus enables detection of infection (Ha et al., 2008).

### 2.1.2 Universal primers for detection of potyviruses

The choice of primers determine the effectiveness of PCR in detection of a single virus species or many members of a group or family of related species (Colinet, Nguyen, Kummert, Lepoivre, & Xia, 1998). Information about the conserved viral sequences helps in designing primers for use in PCR, to help in rapid identification of uncharacterised potyviruses (Langeveld et al., 1991). Use group-specific PCR primers for detection of plant viruses has been demonstrated for several viruses; for example, carlaviruses and luteoviruses (Badge et al., 1996; Robertson, French, & Gray, 1991).

The increase in potyvirus sequence data has permitted the design of genus-specific universal primers for the detection of new and/or uncharacterised potyviruses (Antoniw, 1995; Langeveld et al., 1991). Universal primers are primers that bind to a sequence found within a conserved region (Zheng, B. C. Rodoni, M. J. Gibbs, & A. J. Gibbs, 2010b). Universal primers may provide a way to detect many viral species within a group or family. Such primers target conserved regions within the viral genome, therefore identification of conserved regions among RNA viruses is especially useful as they can be used as target regions for RNA viruses (Langeveld et al., 1991; Malpica et al., 2002). Lack of proofreading during replication of RNA viruses causes high mutation rates (Malpica et al., 2002). This spontaneous mutation creates a high level of genetic diversity, providing the foundation for rapid genomic evolution (Vignuzzi, Stone, Arnold, Cameron, & Andino, 2006). Within the host plant, RNA viruses live as quasispecies; therefore, many variants of the pathogenic virus are present in an infected host plant (Domingo, 2002). This variation may be different between strains of the same virus, at different stages of infection and in different hosts (Drake & Holland, 1999). Therefore, the use of universal primers allows fragments of similar DNA to be amplified from a group of related sequences, for example potyviruses (Antoniw, 1995).

Universal primers for RNA viruses such as potyviruses are often highly degenerate and designed manually finding the conserved sequences by comparing genome sequences. They allow the detection of a range of species with different sequences as well as previously unknown viruses. The conserved regions identified in the core domain of the potyvirus coat protein (CP) gene and NIb as well as the regions are commonly used for designing generic primers for detection of potyviruses (Langeveld et al., 1991). The binding positions for these primers are shown in figure

2.3. Several *Potyvirus* genus-specific universal primer pairs have been developed in the past few years, such as MJ1-MJ2, PVIP1-PV2, U1000-D1000, U335-D335 and U341-D341 (Colinet et al., 1998; Gibbs & Mackenzie, 1997; Grisoni et al., 2006; Langeveld et al., 1991). These primers are some of the most commonly used universal primers that are used for detection and identification of Potyviruses (figure 2.3).

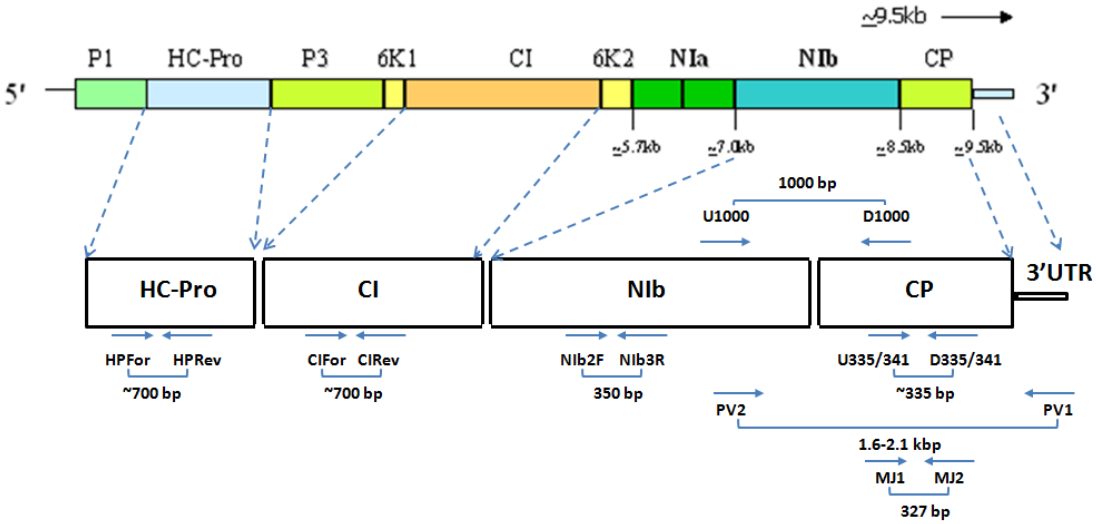


Figure 2.3: The relative positions of the potyvirus universal primer binding sites

The majority of universal primers designed for *Potyvirus* detection target the 3' end of the potyviral genome such as the CP and NIb regions, as these are the most extensively studied regions of potyviruses (Ha et al., 2008). Studies have shown that among potyviruses, a part of the gene encoding the NIb protein is highly conserved among the entire genus of *Potyvirus* (Gibbs & Mackenzie, 1997; Ha et al., 2008). In 2010, Zheng *et al.* carried out an extensive analysis of 17 of the most conserved regions within *Potyvirus* genomes and found that the most conserved region was present within the NIb region (Zheng et al., 2010b). Based on their findings, Zheng *et al.* (2010) designed a pair of universal primers, NIb2F and NIb3R based on two conserved domains within the NIb region (Zheng et al., 2010b) (figure 2.3). This primer pair was reported to amplify a 350 bp product from DsMV (Zheng et al., 2010b). In a similar study, Ha et al. (2008) designed two alternative sets of degenerate primers to amplify sequences from the 5' (HC-Pro) and central (CI) regions of potyviral genomes that can be used for detection of potyviruses (Ha et al., 2008). The CIFor-CIRev universal primers were designed against the conserved region within the CI coding gene (figure 2.3) based on the alignment of 56 complete nucleotide sequences from viruses within the family *Potyviridae* and has been shown to amplify a ~700 bp product (Ha et al.,

2008). They further designed HPFor and HPRev primers based on the alignment of 149 complete *Potyvirus* nucleotide sequences to amplify a ~700 bp fragment from the HC-Pro coding region (Ha et al., 2008) (figure 2.3).

However, as discussed earlier, the design of universal primers depends upon the availability of sequence information of the viruses at any given time. Primers developed using a small set of sequences may not cover genetic variations occurring within each strain of virus (Zheng et al., 2010b). Therefore, as more sequences are determined, the variability within conserved regions becomes more apparent and can explain why many universal primers fail. This decay in consensus explains why *Potyvirus* genus specific primers designed in the past have failed to detect and identify all members of the genus (L. Zheng et al., 2008). Therefore, this study was carried out to test the efficiency and effectiveness of published universal Potyvirus primers, which are often used for detection and identification of DsMV. MJ1 and MJ2 has been shown to detect DsMV (Babu et al., 2011); however, the usefulness of HPFor-HPRev, CIFor-CIRev and NIbFor-NIb to detect DsMV has not been determined. Each of these primer pairs was selected since they target a different region of Potyviral genome. Therefore, these three primer pairs were tested in this study for their effectiveness in amplifying the DsMV genome.

In this study, RT-PCR was used for testing the NIb, CI and HC-Pro primer pairs for their effectiveness in amplifying the DsMV genome. The findings of this study helped in examining the usefulness of universal Potyvirus primers in detection and identification of DsMV infection.

## 2.2 Methods and Materials

### 2.2.1 RNA Extraction

Total RNA was extracted from 100 mg DsMV infected taro leaves using the Spectrum Plant Total RNA kit (Sigma-Aldrich, St Louis, MO, USA) following the manufacturer's instructions. Infected leaves were ground to a fine powder in a sterile mortar and pestle using liquid nitrogen. Lysis solution (500 µl) containing ß-mercaptoethanol was added to the tissue powder and was vortexed immediately for 30 secs. The mixture was incubated at 56˚C for 5 mins, after which it was centrifuged at maximum speed for 3 mins to pellet the cellular debris. The supernatant was transferred to a filtration column and centrifuged at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 1 min to remove the residual debris. Binding solution (500 µl) was mixed with the clarified lysate. This mixture was then transferred into a binding column and centrifuged at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 1 min to allow RNA to bind to the column. Wash solution I (300 µl) was added to the binding column and centrifuged at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 1 min. In order to get rid of genomic DNA, 10 µl of DNase I and 70 µl of DNase digestion buffer was added to the centre of the column and incubated at room temperature for 15 mins. The digested DNA was removed from the column by adding 500 µl of Wash solution I to the column and was centrifuging at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 1 min. Wash solution II containing ethanol (500 µl) was added to the column and the column was centrifuged at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 30 secs. The flow through liquid was discarded and the column was again washed with 500 µl of Wash solution II. The empty column was dried by centrifuging the column at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 1 min. The RNA was eluted from the column by adding 50 µl of elution buffer for 1 min at room temperature and centrifuging at 13,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 1min. The eluted RNA was stored at -80˚C for long-term usage.

### 2.2.2 Assessment of RNA quality

The concentration and purity of the extracted RNA was measured based on the 260/280 nm absorbance ratio of 1.8-2.0 using a GE Nanovue spectrophotometer (GE Biosciences, Auckland, NZ). The integrity of the RNA was determined by electrophoresis in 1.5% agarose/Tris/Borate/EDTA (TBE) gels (Mini-Sub® Cell GT Cell, BioRad, Auckland, NZ). Gels were run for 1.5 hours at 75 volts. 350 ng of 100 bp DNA ladder (Dnature, Gisborne, NZ) was used as a marker. The agarose gels were visualised by exposing to UV and photographing.

Table 2.1: Primers used in cDNA synthesis

| Primer | Primer Sequence (5'-3') | Reference |
|---|---|---|
| PV1/SP6 (FP) | GATTTAGGTGACACTATAG$(T)_{17}$(A/G/C) | (Mackenzie et al., 1998) |
| Oligo dT (FP) | ACTATCTAGAGCGGCCGCTTT$_{(16)}$ | Quanta Biosciences kit |
| HPRev (RP) | GANCCRWANGARTCNANNACRTG | (Ha et al., 2008) |
| CIRev (RP) | ACNCCRTTYTCDATDATRTTNGTNGC | (Ha et al., 2008) |

Single letter code N=A/T/G/C, Y=C/T, R=A/G, W=A/T, D=A/G/T; FP: Forward primer, RP: Reverse primer

Table 2.2: List of all the Potyvirus universal primers tested in this study

| Primer | Primer Sequence (5'-3') | Expected Product size | Target Region | Published annealing temp. (˚C) | Reference |
|---|---|---|---|---|---|
| NIb2F (FP) | GTNTGYGTNGAYGAYTTYAAYAA | 350bp | NIb | 45 | (Zheng, Rodoni, Gibbs, & Gibbs, 2010a) |
| NIb3R (RP) | TCNACNACNGTNGANGGYTGNCC | | | | |
| HPFor (FP) | TGYGAYAAYCARYTNGAYNNNAAYG | 700bp | HC-Pro | 40 | (Ha et al., 2008) |
| HPRev (RP) | GANCCRWANGARTCNANNACRTG | | | | |
| CIFor (FP) | GGNVVNGTNGGNWSNGGNAARTCNAC | 700bp | CI | 40 | (Ha et al., 2008) |
| CIRev (RP) | ACNCCRTTYTCDATDATRTTNGTNGC | | | | |

Single letter code N=A/T/G/C, Y=C/T, R=A/G, W=A/T, V=A/C/G, S=G/C, D=A/G/T; FP: Forward primer, RP: Reverse primer

### 2.2.3 Two-step RT-PCR

#### 2.2.3.1 cDNA synthesis

Single stranded cDNA was synthesised using a qScript Flex cDNA synthesis kit (Quanta Biosciences) according to the manufacturer's instructions and using one of the primers in table 2.1. RNA (1 µg) was added to a mixture containing 2 µl of 10 µM primer, 2 µl of GSP enhancer, and nuclease free water (variable) to make up the reaction volume to 15 µl. In some instances 1 µl of 10 µM random hexamers was added. The mixture was incubated at 65 ˚C for 5 mins to facilitate the primer annealing. Following that, a mixture of reverse transcriptase (1 µl) and qScript reverse-transcriptase reaction mix (4 µl) was added to the mixture to make the final reaction volume to 20 µl. The first strand cDNA synthesis was carried out at 42 ˚C for 60 mins, followed by 5 mins incubation at 85 ˚C and held at 4 ˚C using a TECHNE thermocycler (Straffordshire UK, Model FTGRAD2D). The cDNA was stored at -20 ˚C.

#### 2.2.3.2 PCR

The amplification reaction was performed in 12.5 µl reaction volumes containing 6.25 µl of 1X Go Taq$^{R}$ Green master mix (Promega, Auckland, NZ), 1.0 µl of 10 µM forward and reverse primers (see table 2.2), 2.5 µl cDNA template and 1.75 µl of sterile water. PCRs were carried out in a Techne PCR machine (Straffordshire UK, Model FTGRAD2D). For the NIb primers the conditions published by Zheng et al. (2010) were used as follows: initial denaturation at 94˚C (2 min), denaturation at 95˚C (45 sec), 35 cycles of primer annealing at 45˚C (45 secs), extension at 72˚C (45 secs) and a final extension at 72˚C (5 mins). After PCR, tubes were stored at 4˚C. For the HP and CI primers, PCR conditions published by Ha et al. (2008) were used for the reaction. The setup used for the thermal cycler was as follows: initial denaturation at 94ºC (2 mins), denaturation at 94ºC (30 secs), primer annealing at 40ºC (30 secs), extension at 68ºC (1 min) and final extension at 68ºC (5 mins). The reaction tubes were stored at 10ºC.

### 2.2.4 Touchdown PCR

Reactions were set up as described in 2.2.3.2 with 2.5 µl of cDNA. Touchdown PCR using following cycling conditions: initial denaturation at 95˚C (30 secs), 3 cycles of 95˚C for 30 secs, annealing (T-n) ˚C for 30 secs (where T equals start temperature and n equals cycle number), 72˚C for 1 min; 25 cycles of 95˚C for 30 secs, 45˚C for 30 secs, 72˚C for 1 min and final extension at 72˚C for 5 mins.

### 2.2.5 Gradient PCR

PCR were set up as described in 2.2.3.2 with 2.5 µl of cDNA template. Gradient PCR conditions used for NIb2F and NIb3R were as follows: initial denaturation at 94 ºC (5 min), 35 cycles of denaturation at 95 ºC (45 sec), annealing over the range 35˚C-55˚C (45 sec), extension at 72 ºC (45 sec) and final extension at 72 ºC for 5 min. The following PCR cycling conditions were used for HP and CI primers: initial denaturation at 94 ºC (2 min), 40 cycles of denaturation at 95 ºC (45 sec), annealing over a 10˚C range (30 sec), extension at 68 ºC (1 min) and final extension at 68 ºC for 5 min. For the HP primers the range of annealing temperatures was 50˚C-60˚C and 60˚C-70˚C, and for the CI primers it was 45˚C-55˚C. The PCR products were stored at 10 ˚C.

### 2.2.6 One-step RT-PCR

One-step RT-PCR was carried out using a SuperScript$^{R}$ III One-step RT-PCR system with Platinum$^{R}$ Taq (Life Technologies, Victoria, Australia) according to the manufacturer's instructions. The amplification reaction was carried out in a Techne PCR machine (Straffordshire UK, Model FTGRAD2D), in 25 µl final reaction volume containing 1 µg of RNA template, 0.5 µl of 10 µM forward and reverse primer, 12.5 µl Superscript 2X reaction mix, 1 µl of Superscript III reverse transcriptase and 9.08 µl of sterile distilled water. cDNA synthesis was carried out at 55 ˚C for 30 min. Amplification was carried out as follows: initial denaturation at 94 ˚C (2 min), 40 cycles of denaturation at 94 ˚C (15 sec), annealing at 55 ˚C or 60 ˚C (30 sec), extension at 68

˚C (1 min) and one cycle of final extension at 68 ˚C (5 min). The amplification product was stored at 10 ˚C.

### 2.2.7 Agarose gel electrophoresis

Agarose gel electrophoresis was used for visualising PCR products following amplification. The whole of each amplification reaction was loaded on a 1% agarose/ 1X TBE gel for 1 hour at 75 volts. 100 bp DNA ladder (350 ng) was used as a marker and the gels were visualised under UV light with the help of 10 µg/ml ethidium bromide stain present within the gel.

### 2.2.8 Cloning of PCR product

PCR product obtained using the NIb primers (350 bp) were ligated into pGEM-T EASY (Promega, Madison, WIs., USA) following the manufacturer's protocol and was transformed into *E.coli* strain JM109.

#### 2.2.8.1 Purification of PCR product

After amplification of the expected sized PCR product, it was purified using the protocol of the Zymo Research PCR clean up kit. Two volumes of DNA binding buffer was added to the PCR product, transferred to a binding column and centrifuged at 10,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 30 secs. The flow through liquid was again discarded and the wash step was repeated. After washing, the column was transferred to a fresh collection tube and 10 µl of sterile RNase free water was added to the column and incubated for 1 min. The DNA was eluted by centrifuging the column at 10,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 30 secs. The concentration of the purified PCR product was measured using a Nanovue spectrophotometer (GE Biosciences, New Jersey, DE, USA).

### 2.2.8.2 Ligation to vector

The pGem-T Easy vector system (Promega) was used for cloning of the purified PCR product following the manufacturer's instructions. The ligation to vector was set up as follows: 1.45 µl of 0.122 µg/µl PCR product was added to a mixture containing 2X Rapid ligation buffer (5 µl), 1 µl of pGEM-T Easy vector (50 ng), 1 µl of T4 DNA ligase (3 Weiss units/µl) and 1.55 µl of nuclease-free water to make a final reaction volume to 10 µl. The reaction was incubated at room temperature for 1 hour.

### 2.2.8.3 Transformation

Frozen JM109 high efficiency competent cells were kept in ice bath until just thawed, following which 50 µl of cells were transferred into cold 1.5 ml tubes containing either 2 µl of ligation mixture or 0.1 ng of uncut pUC18 and were incubated on ice for 20 mins. After which they were heat shocked at 42 ˚C for 45-50 secs. The tubes were immediately returned on ice and 950 µl SOC medium was added to the tube containing cells transformed with ligation mixture while 900 µl was added to the tube containing cell transformed with uncut pUC18. The cells were incubated at 37 ˚C for 1.5 hours with shaking at 150 rpm (211DS Shaking Incubator, Labnet International Inc.).

Transformed cells were plated onto duplicate LB (Luria-Bertani) plates containing 100 µg/ml ampicillin/0.5mM IPTG/80µg/ml X-Gal and incubated overnight at 37°C. A minimum of five white colonies were selected and re-plated onto a LB grid plate and incubated overnight at 37 °C. Each colony was then inoculated into 5 mL of LB broth containing 100 µg/ml ampicillin and grown at 37 $^{O}$C overnight with shaking at 150 rpm (211DS Shaking Incubator, Labnet International Inc.). Cultures were then screened for the presence of an insert by colony PCR.

### 2.2.8.4 Colony PCR

Following transformation into *E. coli*, the resulting bacterial colonies were screened by PCR for the correct recombinant vector using the universal primers T7 and SP6 to amplify the insert. Single colonies were picked from the plate and re-suspended in 5 µl of sterile water to which 2 µl each of 10 µM of T7 (2 µl) and SP6 (2 µl) were added. The mixture was then added to 12.5 µl of 1X Go Taq$^{R}$Green master mix (Promega) and the final volume was made to 25 µl with free RNase water. Amplification was carried out in a TECHNE thermocycler (Straffordshire UK, Model FTGRAD2D) as follows: initial cycle (95°C for 5 min), 34 amplification cycles (95 °C for 1 min, 55 °C for 1 min, 68 °C for 1 min) and final extension (68 °C for 10 min) with a holding temperature of 15 °C. PCR product (5 µl) was analysed by electrophoresis in 1.5 % agarose/ 1X TBE gel for 1 hour at 75 volts. 100 bp DNA ladder (350 ng) was used as a marker and the gels were visualised under UV light with the help of 10 µg/ml ethidium bromide stain present within the gel.

### 2.2.8.5 Plasmid extraction

The overnight culture was diluted 1/500 into 5 ml LB medium with 100 µg/ml amplicillin and were incubated at 37 ˚C overnight with shaking (~300 rpm (211DS Shaking Incubator, Labnet International Inc.)). Bacterial cells were harvested by centrifugation at 6000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 15 mins at 4˚C. The plasmid was extracted using Qiagen Plasmid mini kit, following manufacturer's protocol. The cells were resuspended in 300 µl of P1 buffer and P2 buffer (300 µl) was added to the solution and mixed thoroughly by inverting the tubes 4-6 times until it turned blue. The solution was incubated at room temperature for 5 mins. Pre-chilled P3 buffer (300 µl) was added to the solution and was mixed thoroughly by inverting the tubes 4-6 times until it was colourless. The solution was incubated on ice for 5 mins and centrifuged at 15000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 10 mins at 4˚C. The supernatant was transferred to a binding column and allowed to enter the resin by gravity flow. 800 µl of QF buffer was added to the binding column and DNA was eluted into a fresh 2 ml collection tube. The eluted DNA was precipitated by adding room-temperature isopropanol (560 µl) and

centrifuged at 15,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 30 mins at 4˚C. The DNA pellet was washed with 1 ml of room-temperature 70% ethanol and centrifuged at 15,000 rpm (Eppendorf 5417 R Centrifuge, Eppendorf) for 10 mins. The supernatant was discarded and the DNA pellet was air-dried for 5-10 mins and re-suspended in 30 µl of TE buffer. DNA was crudely checked for concentration and purity using GE Nanovue spectrophotometer (GE Biosciences, New Jersey, DE, USA) and 5 µl of the sample was run on 1.5% TBE/agarose-gel electrophoresis containing 10 ng/ml ethidium bromide at 75V for 1.5 hours. The extracted plasmid DNA was then sent to Waikato DNA Sequencing Facility, University of Waikato, Hamilton.

### 2.2.9 Bioinformatics analysis

DNA sequence was analysed by using GeneiousPro ver 5.5 (Drummond et al., 2011). The sequence was edited by removing vector sequence and manually checked for ambiguous nucleotides. The cloned sequence was aligned with a DsMV reference sequence (Accession no. NC_003537). The primer sequences were identified within the cloned sequence by aligning the NIb2F and NIb3R primer sequences with the cloned sequence.

### 2.2.10 DsMV specific primer design

Primers of the DsMV specific sequences were designed for the regions where the universal degenerate primers were expected to bind. Universal HC-Pro and CI primers were aligned with the DsMV reference genome sequence (Accession no. NC_003537) using the Geneious multiple alignment tool in GeneiousPro ver 5.5 (Drummond et al., 2011). The DsMV sequence where these universal primers were found to align were used as the DsMV specific primers. The likelihood of these primers binding templates other than DsMV was estimated by BLASTtn searching of the nucleotide collection at NCBI (www.ncbi.nlm.nih.gov).

## 2.3 Results

### 2.3.1 RNA extraction

In order to test the effectiveness of universal potyvirus primers for detection of DsMV, total RNA was needed from DsMV infected tissue to be used as a template for cDNA RT-PCR. Therefore, total RNA was extracted from taro leaves infected with different strains of DsMV, namely, DsMV-NZ1, DsMV-NZ1.1 and DsMV-B1.2. An example agarose gel is shown in figure 2.4. The 18S and 28S rRNA bands from each plant were obvious with a background smear indicating the presence of mRNA in each sample. The results confirmed that the RNA from DsMV infected tissue was of good quality and could be used as a template for cDNA synthesis and/or one-step RT-PCR.

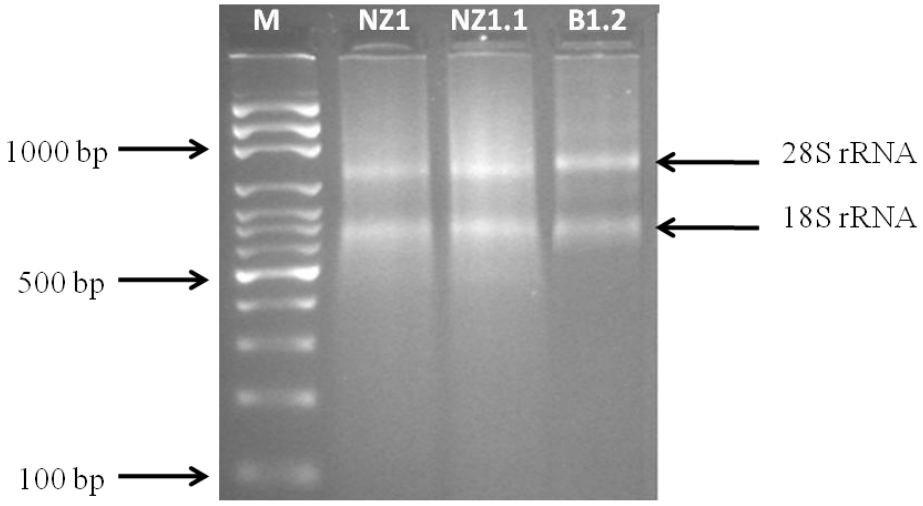


Figure 2.4: Agarose gel electrophoresis of total RNA extracted from taro leaves infected with different strains of DsMV. From left to right are: Lane M: 350 ng of 100 bp DNA ladder; and RNA extracted from leaves infected with DsMV-NZ1, DsMV-NZ1.1 and DsMV-B1.2, respectively. The 28S and 18S rRNA bands are indicated on the right while molecular size markers are indicated on the left.

### 2.3.2 Testing NIb2F and NIb3R primers for amplification of DsMV

Two-step RT-PCR amplification of the NIb region within the DsMV genome was carried out with the NIb2F and NIb3R primers using cycling conditions described by Zheng et al. (2010). cDNA synthesis reaction was primed with the PV1SP6 primer (Mackenzie et al., 1998) using the RNA extracted in 2.3.1 as template. Zheng et al (2010) reported that the NIb primer pair should amplify a 350 bp product. Figure 2.5

shows that multiple bands were amplified from the taro leaves infected with DsMV-NZ1.1 and DsMV-NZ1 using an annealing temperature of 45 ˚C as described by Zheng et al. (2010). The NIb primers produced non-specific products of three different sizes i.e. 100 bp, 200 bp and ~300 bp (figure 2.5) as well as a background smear for both templates. The no template control (NTC) reaction did not show any products, indicating that there was no contamination. Therefore, it was speculated that the difference between the observed and the expected result could be attributed to the different experimental conditions that were used for this experiment. Zheng et al. (2010) used these primers in a one-step RT-PCR using Superscript III Platinum Taq (Invitrogen). However, in this study a two-step RT-PCR was carried out using Zheng's cycling conditions. Therefore, it was concluded that PCR cycling conditions needed to be optimised to amplify the right product.

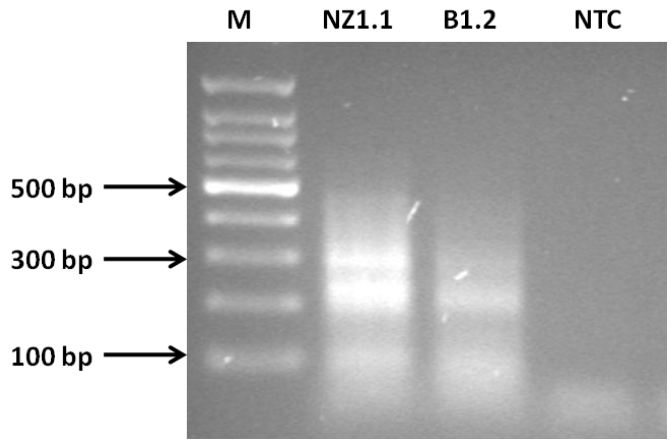


Figure 2.5: Two-step RT-PCR with NIb2F and NIb3R primers. From left to right: Lane M: 350 ng of 100 bp DNA ladder, RT-PCR of DsMV-NZ1.1 and NZ1 and No template control (NTC). Selected markers are indicated on the left.

## 2.3.2.1 Optimisation of PCR conditions for the NIb primers

### 2.3.2.1.1 Optimisation of annealing temperature by using touchdown PCR (45˚C-55˚C)

Touchdown PCR was carried out to determine if the NIb primers could amplify a product. The melting temperature (Tm) of the NIb primers was 40 ˚C (Zheng et al., 2010b). However, due to amplification of non-specific products by the NIb2F and NIb3R primer pair, it was possible that the primer annealing temperature of 45˚C was not optimal. Therefore, touchdown PCR was carried out with an initial annealing

temperature range of 45 ˚C to 55 ˚C. Figure 2.6 shows very faint bands at 100 bp and 300 bp for DsMV-NZ1.1 and DsMV-NZ1. No contamination was observed as the NTC amplification reaction did not produce any product. Thus, it appeared that 45 ˚C – 55 ˚C was not a suitable range for primer annealing under these experimental conditions.

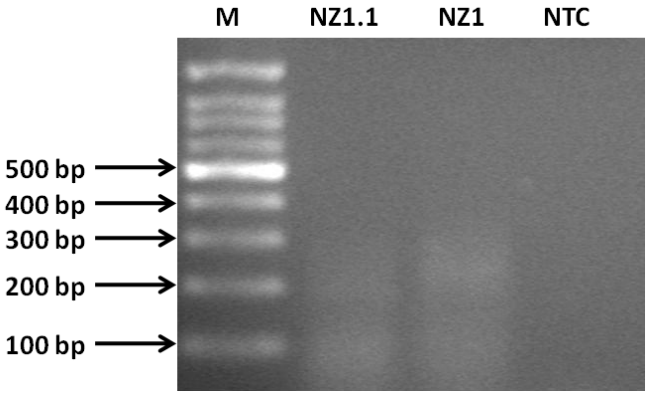


Figure 2.6: Touchdown PCR at 45-55˚C using NIb2F and NIb3R. From left to right: Lane M: 100 bp DNA marker (350 ng); RT-PCR of DsMV-NZ1.1 and DsMV-NZ1, respectively; and Negative control. Selected markers are indicated on the left

#### 2.3.2.1.2 Optimisation of primer concentration

The NIb primers tested in this study were degenerate and one major problem with degenerate primers is that the concentration of some permutations in the mixture is so small, due to their high multiplicity, that amplification is effectively inhibited. Therefore, for any target sequence only a proportion of the primers participate in the amplification reaction (Maher-Sturgess et al., 2008). Since no specific amplification was observed using Touchdown PCR, it was possible that the primer concentration was sub-optimal. Therefore, RT-PCR with NIb2F and NIb3R primers were performed using three different primer concentrations in the amplification reaction i.e. 0.5 µM, 0.8 µM and 1.2 µM final concentration of each primer.

Figure 2.7 shows that the expected 350 bp product was not amplified at any primer concentration tested. Two bands of around 100 bp and 200 bp were seen for both DsMV-NZ1 and DsMV-NZ1.1 at all the concentrations tested. In addition, a smear ranging up to 500 bp was seen for NZ 1.1. There was no contamination in the amplification reactions since the NTC's did not have a product. Occurrence of bands in the NTC reactions indicated the formation of primer dimers. While the expected product

was not obtained from any reaction, the maximum intensity of product amplification was seen when each primer concentration was 0.8 µM, therefore, this concentration was used for further amplifications.

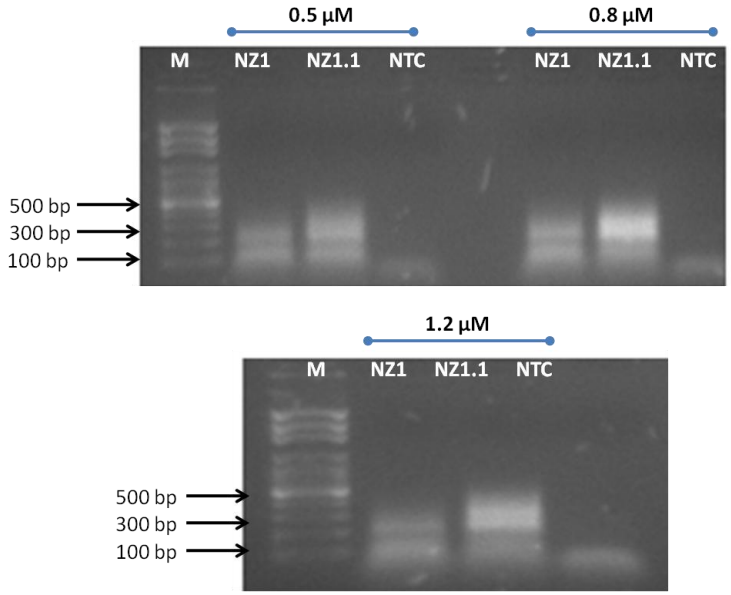


Figure 2.7: Two-step RT-PCR of DsMV-NZ1 and DsMV-NZ1.1 using three different primer concentrations indicated above the relevant lanes. From left to right: Lane M: 350 ng of 100 bp DNA ladder; and No Template Control (NTC). Selected markers are indicated on the left.

#### 2.3.2.1.3 Optimisation of annealing temperature using Touchdown PCR (40˚C - 50˚C)

After optimising primer concentration another Touchdown RT-PCR was carried out at an annealing temperature ranging from 40 ˚C to 50 ˚C. This was done to determine if there was a suitable annealing temperature within the range tested. Figure 2.8 shows that the expected 350 bp product was not amplified. A smear ranging from 100 bp up to 500 bp was observed again for NZ 1.1 within which bands of around 100 bp, ~220 bp and 300 bp were observed. A similar pattern was seen for NZ 1 i.e. multiple bands of around 100 bp, 200 bp and 300 bp were observed. The negative control did not produce a product suggesting that there was no contamination in the reactions. It was concluded that this temperature range under these reaction conditions was also not suitable for amplification.

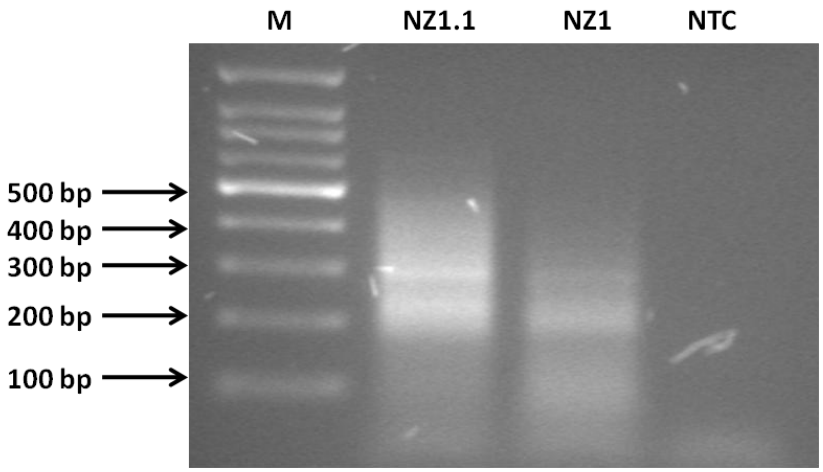


Figure 2.8: Touchdown PCR from 40˚C-50˚C with NIb2F and NIb3R primers. From left to right: Lane M: 100 bp DNA marker (350 ng); RT-PCR of NZ1.1, and NZ1, respectively; and No template control (NTC). Selected markers are indicated on left.

#### 2.3.2.1.4 Optimisation of the annealing temperature of the NIb primers

Gradient RT-PCR was carried out to determine if an optimal annealing temperature for NIb primers could be identified. Based on published literature for the NIb primers, the temperature range of 35˚C-55˚C was selected for gradient PCR. Five different annealing temperatures were used for PCR, which is 35˚C, 40˚C, 45˚C, 50˚C and 55˚C. Figure 2.9 shows that the expected 350 bp fragment was not amplified at any of these annealing temperatures. A smeary pattern was seen at all temperatures with bands at around 200 bp-300 bp. The MJ1/MJ2 primer pair which was known to amplify a 327 bp fragment from coat protein (CP) region of DsMV genome (Babu, Hegde, Makeshkumar, & Jeeva, 2010) was used as a positive control. These primers amplified the expected 327 bp product (figure 2.9). The amplification of the 327 bp product confirmed the presence of DsMV in these samples. There was no contamination in the reactions, as all the negative controls had no products. It is possible that the cDNA primed with PV1/SP6 may not have been optimal for amplification, perhaps being too short and not covering the NIb primer binding regions.

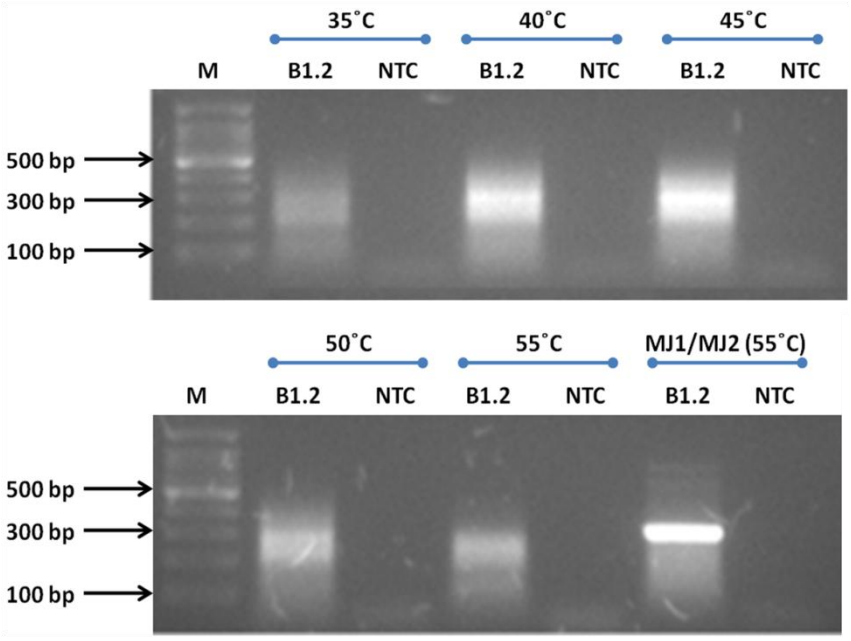


Figure 2.9: Gradient PCR with NIb2F and NIb3R primers using five different annealing temperatures (i.e. 35˚C, 40˚C, 45˚C, 50˚C and 55˚C) and MJ1/MJ2 as positive control indicated above the relevant lanes. From left to right: Lane M: 350 ng of 100 bp DNA ladder; and No template control (NTC). Selectable markers are indicated on the left

#### 2.3.2.1.5 Optimisation of reaction for first strand cDNA synthesis

In all the experiments discussed so far, the cDNA synthesis reaction was primed with PV1SP6 using a qScript Flex cDNA synthesis kit (Quanta Biosciences) following the manufacturer's instructions. This primer is an anchored oligodT primer, which would bind to the very 3'end of the Potyviral genome. However, due to the inability of the NIb primer pair in amplifying the expected sized product, even after series of optimisation steps, it was speculated that either the cDNA synthesis reaction was not efficient or the cDNA was too short in length. In order to ensure cDNA of sufficient length, cDNA synthesis was primed by a combination of PV1SP and random hexamers. Figure 2.10 shows that the 350 bp expected product was amplified by the NIb2F and NIb3R primers under those conditions. There was no contamination in the reaction as no product was observed in the NTC. This PCR product was purified, cloned and sequenced to determine if NIb primers amplified the expected sequence from the NIb gene of the DsMV genome.

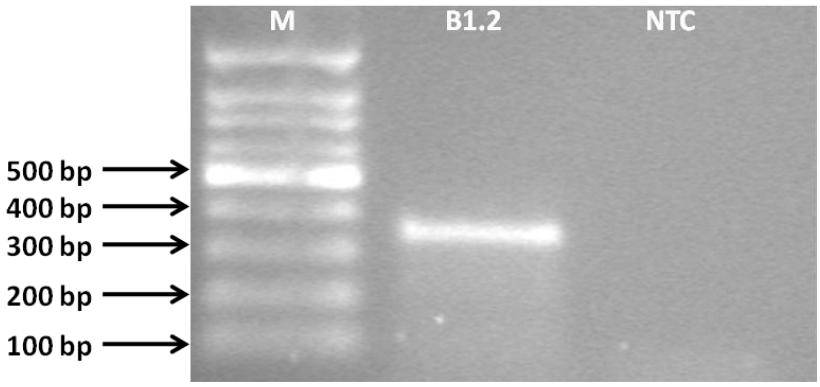


Figure 2.10: RT-PCR using NIb2F and NIb3R primers. From left to right: Lane M: 100 bp DNA ladder (350 ng); RT-PCR of DsMV-B1.2 and No template control (NTC). Selectable makers are indicated on the left.

#### 2.3.2.1.6 Sequence confirmation of PCR product

The product amplified by the NIb primers was purified and ligated into pGEM-T Easy and transformed into *Escherichia coli* JM109 cells. Nine white colonies were tested by colony PCR to confirm the presence of the insert. Figure 2.11 shows that out of the nine colonies only two had the insert (Colony 7 and 8). Plasmid DNA from the seventh colony was purified and sequenced. The alignment of the cloned sequence against the DsMV reference sequence (NCBI Accession no: NC_003537) showed that NIb primers amplified the expected 346 bp fragment from the NIb gene coding region of DsMV (Figure 2.12).

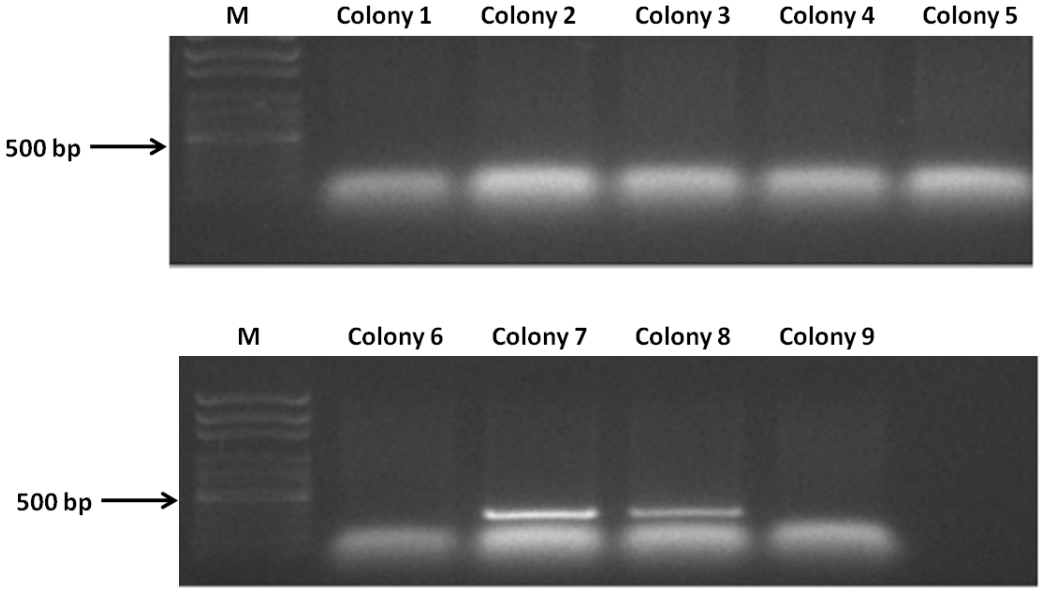


Figure 2.11: Colony PCR of NIb colonies. Lane M: 100 bp DNA ladder (350 ng); lanes 2-6: colonies 1-5, respectively; lanes 8-11: colonies 6-9, respectively; and lane 12: NTC

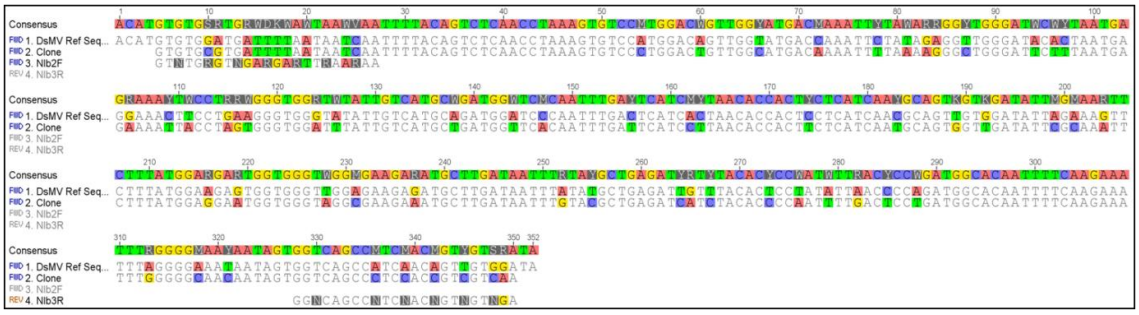

Figure 2.12: Alignment of the cloned sequence ("Clone") and NIb region of DsMV genome ("DsMV Ref Seq") and the NIb primers ("NIb2F" and "NIb3R")

### 2.3.3 Testing HPFor/HPRev and CIFor/CIRev primers for amplification of DsMV

Two-step RT-PCR amplification of the HC-Pro and CI regions within the DsMV genome was carried out with the HPFor/HPRev and CIFor/CIRev primer pairs, respectively, using the cycling conditions published by Ha et al. (2008). First strand cDNA synthesis was primed by a combination PV1SP6 and random primers (qScript Flex cDNA synthesis kit by Quanta Biosciences), followed by PCR with HPFor/HPRev and CIFor/CIRev primers respectively. According to Ha et al, the HP and CI primer pairs were both expected to amplify a ~700 bp product. However, figure 2.13 shows that no amplification of the expected ~700 bp product was observed for either the HP or CI primers. For the HPFor-HPRev primers a faint band of approximately 150 bp was observed. For the CIFor-CIRev primers, a smear was seen between 150 bp to 600 bp with no discrete bands observed. NTCs for both primer pairs showed that there was no contamination in any reaction. No amplification of the expected products suggested that experimental conditions used was not optimal, therefore conditions needed to be optimised to amplify the right product.

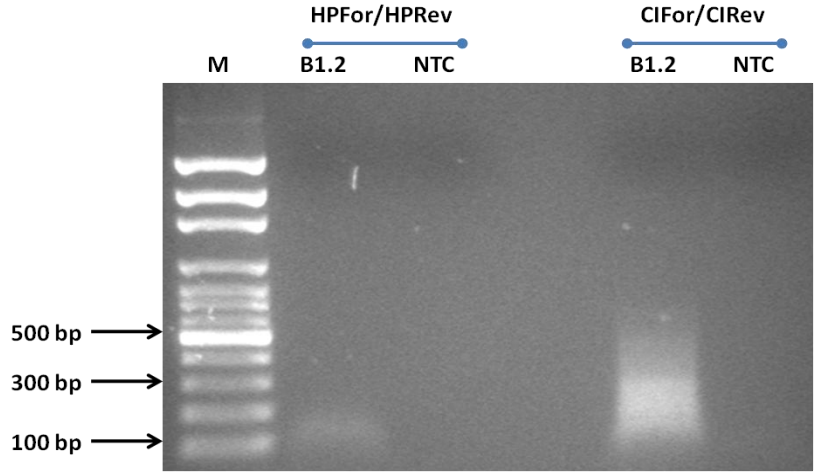


Figure 2.13: RT-PCR of DsMV using HPFor/HPRev and CIFor/CIRev primers. Lane M: 350 ng of 100 bp DNA marker. Selected markers are indicated on the left.

### 2.3.3.1 Testing various parameters to optimise two-step RT-PCR

A series of parameters were altered in attempts to optimise the amplification of the expected 700 bp product using the HC-Pro and CI primer pairs. These parameters and outcomes are listed in table 2.4, but none gave rise to the expected product. No positive control was included as the annealing temperature for the MJ1/MJ2 primers is 55˚C. To overcome the possibility that the amount of cDNA covering the region to be amplified was limiting, the amount of cDNA included in the PCR was increased. Also, priming cDNA synthesis with the reverse PCR primers, HPRev and CIRev, was done to ensure cDNA covering the regions of interest would be present in the PCRs. These primers have a low Tm of 40 ˚C, so the temperature for cDNA synthesis was reduced to 42 ˚C to ensure the primer could anneal to the RNA template. Also, because these primers are degenerate, the primer was increased in an attempt to ensure there was sufficient DsMV specific primer to anneal. None of these parameters gave rise to the expected sized products (figure 2.14). Indeed, very similar results were seen as for the initial amplification results described above.

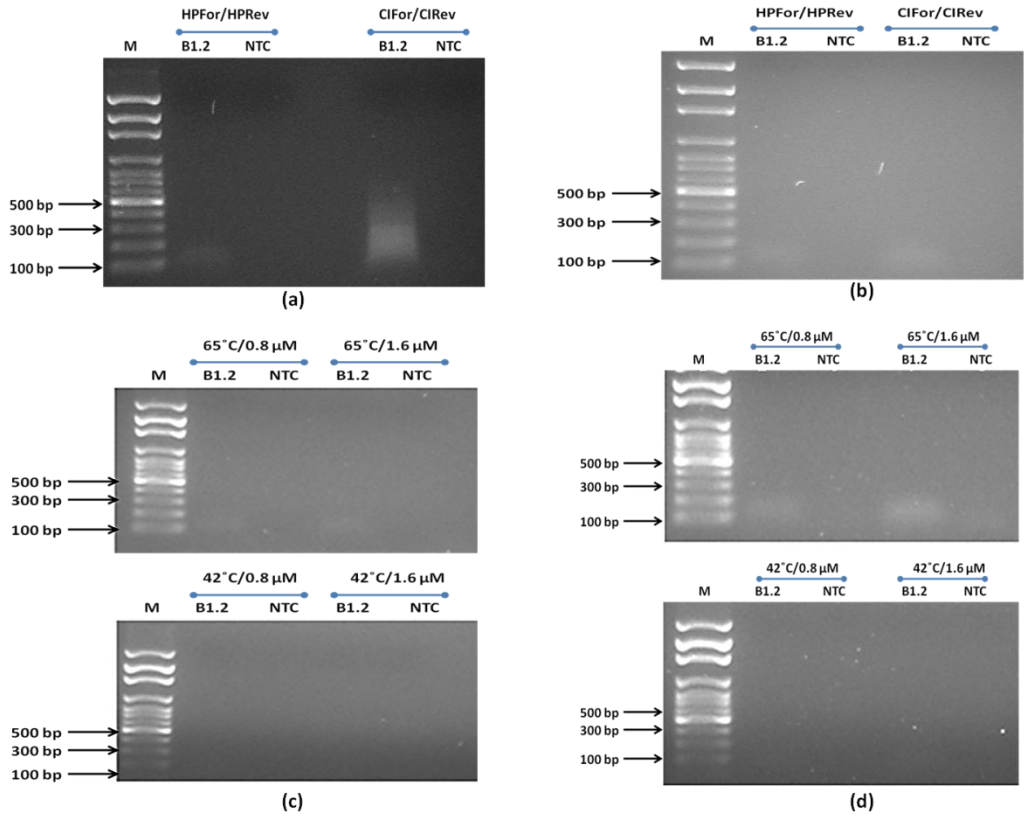


Figure 2.14: RT-PCR of DsMV using HPFor/HPRev and CIFor/CIRev using different parameters, where (a) represent RT-PCR using HC-Pro and CI primers with higher cDNA concentration; (b) represent RT-PCR with HC-Pro and CI primers using reverse primers for cDNA synthesis; (c) and (d) represent RT-PCR using HC-Pro and CI primers, respectively, with cDNA synthesised at 65˚C and 42˚C having different primer concentrations i.e. 0.8 µM and 1.6 µM. Lane M: 100 bp DNA ladder (350 ng). Selected markers are indicated on the left.

### 2.3.4 Designing and testing DsMV specific primers for HP and CI regions

After varying several RT-PCR parameters it was evident that the degenerate HP and CI primers published by Ha et al. (2008) could not amplify the expected products from DsMV under the experimental conditions used. Figure 2.15 and figure 2.16 shows an alignment of these degenerate primers with the DsMV reference genome (Accession no: NC_003537). This suggested that the primers should have been able to anneal to the DsMV template, yet no amplification of the expected products was observed for either primer pair. The reason for this result was unclear at this point. However, to ensure these regions could be amplified, new primers were designed that were the DsMV specific sequences rather than degenerate sequences. These primers are shown in Table 2.3.

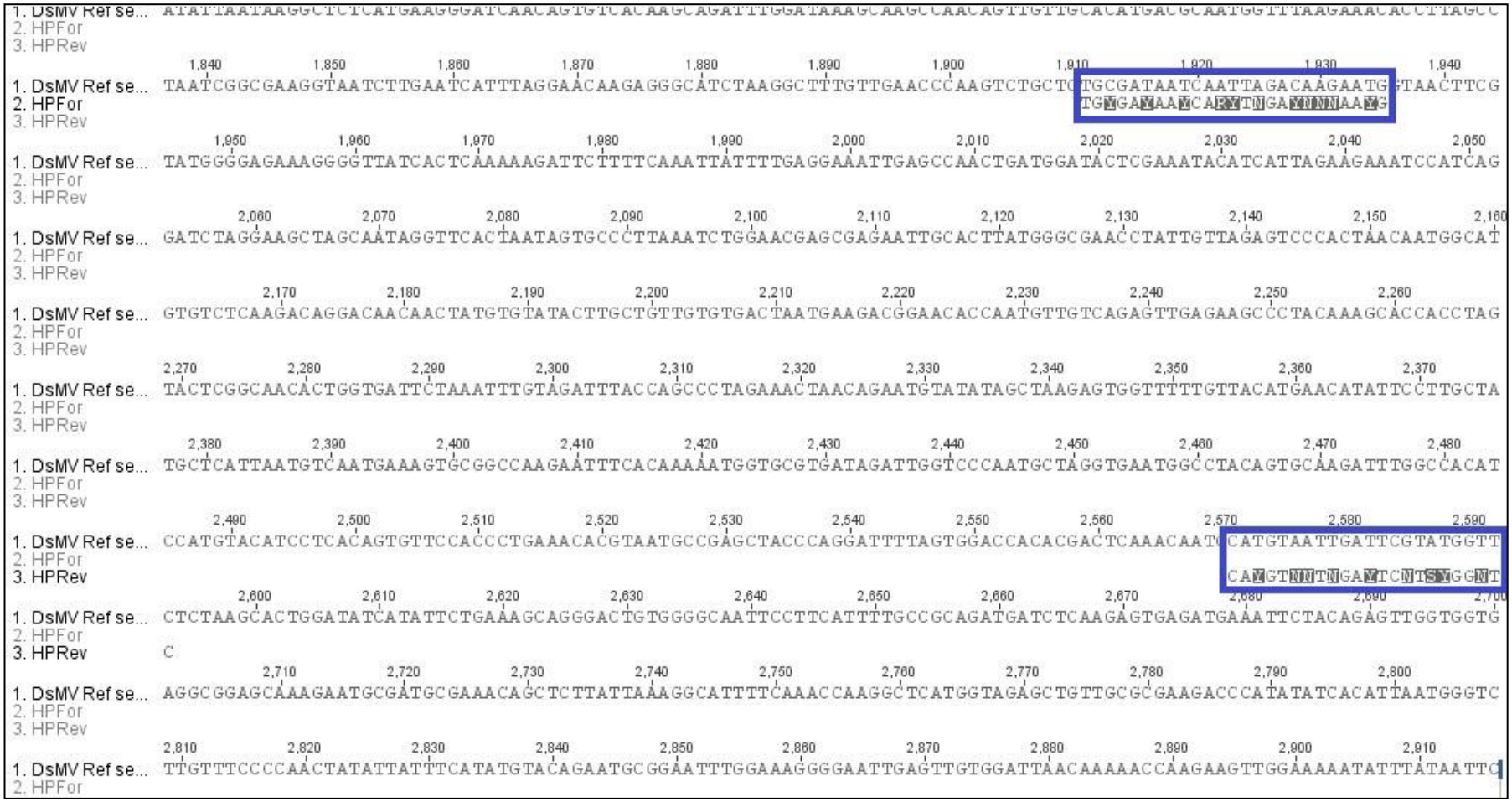


Figure 2.15: Alignment of universal HPFor/HPRev primers with the DsMV reference sequence. Box indicates the primer binding position.

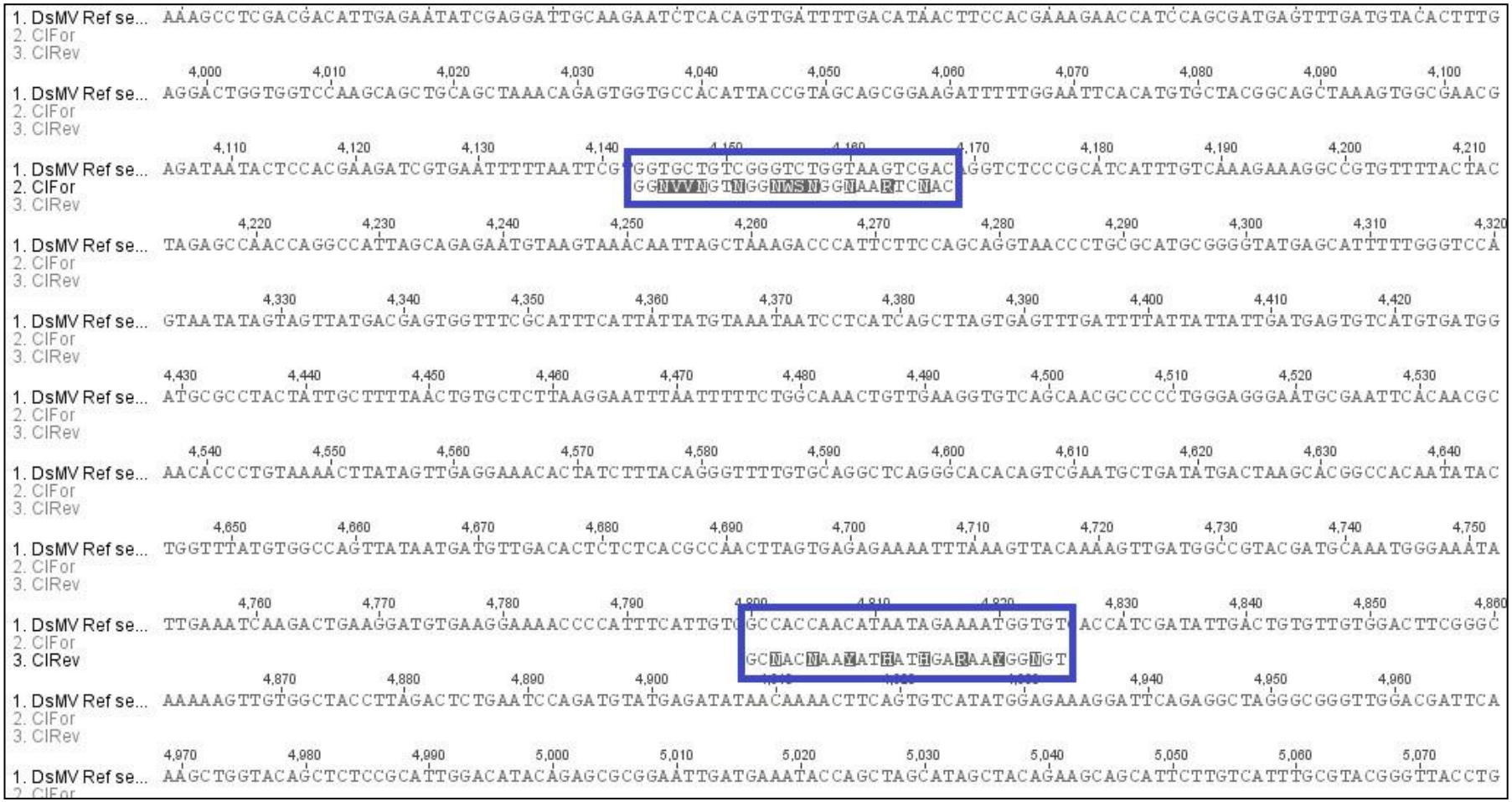


Figure 2.16: Alignment of universal CIFor/CIRev primers with the DsMV reference sequence. Box indicates the primer binding position

Table 2.3: DsMV specific primers designed for amplifying HC-Pro and CI regions of the DsMV genome

| Primer | Sequence (5'-3') | $T_m$ (°C) | Length (bp) | Product size (bp) |
|---|---|---|---|---|
| DsMV-HPFor | TGCGATAATCAATTAGACAAGAATG | 52.6 | 25 | 683 |
| DsMV-HPRev | GAACCATACGAATCAATTACATG | 51.2 | 23 | |
| DsMV-CIFor | TTAATTCGTGGTGCTGTCGG | 56.6 | 20 | 670 |
| DsMV-CIRev | ACACCATTTTCTATTATGTTGGTGGC | 56.9 | 26 | |

### 2.3.4.1 Testing new HC-Pro and CI primers on DsMV

To test the effectiveness of DsMV specific HPFor/ HPRev and CIFor/CIRev in amplifying DsMV, RT-PCR was carried out using new HP and CI primers. cDNA synthesis was primed with either the DsMV specific reverse HP and CI primers or with the VdT primer. The PCR was performed on both cDNAs using the cycling conditions of Ha et al. (2008). Figure 2.17 shows that the HP primers gave the expected 700 bp together with products of 100 bp and 200 bp from cDNA primed the HPRev primer. In contrast, PCR done with cDNA primed with VdT amplified a non-specific product of ~150 bp. PCR using cDNA primed with the CIRev primer was not able to amplify the expected product, but rather a single 200 bp product was amplified. In contrast, PCR with VdT primed cDNA amplified a 100 bp to 500 bp smear. The NTC for the CI primers showed multiple bands of different sizes. The pattern of bands seen in NTC after PCR was different for cDNA made with the CIRev and VdT primers suggesting either the CIRev primers were contaminated or they were binding with each other to produce false bands. Using fresh primer preparations gave the same results, suggesting primer interactions have occurred. The inability of new HC-Pro and CI primers in amplifying just the expected product under the experimental conditions used, suggested that the RT-PCR conditions were not optimal.

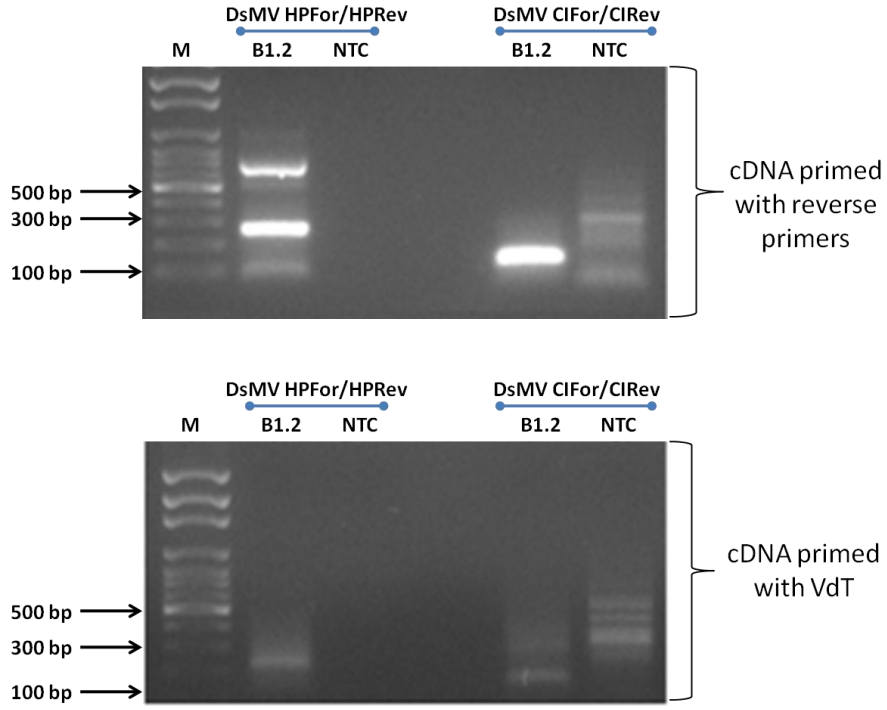


Figure 2.17: RT-PCR using DsMV specific HC-Pro and CI primers. From left to right: Lane M: 350 ng of 100 bp DNA ladder; RT-PCR using DsMV-B1.2; No template control (NTC). Selected markers are indicated on the left

### 2.3.4.2 Gradient RT-PCR using DsMV specific HC-Pro and CI primers

RT-PCR with DsMV specific HP and CI primers using the cycling conditions described by Ha et al. (2008) did not amplify the expected product. Therefore, gradient RT-PCR was carried out for both primer pairs to determine the optimum annealing temperature for amplifying the expected ~ 700 bp product. For both HP and CI primers, cDNA synthesis was primed with HPRev and CIRev primers, respectively, which was used as the templates for gradient PCR. Figure 2.18 shows the gradient RT-PCR result for the DsMV specific HPFor and HPRev primers carried out from 50˚C-65˚C. The same, Multiple products seen in figure 2.17 were observed at all temperatures except at 53.7˚C (this experiment was not repeated due to limited funding). A single discrete band of approximately 850 bp was amplified at 53.7˚C. This false product could be the result of mis-priming during the amplification reaction. Between 50˚C-60˚C, three discrete bands of 100 bp, ~250 bp and ~700 bp were amplified. Out of the three products amplified, only the ~700 bp product was of the expected size and the other two bands were likely the result of non-specific amplification. The intensity of 250 bp fragment was significantly higher than the expected fragment and an increase in the annealing temperature had no significant effect on the intensity of amplification. This suggested

that the primers were binding at several positions on the template and the probability of non-specific primer binding was significantly higher than the probability of specific primer binding under these reaction conditions. Also, between 63˚C-65˚C, the efficiency of the amplification reaction decreased significantly.

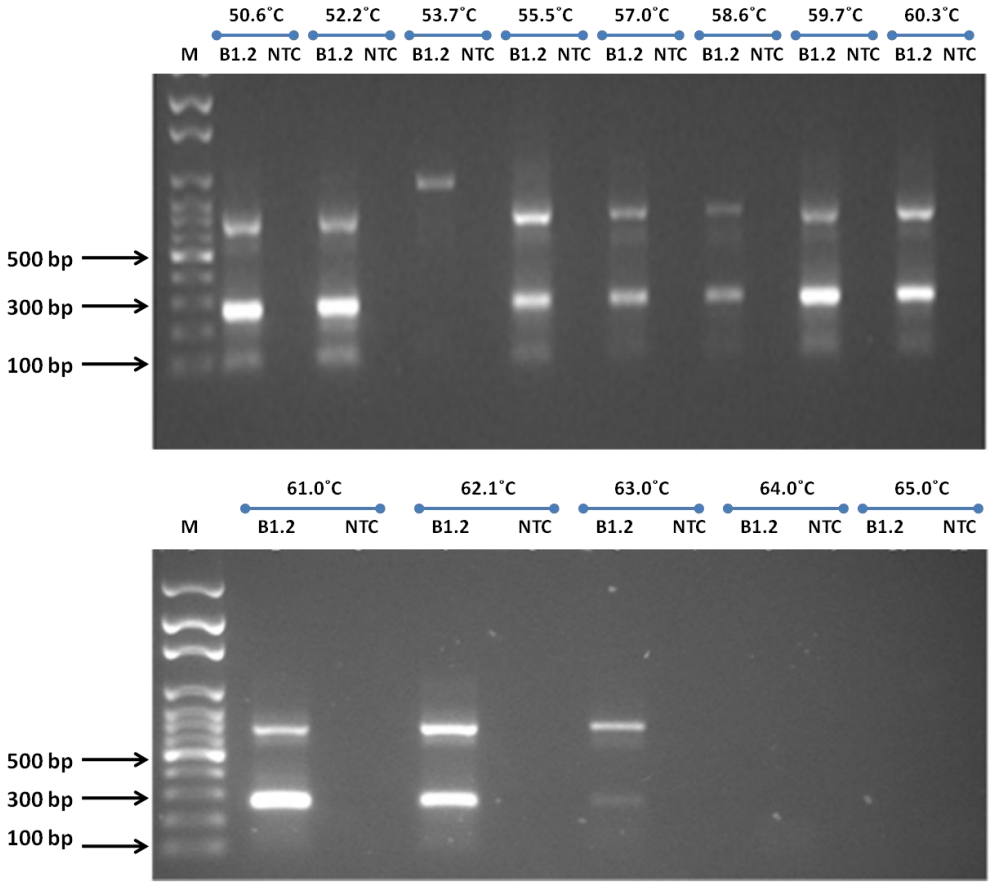


Figure 2.18: RT-PCR of DsMV-B1.2 using DsMV specific HPFor and HPRev primers compared with no template controls (NTC) at annealing temperatures ranging from 50˚C to 65˚C. For each temperature, Lane M: 100 bp DNA ladder (350 ng). Selected markers are indicated on the left.

Figure 2.19 shows RT-PCR results for the CIFor and CIRev primers. For these primers, gradient PCR was carried out from 45 ˚C to 55 ˚C. The expected ~700 bp product was not amplified, instead a non-specific product of 150 bp was amplified at all annealing temperatures. Multiple bands were again observed in all the NTCs. These results suggest that the DsMV specific CI primers are not suitable for amplifying the expected 700 bp product under the conditions used. In contrast, the HC-Pro primers could amplify a product of the expected size; however, they also amplified non-specific products.

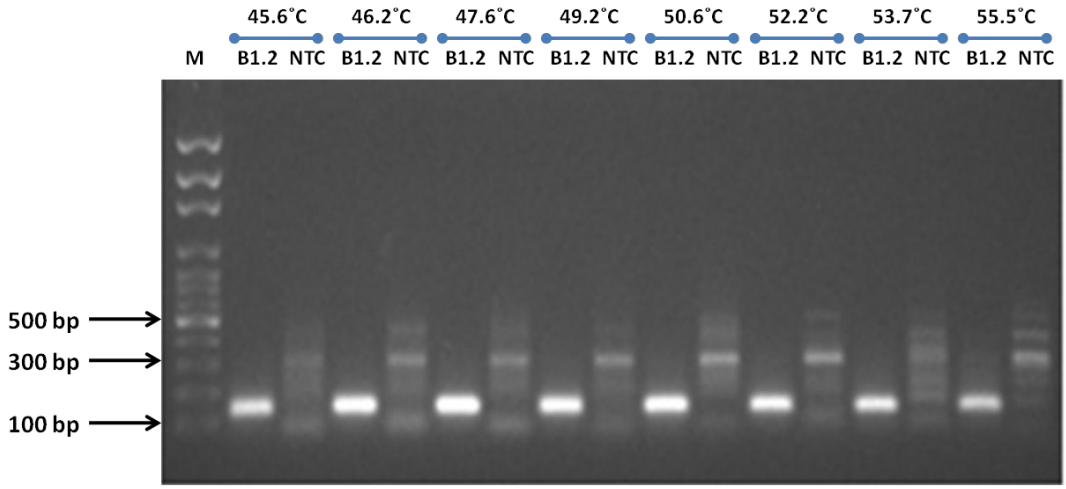


Figure 2.19: RT-PCR of DsMV-B1.2 using DsMV CIFor and CIRev primers compared with no template controls (NTC) using annealing temperatures ranging between 45˚C to 55˚C. For each temperature, Lane M: 100 bp DNA ladder (350 ng). Selected markers are indicated on the left.

## 2.3.5 Comparison of one-step RT-PCR and two-step RT-PCR

Amplification of DsMV with the universal HC-Pro and CI primers reported by Ha et al. (2008) as well as the DsMV specific primers for these regions designed here did not give the expected sized products when a two-step RT-PCR was done using a Quanta Biosciences cDNA synthesis kit followed by PCR using PCR Promega's Go Taq® Green Master Mix. It was noted that Ha et al. (2008) had used a Titan One-step RT-PCR kit for amplifying the HC-Pro and CI regions using the universal HC-Pro and CI primers. Therefore, it was decided to do a comparison of the two step RT-PCR with a one-step kit. For this purpose, SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life Technologies) was tested using both the universal and DsMV specific HC-Pro and CI primers.

### 2.3.5.1 One-step RT-PCR using new DsMV specific HC-Pro and CI primers

One-step RT-PCR was carried out on the DsMV-B1.2 isolate with SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life technologies) using cycling conditions given by Ha et al. (2008). This kit included SuperScript® for cDNA synthesis and the hot-start Platinum®*Taq* polymerase for PCR. Figure 2.20 shows that

no amplification was observed with the DsMV specific HC-Pro primers. In contrast, three faint bands of ~700 bp, 900 bp and 1000 bp were amplified by CI primers. Out of four bands, one was the expected ~700 bp while the other two were the products of non-specific amplification.

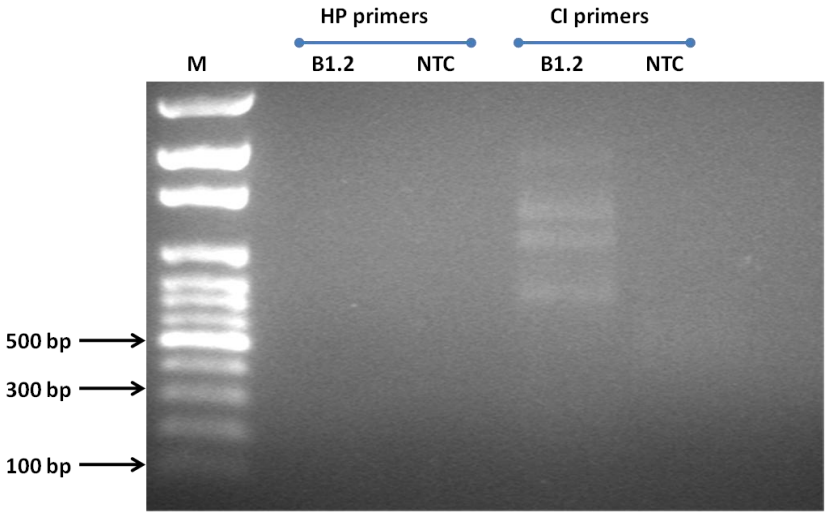


Figure 2.20: RT-PCR with the DsMV specific HP and CI primers using SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life technologies). Lane M: 100 bp DNA marker (350 ng). Selected markers indicated on the left.

### 2.3.5.2 Comparison between the DsMV specific HC-Pro and CI primers and the universal HP and CI primers

The DsMV specific HC-Pro and CI primers did not amplify the expected ~700 bp product using one-step PCR. Ha et al. (2008) had used the universal primers to amplify the expected product using one-step PCR. Therefore, it was decided to do a comparison between DsMV specific primers and universal primers using one-step RT-PCR at two primer concentrations, which is 0.4 µM and 1.6 µM. The reason for using these two primer concentrations was that according to SuperScript® III One-Step RT-PCR System (with Platinum®*Taq* (Life Technologies)), 0.4 µM final concentration of primers was sufficient for amplification, however Ha et al. (2008) used a final primer concentration of 1.6 µM for their experiment. Thus, it was decided to do a comparison between specific and universal primers at the two different primer concentrations. Figure 2.21(a) shows that no product was amplified using the DsMV specific HPFor/HPRev primers at any primer concentration. No amplification was observed using the universal primers at 0.4 µM; however, the expected product was amplified

using 1.6 µM primers. Due to the poor resolution of bands in the gel it appeared that the universal HP primers also amplified an additional non-specific product of ~800 bp; however, a subsequent repeat experiment proved that only a fragment of 700 bp was amplified (figure 2.22). Figure 2.21(b) shows the comparison between the DsMV specific CI primers and the universal CI primers. No product was amplified using the DsMV specific CI primers at either of the primer concentration. The expected ~700 bp product was amplified using the universal CI primers at both primer concentrations, but the efficiency of amplification reaction appeared to be higher at 1.6 µM of final primer concentration since more product was produced. Therefore, it was decided to use this primer concentration for further amplifications. However, similar to the HC-Pro primers, it appeared that the universal CI primer was amplifying an additional non-specific product of ~800 bp, but again this was due to the gel (figure 2.22).

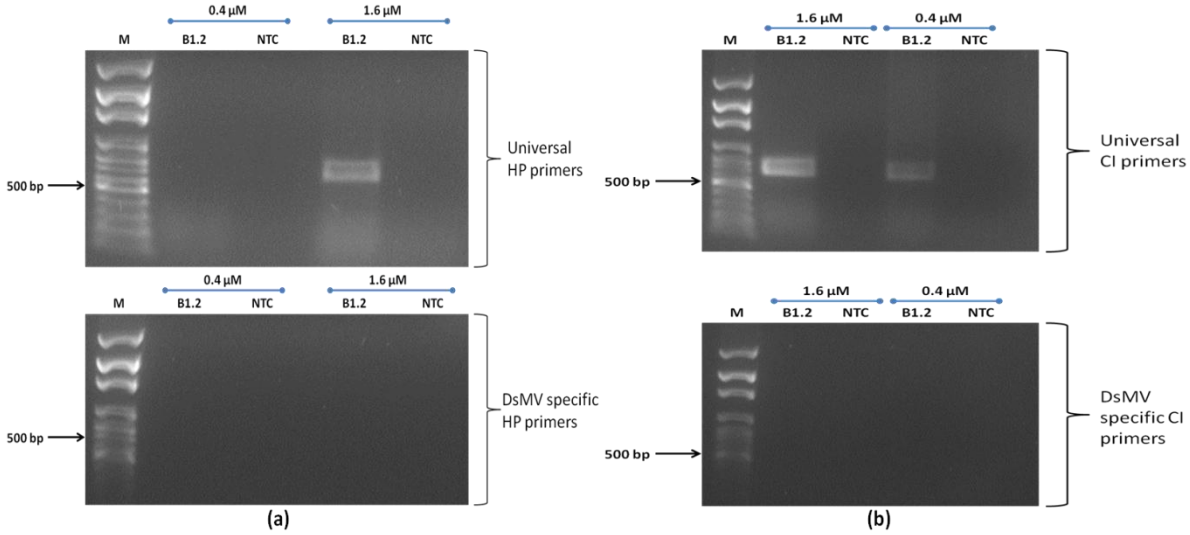


Figure 2.21 (a) :RT-PCR using universal and DsMV specific HC-Pro primers at two primer concentrations, i.e. 0.4 µM and 1.6 µM. Lane M: 100 bp DNA marker (350 ng). Selected markers are indicated on the left. Figure 2.21 (b): RT-PCR using universal and DsMV specific CI primers at two primer concentrations, i.e. 0.4 µM and 1.6 µM. Lane M: 100 bp DNA marker (350 ng). Selected markers are indicated on the left.

### 2.3.5.3 One-step RT-PCR using the universal HC-Pro and CI primers

Figure 2.21 showed that the expected ~700 bp product could be amplified from the HC-Pro and CI regions of the DsMV genome using the universal primers, at a final primer concentration of 1.6 µM. However, due to the poor resolution of the agarose gels, it appeared that HP and CI primers were also amplifying a non-specific product of ~800 bp. Therefore, a repeat of these reactions showed that only a 700 bp fragment was

generated (figure 2.22). Therefore, it was concluded that universal HP and CI primers can be used for detection of the DsMV infection of taro under the specific experimental conditions.

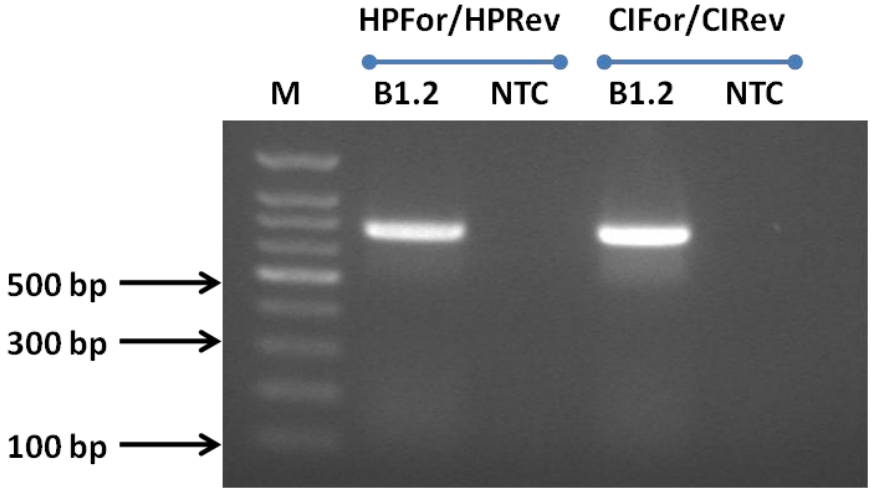


Figure 2.22: RT-PCR with universal HP and CI primers using SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life technologies). Lane M: 100 bp DNA marker (350 ng). Selected markers indicated on the left.

Table 2. 4: Experimental conditions tested on HC-Pro and CI primers using two-step RT-PCR

| Parameter tested | cDNA synthesis | | | PCR | | | | Amplification of expected product? |
|---|---|---|---|---|---|---|---|---|
| | RNA amount | Primer | Temp. | Forward primer (0.8 µM) | Reverse primer (0.8 µM) | cDNA amount | Conditions | |
| Conditions used for amplification with NIb primers | 400 ng | 0.8 µM PV1/SP6 and 0.8 µM random primer | 65˚C | HPFor | HPRev | 2.5 µl | Ha et al. | No |
| | | | | CIFor | CIRev | 2.5 µl | Ha et al. | No |
| Increasing amount of cDNA in PCR | 400 ng | 0.8 µM PV1/SP6 and 0.8 µM random primer | 65˚C | HPFor | 0.8 µM HPRev | 4.25 µl | Ha et al. | No |
| | | | | CIFor | 0.8 µM CIRev | 4.25 µl | Ha et al. | No |
| cDNA synthesis primed with reverse PCR primer | 400 ng | 0.8 µM HPRev | 65˚C | HPFor | 0.8 µM HPRev | 2.5 µl | Ha et al. | No |
| | 400 ng | 0.8 µM CIRev | 65˚C | CIFor | 0.8 µM CIRev | 2.5 µl | Ha et al. | No |
| Optimising temperature and primer concentration cDNA synthesis | 400 ng | 0.8 µM HPRev | 42˚C | HPFor | 0.8 µM HPRev | 2.5 µl | Ha et al. | No |
| | 400 ng | 1.6 µM HPRev | 65˚C | HPFor | 0.8 µM HPRev | 2.5 µl | Ha et al. | No |
| | 400 ng | 0.8 µM CIRev | 42˚C | CIFor | 0.8 µM CIRev | 2.5 µl | Ha et al. | No |
| | 400 ng | 1.6 µM CIRev | 65˚C | CIFor | 0.8 µM CIRev | 2.5 µl | Ha et al. | No |
| Testing DsMV-specific HC-Pro and CI primers | 400 ng | 0.8 µM DsMV-HPRev | 65˚C | DsMV-HPFor | DsMV-HPRev | 2.5 µl | Ha et al. | Yes+ other products |
| | 400 ng | 0.8 µM DsMV-HPRev | 65˚C | DsMV-HPFor | DsMV-HPRev | 2.5 µl | 50-65˚C | No |
| | 400 ng | 0.8 µM DsMV-CIRev | 65˚C | DsMV-CIFor | DsMV-CIRev | 2.5 µl | Ha et al. | No |
| | 400 ng | 0.8 µM DsMV-CIRev | 65˚C | DsMV-CIFor | DsMV-CIRev | 2.5 µl | 45-56˚C | No |
| | 400 ng | 0.8 µM VdT | 65˚C | DsMV-HPFor | DsMV-HPRev | 2.5 µl | Ha et al. | No |
| | 400 ng | 0.8 µM VdT | 65˚C | DsMV-CIFor | DsMV-CIRev | 2.5 µl | Ha et al. | No |

## 2.3.6 Summary of results

The aim of this study was to test the usefulness of universal potyvirus primers in detection and identification of the DsMV infection of taro (figure 2.23). The conditions used are summarised in table 2.4. The universal primers tested in this study were NIb2F/NIb3R, HPFor/HPRev and CIFor/CIRev. From the findings of this study, it was concluded that universal primers NIb2F/NIb3R, HPFor/HPRev and CIFor/CIRev can be used for amplifying the expected products from the NIb, HC-Pro and CI regions of the DsMV genome under specific experimental and cycling conditions. The experimental conditions that can be used for these primers in detection of the DsMV infection within taro are summarised in table 2.5.

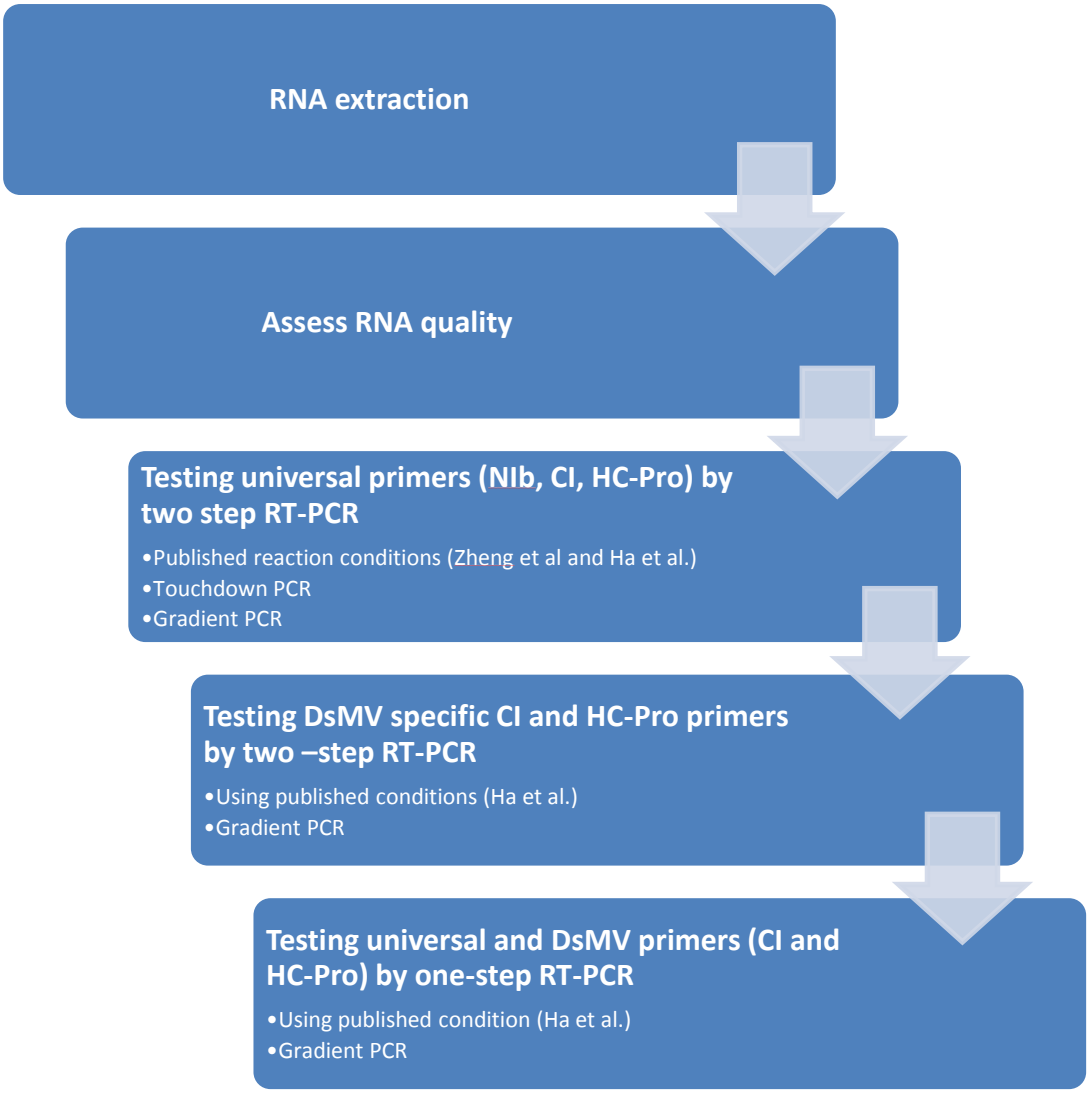


Figure 2. 23: Flow diagram showing the different stages of experiment used for testing potyvirus universal primers

Table 2.5: Recommendations for using tested universal primers for identification of the DsMV infection of taro

| DsMV gene region | RT-PCR system | cDNA primer | cDNA T˚C | PCR primers | Template and volume | Annealing T˚C | Cycling conditions | Expected Product |
|---|---|---|---|---|---|---|---|---|
| NIb | Two-step RT-PCR using Quanta Biosciences Flex cDNA synthesis kit and Promega's Go Taq$^{R}$ Green PCR master mix | PV1/SP6 and random primer (0.8 µM of each primer) | 65˚C | Universal NIb2F and NIb3R (0.8 µM of each primer) | cDNA of DsMV infected taro leaves (2.5 µl) | 45˚C | Zheng et al. (2010) | 350 bp |
| HC-Pro or CI | SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life technologies) | 1.6 µM HPRev or CIRev | 55˚C | Universal HP and CI primers (1.6 µM of each primer) | RNA extracted from DsMV infected taro leaves (2.5 µl) | 40˚C | Ha et al. (2008) | ~700 bp |

## 2.4 Discussion

The aim of this study was to test the effectiveness of universal potyvirus primers for the identification and detection of DsMV. This virus is known to infect a wide variety of cultivated aroids and ornamental plants worldwide (Babu et al., 2011) and belongs to one of the largest family of plant viruses named *Potyviridae* (Chen et al., 2001; McDonald et al., 2010). DsMV belongs to the Bean Common mosaic virus (BCMV) lineage (Gibbs, Trueman, & Gibbs, 2008) and infection may cause 40-60% reduction in total crop yield or around 40% reduction in the size of edible corms (Elliot et al., 1997). Precise and accurate identification of potyvirus infection of plants is essential to monitor the spread of disease and for developing better control mechanisms (Grisoni et al., 2006).

### 2.4.1 Universal primers for the detection of potyviruses

In this study, RT-PCR was used to test universal primers for identifying the DsMV in infected taro leaves. The reason for choosing this technique was that RT-PCR is a sensitive method of detecting plant RNA viruses (Jan et al., 2011). The usefulness of this technique has been significantly enhanced by the design and application of universal primers that broadens the specificity of the reaction (Antoniw, 1995). For this study three pairs of universal Potyvirus primers were selected for analysis; NIb2F/NIb3R, HPFor/HPRev and CIFor/CIRev primer pairs. Universal primers that target the CP (for example MJ1 and MJ2) have been shown to amplify DsMV (Babu et al., 2011), however primers targeting the NIb, HC-Pro and CI have not been tested extensively on DsMV. Based on the literature, NIb2F and NIb3R primers were expected to amplify a 350 bp fragment from the NIb gene coding region of the DsMV genome (Zheng et al., 2010b), while the HP and CI primer pairs were both expected to amplify a ~700 bp fragment (Ha et al., 2008).

### 2.4.2 Amplification of NIb region

Zheng et al. (2010) carried out one-step RT-PCR using Superscript$^{TM}$ III Platinum Taq (Life Technologies) with 0.5 µM NIb2F and NIb3R primers and were able to amplify a 350 bp product from NIb region of the DsMV genome. However, in the work presented here the NIb primers were tested on taro infected with different DsMV strains and isolates (DsMV-NZ 1, DsMV-NZ 1.1 and DsMV-B 1.2) using two-step RT-PCR based on cDNA synthesis using the Quanta Biosciences Flex cDNA synthesis kit and Promega's Go Taq$^{R}$Green PCR Master Mix. The cDNA synthesis was primed with PV1/SP6 primer which is known as a general potyvirus primer (Mackenzie et al., 1998). The reason for using a combination of PV1/SP6 primers for cDNA synthesis was because of its ability in synthesising a 1.7-2.1 kb fragment (Mackenzie et al., 1998) which covers the NIb gene, CP gene and 3' UTR of potyviruses. Therefore, it was assumed that cDNA synthesis primed with PV1/SP6 would produce cDNA of sufficient length to cover NIb region of the DsMV genome (Wei, Pearson, & Fletcher, 2006). However, the expected 350 bp product was not amplified even after a series of PCR optimisation steps. The results suggested that PV1/SP6 was not efficiently producing the cDNA or the cDNA synthesised was not of the right length for PCR primers to anneal. To solve this problem, random primers were used in combination with PV1/SP6 to facilitate the efficient synthesis of cDNA covering the NIb region, which led to amplification of the expected 350 bp product in the presence of the NIb2F and NIb3R primers. Based on the findings of this study, it was concluded that two-step RT-PCR with the NIb2F and NIb3R primers can be used for identification of DsMV infection of taro if the cDNA is synthesised using a combination of PV1/SP6 with random primers.

### 2.4.3 Amplification of the HC-Pro and CI regions

Based on the analysis of conserved regions within potyviruses, Ha et al. (2008) designed the HPFor/HPRev and CIFor/CIRev primer pairs. Since these primers anneal to reportedly conserved regions, it was expected that these primers would also amplify these regions of the DsMV genome. Ha et al. (2008) carried out one-step RT-PCR with a Titan One Tube RT-PCR system (Roche Applied Science) using 1.6 µM of final

primer concentration. They were able to amplify the expected ~700 bp product from the HC-Pro and CI regions of the potyvirus genome. However, in this study, two-step RT-PCR used with NIb primers was initially used to test the universal HC-Pro and CI primers, no amplification of the expected products for the HC-Pro and CI regions was observed. Several parameters were tested; including designing DsMV specific primers, no amplification was observed with the universal primers, and the DsMV specific CI primers while the expected product was amplified with the DsMV specific HC-Pro primers together with other products.

These non-specific products could not be eliminated by increasing the annealing temperature. Even though "in silico" analysis predicted that-these primers should not bind elsewhere in the DsMV genome or to any other sequence the probability of DsMV HC-Pro primers annealing to non-specific regions on DsMV or other templates was higher than expected. No product of the expected size was amplified for the CI region using the DsMV specific CI primers, which may be due to the decay in the consensus sequences for HC-Pro and CI primers or the experimental conditions that were used for this analysis.

Ha et al. (2008) had used a Titan One-step RT-PCR kit (Roche). Thus, a one-step RT-PCR kit was tested. For this purpose, the SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life Technologies) was tested using the universal and DsMV specific primers having a final primer concentration of 1.6 µM for each primer. While the expected ~700 bp product was not amplified by the DsMV HC-Pro and CI primers, it was amplified by the universal HC-Pro and CI primer pairs. The inability of DsMV specific HP and CI primers to amplify the expected product indicates either the decay in the consensus sequence of the DsMV reference sequence (Accession no: NC_003537), alterations in the template-to-product ratio, primer design, concentrations of template, enzymes, primers and nucleotides, proportion of reaction mixture, or their combination. In order to understand the high rate of consensus decay among potyviruses, Zheng et al. (2008) carried out an analysis of conserved regions identified within potyvirus genome over the past 20 year. They found that with the accumulation of variants with the increase in available sequences causes the decay of the consensus at the sites; therefore, the level of known conservation changes from time to time. The present study support the findings of Zheng et al. (2008). Therefore, the DsMV specific HC-Pro and CI primers cannot be used for identification of the DsMV infection under the experimental conditions tested. However, universal HC-Pro and CI primers can be

successfully used for detection and identification of the DsMV infection of taro. It was also concluded that, two-step RT-PCR cannot be used for detection of the DsMV infection of taro under the tested experimental conditions, whereas one-step RT-PCR can be.

The findings of this study showed that, NIb primers were efficient in detection and identification of DsMV under specific experimental conditions, that is using a two-step RT-PCR system based on cDNA synthesis primed by a combination of random primer and PV1/SP6 primer using the Quanta Biosciences Flex cDNA synthesis kit and Promega's Go Taq$^{R}$ Green PCR master mix, using the cycling conditions described by Zheng et al. (2010). It was also found that the universal HC-Pro and CI primers described by Ha et al (2008) can be used for identification of the DsMV infection of taro using experimental conditions as follows: One-step RT-PCR SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life Technologies) having a final concentration of 1.6 µM for both forward and reverse primers using the cycling conditions published by Ha et al. (2008). The inability of the DsMV specific HC-Pro and CI primers in amplifying the expected product suggested that identifying conserved regions and monitoring consensus decay at these conserved regions from time to time is important when designing primers for detecting both known and unknown viruses (Zheng et al., 2008).

# Chapter 3

# Genetic variability in the *Dasheen mosaic virus* genome

## 3.1 Introduction

Unlike DNA viruses, RNA viruses lack proofreading activity i.e. the ability to detect and repair mistakes during the replication process. As a result, mutation rates among RNA viruses can be as high as 1 mutation per each 1,000-100,000 bases copied per replication cycle (Astrovskaya et al., 2011; Malpica et al., 2002). This spontaneous mutation leads to the great genetic diversity observed among RNA viruses which provides the foundation for rapid genomic evolution (Vignuzzi et al., 2006). Many mutations are apparently well accepted and passed down to descendants, producing a family of co-existing related variants of the original viral genome referred to as quasispecies (i.e. population of sequences), a concept that is originally described as mutation-selection balance (Astrovskaya et al., 2011). Therefore, many variants of a pathogenic virus are present in an infected host plant (Domingo, 2002). The population of viral sequences (i.e. variants) vary from one infected host plant to another with a different set of sequences potentially dominating in each infected plant. Therefore, the severity of infection can be different for separate plants. As the virus continues to replicate the population of one variant may also change with time within an infected plant. Therefore, one infected plant may have two different sets of dominant sequences at two separate time points. If the sequences change over time, this may influence the direction of evolution (Eigen, 1996; Higgins et al., 1998).

The increase in mutation rate can facilitate the increase in the population of some variants and therefore those variants will become dominant among the populations of different variants. The diffuse 'cloud-like' nature of potyviral populations allows them to adapt rapidly to changing replicative environments by selecting pre-existing variants with better fitness. Therefore, several important virus properties cannot be explained by a single consensus sequence, but require knowledge about the microvariants present in viral populations. These sequence variants may be critically relevant for the efficiency of processes such viral translation and replication, host selection, viral evolution, spread and virulence (Barzon, Lavezzo, Militello, Toppo, & Palù, 2011; Domingo et al., 1996; Drake & Holland, 1999). Therefore, it is important to understand the pattern of genetic diversity of plant RNA viruses.

Most studies on viral genetic diversity have relied on techniques such as reverse transcriptase- polymerase chain reaction (RT-PCR) and sequencing of PCR products

using traditional dideoxy sequencing methods. However, this approach can only detect one or two variants among the viral population (Coetzee et al., 2010). In addition, PCR amplification can introduce errors depending on the polymerase used; variations may be misinterpreted as mutation or polymorphism which could decrease the accuracy of analysis. Due to these technical limitations it is not feasible to carry out an in-depth analysis of genetic variation using PCR (Barzon et al., 2011).

Next-generation sequencing (NGS) provides opportunities to determine the true extent of sequence variation. There are several reasons for choosing NGS, such as the ability to generate massive amounts of data for multiple samples in a single run, low per base sequencing cost, reduced dependence on other techniques such as cloning and PCR, and the ability to identify rare variant sequences among a population of sequences (Barzon et al., 2011; Jan et al., 2011). The throughput of next generation sequencers allows for more reads and thus more opportunities to correct innate errors in raw accuracy. Even a sequencing method of 90% accuracy can attain 99.999% consensus accuracy by providing five-fold coverage (Lander et al., 2001; Venter, Adams, Myers, & Li, 2001). The number of sequence reads generated with the Illumina platform ensures the genome coverage is very high (table 3.1), thus this approach was more suitable for this study (Wheeler et al., 2008). Single-nucleotide polymorphism (SNP) discovery undoubtedly benefits from the enormous data sets generated by Illumina sequencing technology and it has been shown that the Illumina platform can be successfully used for carrying out this kind of analysis (Hillier et al., 2008; Wheeler et al., 2008).

Table 3.1: Comparison of next generation sequencing (NGS) platforms (Metzker, 2010)

| Platform | Library/ template preparation | NGS chemistry | Read length (bases) | Run time (days) | Gb per run | Machine cost (US$) | Pros | Cons | Biological applications |
|---|---|---|---|---|---|---|---|---|---|
| Roche/454's GS FLX Titanium | Frag, MP/ emPCR | PS | 330* | 0.35 | 0.45 | 500,000 | Longer reads improve mapping in repetitive regions; fast run times | High reagent cost; high error rates in homo-polymer repeats | Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics |
| Illumina/ Solexa's $GA_{II}$ | Frag, MP/ solid-phase | RTs | 75 or 100 | 4‡, 9§ | 18‡, 35§ | 540,000 | Currently the most widely used platform in the field | Low multiplexing capability of samples | Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics |
| Life/APG's SOLiD 3 | Frag, MP/ emPCR | Cleavable probe SBL | 50 | 7‡, 14§ | 30‡, 50§ | 595,000 | Two-base encoding provides inherent error correction | Long run times | Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics |
| Polonator G.007 | MP only/ emPCR | Non-cleavable probe SBL | 26 | 5§ | 12§ | 170,000 | Least expensive platform; open source to adapt alternative NGS chemistries | Users are required to maintain and quality control reagents; shortest NGS read lengths | Bacterial genome resequencing for variant discovery |

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

NGS generates hundreds of megabases to gigabases of nucleotide-sequence output in a single instrument run, depending upon the platform. The Illumina sequencing platform is currently one of the more commonly used NGS platform. The principle behind Illumina sequencing is shown in figure 3.1. Illumina's Genome analyser uses a flow cell consisting of an optically transparent slide with eight individual lanes on the surface of which are bound oligonucleotide anchors (Illumina, 2012). The RNA is converted into cDNA using random primers, the cDNA is sheared to the appropriate size using a compressed air device known as a nebuliser. The ends of the fragmented DNA are polished and adapters are ligated to the fragments (figure 3.1a) (Voelkerding, Dames, & Durtschi, 2009). Ligated fragments of the size range 150-200 bp are isolated via gel electrophoresis and amplified using limited cycles of PCR. The Illumina platform utilises a 'bridged' amplification reaction that occurs on the surface of the flow cell. The flow cell surface is coated with single stranded oligonucleotides that correspond to the sequences of the adapters ligated during the sample preparation stage. Single-stranded, adapter-ligated fragments are bound to the surface of the flow cell exposed to reagents for polymerase-based extension. Priming occurs as the free/distal end of a ligated fragment "bridges" to a complementary oligo on the surface. Repeated denaturation and extension results in localized amplification of single molecules in millions of unique locations across the flow cell surface. This process occurs in what is referred to as Illumina's "cluster station", an automated flow cell processor (figure 3.1b) (Voelkerding et al., 2009).

A flow cell containing millions of unique clusters is now loaded into the Genome sequencer for automated cycles of extension and imaging. For sequencing, the clusters are denatured, and a subsequent chemical cleavage reaction and wash leave only forward strands for single-end sequencing. Sequencing of the forward strands is initiated by hybridizing a primer complementary to the adapter sequences, which is followed by addition of polymerase and a mixture of four differently coloured fluorescent reversible dye terminators. The terminators are incorporated according to sequence complementarity in each strand in a clonal cluster. After incorporation, excess reagents are washed away, the clusters are optically interrogated, and the fluorescence is recorded (Voelkerding et al., 2009). These images represent the data collected for the first base. Any signal above background identifies the physical location of a cluster, and the fluorescent emission identifies which of the four bases was incorporated at that position. This cycle is repeated, one base at a time, generating a series of images each

representing a single base extension at a specific cluster (figure 3.1b). Base calls are derived with an algorithm that identifies the emission colour over time. At this time reports of useful Illumina reads range from ~75-100 bases (Olivares, 2011; Voelkerding et al., 2009). Illumina and other NGS technologies have devised strategies to sequence both ends of template molecules. Such "paired-end" sequencing provides positional information that facilitates alignment and assembly, especially for short reads (Campbell et al., 2008; Korbel et al., 2007).

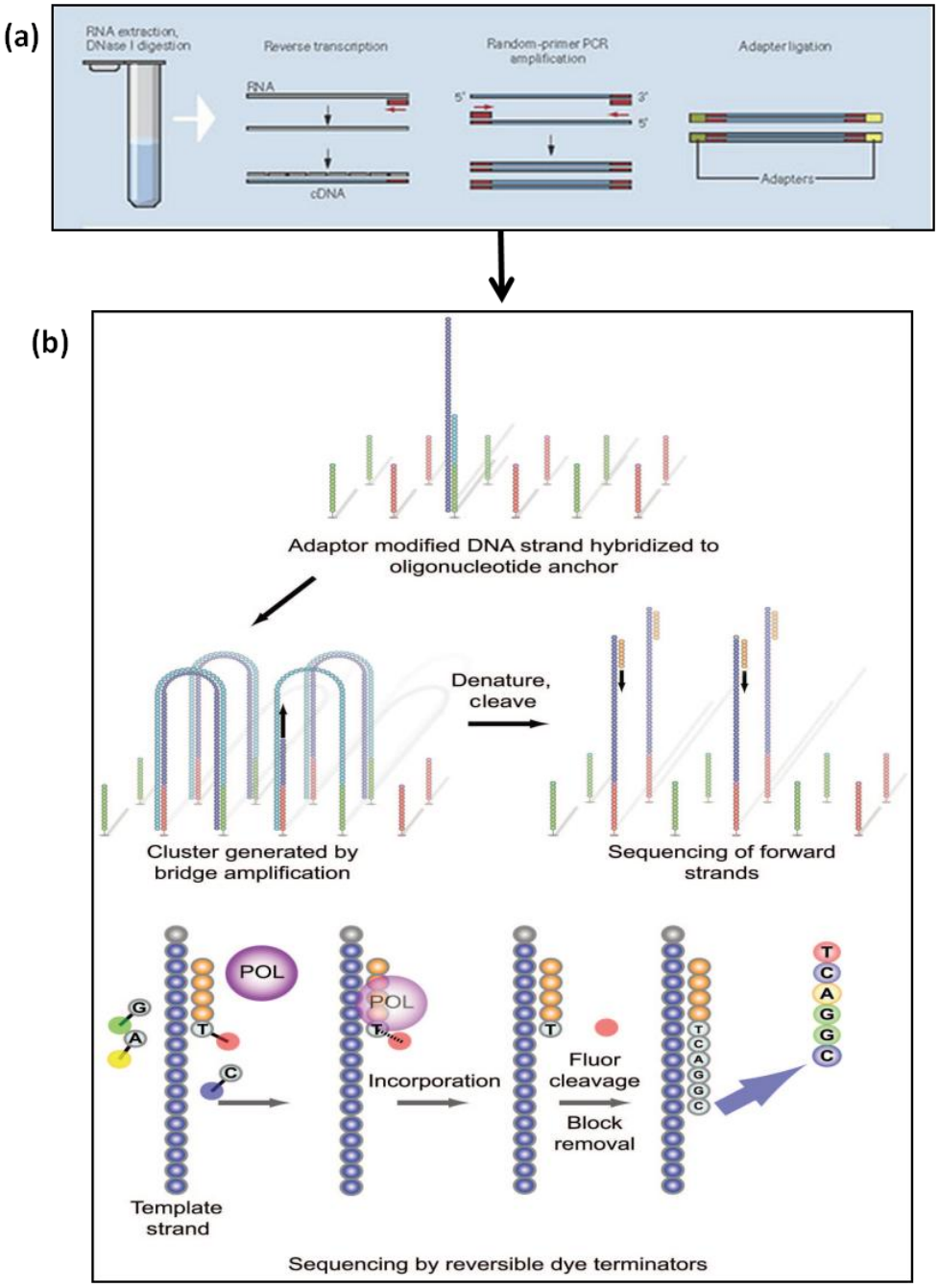


Figure 3.1: Pictorial representation of the Illumina sequencing process, where (a) shows preparation of cDNA template and (b) represents bridge PCR and sequencing process (Pallen, Loman, & Penn, 2010; Voelkerding et al., 2009).

However, each sequencing platform has its own limitations. For example, pyrosequencing is suitable for *de novo* assembly projects and amplicon studies because of the long read length while the Illumina sequencing platform is not affected by homo-polymers but shows lower coverage of AT rich regions (Barzon et al., 2011).

Another parameter that should not be ignored while working with NGS data is the use of bioinformatics tools and issues concerning sequencing output. For example, analysis of a large volume of sequencing data (70-80 gigabytes) can become computationally challenging. (Schadt, Linderman, Sorenson, Lee, & Nolan, 2010). However, increase in the number of studies using deep sequencing data for mutation analysis in viral genomes has led to the development of computational methods for estimation of the quality of sequences and for error corrections, algorithms for sequence alignments and SNP's reconstruction, statistical models to infer the frequencies of the SNP in the population, for comparative analysis, and for their visualisation (Barzon et al., 2011). Despite current challenges and limitations regarding the generation and handling of massive amount of sequence data, studies have shown that NGS can be used as an effective tool for studying the genetic diversity among plant viruses (Adams et al., 2009; Cann, Fandrich, & Heaphy, 2005).

The aim of this study was to understand the pattern of genetic variability in potyviruses and to determine if the level of genetic variation differs between the isolates or strains of potyviruses. For studying the pattern of genetic variability in potyviruses, two different strains of *Dasheen mosaic virus* (DsMV) i.e. DsMV-NZ and DsMV-B and their isolates NZ1.1, NZ1.2, B1.1 and B1.2 were used models. The isolates were produced by the vegetative propagation of host plants (taro) infected with DsMV-NZ and DsMV-B. The total RNA was extracted from each infected plant and was sent for Illumina deep sequencing, which generated the deep sequencing profiles for each virus sample. These profiles were analysed using bioinformatics tools such as GeneiousPro ver 5.5 (Drummond et al., 2011) to determine the genetic variation within each sample and the variation observed was further compared between the two strains of DsMV i.e. NZ and B. This study of genetic diversity of DsMV helped in developing better understanding of potyviral genome, which could be helpful in identification of conserved regions within the potyviral genome. This may help to explain the infectivity and host-selection process of potyviruses.

## 3.2 Methods and Materials

### 3.2.1 FastQC analysis

Total RNA extracted from taro infected with different strains of DsMV i.e. DsMV-NZ1, DsMV-NZ1.1, DsMV-NZ1.2, DsMV-B1, DsMV-B1.1 and DsMV-B1.2 was sent to Macrogen for Illumina deep sequencing using the Hiseq 2000 platform. The paired end reads returned after sequencing were checked for quality by using the bioinformatics software called FastQC (www.bioinformatics.babraham.ac.uk) (Andrews, 2011). The sequence files for each sample were imported into FastQC and the run command was used to analyse the quality of sequence data. FastQC checked the sequence quality based on several predefined parameters namely, basic statistic, per-base sequence quality, per sequence quality score, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and K-mer content. Quality reports of each sample were generated in HTML format, which were analysed to determine the quality of reads. Reports for each strain and isolate were compared to each other.

### 3.2.2 Assembly preparation

The paired end sequence files for each sample were imported into GeneiousPro ver 5.5 (Drummond et al., 2011). Paired reads were created using the default parameters and setting the distance between the two reads to 101 bp. The DsMV genome sequence (Accession no: NC_003537) was imported from Genbank (www.ncbi.nlm.nih.gov) into GeneiousPro and was used as the reference genome for creating sequence assemblies for each sample. Assemblies were created using the Assembly tool in GeneiousPro which aligned the reads against the DsMV reference genome sequence using the default parameters.

### 3.2.3 Assembly optimisation

Several parameters were tested to assemble the genomes of each strain of DsMV such as different sensitivity (medium and highest), with or without gaps with a percentage mismatch score ranging from 15% (default) to 30%. After testing several parameters to maximise the number of reads assembling against the DsMV reference genome, optimal conditions were determined and were used for creating assemblies. From the sequence library of each sample, ~10 nucleotides were trimmed from the 3' end of the reads to improve the quality of reads. The trimmed libraries were then used for creating the assemblies against the DsMV reference sequence using the highest sensitivity without allowing any gaps with the percentage mismatch score of 20%. A consensus genome sequence was generated for each virus strain, which was extracted from the assembly and used for further analysis.

### 3.2.4 Consensus sequence alignment

Multiple sequence alignment of each DsMV consensus genome sequence and the reference DsMV genome (Accession no: NC_003537) were carried out using Geneious_ConsensusAlign within GeneiousPro version 5.5 (Drummond et al., 2011).

### 3.2.5 Phylogenetic analysis

All complete genomes of viruses in the *Bean common mosaic virus* (BCMV) family were downloaded from GenBank (Table 3.1). They were aligned with the consensus genome sequence for each DsMV strain and isolate using the GeneiousAlign algorithm for creating multiple alignments in GeneiousPro version 5.5 (Drummond et al., 2011). The multiple sequence alignment generated was used for constructing Neighbour Joining (NJ) and Maximum Likelihood (ML) trees using the default parameters in GeneiousPro version 5.5. The NJ tree was created using the Jukes-Cantor model of substitution while the ML tree was based on the HYK algorithm. The robustness of each tree was determined by bootstrapping 100 replicates. The whole genome sequences of *Beet mosaic virus* (Accession no: NC_005304) and *Potato virus Y*

(PVY) (Accession no: NC_001616) was used as an outgroup having a bootstrap value of 100.

Figure 3. 2: List of all complete viral genomes used for phylogenetic analysis

| Accession Number | Name | Reference |
|---|---|---|
| **NC_004047** | Bean common mosaic necrosis virus | (Fang, Allison, Zambolim, Maxwell, & Gilbertson, 1995) |
| **NC_003397** | Bean common mosaic virus | (Zheng, Chen, Adams, & Hou, 2002) |
| **NC_004013** | Cowpea aphid-borne mosaic virus | (Mlotshwa et al., 2002) |
| **NC_007728** | East Asian Passiflora virus | (Iwai, Terahara, Yamashita, Ueda, & Nakamura, 2006) |
| **NC_010954** | Fritillary virus Y | (Chen et al., 2006) |
| **NC_015394** | Hardenbergia mosaic virus | (Wylie & M. G. Jones, 2011) |
| **NC_014790** | Passion fruit woodiness virus | (Wylie & M. Jones, 2011) |
| **NC_002634** | Soybean mosaic virus | (Eggenberger, Stark, & Beachy, 1989) |
| **NC_009742** | Telosma mosaic virus | (Ha et al., 2008) |
| **NC_006262** | Watermelon mosaic virus | (Desbiez & Lecoq, 2004) |
| **NC_007216** | Wisteria vein mosaic virus | (Liang, Song, Tian, Li, & Fan, 2006) |
| **NC_011560** | Zantedeschia mild mosaic virus | (Huang, Hu, Yang, & Chang, 2007) |
| **NC_003224** | Zucchini yellow mosaic virus | (Lin, Hou, & Yeh, 2001) |
| **NC_005304** | Beet Mosaic Virus | (Nemchinov, Hammond, Jordan, & Hammond, 2004) |
| **NC_001616** | Potato Virus Y | (Robaglia et al., 1989) |

### 3.2.6 Single nucleotide polymorphism

Single nucleotide polymorphism (SNP) detection for each sample was carried out using GeneiousPro version 5.5 (Drummond et al., 2011). The genome assembly of each sample created in GeneiousPro was used for SNP detection. The SNP_Finder tool was used for determining SNPs in each sample and a table was generated containing the information about the type of SNP, frequency of SNP, absolute counts and coverage. These data were used for calculating the level of percentage variation within each sample by comparing the SNPs identified for each sample with the DsMV reference sequence. This data was further used for determining the percentage variation within

each gene of the DsMV strains used in this study by using the consensus sequence of each gene and comparing it with the DsMV reference gene sequence.

## 3.3 Results

### 3.3.1 Analysis of library quality

FastQC software (Andrews, 2011) was used for determining the quality of Illumina deep sequencing data. After sequencing, Macrogen (Korea) gave one fastq file for B1.1, NZ1 and NZ1.2 and provided two files each for B1, B1.2 and NZ1.1. The quality of the Illumina data set for each sample was determined based on several predefined parameters namely, per-base sequence quality, per sequence quality score, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and K-mer content. The results for each of these parameters are described below.

#### 3.3.1.1 Per base sequence quality

This module shows an overview of the range of quality values across all bases at each position in the FastQ file. This module gave a plot on which the X-axis represented all of the individual bases in the reads and the Y-axis on the graph shows the quality scores (figure 3.2). The higher the quality score the better the base call. The background of the graph divides the Y-axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). For a good library, the mean quality score was expected to stay above the threshold quality score of 24 (red to orange transition). A quality score above 24 indicates a good quality base call while a quality score below 20 indicates the library is of poor quality and requires trimming to remove bad quality sequences. Figure 3.2 shows that for all the sequence libraries per base quality score for up to ~80 bases was above the threshold quality score of 24. However, decrease in per base quality score (i.e. score below 20) was observed after ~80 bases for all the sequence libraries. This decrease in base call efficiency with increasing read lengths is commonly observed in Illumina sequencing data (Voelkerding et al., 2009). Therefore, it was suggested that trimming the 3'end of the reads would be helpful in overcoming this bias.

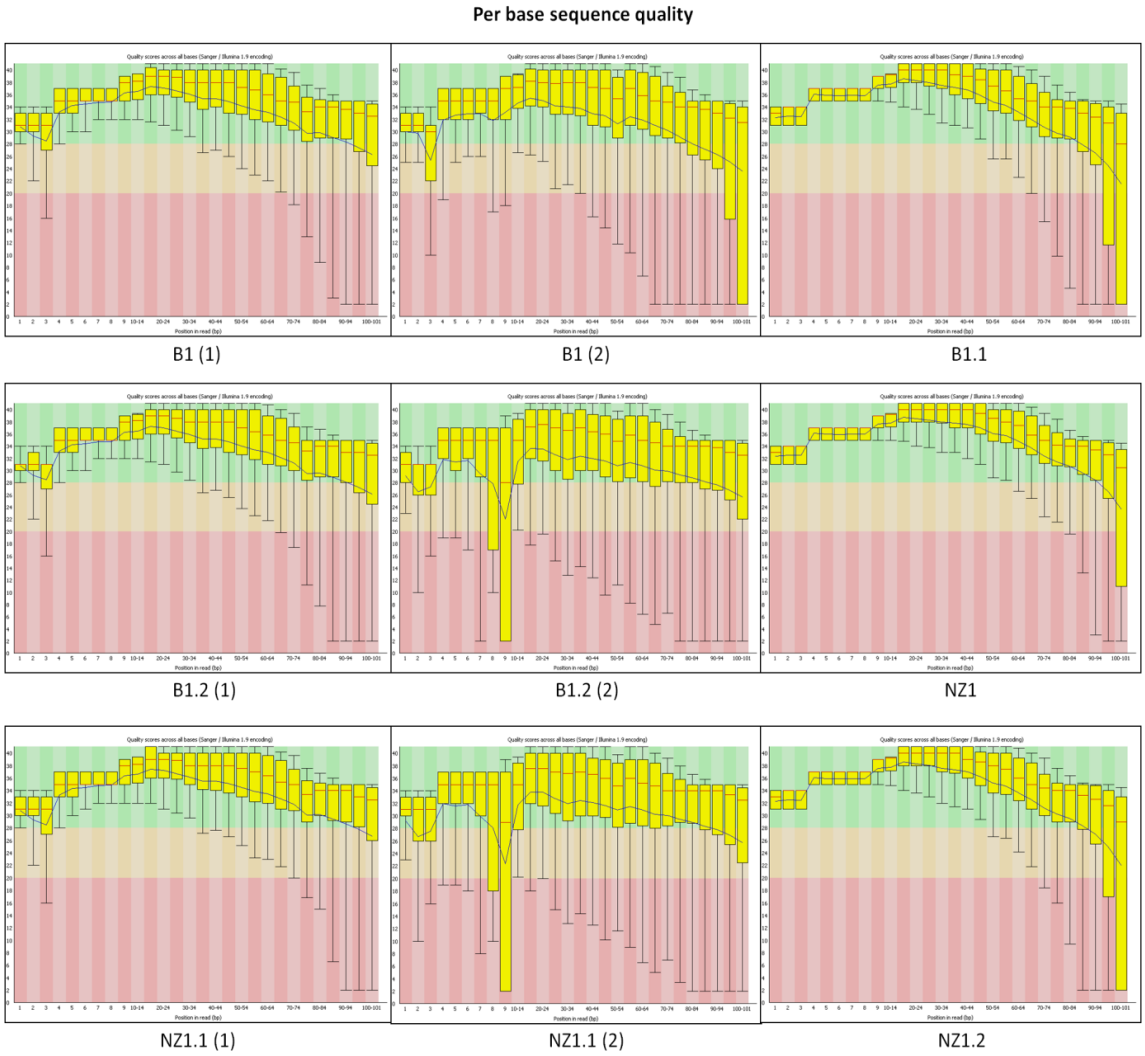


Figure 3.3: Per base sequence quality measurements for each Illumina data set. Yellow boxes show the inter-quartile range (i.e. from 25-75%), the red line shows the median value, and black whiskers (that goes from 10-90%) and the blue line represents the mean quality score (Andrews, 2011)

### 3.3.1.2 Per sequence quality score

The per sequence quality score helps in detecting if a subset of sequences within a library have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged, however, these should represent only a small proportion of total sequences. Unlike per base sequence quality module that calculates the quality score for individual bases within a read, per sequence quality score calculates the average quality score for a whole read and plots out the distribution of those means. For a good sequence library, it was expected that the majority of sequences within a library would have a high average

quality score i.e. majority of sequences were expected to cluster together to form a single peak. For a bad quality library, more than one peak would be observed, where the appearance of secondary peaks (i.e. a second population of sequences) was an indicator of a positional bias of the flow cells used for sequencing or indicates a subset of bad quality sequences.

Figure 3.3 shows that each library had a per sequence quality score of over 20 (threshold quality score) indicating each library was of good quality. Though the distribution of means did not formed a tight peak; however, the majority of sequences were still within the threshold quality score and formed a single peak. Minor secondary peaks were observed in B1 (2), B1.2 (1), B1.2 (2), NZ1.1 (2) sequence library (figure 3.3), which suggested that a small proportion of sequences within these libraries were of bad quality; however, due to their small population size it was concluded that these bad sequences would not affect the data analysis.

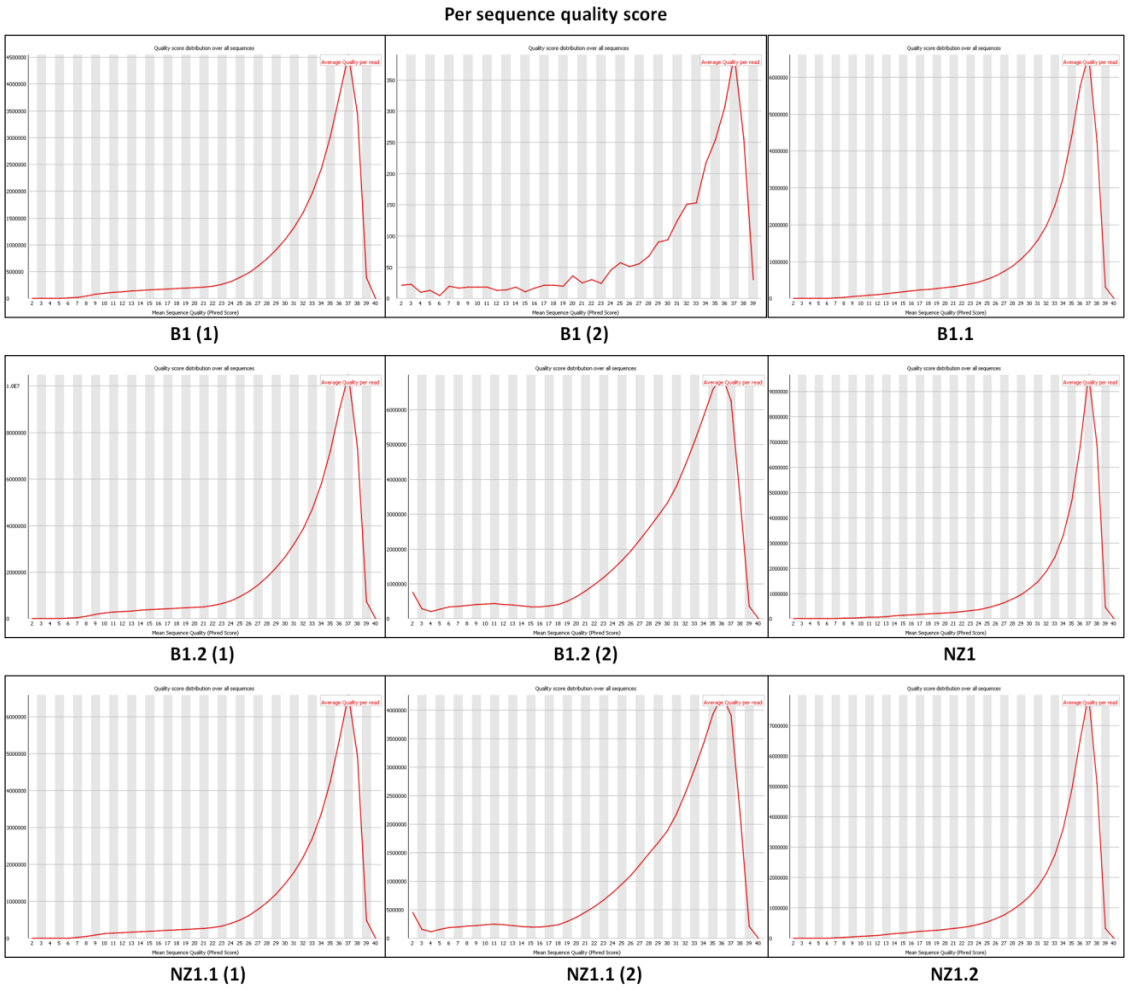


Figure 3.4: Per sequence quality score for each Illumina data set, Y-axis represents number of reads and X-axis represents the mean sequence quality score

### 3.3.1.3 Per base sequence content

Analysis of the per base sequence content gave the distribution of the four bases within each library. In a good library, it was expected that there would be an even distribution of the four bases that does not change with the base position, that is, parallel lines going across the plot. This would suggest that the position of the base within a read did not influence the base call. For a bad library, it was expected to observe wobbling across the plot with peaks going as high as 80-90%, which would suggest that the base call was influenced by the base position (Andrews, 2011).

Figure 3.4 shows that sequence libraries for all the samples failed this test. In none of the libraries, parallel lines were observed in the plot that suggested that the distribution of four bases was not even in any of the libraries. For all the samples, bias in nucleotide composition was observed at the beginning of reads. This bias was strong at the first few positions of the reads (i.e. up to 10-14 bases). This difference in the distribution of bases across the reads are mostly observed in RNA sequencing data because of the primers that are generally used for RNA sequencing (Andrews, 2011; Schwartz, Oren, & Ast, 2011). Therefore, it was concluded that for this module the anomaly observed in each library was not of any concern and would not affect the results in during sequence data analysis.

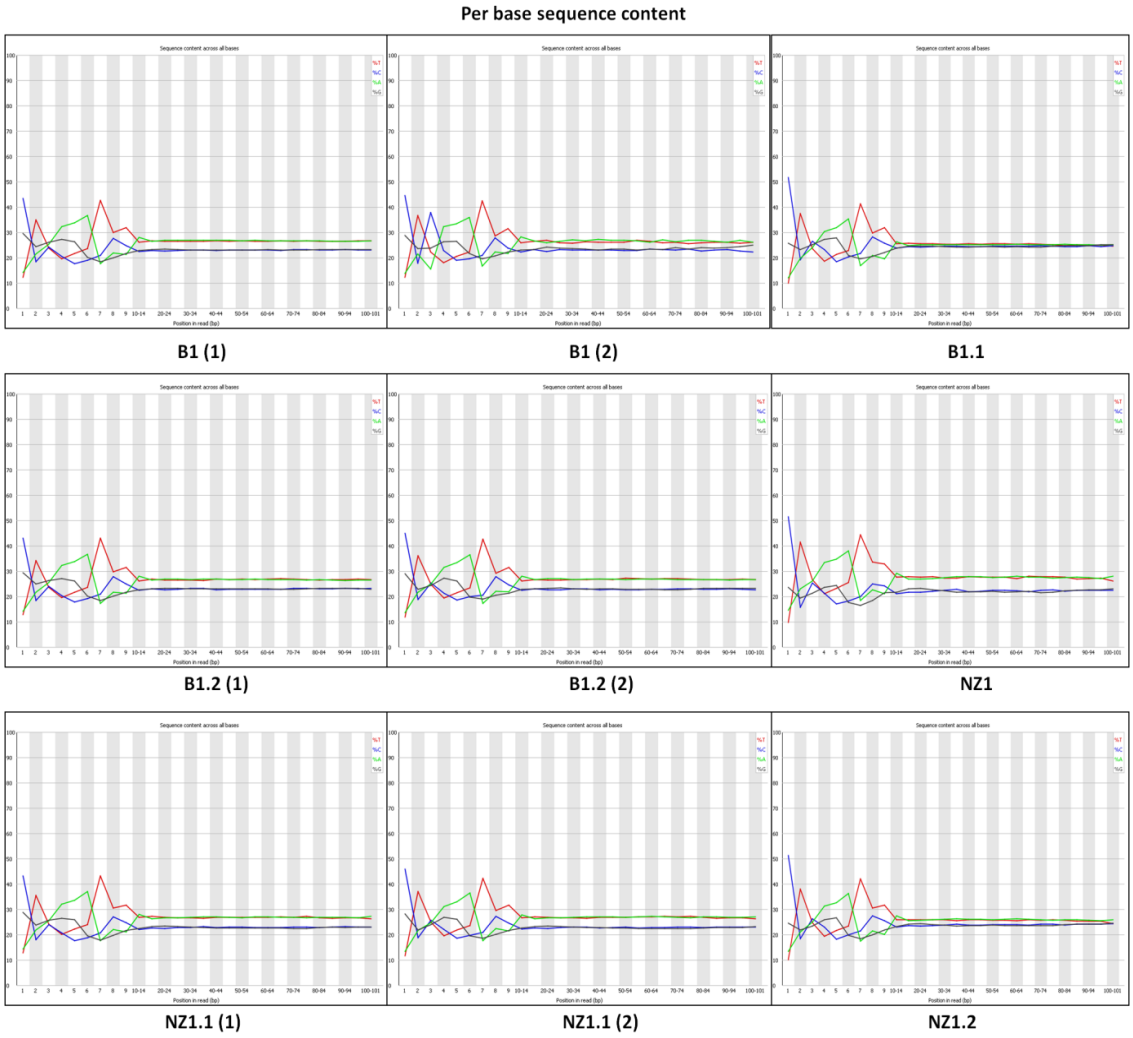


Figure 3.5: Plots showing the per base sequence content of all sequence libraries where the X-axis represents base position within a read (with left to right being 5' to 3') and the Y-axis represents percentage bias (Andrews, 2011).

### 3.3.1.4 Per base GC content

Per base GC content was similar to per base sequence content, this module was used to analyse the distribution of GC content across the read. The overall GC content should reflect the GC content of the underlying genome. In a good quality library it was expected to see a straight line running across the plot, which indicated that GC content was uniform across the read and was not influenced by the position of the base. For a bad quality library, it was expected to see a wobble or small peaks across the plot instead of a straight line, indicating that distribution of GC content was uneven across the read and was influenced by the base position.

Figure 3.5 shows that for each sequence library a wavy line was observed instead of the expected straight line. However, the wavy pattern was observed for only first few bases i.e. up to 14 bases, followed by a straight line going across the reads (figure 3.5). This pattern suggested that there was a bias in distribution of GC content in the first few bases of the reads and distribution of GC was influenced by base position. However, this kind of pattern is commonly observed in RNA sequencing libraries, the reason behind this pattern are the primers and the adapters that are generally used for RNA sequencing and does not influence the results during the sequence analysis (Andrews, 2011). Therefore, it was concluded that the variation observed in the distribution of GC content within the libraries would not contribute to any bias in the results.

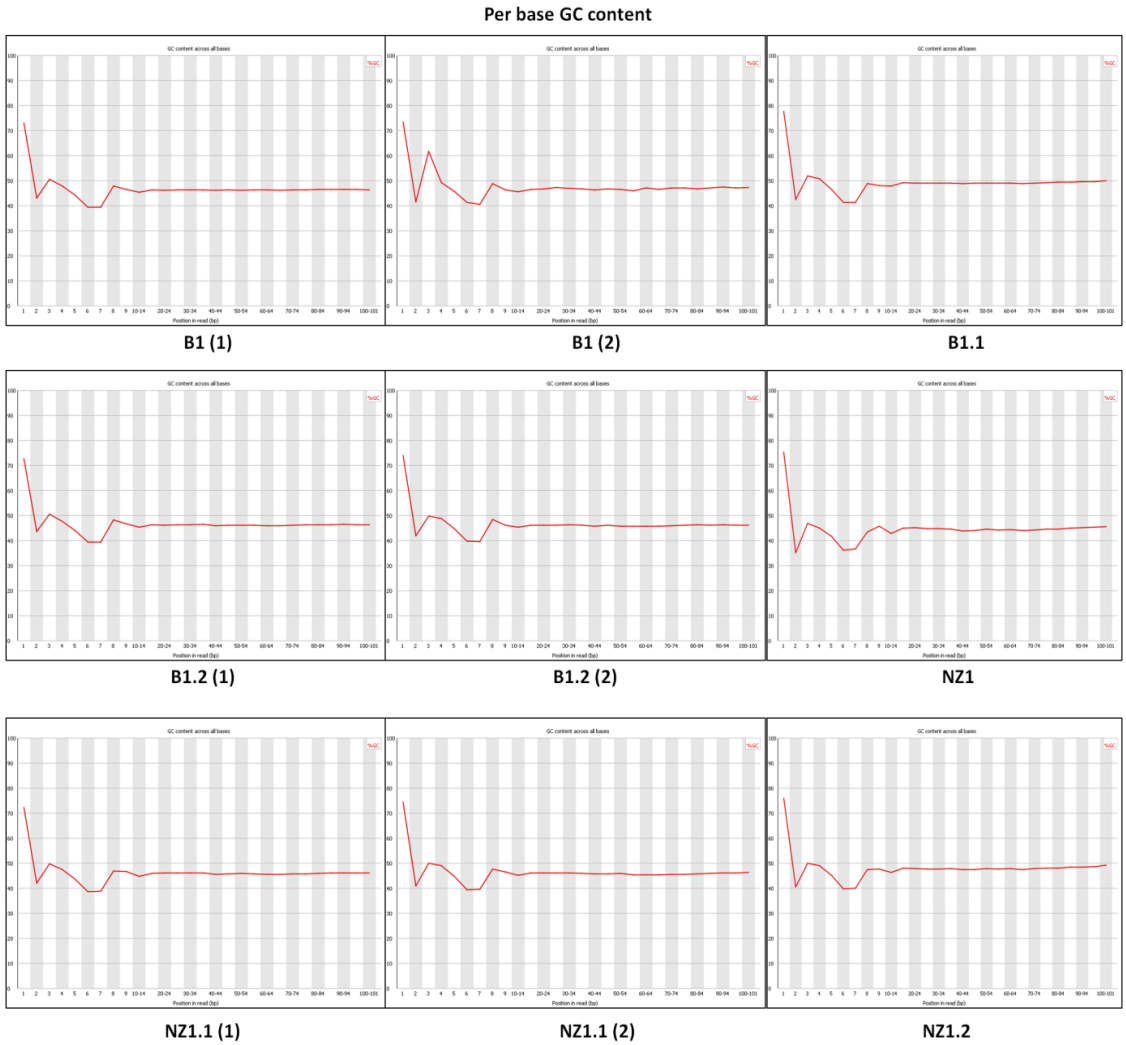


Figure 3.6: Plot showing the per base GC content of all sequence libraries. The X-axis represents base position in reads (left to right being 5' to 3') and the Y-axis represents percentage GC content

### 3.3.1.5 Per sequence GC content

Per base GC content described the distribution of GC content across all the sequences within a library. This module plot out the distribution of GC content across all the sequences within a library (Andrews, 2011). For a good quality library, a normal distribution curve was expected. In a good library, there should be a good overlap between the actual and the expected distribution curve. For a bad quality library, it was expected that there would be wide overlap between the actual and the expected curve. Also, in a bad quality library, secondary peaks would be seen, which is an indicator of some kind of contamination within the library (Andrews, 2011).

Figure 3.6 shows that the distribution of GC content was consistent across majority of reads within each sequence library as good overlap between expected and actual curve was observed for each of the library. However, within each library secondary peaks were also observed, which indicated the presence of some contaminants within the libraries. These contaminants could be the residues of the primers, adapters or reagents that were used for RNA sequencing (Andrews, 2011; Schwartz et al., 2011). It is known that GC content is positively related to the read coverage and therefore the variations observed in the GC distribution within each library could contribute to the variations in the read coverage during assembly preparation.

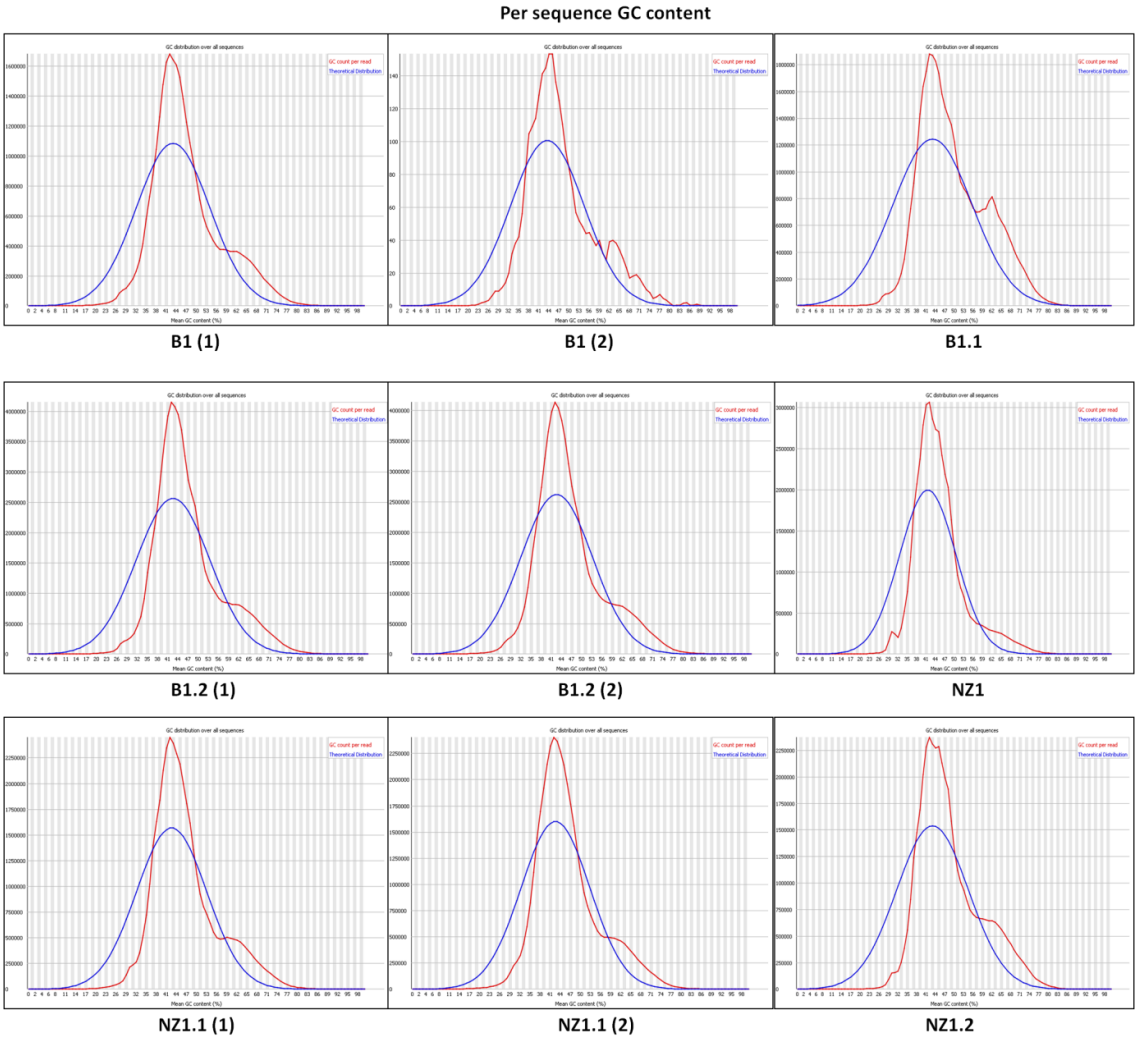


Figure 3.7: The GC distribution curve for each sequence library. The X-axis represents the mean GC content (%) and the Y-axis represents the number of reads within a library. Blue and red line represents the expected and the observed curve, respectively.

### 3.3.1.6 Per base N content

If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called. This module was used to determine whether there were any uncalled bases within the libraries or not. In a good library, it was expected that no uncalled bases would be present in the library and a straight red line should be observed, overlapping with the X-axis. In a bad quality library, it was expected to observed many secondary peaks instead of a straight line, which indicate the presence of contaminants in the library (Andrews, 2011).

Figure 3.7 shows that no uncalled bases were present in all sequence libraries, except in one of the libraries of B1.2 (2) and NZ 1.1 (2). However, the uncalled bases were observed only at few base positions and the percentage of uncalled bases observed for both libraries were below the threshold value (i.e. 30%) (Andrews, 2011). Therefore, it was concluded that these anomalies would not affect the sequence analysis.

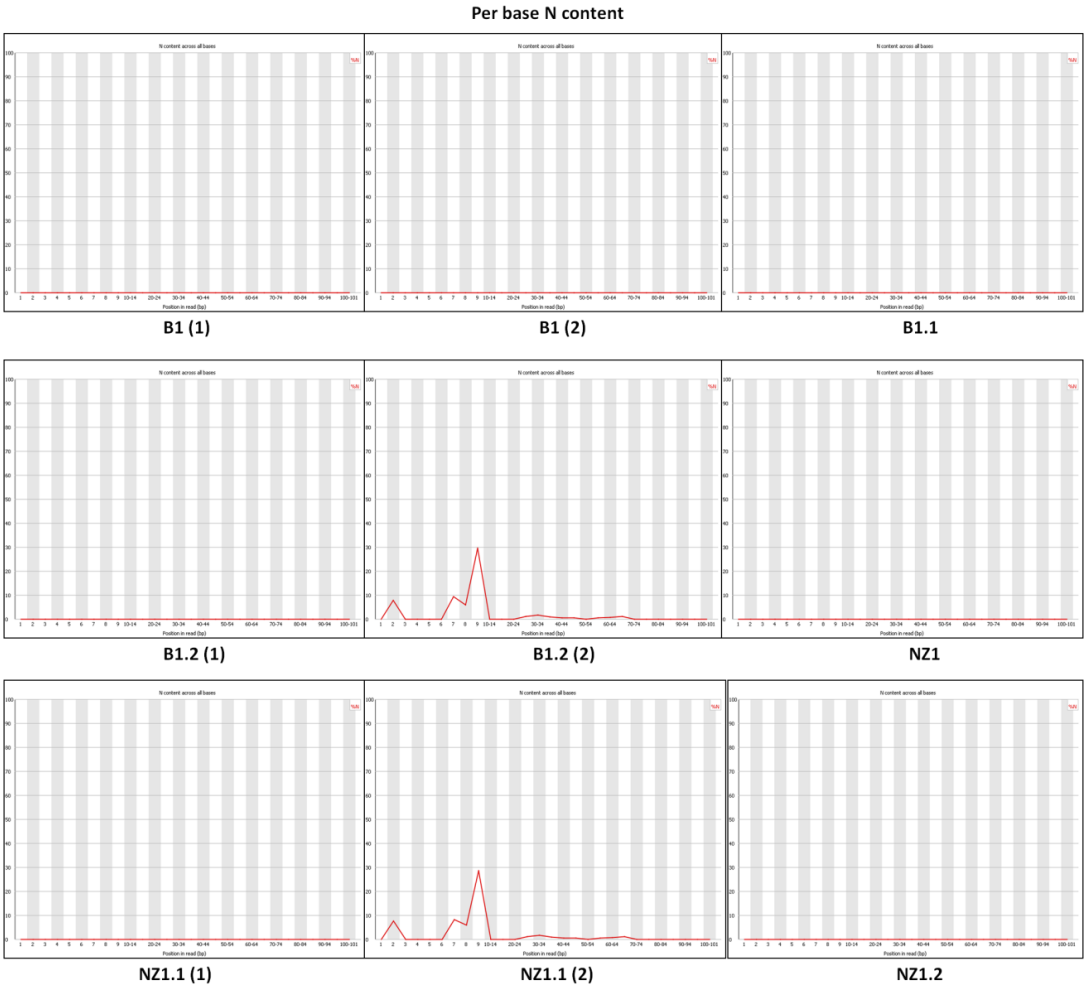


Figure 3.8: Per base N content measurements for each Illumina data set, where X-axis shows the position within a read and Y-axis shows the percentage N content (Andrews, 2011)

### 3.3.1.7 Sequence length distribution

Sequence length distribution module is only used as an informative tool that tells whether all the reads within a library are of same length or not. As some high throughput sequencers generate reads of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries, some pipelines will trim

sequences to remove poor quality base calls from the end. This module generates a graph showing the distribution of fragment sizes in the file, which was analysed (Andrews, 2011; Olivares, 2011). For Illumina HiSeq 2000 sequencing, the average read length is expected to be ~75-100 bp (Illumina, 2012; Metzker, 2010). Figure 3.8 shows that all the reads within each sequence library was 101 bp in length, which was expected for Illumina HiSeq data.

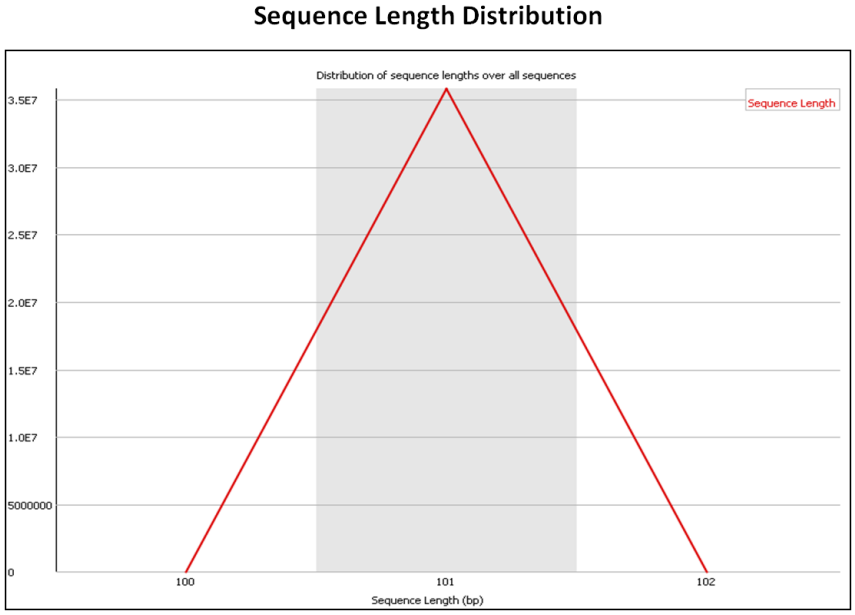


Figure 3.9: Plot showing the sequence length distribution across all sequence libraries. The X-axis represents sequence length (bp) and the Y-axis represents distribution of length over all sequences (Andrews, 2011)

### 3.3.1.8 Sequence duplication level

This module was used to test how unique the reads were within a library. The sequence duplication level does not tell whether a library is good or bad, it only tells if the library looks unusual in any way. The results are plotted on a relative scale, with the number of sequences occurring exactly once being set as 100%, and everything else would be shown relative to that number. The plot value for duplicate 1 is therefore always 100%, and higher duplication levels can produce values either higher or lower than this. In an ideal plot for a diverse library, the values for duplicate levels above 1 should quickly decay to zero and stay there. However, there are several ways in which the plot can indicate other types of problem in the library (Andrews, 2011).

Figure 3.9 shows that average sequence duplication level for all sequence libraries was over 70%. This kind of pattern was expected because total RNA extracted from taro leaves infected with DsMV was used for Illumina sequencing and viral RNA represents a significant proportion of total RNA. Therefore, high duplication level would be expected. In a similar kind of study on *Passion fruit woodiness virus* (PWV), Wylie and Jones (2011) found that 7% of the total RNA belonged to PWV virus (Wylie & M. Jones, 2011). The other reason for high duplication level could be contamination of the library due to the adapter sequences.

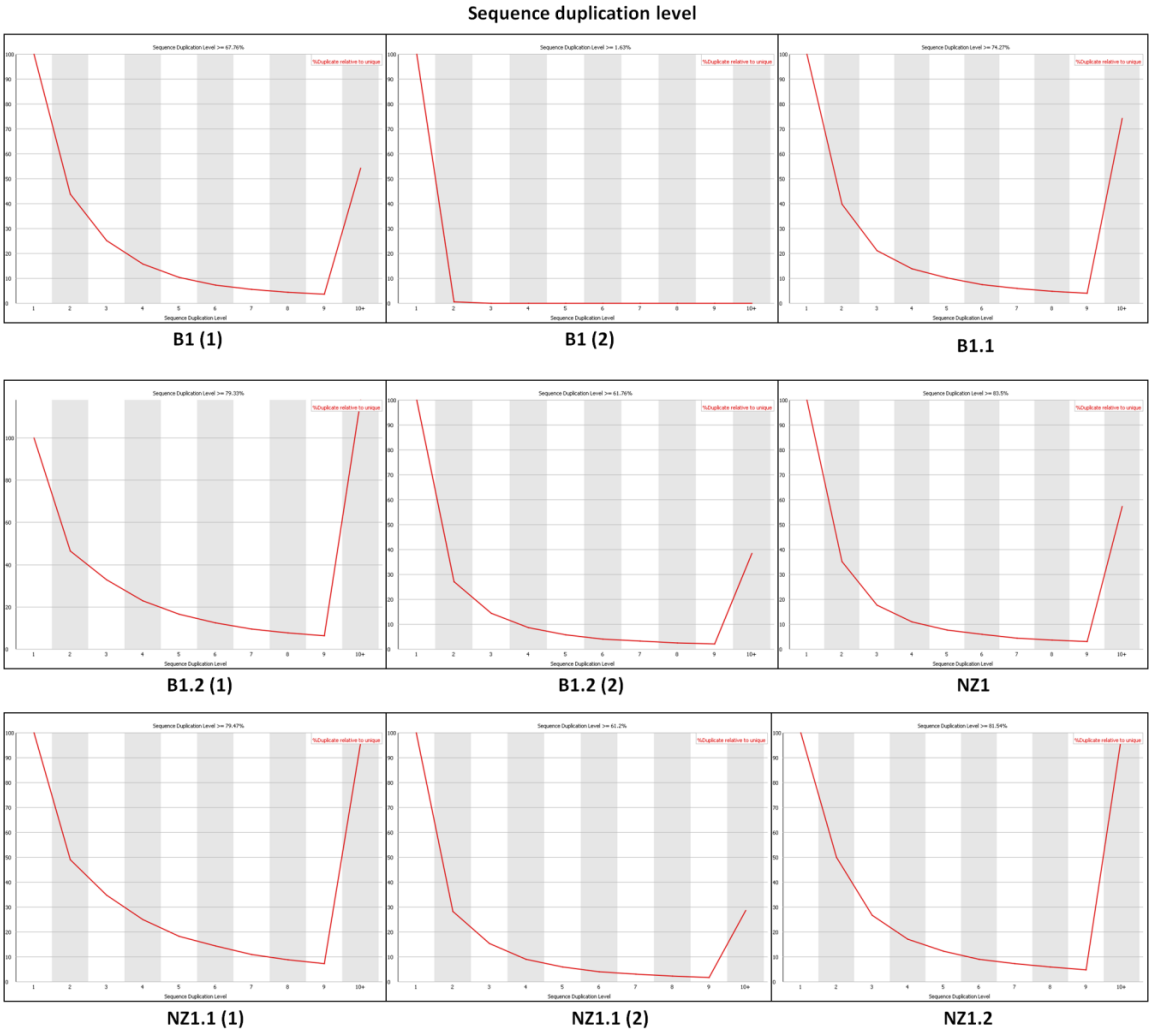


Figure 3.10: Plots showing the sequence duplication level within each sequence library. The X-axis represents duplication level and the Y-axis represents the duplication percentage

### 3.3.1.9 Overrepresented sequences

As high duplication level was observed for each sequence library, it was necessary to determine the sequences that were contributing to the high level of duplication within the libraries. This module looked at the individual sequences that were overrepresented in each library, that is representing more than 0.1% of the sequences in a library (Andrews, 2011). In a good sequence file, it was expected that no overrepresented sequences would be identified in the libraries. Finding overrepresented sequences in a library suggested that either it was highly biologically significant or is an indicator that the library was contaminated, or the library was not as diverse as expected (Andrews, 2011). No overrepresented sequences were identified in NZ1.1 (2), B1 (1), and B1.2 (1 & 2) sequence libraries, suggesting that each of these sequence libraries contain a diverse set of sequences. However, overrepresented sequences were identified in NZ1, NZ1.1 (1), NZ1.2, B1 (2) and B1.1 sequence libraries. FastQC was unable to identify the possible source of these sequences; therefore, BLASTn was used for determining the possible identity of each overrepresented sequence that was identified in the data set. Table 3.2 shows the list of overrepresented sequences identified in each sequence library.

BLASTn results revealed that all the overrepresented sequences identified in each library belonged to DsMV (Table 3.2), which suggested that libraries were not contaminated and were less diverse than expected. Therefore, it was concluded that the pattern observed was not unusual for virus infected transcriptomes and would not affect the data analysis.

Table 3.2: List of overrepresented sequences identified in each sequence library

| Library | Sequence | Count | Source |
|---|---|---|---|
| NZ1 | CTCAAATTCATTCTTTGTGGGTTTGTTGCACCATTGTTTGATATAAGAT | 229336 | DsMV (NIa-Pro gene) |
| NZ1 | CTTTTTCATTAACTGGGATAGTGGAAGGGATTAGAAAGCGATATTTGAAG | 161404 | DsMV (CI gene) |
| NZ1 | CTACAATTAAATCATTAGTGAATTTACTGCCATCCCATTTACCCTTCAAG | 102847 | DsMV (CI gene) |
| NZ1 | CAAATTTTCAAGAAAAGAGCATGCGTGCCACAATTTCTGAAGCATCCTTG | 88725 | DsMV (NIa-Pro gene) |
| NZ1 | CCTAGTTCTGCAAAAGAGTTCTGAAAAGTGTCCAGCACAATCTCATCATC | 76688 | DsMV (HC-Pro gene) |
| NZ1 | CTAATATATGAGTATTTTTCGCGCACTATGCGCGACCCTGTTGTGGCGCA | 75823 | DsMV (6k2 gene) |
| NZ1 | CATAAATCCAAATTTACTTCATTTCGTAGGCGCCTACCACGCCGTTCATG | 69515 | DsMV (HC-Pro gene) |
| NZ1 | GTCAGAAAGAGTTTGAGTATTTTATTAGGGGCGTTTAACAAGTGTTATAG | 63493 | DsMV (P3 gene) |
| NZ1 | GAAAGAGTTTGAGTATTTTATTAGGGGCGTTTAACAAGTGTTATAGTGAT | 60550 | DsMV (P3 gene) |
| NZ1 | CTGGTTATTCTTTTTCATTAACTGGGATAGTGGAAGGGATTAGAAAGCGA | 59135 | DsMV (CI gene) |
| NZ1 | CAAAGAAATAATGCATAGAACAGAGGCCATATGCGCAGCCATGATAGAAG | 57067 | DsMV (NIb gene) |
| NZ1 | CTCGAAATCATCAGTAAAGGACACAAAATAGTTTCTTTCAGATATATTTG | 56166 | DsMV (NIa gene) |
| NZ1 | ATTCTTTTTCATTAACTGGGATAGTGGAAGGGATTAGAAAGCGATATTTG | 50039 | DsMV (CI gene) |
| NZ1 | CCCAGCTCCGCCTGCACCCAAAACCACAGAAACTCCGGTAGTGAAGGATG | 49734 | DsMV (CP gene) |
| NZ1 | CTCATATATTAGCCACCCACCTCCTAGTAATGTTATAGCTCCTACAATTA | 48081 | DsMV (CI gene) |
| NZ1 | GTTTGCTCGAAATCATCAGTAAAGGACACAAAATAGTTTCTTTCAGATAT | 47223 | DsMV (HC-Pro gene) |
| NZ1 | CTCAAATTCATTCTTTGTGGGTTTGTTGCACCATTGTTTGATATAAGAT | 229336 | DsMV (NIa-Pro gene) |
| NZ1.1 (1) | CTACAATTAAATCATTAGTGAATTTACTGCCATCCCATTTACCCTTCAAG | 67624 | DsMV (CI gene) |
| NZ1.1 (1) | CTCAAATTCATTCTTTGTGGGTTTGTTTGCACCATTGTTTGATATAAGAT | 48161 | DsMV (NIa-Pro gene) |
| NZ1.1 (1) | CTTTTTCATTAACTGGGATAGTGGAAGGGATTAGAAAGCGATATTTGAAG | 41640 | DsMV (CI gene) |
| NZ1.2 | CTTTTTCATTAACTGGGATAGTGGAAGGGATTAGAAAGCGATATTTGAAG | 229336 | DsMV (NIa-Pro gene) |
| NZ1.2 | CTCAAATTCATTCTTTGTGGGTTTGTTTGCACCATTGTTTGATATAAGAT | 161404 | DsMV (HC-Pro gene) |
| NZ1.2 | CTACAATTAAATCATTAGTGAATTTACTGCCATCCCATTTACCCTTCAAG | 102847 | DsMV (6k2 gene) |
| B1 (2) | AGCAAACCGACTACAAGCAATGGATGGAAACGAACAAATTGAATACCCAT | 5 | DsMV (CP gene) |
| B1 (2) | CAGAAATGTTATTGAACAAACTTGCAATCCCCAACCAGTACGTGAATCAA | 3 | DsMV (CI gene) |
| B1 (2) | CACATTTTCAAGAAAAGAGCATGCGTGCCACGCTTTCTGAAGCATCCTTA | 3 | DsMV (NIa-Pro gene) |
| B1.1 | GGAAAATTTTCATTGTGTTGGAGCAATTGGTGAGGAAAATGGCAATGAAT | 47177 | DsMV (P3 gene) |
| B1.1 | CTCGTACGGTCCTCCTCCTCATGGGTGTGCTCCTCGCCACCTTGCTGCTC | 43215 | DsMV (P3 gene) |

Count: Number of times a particular sequence appears in the library; Source: Possible identification of each sequence identified using BLASTn

### 3.3.1.10 K-mer content

K-mer is a motif (or a small word) of length k, observed more than once in a genomic or sequenced data. The order of a K-mer is defined by its word size. The K-mer content is the number of times the K-mer occurs in the sequence and the distribution is related with the enrichment of a genomic sequence based on a particular K-mer (Olivares, 2011). The K-mer content module counts the enrichment of every 5-mer within the sequence library. It calculates an expected level at which this k-mer should have been seen based on the base content of the library as a whole and then uses the actual count to calculate an observed/expected ratio for that k-mer. In addition to reporting a list of hits it draws a graph for the top six hits to show the pattern of enrichment of that K-mer across the length of the reads. This will show if you have a general enrichment, or if there is a pattern of bias at different points over your read length (Andrews, 2011; Olivares, 2011).

For each library, the six most commonly occurring 5-mers were identified and enrichment over length analysed. Each sequence library showed an unusual pattern for this module. This suggested that a K-mer was either enriched more than three-fold overall, or more than five-fold at any individual position within a read. Figure 3.10 shows a graph for the top six hits to show the pattern of enrichment of that K-mer across the length of the reads. The top six K-mers (or motifs) identified in each library is listed in table 3.3.

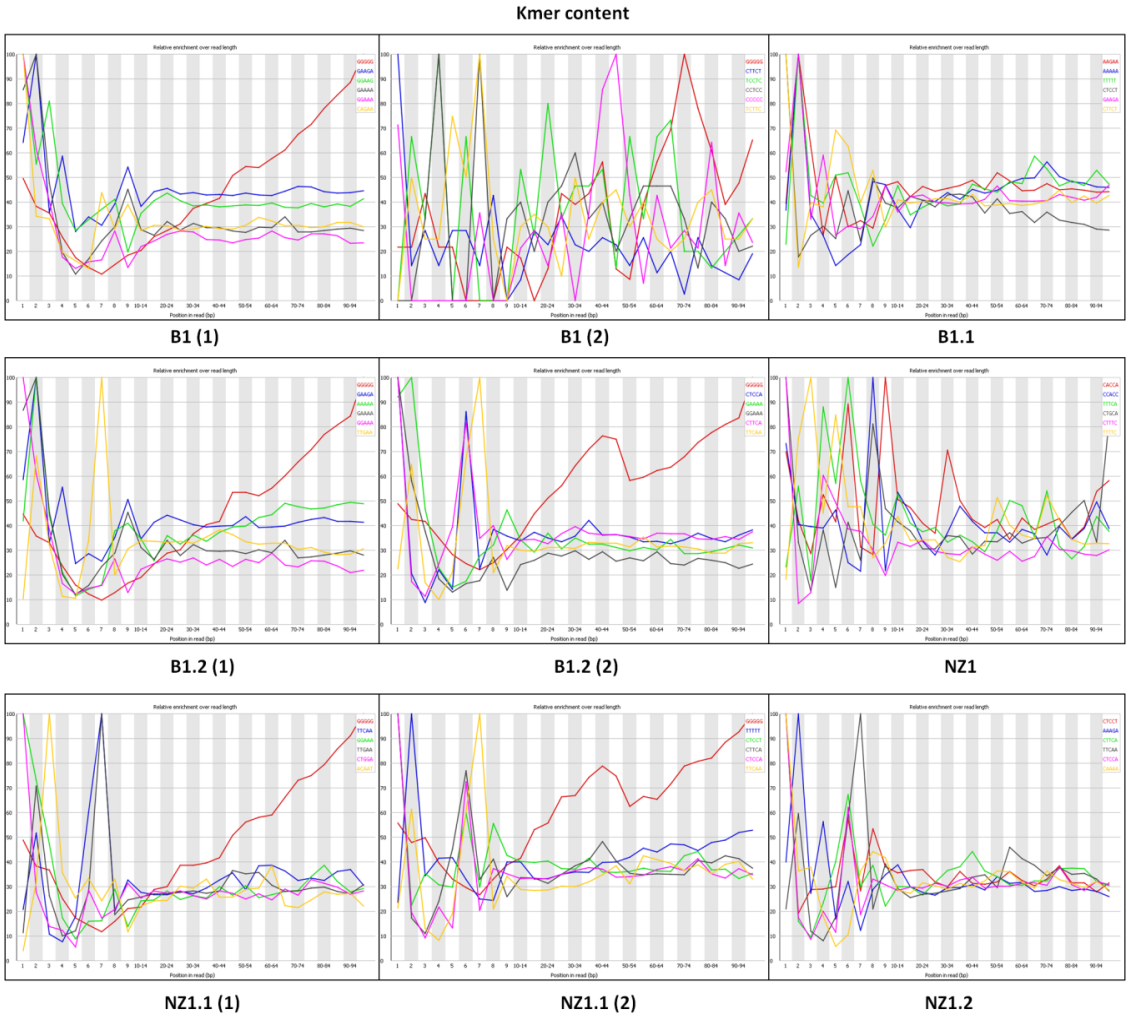


Figure 3.11: Plot showing the K-mer content within each library. The X-axis represents base position in a read and the Y-axis represents relative enrichment over read length (Andrews, 2011)

Table 3.3: List of top six k-mers identified within each sequence library (Andrews, 2011)

| Library | K-mer |
|---|---|
| B1 (1) | GGGGG, GAAGA, GGAAG, GGAAA, GAAAA, CAGAA |
| B1 (2) | GGGGG, CTTCT, TCCTC, CCTCC, CCCCC, TCTTC |
| B1.1 | AAGAA, AAAAA, TTTTTT, CTCCT, GAAGA, CTTCT |
| B1.2 (1) | GGGGG, GAAGA, AAAAA, GAAAA, GGAAA, TTGAA |
| B1.2 (2) | GGGGG, CTCAA, GAAAA, GGAAA, CTTCA, TTCAA |
| NZ1 | CACCA, CCACC, TTTCA, CTGCA, CTTTC, TTTTC |
| NZ1.1 (1) | GGGGG, TTCAA, GGAAA, TTGAA, CTGGA, ACAAT |
| NZ1.1 (2) | GGGGG, TTTTT, CTCCT, CTTCA, CTCCA, TTCAA |
| NZ1.2 | CTCCT, AAAGA, TTCAA, CTTCA, CTCCA, CAAA |

After analysing all the FastQC reports on the Illumina deep sequencing data for all samples, it was concluded that the average quality of all the sequence libraries was above the threshold level. Therefore, these sequences libraries were used for carrying out further analyses.

### 3.3.2 Assembly preparation

Genome assemblies were created for each strain and isolate of DsMV i.e. B1, B1.1, B1.2, NZ1, NZ1.1 and NZ1.2, using the published full-length DsMV genome (Accession no: NC_003537) as the reference genome. It was found that using the default parameters (i.e. medium sensitivity with gaps and mismatch score of 10%) for creating genome assemblies of each sample resulted in low genome coverage of reads. Figure 3.11 shows that the overall coverage for each genome assembly was low and big coverage gaps were observed in each assembly. High coverage was observed for the 6K2, NIa-VPg and CP regions of the DsMV genome. The overall coverage was also higher towards the 3' end of the DsMV genome in comparison to the 5'end of the genome. It was also observed that each assembly was of different length due to introduction of large number of gaps in the assemblies. The regions where coverage was higher was consistent between isolates of the same strain, that is, all B's showed similar coverage patterns and all NZ isolates showed a similar pattern. However, it was found that the pattern observed for DsMV-B's was different from the pattern observed for DsMV-NZ.

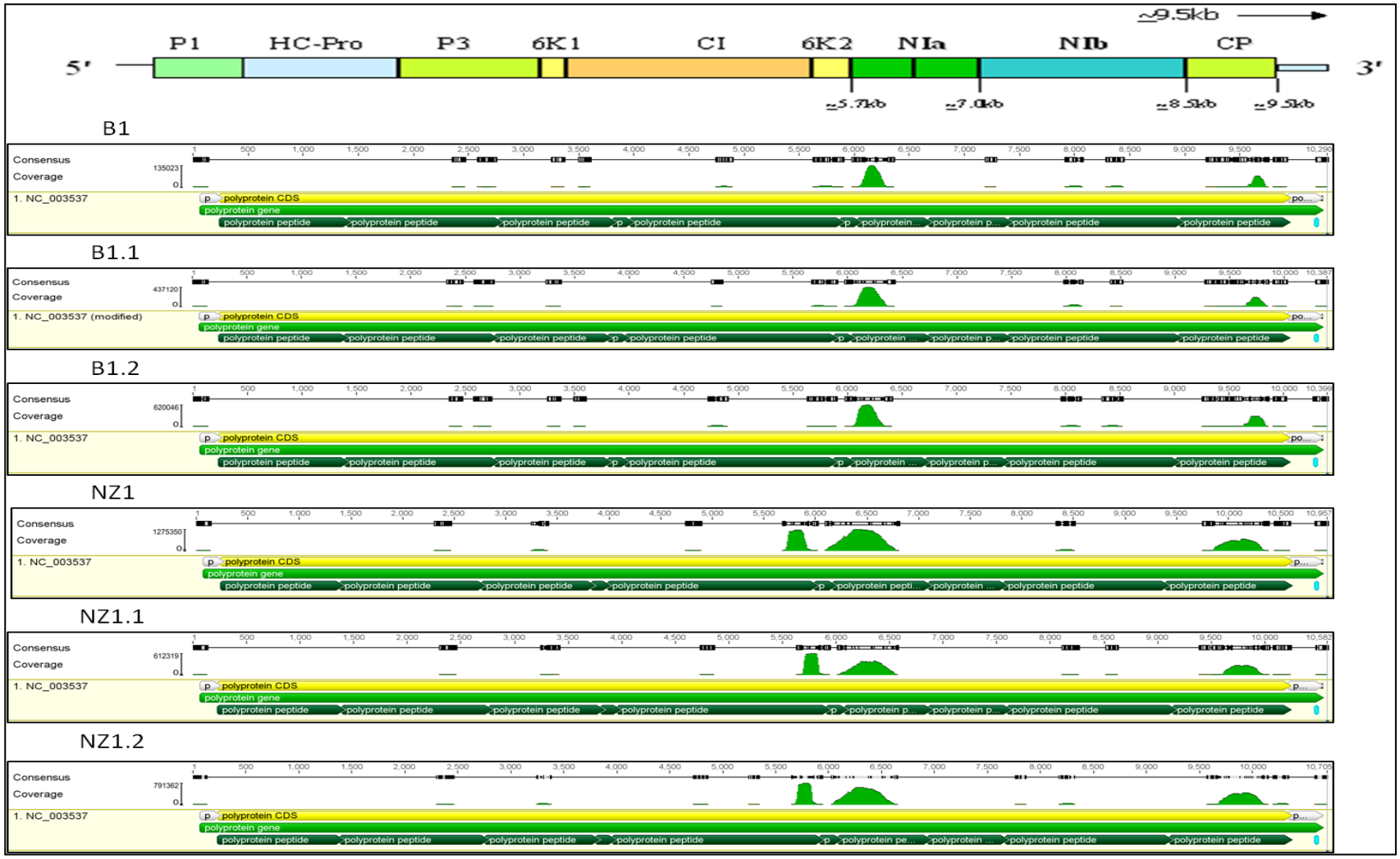


Figure 3.12: Genome assemblies created for each sample of DsMV using the default parameters in GeneiousPro. DsMV reference genome was used for creating each assembly. The breakup of DsMV genome into the 10 different coding regions is shown at the top of reads.

The genome coverage is defined as the percentage of bases in the reference covered by the reads that were assembled into contiguous sequences (contigs); this can only be computed if a reference genome exists. The contigs aligning with large internal gaps or low confidence indicate either structural variants or misassemblies, and should be carefully monitored. However, if paired reads are available, then large deviations in the mapping distances or read pairs in which only one read maps against the contig hint at misassemblies rather than structural variants (Illumina, 2012). Therefore, the low genome coverage and big coverage gaps observed for each sample suggested that reads were not assembling as efficiently as expected, and each assembly needs to be optimised for carrying out further analysis.

### 3.3.3 Assembly optimisation

Due to the low coverage of reads across the DsMV reference genome, different parameters were used for optimising the assembly generation process. The parameters used for genome assembly were optimised using the B1 paired end libraries. Changing from the default sensitivity to the highest sensitivity increased the coverage across the length of the genome (figure 3.12). It was also found that using the highest sensitivity with an increased mismatch percentage score to 20% resulted in more reads being assembled. The highest overall genome coverage at these settings was 469437 while for the default settings it was 135023. However, when the mismatch percentage score was increased to 30%, there was a reduction in the number of reads being assembled (9,203,344) against the DsMV reference genome. This suggested that the highest sensitivity with 20% mismatch score was optimal for library preparation.

However, the per base sequence quality module of FastQC revealed that the quality of 3' end of the reads was not good due to a decrease in base call efficiency with increase in the read length (figure 3.2). It was recommended to trim the 3'end of the reads to remove the 3'end base call bias. Therefore, it was decided to assess the effect of trimming 10 bases from the 3'end of each read. Figure 3.13 shows that after trimming, the number of reads being assembled increased from 10,359,693 to 10,424,397. However, due to increased stringency in the assembly generation parameters, the overall coverage decreased from 469,437 to 295,710, but the coverage increased across the length of the genome and gaps in the coverage were also reduced. After this series of

optimisation steps, creating an assembly on sequences trimmed at the 3’end using the highest sensitivity without gaps with a mismatch score of 20% appeared to be the optimal conditions.

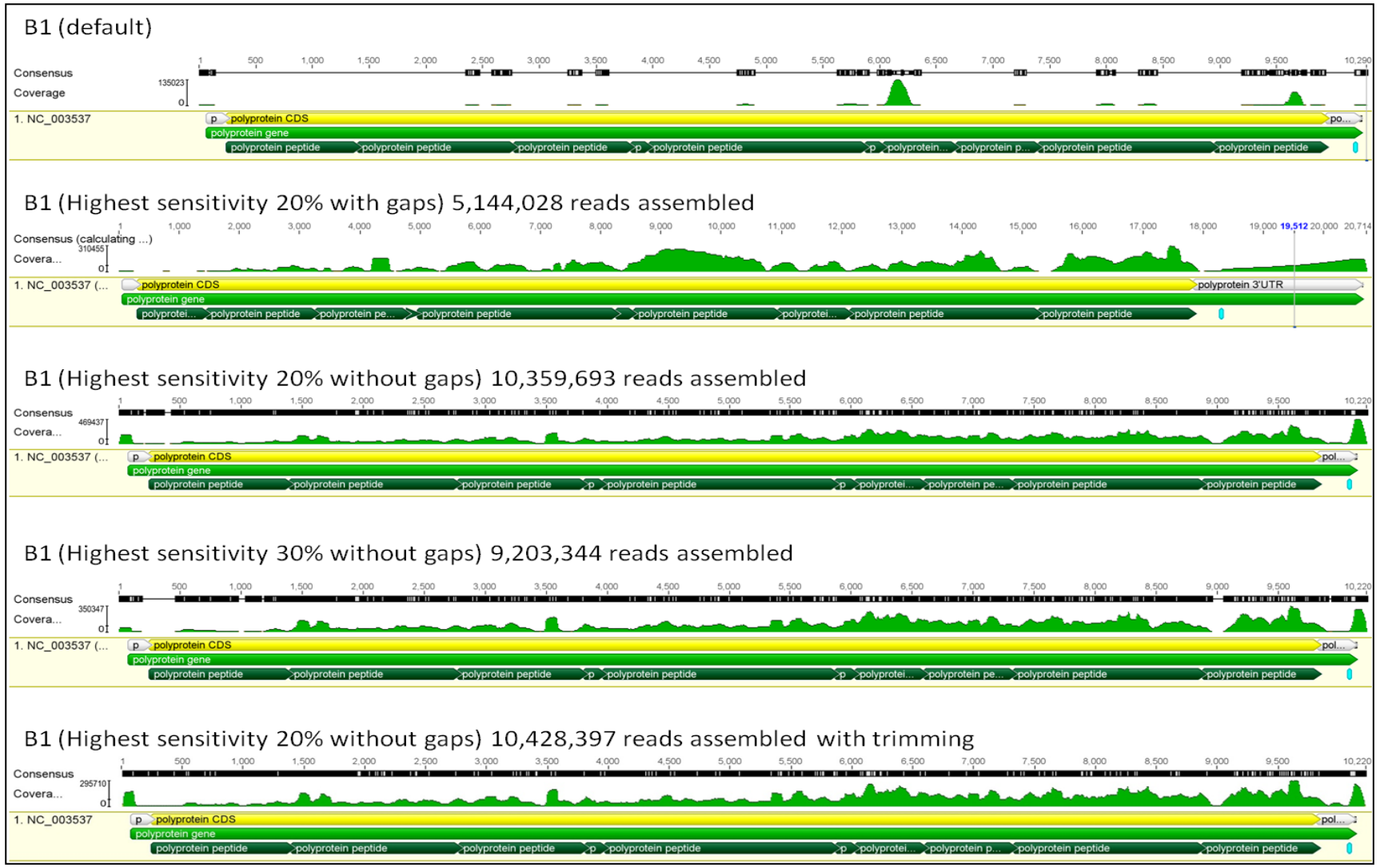


Figure 3.13:Comparison of parameters for assembly of DsMV-B1. The published DsMV genome sequence (Accession no: NC_003537) was used as the reference genome for each assembly.

### 3.3.4 Assembly of DsMV genomes

The genome of each DsMV strain and isolate was assembled using the optimal conditions, that is, trimming the 3'ends of the reads and using the highest sensitivity without gaps with a mismatch score of 20% for assembly preparation. GeneiousPro used around 35% of the total reads from each paired library for creating each assembly; however, only ~14% of those 35% reads were successfully assembled to the DsMV genome. Therefore, based on this finding, it was estimated that 5% of the sequences within each library represents DsMV. In a similar study, Wylie and Jones (2011) found that out of the total reads only ~7% of the reads belonged to PWV (Wylie & M. Jones, 2011). Figure 3.13 shows that the assemblies created using the optimal conditions showed a uniform and a higher read coverage than using the default assembly parameters. Still, within each library there were certain regions of the DsMV genome that had lower read coverage, such as the 5'UTR, P3, N-terminal of CP and the 3'UTR. It was observed that the coverage pattern was consistent between isolates of the same strain, but was different between the strains. DsMV-NZ strain and its isolates showed an overall higher coverage than DsMV-B's.

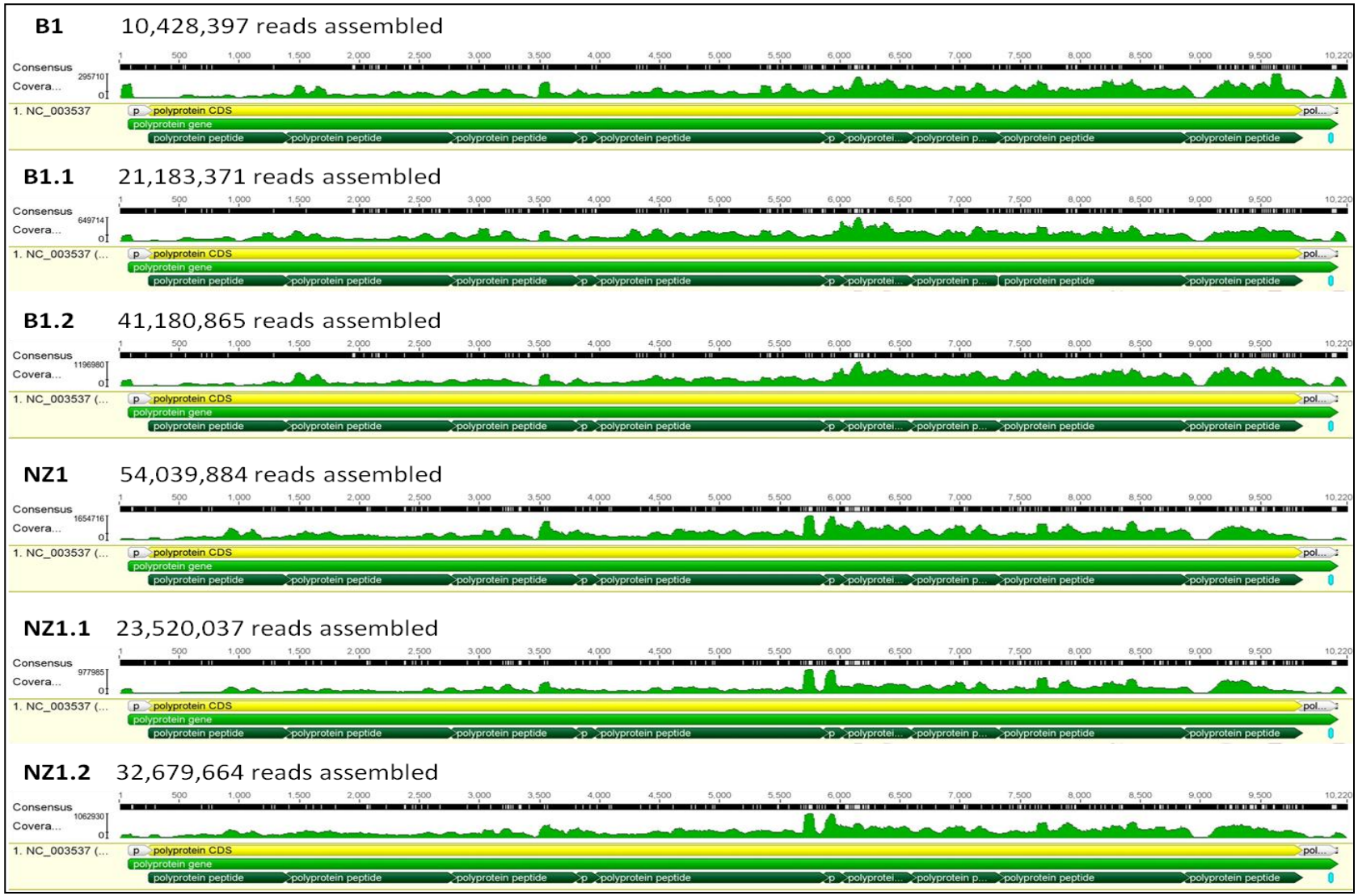


Figure 3.14: The assembly of each sample prepared in GeneiousPro using highest sensitivity without gaps with a mismatch score of 20% after trimming the 3'end of the reads.

### 3.3.5 Phylogenetic analysis of DsMV

The evolutionary relationships of each DsMV sequence were determined. Gibbs et al. (2008) reported that DsMV is a species within the BCMV group of potyviruses. The relatedness of the DsMV sequences from this study with other members of the BCMV group was determined by multiple sequence alignment of all available full-length genome sequences and phylogenetic analysis. Each consensus nucleotide sequence was extracted from their respective assemblies and aligned with full length genomes of other potyviruses belonging to the BCMV (Gibbs et al., 2008). Figure 3.15 shows a screenshot of the multiple alignment generated with GeneiousAlign in GeneiousPro ver 5.5 (Drummond et al., 2011). Figure 3.14 shows, that the average genome length of the potyviruses belonging to the BCMV was around 10-11 kb and shows a high level of sequence conservation. It was also found that the 5'end of the genome was variable in comparison to other regions in the alignment (figure 3.14).

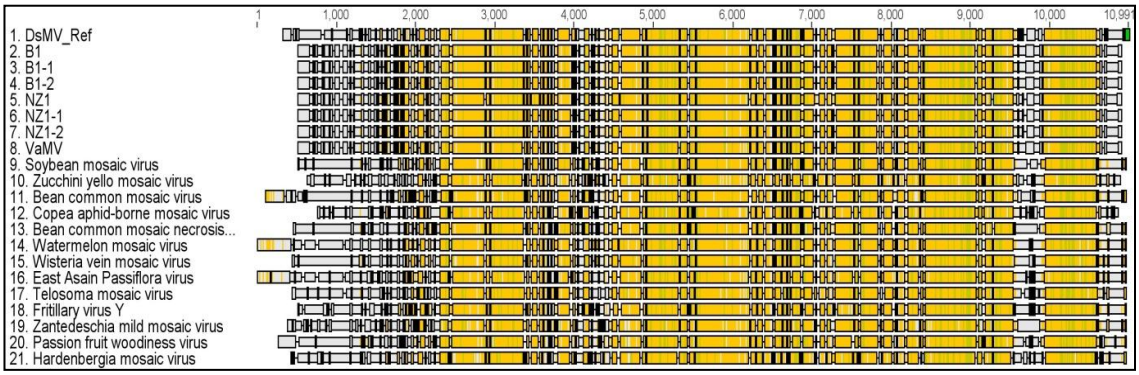


Figure 3.15: Screenshot of the alignment of each consensus sequence with other sequences belonging to the BCMV lineage

Phylogenetic analysis of this alignment was carried out using neighbour joining (NJ) and maximum likelihood (ML). Based on the phylogenetic analysis of potyviruses using the CP sequences, potyviruses has been classified under four phylogentically distinct groups namely, the PVY group, the SCMV group, the BYMV group, and the BCMV group (Gibbs & K. Ohshima, 2010). Therefore, for this study BtMV (Accession no: NC_005304) and PVY (NC_001616) was used as the outgroup because they belonged to different group of potyviruses and were phylogenetically distinct from the BCMV lineage.

Both trees showed that all viruses within the BCMV group formed a monophyletic group, indicating that they all have a common ancestor (figure 3.15 and 3.16). All DsMV sequences were most closely related to each other and to VanMV than to other potyviruses belonging to the BCMV group. It was also observed that the level of nucleotide variation was greater between DsMV-B's than DsMV-NZ. DsMV-B appeared before DsMV-NZ and therefore has had longer to develop greater diversity within its population. VanMv also appeared to be closer to DsMV than other BCMV species, supporting it being a strain of DsMV (Farreyrol, Pearson, Grisoni, Cohen, & Beck, 2006). Two major lineages were observed in BCMV, group 1 including DsMV and VanMV, while the other included all other BCMV viruses. This may be due to their being more DsMV sequences than any other (Gibbs et al., 2008).

NJ and ML analyses showed similar branching patterns. Major branching patterns were well supported with bootstrap values above 50 in NJ true. In the ML tree, some values were below 50, suggesting little support for these branch patterns. The difference in the branching pattern of both the trees was due to the difference in their algorithms. NJ is based on the Jukes-Cantor model of sequence evolution that assumes that each site evolves independently of each other and every base has equal probability of occurrence. Whereas ML uses the HKY model for tree building, this is a generalised form of the Kimura model. This model distinguishes between transition/transversion mutational events (Allen & Salzberg, 2006). Therefore, subtly different branching patterns were expected for these trees but the major groupings were the same in both the trees leading to a same conclusion that all DsMV was a member of BCMV group.

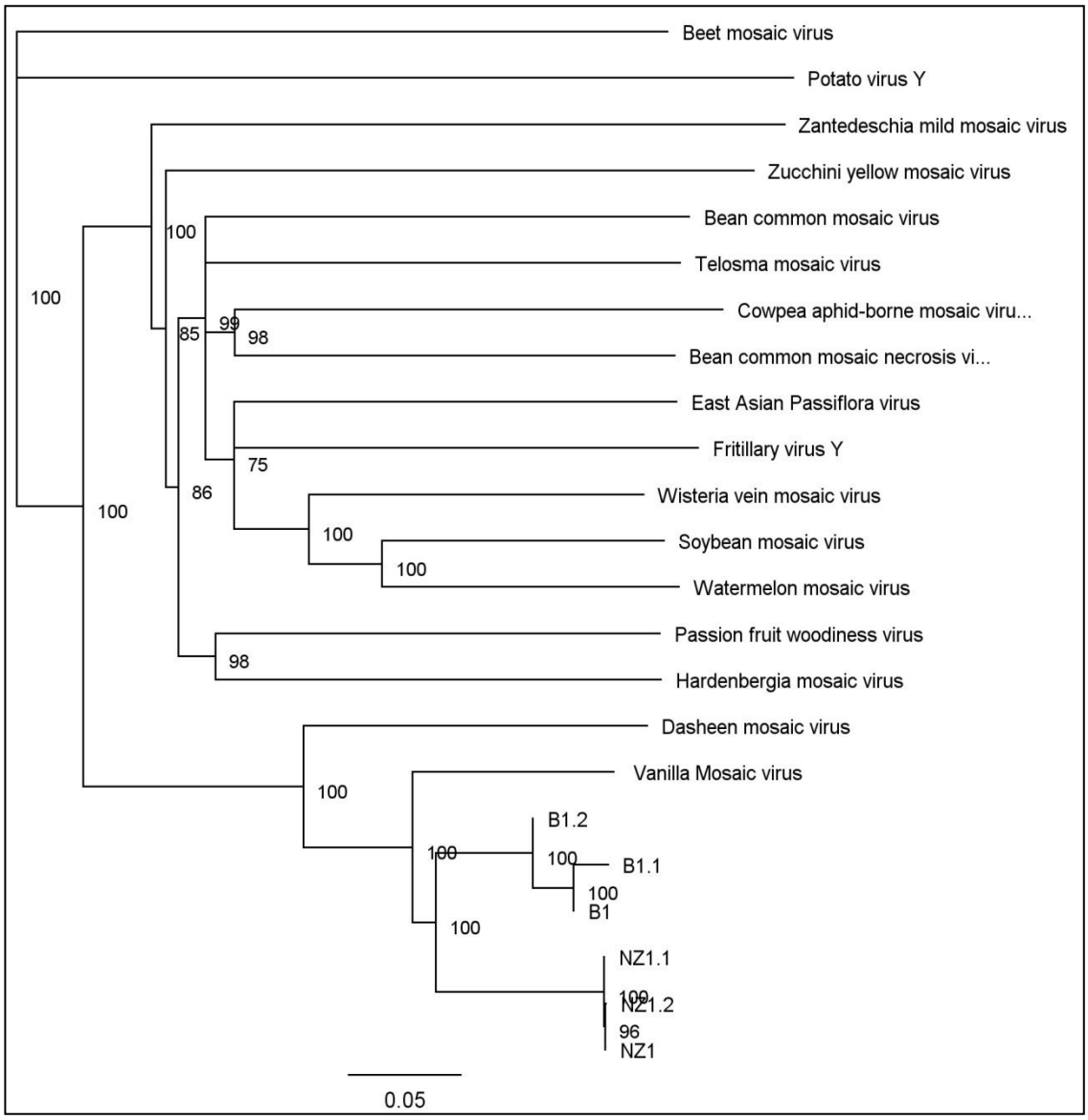


Figure 3.16: Neighbour Joining (NJ) tree consisting of consensus sequences of all the DsMV samples and publicly available whole genome sequences of BCMV lineage, using Beet mosaic virus and Potato virus Y as an outlier. Scale bar represents the number of nucleotide substitutions per site. See appendix for multiple alignment.

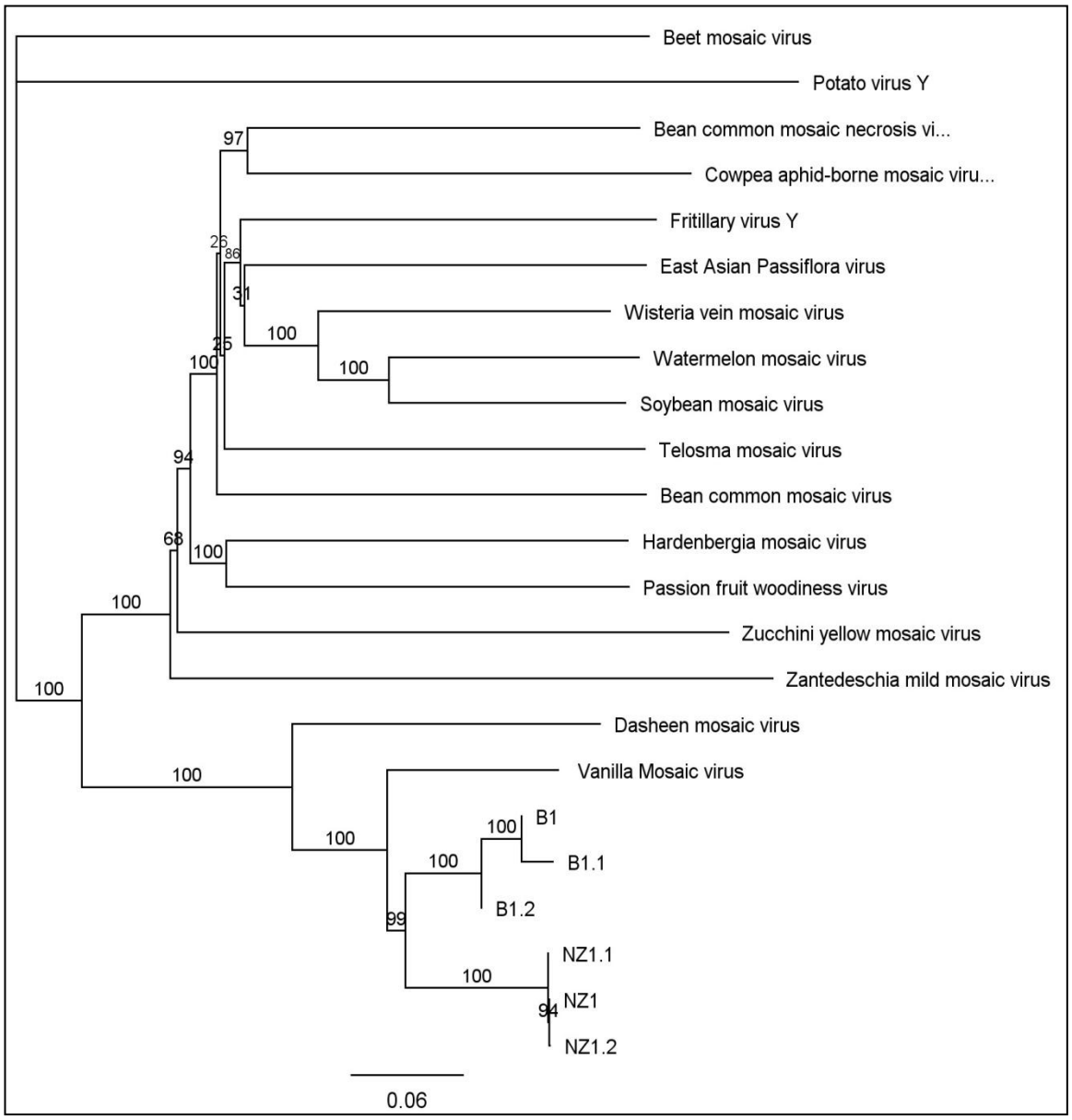


Figure 3.17: Maximum likelihood (ML) tree of DsMV consensus sequences and publicly available whole genome sequences of BCMV lineage. Beet mosaic virus and Potato virus Y were used as outgroup. Scale bar represents the number of nucleotide substitutions per site. See appendix for multiple alignment.

### 3.3.6 Variation detection

The alignment used for carrying out the phylogenetic analysis of DsMV, was further used for calculating the percentage identities between all the sequences included in the alignment. The distance calculator tool embedded in GeneiousPro alignment tool computed the overall percentage nucleotide variation within and between each strain/isolate of DsMV. It also calculated the overall nucleotide identity within and between each sample (table 3.4).

Table 3.4 shows that all the sequences selected for phylogenetic analysis belonged to the same genus of virus, that is, *Potyvirus.* This conclusion was based on the percentage identities between the sequences; it was found that all the sequences were under the nucleotide threshold cut-off limit for genera classification, i.e. >46% nucleotide identity (Adams, Antoniw, & Fauquet, 2005b). Table 3.4 also shows that all the six samples used in this study belonged to the same species of virus, as the nucleotide identities between the six samples was above the 76% cut-off limit. Due to the low nucleotide identities between the BCMV group and BtMV (51%) and the BCMV group and Potato Virus Y (48%), it was concluded that were genetically distinct from the BCMV group. In addition, the low nucleotide identity (48%) between BtMV and PVY showed that they were also genetically distinct from each other.

In order to determine the overall level of genetic variation within and between the samples, a multiple sequence alignment was created in GeneiousPro that consisted of DsMV reference sequence and the consensus sequence of each sample generated from their respective assemblies. As all the samples used in this study was DsMV, it was found that each sequence was of approximately the same length i.e. ~10 kb and an overall high level of sequence conservation was observed across the DsMV genome, with some variation observed at the 5'end and the CP regions of the DsMV genome (figure 3.17).

Table 3.4: Percentage identities between the sequences in alignment of each consensus sequence with other sequences beloning to the BCMV lineage, using BtMV and PVY as outliers.

| | DsMV Ref | B1 | B1-1 | B1-2 | NZ1 | NZ1-1 | NZ1-2 | VarMV | BCMV | SMV | ZYMV | CAMV | BCMNV | WMV | WVMV | EAPV | TeMV | FVY | ZaMMV | PWV | HarMV | BtMV | PVY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B1 | 76.3% | | | | | | | | | | | | | | | | | | | | | | |
| B1-1 | 76.2% | 96.8% | | | | | | | | | | | | | | | | | | | | | |
| B1-2 | 76.2% | 95.1% | 93.9% | | | | | | | | | | | | | | | | | | | | |
| NZ1 | 76.4% | 85.8% | 86.1% | 85.0% | | | | | | | | | | | | | | | | | | | |
| NZ1-1 | 76.6% | 85.8% | 86.1% | 85.0% | 99.2% | | | | | | | | | | | | | | | | | | |
| NZ1-2 | 76.6% | 85.7% | 86.1% | 85.0% | 99.3% | 99.5% | | | | | | | | | | | | | | | | | |
| VarMV | 76.1% | 82.5% | 82.3% | 82.3% | 82.8% | 82.9% | 82.9% | | | | | | | | | | | | | | | | |
| BCMV | 60.7% | 59.1% | 58.9% | 59.0% | 59.0% | 59.0% | 59.0% | 58.4% | | | | | | | | | | | | | | | |
| SMV | 60.8% | 59.0% | 58.9% | 59.1% | 58.9% | 58.9% | 58.9% | 58.6% | 63.6% | | | | | | | | | | | | | | |
| ZYMV | 58.4% | 57.8% | 57.6% | 57.8% | 57.5% | 57.5% | 57.5% | 57.1% | 59.9% | 61.4% | | | | | | | | | | | | | |
| CAMV | 59.1% | 57.8% | 57.7% | 57.8% | 57.7% | 57.7% | 57.7% | 57.8% | 61.9% | 64.5% | 60.6% | | | | | | | | | | | | |
| BCMNV | 60.1% | 58.5% | 58.3% | 58.5% | 57.9% | 58.0% | 58.0% | 57.9% | 63.4% | 66.4% | 61.3% | 65.5% | | | | | | | | | | | |
| WMV | 59.7% | 58.6% | 58.4% | 58.6% | 58.4% | 58.4% | 58.3% | 58.1% | 65.7% | 73.6% | 59.3% | 62.1% | 63.7% | | | | | | | | | | |
| WVMV | 60.7% | 60.0% | 59.9% | 60.1% | 60.0% | 59.9% | 59.9% | 59.4% | 64.8% | 72.3% | 61.1% | 64.8% | 66.2% | 70.2% | | | | | | | | | |
| EAPV | 59.8% | 58.4% | 58.2% | 58.5% | 58.7% | 58.7% | 58.7% | 57.8% | 64.5% | 65.0% | 59.6% | 61.8% | 63.7% | 65.9% | 66.2% | | | | | | | | |
| TeMV | 59.8% | 58.2% | 58.1% | 58.4% | 58.2% | 58.2% | 58.2% | 57.8% | 63.5% | 66.2% | 61.6% | 63.5% | 65.7% | 63.5% | 66.2% | 63.8% | | | | | | | |
| FVY | 59.7% | 58.1% | 57.9% | 57.9% | 57.9% | 57.9% | 57.9% | 57.6% | 63.4% | 66.9% | 60.8% | 62.9% | 65.3% | 64.7% | 66.9% | 64.2% | 65.7% | | | | | | |
| ZaMMV | 59.2% | 58.3% | 58.3% | 58.3% | 58.1% | 58.1% | 58.1% | 57.2% | 58.7% | 59.4% | 57.9% | 58.6% | 59.3% | 58.5% | 60.6% | 59.2% | 58.7% | 59.2% | | | | | |
| PWV | 60.8% | 58.9% | 58.7% | 59.0% | 59.1% | 59.2% | 59.2% | 58.8% | 62.7% | 65.1% | 62.6% | 63.5% | 64.1% | 62.4% | 65.8% | 63.0% | 64.4% | 64.3% | 60.0% | | | | |
| HarMV | 60.6% | 59.5% | 59.4% | 59.3% | 59.0% | 59.1% | 59.1% | 58.8% | 62.8% | 65.2% | 61.4% | 62.9% | 64.0% | 62.0% | 65.7% | 62.4% | 63.7% | 64.0% | 60.2% | 66.8% | | | |
| BtMV | 53.1% | 52.0% | 52.2% | 51.9% | 52.2% | 52.2% | 52.2% | 51.3% | 53.4% | 54.9% | 53.5% | 54.6% | 54.3% | 53.0% | 54.9% | 53.0% | 53.9% | 54.2% | 51.9% | 54.6% | 53.8% | | |
| PVY | 50.4% | 49.2% | 49.1% | 49.3% | 49.2% | 49.2% | 49.1% | 48.6% | 49.1% | 50.9% | 50.4% | 50.7% | 51.6% | 49.2% | 51.4% | 49.3% | 50.3% | 50.5% | 48.9% | 51.1% | 51.4% | 50.7% | |

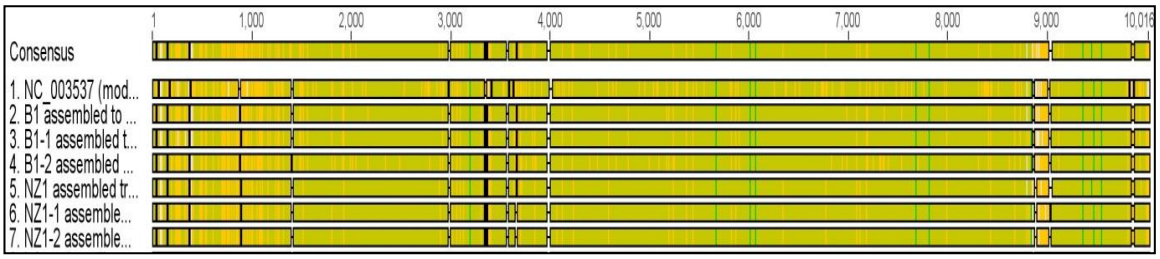


Figure 3.18: Snapshot of multiple sequence alignment consisting of DsMV reference sequence and consensus sequence of B1, B1.1, B1.2, NZ1, NZ1.1 and NZ1.2 extracted from their respective assemblies. See appendix for multiple alignment

Table 3.5: Percentage genetic variation (above the diagonal) and percentage sequence identity (below the diagonal) identified between different samples.

% Variation ←

| | **DsMV Ref** | **B 1** | **B 1.1** | **B 1.2** | **NZ 1** | **NZ 1.1** | **NZ 1.2** |
|---|---|---|---|---|---|---|---|
| **DsMV Ref** | | 21.5% | 21.6% | 21.6% | 21.0% | 20.9% | 21.0% |
| **B 1** | 78.5% | | 3.5% | 5.1% | 13.8% | 13.8% | 13.8% |
| **B 1.1** | 78.4% | 96.5% | | 6.2% | 13.4% | 13.4% | 13.3% |
| **B 1.2** | 78.4% | 94.9% | 93.8% | | 14.5% | 14.4% | 14.5% |
| **NZ 1** | 79.0% | 86.2% | 86.6% | 85.5% | | 0.6% | 0.4% |
| **NZ 1.1** | 79.1% | 86.2% | 86.6% | 85.6% | 99.4% | | 0.5% |
| **NZ 1.2** | 79.0% | 86.2% | 86.7% | 85.5% | 99.6% | 99.5% | |

→ % identity

Table 3.5 shows, that when compared to the DsMV reference genome, B1.1 and B1.2 isolates of DsMV showed the highest percentage of overall genetic variation i.e. 21.6% while NZ1.1 showed the least amount of variation i.e. 20.9%. The overall level of variation determined in other samples, NZ1, NZ1.2 and B1 were 21.0%, 21.0%, and 21.5%, respectively. It was also found that the level of variation within the isolates of same strain was lower than level of variation found between the strains. The variation between the B and NZ strains was 13.3-14.5%. This was less than the variation from the DsMV reference genome, supporting the more recent common ancestry for the B and NZ strains (figures 3.15 and 3.16). It was also found that the level of variation within the B strain (3.5%-6.2%) was greater than that observed within the NZ strain (0.4-0.6%). This supports the phylogenetic analysis, which showed that the B strain emerged

earlier from the common ancestor than NZ, indicating that DsMV-B strain had longer time to generate more sequence variants.

SNP finder tool embedded in GeneiousPro ver 5.5 was used to identify all SNP's present along the length of each assembled genome (Drummond et al., 2011). The level of variation within each gene was then determined for all the genomes. Table 3.6 shows the percentage genetic variation identified within each gene for each genome assembled in this study.

Table 3.6: Percentage genomic variation identified for each gene of the samples using consensus sequences compared with reference genome (Adams et al., 2005b)

| | **NZ1** | **NZ1.1** | **NZ1.2** | **B1** | **B1.1** | **B1.2** | **Variation for species discrimination** |
|---|---|---|---|---|---|---|---|
| **5'UTR** | 30.8 | 30.5 | 30.2 | 23.7 | 27.5 | 23.7 | - |
| **P1** | 30.8 | 30.7 | 30.8 | 30.9 | 32 | 31.8 | 42% |
| **HC-Pro** | 18.9 | 18.9 | 18.2 | 19.5 | 19 | 19.5 | 24% |
| **P3** | 22.1 | 22 | 22.1 | 24.1 | 24.4 | 23.7 | 26% |
| **PIPO** | 16.1 | 16.1 | 16.1 | 17.2 | 17.2 | 16.9 | - |
| **6K1** | 19.2 | 19.2 | 19.2 | 20.5 | 15.4 | 20.2 | - |
| **CI** | 18.4 | 18.4 | 18.4 | 18.7 | 18.9 | 18.7 | 21.7% |
| **6K2** | 20.1 | 20.1 | 20.1 | 23 | 23.9 | 23.3 | - |
| **NIa-VPg** | 15.8 | 15.8 | 15.9 | 16.8 | 16.8 | 16.7 | 24% |
| **NIa** | 20.3 | 20.3 | 20.3 | 20.5 | 20.6 | 20.5 | 23.5% |
| **NIb** | 19.9 | 19.9 | 20 | 20.9 | 20.2 | 21 | 25% |
| **CP** | 19.1 | 18.1 | 18.9 | 17.9 | 19 | 18.4 | 24% |
| **3'UTR** | 28.5 | 28.5 | 28.7 | 28.8 | 28 | 28.1 | - |

The level of variation identified within each gene was consistent in both strains of DsMV. Table 3.6 shows that the pattern of variation across all the genomes was similar and the level of variation for most of genes was about 15.0%-30%. The 5'UTR and 3'UTR showed high variation at 23.7-30.8% and 28.0-28.8%, respectively. Within the protein-coding region, P1 showed the highest variation at 30.7-31.8%, while the NIa-VPg gene showed the lowest variation at 15.8-16.8% (figure 3.18). According to Adam et al. (2005), the optimal nucleotide identity threshold (% identity) for species discrimination within genes of potyviruses are as follows, 58% (P1 gene), 76% (HC-Pro gene), 74% (P3 gene), 78.3% (CI gene), 76% (NIa-VPg gene), 76.5% (NIa-Pro gene),

75% (NIb gene), and 76% (CP gene). Comparing these values to the variation identified within each gene of the genomes, it was found that the variation observed within each gene was below the threshold limit; therefore, it was concluded that all the samples used in this study belong to same species and were different strains/isolates of DsMV.

The figure 3.18 show the graphical representation of variation identified within each DsMV genomes. Each point represents the average variation for DsMV-B1, B1.1, and B1.2 compared with that for DsMV-NZ1, NZ1.1 and NZ1.2. The variation in each region is similar, with overlapping values except for the 5'UTR and 6K2. Thus, the variation within the 5'UTR between DsMV-NZ1, NZ1.1 and NZ1.2 was greater than that observed for the DsMV-B sequences, while they showed less variation in the 6K2 region.

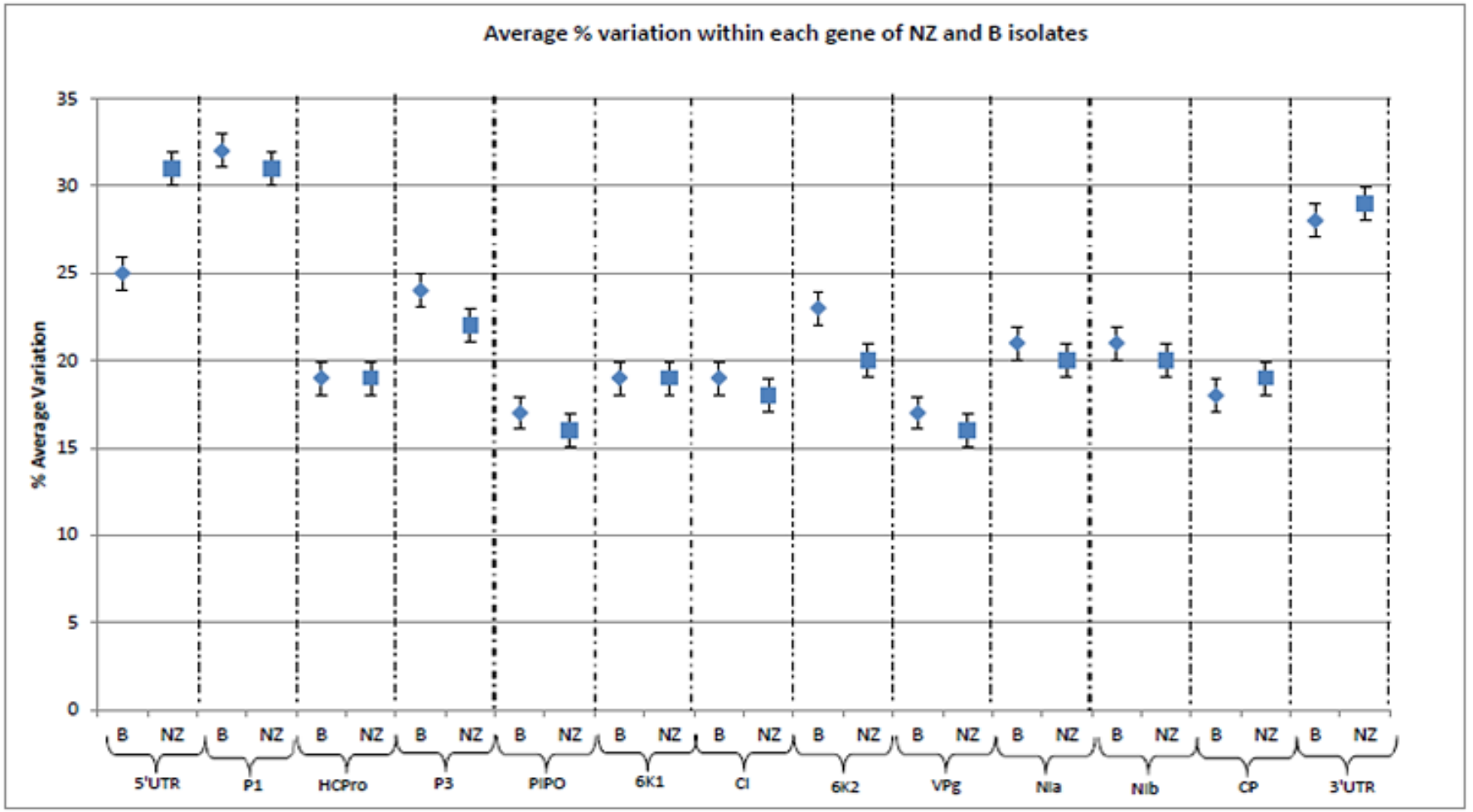


Figure 3.19: Plot showing the percentage average variation identified within DsMV-B and DsMV-NZ strains/isolates. X-axis represents the genes present within the DsMV-B and DsMV-NZ genomes and Y-axis represents percentage average variation. Blue boxes represent the mean value.

### 3.3.7 Summary of results

The aim of this study was to determine the level of variation within and between the DsMV-B and DsMV-NZ strains/isolates. The total RNA extracted from taro leaves infected with DsMV-B and DsMV-NZ was sent for Illumina deep sequencing. The Illumina sequencing data was analysed using GeneiousPro bioinformatics software. During the analyses, it was found that the overall level of variation within isolates of same strain was lower than the variation between the strains. It was also found that the variation within B's was higher than NZ. The phylogenetic analysis using the full-length genome sequences confirmed that DsMV belongs to the BCMV lineage. It was also concluded that the B and NZ strains share a common ancestor. The level of variation within each gene of the DsMV genomes revealed that the variation was consistent across B's and NZ strain. It was also found that within the coding region, P1 gene showed the highest level of variation, around 30%. However, the level of variation observed in all the genes including P1 was within the species cut-off limit. The variation identified within 5'UTR was higher in DsMV-NZ isolates than DsMV-B, while the variation identified within the 6K2 gene was higher in DsMV-B isolates than that of DsMV-NZ isolates. The level of variation identified within each gene was below the threshold variability generally observed among potyviruses. This study confirmed that variability is an important characteristic of potyviruses and certain level of variability can be accommodated within each gene of potyviruses. The knowledge of this genetic plasticity is important for understanding the functionality of Potyviral genome and can help in developing better diagnostics that could aid in accurate and precise detection of Potyviral infection among agricultural important plants.

## 3.4 Discussion

The genus *Potyvirus* is one of the largest groups of RNA viruses (McDonald et al., 2010). These viruses do not possess a single representative genome; instead, they have a set of related variants occurring as quasi-species. These genetic variants co-exist within a population and may have a significant effect on the infectivity or the host selection process among other effects of *Potyviruses* (Stewart, Watts, & Litwin, 2001). Therefore, it is important to study the genetic diversity of viruses to understand the host selection process (i.e. host specificity) of viruses, infectivity, replication, movement etc. (Drake & Holland, 1999).

The advancements in the field of next generation sequencing (NGS) have allowed transcriptome analysis (Glenn, 2011; Voelkerding et al., 2009). Massively parallel (also known as NGS) technologies allow the profile of all genetic variants of viruses to be identified within a host plant (Tucker, Marra, & Friedman, 2011; Voelkerding et al., 2009). There are several NGS platforms such as Illumina, SOLID and 454 pyrosequencing etc. However, due to the broadest utility, lowest cost per read and overall lower error rate ($\leq 0.1\%$) (Glenn, 2011) the Illumina sequencing platform was selected for this study. The aim of this study was to understand the extent and pattern of genetic diversity among strains and isolates of the potyvirus *Dasheen mosaic virus* (DsMV). Understanding the pattern of diversity was also expected to allow the discovery of conserved regions within the DsMV genome that may provide insights into potyvirus translation and replication. In this study, two different strains of DsMV, NZ1 and B and the isolates NZ1.1, NZ1.2, B1.1 and B1.2 were analysed. Total RNA extracted from taro plants infected with these viruses were sent to Macrogen for sequencing using the Illumina deep sequencing platform.

### 3.4.1 Library Quality

FastQC software was used to determine the quality of reads generated by Illumina sequencing platform from each library. The quality of each library was assessed by FastQC analysis. This analysis was based on several predefined parameters such as per-base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and K-mer

content. While for some parameters anomalies were identified within the sequence libraries, those were generally expected for RNA sequencing due to the primers and adapters used during this sequencing (Andrews, 2011; Schwartz et al., 2011). Overall FastQC analyses showed that all the sequence libraries were above the threshold quality score and therefore could be used for this study.

### 3.4.2 Genome assembly

Paired end libraries were generated for each virus and these were linked and assembled against the DsMV reference sequence, this was done to separate viral RNA sequences from the host plant RNA. Due to the low coverage of reads across the reference, each assembly was optimised to get a uniform read coverage with minimal gaps. After a series of optimising steps, it was concluded that after trimming the reads from the 3' end, using the highest sensitivity without gaps and percentage mismatch score of 20% was optimal for library preparation. The reason for trimming the 3' end of reads was due to the low quality score towards the 3' end of the reads, this is because of the decrease in base-call accuracy with increasing read length (Voelkerding et al., 2009). This pattern was observed mainly due to "dephasing-noise" (Erlich, Mitra, Delabastide, McCombie, & Hannon, 2008). During a given sequencing cycle, nucleotides can be under- or over-incorporated. With successive cycles, these aberrations lead to the accumulation of reads of varying lengths. This heterogeneity in the read length decreases the signal purity and reduces precision in base calling, especially at the 3' end of reads (Erlich et al., 2008).

Even after optimising each assembly, low read coverage was observed towards the 5' end in comparison to the 3' end. This bias may have been due to the formation of secondary structures of the template RNA used for generating the initial PCR products. As a result, shorter reads are produced by GC- rich sequences which leads to variability in read coverage against the reference genome (Astrovskaya et al., 2011). The other reasons that contribute to this coverage variability include differential ligation of adapters to template sequences and differential amplification during clonal template generation (Bentley et al., 2008; Hernandez, Francois, Farinelli, Osteras, & Schrenzel, 2008). In addition, very small gaps were observed within the assemblies, these gaps in the coverage were due to high level of sequence variation between reads and the DsMV

reference sequence. Therefore, to remove this bias a minimum of 10 bases were trimmed from each read (Voelkerding et al., 2009).

On average 5% of reads from each sequence library assembled against the DsMV reference genome (Accession no: NC_003537). This pattern was expected because total RNA was used for sequencing and therefore majority of RNA extracted from each plant was plant RNA and only a small proportion of total RNA would represent viral RNA (Wylie & M. Jones, 2011). On average, the overall read coverage was higher in DsMV-NZ1 than DsMV-B, with a few exceptions, for instance the read coverage in the CP and 3'UTR regions of the DsMV reference genome was higher for B strain in comparison to NZ strain. However, a significant number of reads remained unaligned due to differences between the reference genome and sequences in the viral sample. Therefore, for a better representation of all the reads aligned within each assembly, a consensus genome sequence for each sample was generated from the aligned reads.

### 3.4.3 Phylogenetic analysis

The consensus sequence of B1, B1.1, B1.2, NZ1, NZ 1.1 and NZ1.2 were used for phylogenetic analysis of DsMV. Based on the amino acid composition of virion coat protein (CP), potyviruses are classified under four groups: the PVY group, SCMV group, the BYMV group, and the BCMV group (Gibbs & K. Ohshima, 2010). However, many relationships among potyviruses remain unclear because the majority of phylogenetic studies have been carried out using the CP sequences. This only provides limited phylogenetic information because it is only ~750 nucleotides long (A. Gibbs & K. Ohshima, 2010). Another reason is the CP region is highly variable towards its N-terminal domain (Shukla et al., 1991). The CP is of different lengths in different species and evolves in a hierarchical way by point mutations, often by repetitive duplication (Shukla & Ward, 1988; Ward, Weiller, Shukla, & Gibbs, 1995). The N-terminal end of the CP gene is highly repetitive, whereas that encoding the core and C-terminal of the CP is not, and seems to have evolved in a coherent hierarchical way by point mutations and by occasional homologous recombination. This region is called the "coherently-evolving CP" (cCP) region and only this region of the CP gene is used for phylogenetic studies (A. Gibbs, Ohshima, Phillips, & Gibbs, 2008).

Therefore, in this study full genome sequences available in the public domain were used for carrying out phylogenetic analysis of DsMV. Gibbs and Gibbs (2008) carried out an extensive phylogenetic analysis of the BCMV group using the CP gene sequences and classified DsMV as a member of the BCMV group. Thus, the consensus genome sequence generated for each sample was used to reaffirm the phylogeny of DsMV by both neighbour joining (NJ) and maximum likelihood (ML). This study confirmed the results of Gibbs and Gibbs (2008) showing that DsMV is a member of the BCMV group, since these sequences and other members of the group shared a common ancestor. The results also confirmed the findings of Chang (personal communication, 2012), who studied a 327 bp region of the CP gene and concluded that the NZ and B sequences had a common ancestor. While the NZ strain was first identified in New Zealand, Chang showed that these sequences probably originated in Samoa. The results also showed that *Vanilla mosaic virus* (VanMV) shares a common ancestor with DsMV. This finding supports the hypothesis that VanMV is a strain of DsMV that exclusively infects vanilla (Farreyrol et al., 2006). With more complete viral genomes, phylogenetic analysis of the Potyviridae will provide important information about the evolution and relatedness of genomes overall.

### 3.4.4 Genetic variation

On average DsMV-B and DsMV-NZ showed a percentage genetic variation of 21.5% and 20.9%, respectively, from the DsMV reference genome. This level of genetic variation from the reference genome could be influenced by the different host and geographic origins of these viruses. As the reference sequence was generated by sequencing the DsMV isolated from *Philodendron selloum* (China) while the DsMV-NZ and DsMV-B were isolated from taro. It was also found that there was significant genetic variation between DsMV-B and DsMV-NZ. However, the level of variation identified among the isolates of the same strain was not as great. According to Adams et al. (2005b), currently accepted values that discriminate between the species are 75-76% nucleotide identity across the length of the genome and any value greater than this represents a comparison between different strains of the same species. Therefore, the individual percentage identity shown by all the genomes with the DsMV reference fell

within the threshold range and could be classified as different strains of DsMV. The result of this study confirmed that B and NZ are two distinct strains of DsMV.

One aim of this study was to determine the pattern of genetic variability among potyviruses, the pattern of variability observed within B and NZ strains of DsMV suggests that potyviruses undergoes finite mutation. This means that even though different potyvirus species show high levels of mutation, there is an evolutionary pressure to maintain the structural and functional features of the potyviral genome. Several experimental results also support the concept that RNA viruses replicate with copying fidelity that is balanced near the error threshold (Domingo & Holland, 1997). Therefore, potyviruses maintain an error threshold during their replication process, otherwise the genetic information will be irreversibly lost and this can be critical for the proper functioning of the potyviral genome (Domingo & Holland, 1997).

Bioinformatics analysis was also carried out to determine the genetic variation within each gene of the samples with respect to the DsMV reference genome. This was done to develop a better understanding of the functions of different genes of DsMV and to help in identification of conserved regions within the DsMV genome. The observed pattern for gene level variation was similar between DsMV-B and DsMV-NZ, except for 5'UTR and 6K2 gene. The variation found within each gene was compared with the threshold value (% identity) for species discrimination in the different genes of potyviruses and it was observed that the percentage of identity within each gene of the DsMV was above the threshold limit. This analysis again confirmed that all the samples tested in the study were DsMV (Adams et al., 2005b). In this discussion, the genetic variability identified within the UTRs, P1 gene, 6K2 gene and NIa-VPg gene are discussed in detail.

In both B and NZ, the 5'UTR, P1 and 3'UTR showed the highest level of variation from the reference, that is average ~27.5%, 30% and 28%, respectively. In contrast to P1 gene, HC-Pro, 6K1, CI, NIb, NIa and CP gene showed relatively lower level of genetic variation and NIa-VPg showed the lowest level of genetic variation in comparison to other genes. In this study, high level of genetic variation was also determined for both the 5' and 3' UTR. However, few conserved regions were observed within both the UTRs. There is speculation that the UTRs may be involved in potyviral replication. For instance studies have suggested that the 5'UTR may have a role in stimulating the interaction between the host cell ribosomes and viral RNA in order to

access the translation machinery of the hosts (Lewsey & Carr, 2009). Similarly, the highly variable 3'UTR also possesses some regions of conservation and there is speculation that these regions may have some important role potyviral translation. For instance, the 3'UTR of the potyvirus genome forms secondary structures resembling tRNA, which seem to be associated with translation of viral RNA (Lewsey & Carr, 2009). It may also play an important role in RNA replication (Lewsey & Carr, 2009). The level of variation identified in the 3' UTR for both DsMV-B and DsMV-NZ was ranging between 28%-28.8%. In the 5'UTR, the average variation identified in the DsMV-NZ (30%) was higher than the variation in the DsMV-B's (23%). The difference in the level of genetic diversity between two strains could be a result of low read coverage in the 5'UTR of DsMV-B, which would mean that due to the low read coverage, all the SNPs in the 5'UTR of DsMV-B were not identified (Olivares, 2011). Due to computational limitations of GeneiousPro, it was not possible to determine the exact number of reads assembling within each gene of the genome assemblies (Shane Sturrock, personal communication, 2012). Therefore, it was not confirmed, whether the difference in the level of variation in the 5'UTR of two strains was real or was due to difference in the coverage.

Within the coding region, P1 gene showed the highest genetic variation (~30%) of all genes present in the DsMV genome, and the variation was consistent across different isolates of DsMV. The P1 gene is located at the N-terminal of the potyviral genome and encodes for the P1 protein, which allows its own cleavage from the polyprotein and exhibits single-stranded RNA binding activity *in*-vitro (Verchot & Carrington, 1995b). P1 proteinase from different potyvirus species has been found to be highly variable in its size, ranging from 30 kDa to 60 kDa (Verchot & Carrington, 1995b). In this study, the P1 gene was found to have a high level of genetic variation i.e. average ~30%, which is consistent with the variability observed within this region in PVY and *Zucchini yellow mosaic virus* (Tordo et al., 1995; Wisler, Purcifull, & Hiebert, 1995). The P1 protein is considered the least conserved protein among potyviruses (Valli, López-Moya, & García, 2007), except for the conserved amino acids found at the C-terminus of all potyvirus P1 proteins, which is the proteinase catalytic domain. The nucleotide alignment between the consensus P1 gene sequence of the samples and the DsMV reference P1 sequence showed that the N-terminus region of P1 gene was more variable and the level of sequence conservation was higher at the C-terminus of the P1 gene.

Deletion and mutational analysis of the P1 gene has shown that P1 is not crucial for the potyvirus infectivity, even though it enhances the viral replication and movement. However, cleavage at the boundary between P1 and HC-Pro is important for viability (Klein et al., 1994). Two P1 insertional mutants of *Tobacco vein mottling virus* (TVMV) were used to show that insertion in the C-terminus region between the conserved Asp and Ser of the catalytic triad was lethal, whereas insertional mutation at the N-terminus had no effect on virus mobility (Klein et al., 1994). In a another mutational study using the *Tobacco etch virus* (TEV), a ß-glucuronidase (GUS) insertion mutant with the entire P1 gene replaced, the results showed that the replication efficiency was highly reduced and the rate of cell to cell movement and through the host was also reduced. It was concluded that P1 helps in stimulating potyviral genome replication (Verchot & Carrington, 1995a, 1995b).

It has been established that the high rate of recombination within the P1 gene leads to diversification of P1 (Valli et al., 2007). However, the cleavage site between P1 and the HC-Pro is well conserved (Adams, Antoniw, & Beaudoin, 2005a). As discussed earlier, the C-terminal region of P1 gene is conserved, while the N-terminal is highly variable (Verchot & Carrington, 1995b). It has been speculated that the high genetic variability within the N- terminal of P1 gene helps in broadening the host range of potyviruses (Rohožková & Navrátil, 2011; Valli et al., 2007). For example, *Watermelon mosaic virus* (WMV) has a wider host range than *Soybean mosaic virus* (SMV); it is suggested that the N-terminal region of P1, the primary feature that distinguishes between these two viruses, is especially relevant for host-virus interaction (Desbiez & Lecoq, 2004). Therefore, it was suggested that high variation found in the P1 gene of each DsMV genome might be crucial for host-range selection of DsMV being species in at least 12 aroid genera (Zettler & Hartman, 1987).

The analyses of HC-Pro and CI region showed that the target regions for designing DsMV specific HC-Pro and CI primers for efficient detection and amplification of the DsMV were not as conserved as expected. The comparison of the consensus sequences of each DsMV genome with the reference genome showed that the regions used for designing the DsMV specific primers were highly variable across all the isolates, therefore specific primers cannot be used for detection of the DsMV infection of taro. However, universal primers can be efficiently used for amplification of HC-Pro and CI regions because universal primers can account for all the variability within those regions. These observations supports the phenomenon of consensus decay

described by Zheng et al. (2008) and suggests that universal primers that are currently used for detection of potyviral infection are more efficient than the specific primers. However, the target regions of these universal potyvirus primers also needs to be investigated from time to time, to check the decay in consensus (Zheng et al., 2008).

In the DsMV-B isolates, high variability (~23%) was identified in the 6K2 gene than the DsMV-NZ isolates (~20%). This difference in the variability between the two strains can be explained with low read coverage in the 6K2 region of DsMV-NZ isolates than DsMV-B isolates. It can be speculated that due to low read coverage in the 6K2 gene of DsMV-NZ isolates, GeneiousPro was not able to identify all the SNPs in that region, as a result it show low variability in the 6K2 gene of DsMV-NZ isolates (Olivares, 2011). As discussed earlier, due to computational limitations of GeneiousPro, it was not possible to determine if the coverage between the two strains was significantly different or not (Shane Sturrock, personal communication, 2012). Though the role of 6K2 protein is not clear, it has been suggested that it is a membrane binding protein. It has been proposed that the 6K2 peptide is required for viral replication and that it anchors the replication apparatus to endoplasmic reticulum (ER) -like membranes (Schaad et al., 1997). Due to these roles of 6K2 protein, it is hypothesised that high variability in 6K2 gene can interfere with the functionality of 6K2 gene, therefore, it can accommodate only a finite amount of genetic plasticity.

Within the DsMV coding region, the NIa-VPg gene showed the lowest variability, that is around 16%. The NIa proteinase catalyses most cleavage reactions of the viral protein while the N-terminal half of the NIa functions as the genome liked protein (VPg) (Riechmann et al., 1992). The VPg domain of NIa has essential functions in viral replication and host genotype specificity. The covalent attachment of VPg to the potyviral RNA has shown to involve a Tyr residue. Mutations of the Tyr linking the VPg to the viral TVMV has shown to halt the viral replication process in protoplasts (Murphy, Klein, Hunt, & Shaw, 1996; Riechmann et al., 1992). It was speculated that high genetic variability within the VPg would interfere with the process of viral replication in protoplasts and therefore, it cannot show very high sequence plasticity.

The brief outline of the functionality P1, HC-Pro, 6K2, and VPg proteins encoded by potyviral genome, helps in developing a better understanding of the functionality of the potyviral genome. The lower level of variation observed within these genes is an indicator that there is evolutionary pressure to maintain their

nucleotide sequences, which would be crucial for maintaining the functionality of the proteins encoded by potyviral genes. Therefore, the variation observed among the isolates of the same potyvirus species indicated that certain genomic regions within potyviruses are flexible and can accommodate a certain level of variability. The existence of genetic variants within a viral population increases the probability of survival and the ability to adapt to different hosts. Therefore, variability is one of the major elements of potyvirus evolution.

The aim of this study was to determine the level of variation within each strain of DsMV and comparing that variation between the strains. It was found that all the samples tested in this study were strains of DsMV. Sequence comparisons showed that the level of variation between the strains (DsMV-NZ and DsMV-B) was higher than the level of variation between different isolates. The phylogenetic analysis of DsMV using full genome sequences confirmed the placement of DsMV in the BCMV group. Studying the level of genetic variation suggested that there is a need for each gene of DsMV to maintain their functionality. The other aim of this study was to determine the regions of conservation within the potyviral genome. The reason for looking at conserved regions was to develop a better understanding of potyviral genome and to help develop better diagnostic tools for efficient detection and identification of potyviruses. In this study, few conserved domains were identified within the DsMV; however due to time constraints it was not possible to carry out an extensive analysis on the conserved regions and to compare those regions with the protein sequences of potyviruses. Therefore, in the future it would be interesting to look at the DsMV conserved regions and compare them against all the publicly available potyvirus full genome nucleotide and protein sequences, to determine if the regions identified as conserved in DsMV were also conserved across the *Potyvirus* genus or not.

# Chapter 4

# General Discussion

# 4. General Discussion

The genus *Potyvirus* is one of the largest groups of plant viruses having RNA genomes (Shukla et al., 1991). RNA viruses show high genetic diversity which can influence virus infectivity and host selection (Vignuzzi et al., 2006). Around ~30% of all known plant viruses belongs to the genus *Potyvirus* and causes significant losses in agricultural, pasture, horticultural and ornamental crops (Shukla et al., 1991). Therefore, precise and accurate detection of potyviral infection in plants is essential to ensure safe and sustainable agriculture (Jan et al., 2011). RT-PCR has emerged as one of the most sensitive techniques for detection of potyviruses (Antoniw, 1995) and the recent increase in potyvirus sequence data has permitted the design of genus-specific universal primers for the detection of new and/or uncharacterised potyviruses (Antoniw, 1995; Langeveld et al., 1991). Several studies have shown the effectiveness of group-specific universal potyvirus primers in detection and identification of potyviruses (Babu et al., 2011; Grisoni et al., 2006). One objective of this study was to test the effectiveness of universal primers in detection and identification of potyviruses. In this study, DsMV was used as a model organism, which infects a wide variety of cultivated aroids and ornamental plants worldwide, for example, *Caladium*, *Dieffenbachia*, *Colocasia* and *Xanthosoma*, to name a few and causes significant reduction in crop yield (Elliot et al., 1997). Therefore, precise and accurate detection of the DsMV is important for developing better control mechanisms (Grisoni et al., 2006).

The first aim of this study was to test the usefulness of the universal potyvirus primers, NIb2F/NIb3R, HPFor/HPRev and CIFor/CIRev in amplification of the DsMV genome. This study showed that a two-step RT-PCR using a Flex cDNA synthesis kit (Quanta Biosciences) and GoTaq$^{R}$ Green PCR master mix (Promega) can be used with NIb2F/NIb3R for detection of the DsMV infection of taro, if the cDNA synthesis is primed with PV1/SP6 and random hexamers. This study also showed that a one-step RT-PCR using a SuperScript® III One-Step RT-PCR System with Platinum®*Taq* (Life technologies) could be used with HPFor/HPRev and CIFor/CIRev primers for detection of the DsMV infection of taro.

This study also showed that the DsMV specific primers designed against the same binding sites as the HPFor/HPRev and CIFor/CIRev primers could not be used for detection and amplification of the DsMV genome. This could be due to the

phenomenon called consensus decay (Zheng et al., 2008). The design of universal primers for any viral genera target the conserved sites within those genera. The conserved sites are identified by comparing sequences and therefore the usefulness of a set of primers depends on how well the available sequences represent the target viruses. At any given time point, the available sequences represent only a fraction of the future target sequences, therefore, design of universal primers can be biased. Zheng et al. (2008) found that due to accumulation of genetic variants, several of the conserved sites targeted in early work on potyviruses were no longer conserved. This is likely because as the number of variants increase, the consensus sequence will fail to cover all the sequence variants present within a population (Zheng et al., 2008). Zheng et al. (2008) also found that all conserved sites in potyvirus genomes suffer consensus decay, however, different sites asymptote to different levels of variability. Therefore, it becomes important to analyse the pattern of genetic variability among potyviruses using whole genome sequences to help identify the regions of conservation within potyviral genomes, which could be used for designing better universal primers for detection and identification of potyviruses.

The second aim of this study was to understand the pattern of genetic diversity in potyviruses. In this study, deep sequencing profiles of DsMV were analysed using bioinformatics to determine the level of genetic variation occurring in two different strains of DsMV, that is DsMV-NZ and DsMV-B and their isolates, NZ1.1, NZ1.2, B1.1 and B1.2. The main objective of this aim was to determine the level of genetic variation within and between each strain/isolate of the DsMV. Illumina deep sequencing profiles were created for each sample and were analysed using GeneiousPro bioinformatics software.

The study revealed that, on average, DsMV-B and DsMV-NZ showed a percentage genetic variation of 21.5% and 20.9%, respectively, from the DsMV reference genome (Accession no: NC_003537). It was also concluded that the level of variation observed within isolates of the same strain was lower than the variation found between the strains. It was also established that DsMV-B isolates showed a higher level of genetic variation than DsMV-NZ isolates. For the entire ORF, the cut-off percentage of nucleotide identity for differentiating different species of potyviruses is 76% (Adams et al., 2005b).Values greater than this represent comparison between sequences of different isolates of the same virus. Analysis of each sample revealed that the level of genetic variation observed in each of them was lower than cut-off limit indicating that

each isolate belongs to same species of potyvirus, DsMV. The genetic variation in each genome suggested that they were different isolates of same virus. The level of nucleotide identity for the entire ORF in each isolate of DsMV was around ~79%. This kind of pattern is commonly observed for members of the genus *Potyvirus*. The level of nucleotide identity ranging from 49-58% is usually observed between different species in the same genus (Adams et al., 2005b). Therefore, it was concluded that the level of variation identified within each isolate of DsMV was not unusual and this level of variation is commonly observed among different members of the genus *Potyvirus.*

The level of nucleotide identity determined in the ORF of each DsMV genomes was within the cut-off limit described for members of BCMV group (65-68%) (Adams et al., 2005b), therefore, it was suggested that DsMV is a member of the BCMV group (Gibbs et al., 2008). The phylogenetic analysis of each genome also showed that the DsMV was more closely related to VanMV than to any other virus in the BCMV group (Farreyrol et al., 2006). The phylogenetic analysis of each genome confirmed that DsMV-B and DsMV-NZ shared a common ancestor and DsMV-B has been around longer, therefore having more time for creating its genetic variants in comparison to DsMV-NZ (Chang, 2012; Gibbs et al., 2008). However, not many full genomes from BCMV were available for carrying out this study. As a result, only limited knowledge about the evolution of BCMV group was obtained from this analysis. Therefore, in order to get the full picture of the BCMV group phylogeny, extensive phylogenetic analysis needs to be carried out using all the members of the *Potyvirus* genus and Potyviridae.

The third aim of this study was to determine the level variation within each gene of DsMV genome. It was confirmed that 5'UTR and 3'UTR show a high level of genetic variation with few regions of conservation (Kneller, Rakotondrafara, & Miller, 2006; Lewsey & Carr, 2009). The P1 gene of each DsMV showed the highest level of genetic variation. This was expected for the P1 gene as it is considered to be the least conserved region of potyviruses (Wisler et al., 1995). It is speculated that high genetic variability in the P1 gene helps in the successful adaptation of potyviruses into a wide variety of hosts (Valli et al., 2007). The DsMV-B 6K2 gene showed higher genetic variation than DsMV-NZ, the low coverage of reads at the 6K2 region of DsMV-NZ suggested that the low variation observed in the 6K2 of DsMV-NZ may be due to low read coverage, and needs to be investigated in detail. In contrast, the DsMV-NIaVPg showed the least amount of genetic variation. It has been found that the VPg domain of

NIa has essential functions in viral replication and host genotype specificity. High variability in the VPg domain has shown to halt the viral replication process (Murphy et al., 1996). Therefore, it can be speculated that low variability in the NIaVPg domain could be related to its functionality. Overall the gene level nucleotide variation of each DsMV gene was within the gene-level cut-off for species (Adams et al., 2005b), therefore these findings suggest that the level of mutation in potyviruses that can be accommodated is finite and is constrained by several factors, such as host-virus interaction (Schneider & Roossinck, 2001).

The fourth aim of this study was to determine if there are any regions of conservation within the DsMV genome. The DsMV specific HC-Pro and CI Primers were designed against the reference genomes using the universal primer binding sites. However, the comparison between the consensus sequences of the HC-Pro and CI genes of each DsMV genome with the reference genome showed that the regions used for designing DsMV specific HC-Pro and CI primers were not conserved, certain base positions were variable in each primer pair. In contrast, universal primers account for all variation and more likely to bind to the target sequences than the specific ones. The analysis showed that each DsMV genome differs markedly in these regions, thus the virus specific primers are not useful for DsMV-NZ or DsMV-B isolates whereas the universal ones are. These findings support the phenomenon of consensus decay in potyviruses, with increase in the number of available sequences (Zheng et al., 2008). Therefore, for accurate and precise detection of potyviruses, the usefulness of universal primers needs to be evaluated from time to time. In this study, few conserved domains were identified within each DsMV genome; however due to time constrains, it was not possible to do an extensive analysis for these regions. Therefore, in the future, more extensive analysis needs to be carried out for identification of conserved regions within the potyviruses, which could help in designing better potyvirus genus-specific universal primers for the precise and accurate detection of potyvirus infection in commercially important crops.

# References

# References


Abo El-Nil, M. M., Zettler, F. W., & Hiebert, E. (1977). Purification, serology, and some physical properties of Dasheen mosaic virus. *Phytopathology, 67*(12), 1445-1450.

Adams, Antoniw, J. F., & Beaudoin, F. (2005a). Overview and analysis of the polyprotein cleavage sites in the family Potyviridae. *Molecular Plant Pathology, 6*(4), 471-487. doi:10.1111/j.1364-3703.2005.00296.x

Adams, Antoniw, J. F., & Fauquet, C. M. (2005b). Molecular criteria for genus and species discrimination within the family Potyviridae. *Archives of Virology, 150*(3), 459-479. doi:10.1007/s00705-004-0440-6

Adams, Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., ... Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology, 10*(4), 537-545. doi:10.1111/j.1364-3703.2009.00545.x

Aleman-Verdaguer, M. E., Goudou-Urbino, C., Dubern, J., Beachy, R. N., & Fauquet, C. (1997). Analysis of the sequence diversity of the P1, HC, P3, NIb and CP genomic regions of several yam mosaic potyvirus isolates: implications for the intraspecies molecular diversity of potyviruses. *Journal of General Virology, 78* (6), 1253-1264.

Allen, J., & Salzberg, S. (2006). A phylogenetic generalized hidden Markov model for predicting alternatively spliced exons. *Algorithms for Molecular Biology, 1*(1), 14.

Alvarez, A. M. (2004). Integrated approaches for detection of plant pathogenic bacteria and diagnosis of bacterial diseases. *Annual Review of Phytopathology, 42*, 339-366.

Andrews, S. (2011). *FastQC*. Retrieved from http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Antoniw, J. (1995). A new method for designing PCR primers specific for groups of sequences and its application to Plant viruses. *Molecular Biotechnology, 4*, 111-119.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., & Ben-Tal, N. (2010). Consurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research, 38*, W529-W533.

Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., & Zelikovsky, A. (2011). Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics, 12*(SUPPL.6).

Babu, B., Hegde, V., Makeshkumar, T., & Jeeva, M. L. (2010). Rapid detection and identification of Potyvirus infecting Colocasia esculenta (L.) Schott by Reverse Transcription-Polymerase Chain Reaction *Journal of Root Crops, 36*(1), 88-94.

Babu, B., Hegde, V., Makeshkumar, T., & Jeeva, M. L. (2011). Molecular detection and identification of Dasheen mosaic virus infecting Amorphophallus paeoniifolius. *Archives Of Phytopathology And Plant Protection, 44*(13), 1248-1260. doi:10.1080/03235408.2010.490398

Badge, J., Brunt, A., Carson, R., Dagless, E., Karamagioli, M., Phillips, S., ... Foster, G. D. (1996). A carlavirus-specific PCR primer and partial nucleotide sequence provides further evidence for the recognition of cowpea mild mottle virus as a whitefly-transmitted carlavirus. *European Journal of Plant Pathology, 102*(3), 305-310. doi:10.1007/bf01877970

Bailey-Serres, J. (1999). Selective translation of cytoplasmic mRNAs in plants. *Trends in Plant Science, 4*(4), 142-148. doi:10.1016/s1360-1385(99)01386-2

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., & Palù, G. (2011). Applications of Next-Generation sequencing technologies to diagnostic virology. *International Journal of Molecular Sciences, 12*(11), 7861-7884.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59. doi:10.1101/gr.078212.108

Bienz, K., Egger, D., Pfister, T., & Troxler, M. (1992). Structural and functional characterization of the poliovirus replication complex. *Journal of Virology, 66*(5), 2740-2747.

Bos, L., Kowalska, C., & Maat, D. Z. (1974). The identification of bean mosaic, pea yellow mosaic and pea necrosis strains of bean yellow mosaic virus. *European Journal of Plant Pathology, 80*(6), 173-191. doi:10.1007/bf01976698

Brown, T. (2010). *Gene Cloning and DNA Analysis : An Introduction* (6 ed.). Retrieved from http://AUT.eblib.com.au/patron/FullRecord.aspx?p=530049

Browning, K. S. (1996). The plant translational apparatus. *Plant Molecular Biology, 32*(1), 107-144. doi:10.1007/bf00039380

Brunt. (1992). The general properties of potyviruses. *Archives of virology. Supplementum, 5*, 3-16.

Brunt. (2000). Plant Viruses, Unique and Intriguing Pathogens – A Textbook of Plant Virology. *Journal of Phytopathology, 148*(11/12), 637-637. doi:10.1111/1439-0434.ep5530431

Brunt, Crabtree, Dallwitz, Gibbs, Watson, & Zurcher. (2012a). Dasheen mosaic *Potyvirus*. Virus identification data exchange http://www.agls.uidaho.edu/ebi/vdie/descr289.htm

Brunt, Crabtree, Dallwitz, Gibbs, Watson, & Zurcher. (2012b). Plant Viruses Online (Electron micrograph). VIDE http://www.agls.uidaho.edu/ebi/vdie/images/a1.jpg

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., ... Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics, 40*(6), 722-729.

Cann, A. J., Fandrich, S. E., & Heaphy, S. (2005). Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes, 30*(2), 151-156. doi:10.1007/s11262-004-5624-3

Carrington, J. C., & Freed, D. D. (1990). Cap-independent enhancement of translation by a plant potyvirus 5' nontranslated region. *Journal of Virology, 64*(4), 1590-1597.

Chen, Chen, J., Chen, J., & Adams, M. J. (2001). Molecular characterisation of an isolate of Dasheen mosaic virus from Zantedeschia aethiopica in China and comparisons in the genus Potyvirus. *Archives of Virology, 146*, 1821-1829.

Chen, Zheng, H. Y., Shi, Y. H., Adams, M. J., Wei, C. B., Lin, L., & Chen, J. P. (2006). Detection and characterisation of a second potyvirus from Thunberg fritillary in China. *Archives of Virology, 151*(3), 439-447. doi:10.1007/s00705-005-0678-7

Chung, B. Y., Miller, W. A., Atkins, J. F., & Firth, A. E. (2008). An overlapping essential gene in the Potyviridae. *Proceedings of the National Academy of Sciences, 105*(15), 5897-5902.

Coetzee, B., Freeborough, M.-J., Maree, H. J., Celton, J.-M., Rees, D. J. G., & Burger, J. T. (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of vineyard. *Virology, 400*, 157-163.

Colinet, D., Nguyen, M., Kummert, J., Lepoivre, P., & Xia, F. Z. (1998). Differentiation among Potyviruses infecting Sweet potato based on genus-and virus-specific reverse transcription polymerase chain reaction. *Plant Disease, 82*(2), 223-229. doi:10.1094/pdis.1998.82.2.223

Daròs, J.-A., Schaad, M. C., & Carrington, J. C. (1999). Functional analysis of the interaction between VPg-Proteinase (NIa) and RNA Polymerase (NIb) of *Tobacco Etch Potyvirus*, using conditional and suppressor mutants. *Journal of Virology, 73*(10), 8732-8740.

Davis. (2012). *Fresh dasheen*. Retrieved from http://www.21food.com/products/fresh-dasheen-216797.html

Desbiez, C., & Lecoq, H. (2004). The nucleotide sequence of Watermelon mosaic virus (WMV, Potyvirus) reveals interspecific recombination between two related potyviruses in the 5' part of the genome. *Archives of Virology, 149*(8), 1619-1632. doi:10.1007/s00705-004-0340-9

Domingo. (2002). Quasispecies theory in virology. *Journal of Virology, 76*(1), 463-465.

Domingo, Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., ... Holland, J. J. (1996). Basic concepts in RNA virus evolution. *The FASEB Journal, 10*(8), 859-864.

Domingo, & Holland. (1997). RNA Virus Mutations and fitness for survival. *Annual Review of Microbiology, 51*(1), 151-178. doi:10.1146/annurev.micro.51.1.151

Drake, J. W., & Holland, J. J. (1999). Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences USA 96*(24), 13910-13913.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., ... Wilson, A. (2009). Genome sequences: Types of data and bioinformatic tools. In M. Schaechter (Ed.), *Encylcopedia of Microbiology* (pp. 211-236): Elsevier.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., ... Wilson, A. (2011). GeneiousPro [Retrieved from http://www.geneious.com/

Egerton. (2005). Physical principles of electron microscopy. *Materials Today, 8*(12), 60. doi:10.1016/s1369-7021(05)71232-3

Eggenberger, A. L., Stark, D. M., & Beachy, R. N. (1989). The nucleotide sequence of a soybean mosaic virus coat protein-coding region and its expression in Escherichia coli, Agrobacterium tumefaciens and tobacco callus. *Journal of General Virology, 70*(7), 1853-1860.

Eigen, M. (1996). On the nature of virus quasispecies. *Trends in Microbiology, 4*, 216-218.

Elliot, M. S., Zettler, F. W., & Brown, L. G. (1997). Dasheen mosaic Potyvirus of edible and ornamental aroids. *Plant Pathology Circular, 384*, 1-4.

Erlich, Y., Mitra, P. P., Delabastide, M., McCombie, W. R., & Hannon, G. J. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods, 5*(8), 679-682. doi:10.1038/nmeth.1230

Erni, R., Rossell, M. D., Kisielowski, C., & Dahmen, U. (2009). Atomic-resolution imaging with a sub-50-pm electron probe. *Physical Reviews Letters 102*(9), 096-101.

Fang, G. W., Allison, R. F., Zambolim, E. M., Maxwell, D. P., & Gilbertson, R. L. (1995). The complete nucleotide sequence and genome organization of bean common mosaic virus (NL3 strain). *Virus Research 39*(1), 13-23.

Farreyrol, K., Pearson, M. N., Grisoni, M., Cohen, D., & Beck, D. (2006). Vanilla mosaic virus isolates from French Polynesia and the Cook Islands are *Dasheen mosaic virus* strains that exclusively infect vanilla. *Archives of Virology, 151*(5), 905-919. doi:10.1007/s00705-005-0680-0

Finegardening. (2012). *Colocasia esculenta* Retrieved from http://www.finegardening.com/plantguide/colocasia-esculenta-black-magic-elephant-ear.aspx

Fuji, S., & Nakamae, H. (2000). Complete nucleotide sequence of the geonomic RNA of a mild strain of Japanese yam mosaic potyvirus. *Archives of Virology, 145*(3), 635-640. doi:10.1007/s007050050052

Gallie, D. R., Tanguay, R. L., & Leathers, V. (1995). The tobacco etch viral 5′ leader and poly(A) tail are functionally synergistic regulators of translation. *Gene, 165*(2), 233-238. doi:10.1016/0378-1119(95)00521-7

García-Arenal, F., Fraile, A., & Malpica, J. (2003). Variation and evolution of plant virus populations. *International Microbiology, 6*(4), 225-232. doi:10.1007/s10123-003-0142-z

García-Arenal, F., Fraile, A., & Malpica, J. M. (2001). Variability and genetic structure of plant virus populations. *Annual Review of Phytopathology, 39*(1), 157-186. doi:10.1146/annurev.phyto.39.1.157

Garger, S., Turpen, T., Carrington, J., Morris, T., Jordan, R., Dodds, J., & Grill, L. (1983). Rapid detection of plant RNA viruses by dot blot hybridization. *Plant Molecular Biology Reporter, 1*(1), 21-25. doi:10.1007/bf02680258

German-Retana, S., Walter, J., & Le Gall, O. (2008). Lettuce mosaic virus: from pathogen diversity to host interactors. *Molecular Plant Pathology, 9*(2), 127-136.

Gibbs, & Mackenzie, A. (1997). A primer pair for amplifying part of the genome of all potyvirids by RT-PCR. *Journal of Virological Methods, 63*(1–2), 9-16. doi:10.1016/s0166-0934(96)02103-9

Gibbs, & Ohshima, K. (Eds.). (2010). Potyviruses and the digital revolution. *48*, 205-223. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-77956122421&partnerID=40&md5=38c1f477e4b9a99ee611e5c852c9e9f1

Gibbs, Trueman, J., & Gibbs, M. (2008). The bean common mosaic virus lineage of potyviruses: where did it arise and when? *Archives of Virology, 153*(12), 2177-2187. doi:10.1007/s00705-008-0256-x

Gibbs, A., & Ohshima, K. (2010). Potyviruses and the digital revolution. *Annual Review of Phytopathology, 48*(1), 205-223. doi:10.1146/annurev-phyto-073009-114404

Gibbs, A., Ohshima, K., Phillips, M., & Gibbs, M. (2008). The Prehistory of Potyviruses: Their initial radiation was during the dawn of agriculture. *PLoS ONE, 3*(6), e2523. doi:10.1371/journal.pone.0002523

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources, 11*(5), 759-769. doi:10.1111/j.1755-0998.2011.03024.x

Grisoni, M., Moles, M., Farreyrol, K., Rassaby, L., Davis, R., & Pearson, M. (2006). Identification of potyviruses infecting vanilla by direct sequencing of a short RT-PCR amplicon. *Plant Pathology, 55*(4), 523-529. doi:10.1111/j.1365-3059.2006.01397.x

Ha, C., Coombs, S., Revill, P. A., Harding, R. M., Vu, M., & Dale, J. L. (2008). Design and application of two novel degenerate primer pairs for detection and complete genomic charcaterization of potyviruses. *Archives of Virology, 153*, 25-36.

Haldeman-Cahill, R., Daros, J. A., & Carrington, J. C. (1998). Secondary structures in the capsid protein coding sequence and 3' nontranslated region involved in amplification of the tobacco etch virus genome. *Journal of Virology, 72*(5), 4072-4079.

Harding, R. (2012). *Dasheen mosaic potyvirus*. Retrieved from http://www.ediblearoids.org/Portals/0/TaroPest/LucidKey/TaroPest/Media/Html/Viruses/DasheenMV/DasheenMV6.htm#fig1

Hedskog, C., Mild, M., Jernberg, J., Sherwood, E., Bratt, G., Leitner, T., ... Albert, J. (2010). Dynamics of HIV-1 Quasispecies during antiviral treatment dissected using Ultra-Deep Pyrosequencing. *PLoS One, 5*(7), e11345.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., & Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research, 18*(5), 802-809. doi:10.1101/gr.072033.107

Higgins, C. M., Cassidy, B. G., Teycheney, P. Y., Wongkaew, S., & Dietzgen, R. G. (1998). Sequences of the coat protein gene of five peanut stripe virus (PStV) strains from Thailand and their evolutionary relationship with other Bean common mosaic virus sequences. *Archives of Virology, 143*(9), 1655-1667.

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., ... Mardis, E. R. (2008). Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods, 5*(2), 183-188. doi:10.1038/nmeth.1179

Hong, Y., & Hunt, A. G. (1996). RNA polymerase activity catalyzed by a potyvirus-encoded RNA-dependent RNA polymerase. *Virology, 226*(1), 146-151. doi:10.1006/viro.1996.0639

Hong, Y., Levay, K., Murphy, J. F., Klein, P. G., Shaw, J. G., & Hunt, A. G. (1995). A potyvirus polymerase Interacts with the viral coat protein and VPg in yeast cells. *Virology, 214*(1), 159-166. doi:10.1006/viro.1995.9944

Huang, C. H., Hu, W. C., Yang, T. C., & Chang, Y. C. (2007). Zantedeschia mild mosaic virus, a new widespread virus in calla lily, detected by ELISA, dot-blot hybridization and IC-RT-PCR. *Plant Pathology, 56*(1), 183-189.

Illumina. (2012). *RNA sequencing* Retrieved2012, from http://www.illumina.com/applications/sequencing/rna_seq.ilmn

Iwai, H., Terahara, R., Yamashita, Y., Ueda, S., & Nakamura, M. (2006). Complete nucleotide sequence of the genomic RNA of an Amami-O-shima strain of East Asian Passiflora potyvirus. *Archives of Virology, 151*(7), 1457-1460. doi:10.1007/s00705-006-0779-y

Jan, A. T., Azam, M., Warsi, M. K., Ali, A., & Haq, Q. M. R. (2011). Technical advancement in plant virus diagnosis – an appraisal. *Archives of Phytopathology and Plant Protection, 45*(8), 909-921. doi:10.1080/03235408.2011.599156

JIC. (2012). *What is electron microscopy?* Retrieved2012, from http://www.jic.ac.uk/microscopy/intro_em.html

Karam, S., & Heinze, F. (2003). *Investigation of the Potyviridae on secondary structures*. Retrieved2012, from http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS02/punch/

Kawarasaki, Y., Kasahara, S., Kodera, N., Shinbata, T., Sekiguchi, S., Nakano, H., & Yamane, T. (2000). A trimmed viral cap-independent translation enhancing sequence for rapid in vitro gene expression. *Biotechnology Progress, 16*(3), 517-521.

Kekarainen, T., Merits, A., Oruetxebarria, I., Rajamäki, M. L., & Valkonen, J. P. T. (1999). Comparison of the complete sequences of five different isolates of Potato virus A (PVA), genus Potyvirus. *Archives of Virology, 144*(12), 2355-2366. doi:10.1007/s007050050649

Khan, S., Jan, A. T., Aquil, & Haq. (2011). *Coat Protein Gene based Characterization of Cucumber Mosaic Virus Isolates Infecting Banana in India*. Retrieved from http://journal-phytology.com/index.php/phyto/article/view/6332

Kitajima, E. W. (2004). Electron Microscopy in Plant Virology: Past, Present and Future [10.1017/S1431927604881467]. *Microscopy and Microanalysis, 10*(Supplement S02), 212-213.

Klein, P. G., Klein, R. R., Rodriguez-Cerezo, E., Hunt, A. G., & Shaw, J. G. (1994). Mutational analysis of the *Tobacco vein mottling* virus genome. *Virology, 204*(2), 759-769. doi:10.1006/viro.1994.1591

Kneller, E. L. P., Rakotondrafara, A. M., & Miller, W. A. (2006). Cap-independent translation of plant viral RNAs. *Virus Research, 119*(1), 63-75.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., ... Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science, 318*(5849), 420-426.

Lain, S., Riechmann, J. L., & Garcia, J. A. (1990). RNA helicase: a novel activity associated with a protein encoded by a positive strand RNA virus. *Nucleic Acids Research 18*(23), 7003-7006.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Langeveld, S. A., Dore, J. M., Memelink, J., Derks, A. F., van der Vlugt, C. I., Asjes, C. J., & Bol, J. F. (1991). Identification of potyviruses using the polymerase chain reaction with degenerate primers. *The Journal of General Virology, 72*(7), 1531-1541.

Lellis, A. D., Kasschau, K. D., Whitham, S. A., & Carrington, J. C. (2002). Loss-of-susceptibility mutants of *Arabidopsis thaliana* reveal an essential role for eIF(iso)4E during potyvirus infection. *Current Biology, 12*(12), 1046-1051. doi:10.1016/s0960-9822(02)00898-9

Lewsey, M. G., & Carr, J. P. (2009). Plant Pathogens:RNA viruses. In M. Schaechter (Ed.), *Encyclopedia of Microbiology* (pp. 443-458): Oxford Academic Press.

Liang, W. X., Song, L. M., Tian, G. Z., Li, H. F., & Fan, Z. F. (2006). The genomic sequence of Wisteria vein mosaic virus and its similarities with other potyviruses. *Archives of Virology, 151*(11), 2311-2319. doi:10.1007/s00705-006-0780-5

Lin, S., Hou, R., & Yeh. (2001). Complete genome sequence and genetic organization of a Taiwan isolate of Zuccini yellow mosaic virus. *Botanical Bulletin of Academia Sinica, 42*(4), 243-250.

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.

Madeley, C. R. (1997). Electron microscopy and viral diagnosis. *Journal of Clinical Pathology, 50*(6), 454-456.

Maher-Sturgess, S., Forrester, N., Wayper, P., Gould, E., Hall, R., Barnard, R., & Gibbs, M. (2008). Universal primers that amplify RNA from all three flavivirus subgroups [10.1186/1743-422X-5-16]. *Virology Journal, 5*(1), 16.

Maia, I. G., Haenni, A.-L., & Bernardi, F. (1996). Potyviral HC-Pro: a multifunctional protein. *Journal of General Virology, 77*(7), 1335-1341.

Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W., & Garcia-Arenal, F. (2002). The rate and character of spontaneous mutation in an RNA virus. *Genetics, 162*, 1505-1511.

Marie-Jeanne, V., Ioos, R., Peyre, J., Alliot, B., & Signoret, P. (2000). Differentiation of Poaceae potyviruses by reverse transcription-polymerase chain reaction and restriction analysis. *Journal of Phytopathology, 148*(3), 141-151. doi:10.1046/j.1439-0434.2000.00473.x

Martin, M. T., Cervera, M. T., & Garcia, J. A. (1995). Properties of the active plum pox potyvirus RNA polymerase complex in defined glycerol gradient fractions. *Virus Research, 37*(2), 127-137. doi:10.1016/0168-1702(95)00028-o

Martin, R. R., James, D., & Lévesque, C. A. (2000). Impacts of molecular diagnostic technologies on plant disease management. *Annual Review of Phytopathology, 38*(1), 207-239. doi:10.1146/annurev.phyto.38.1.207

Matthews, R. E. F. (1991). *Plant Virology* (3rd ed. ed.). San Diego, Calif.: Academic Press.

Maule, A. J., Hull, R., & Donson, J. (1983). The application of spot hybridization to the detection of DNA and RNA viruses in plant tissues. *Journal of Virological Methods, 6*(4), 215-224. doi:10.1016/0166-0934(83)90048-4

McDonald, M., Kendall, A., Bian, W., McCullough, I., Lio, E., Havens, W. M., ... Stubbs, G. (2010). Architecture of potyviruses. *Virology, 405*, 309-313.

Merits, Guo, & Saarma. (1998). VPg, coat protein and five non-structural proteins of potato virus A potyvirus bind RNA in a sequence-unspecific manner. *Journal of General Virology, 79*(12), 3123-3127.

Merits, Guo, D., Järvekülg, L., & Saarma, M. (1999). Biochemical and genetic evidence for interactions between Potato A potyvirus-encoded proteins P1 and P3 and proteins of the putative replication complex. *Virology, 263*(1), 15-22. doi:10.1006/viro.1999.9926

Metzker. (2010). Sequencing technologies-the next generation. *Nature, 11*, 31-46.

Mlotshwa, S., Verver, J., Sithole-Niang, I., Van Kampen, T., Van Kammen, A., & Wellink, J. (2002). The genomic sequence of cowpea aphid-borne mosaic virus and its similarities with other potyviruses. *Archives of Virology, 147*(5), 1043-1052. doi:10.1007/s00705-002-0800-z

Murphy, J. F., Klein, P. G., Hunt, A. G., & Shaw, J. G. (1996). Replacement of the tyrosine residue that links a potyviral VPg to the viral RNA is lethal. *Virology, 220*(2), 535-538. doi:10.1006/viro.1996.0344

NCBI. (2012). PCR. Available from NCBI Probe DB, from National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechPCR.shtml

Nelson, S. (2008). Dasheen Mosaic of Edible and Ornamental Aroids . *Plant Disease*, 9.

Nelson, S. (2012). *Leaf curling*. Retrieved from http://www.hawaiiplantdisease.net/glossary/images/Leaf_curling/ChuukDMV2.jpg

Nemchinov, L. G., Hammond, J., Jordan, R., & Hammond, R. W. (2004). The complete nucleotide sequence, genome organization, and specific detection of Beet mosaic virus. *Archives of Virology, 149*(6), 1201-1214. doi:10.1007/s00705-003-0278-3

Nicolas, O., & Laliberté, J.-F. (1992). The complete nucleotide sequence of turnip mosaic potyvirus RNA. *Journal of General Virology, 73*(11), 2785-2793.

Niepel, M., & Gallie, D. R. (1999). Identification and characterization of the functional elements within the tobacco etch virus 5' leader required for cap-independent translation. *Journal of Virology, 73*(11), 9080-9088.

Olivares, E. (2011). The next generation sequencing community forum. Retrieved from http://www.seqanswers.com/

Owens, R. A., & Diener, T. O. (1981). Sensitive and rapid diagnosis of potato spindle tuber viroid disease by nucleic acid hybridization. *Science, 213*(4508), 670-672.

Pallas, V., Mas, P., & Sanchez-Navarro, J. A. (1998). Detection of plant RNA viruses by nonisotopic dot-blot hybridization. *Methods in Molecular Biology, 81*, 461-468. doi:10.1385/0-89603-385-6:461

Pallen, M. J., Loman, N. J., & Penn, C. W. (2010). High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Current Opinion in Microbiology 13*(5), 625-631. doi:10.1016/j.mib.2010.08.003

Pelletier, J., & Sonenberg, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature, 334*(6180), 320-325. doi:10.1038/334320a0

Pestnet. (2012). *Dasheen mosaic potyvirus* Retrieved from http://portal.cbit.uq.edu.au/pestnet/SummariesofMessages/Crops/Rootstubers/Taro/Viruses/DasheenmosaicpotyvirusSamoa.aspx

PNW-Handbook. (2012). *Dieffenbachia-Dasheen Mosaic*. Retrieved from http://pnwhandbooks.org/plantdisease/host-disease-descriptions/dieffenbachia-dasheen-mosaic

Restrepo-Hartwig, M. A., & Carrington, J. C. (1994). The tobacco etch potyvirus 6-kilodalton protein is membrane associated and involved in viral replication. *Journal of Virology, 68*(4), 2388-2397.

Riechmann, Cervera, M. T., & García, J. A. (1995). Processing of the plum pox virus polyprotein at the P3-6K1 junction is not required for virus viability. *Journal of General Virology, 76*(4), 951-956.

Riechmann, Laín, S., & García, J. A. (1989). The genome-linked protein and 5′ end RNA sequence of Plum Pox potyvirus. *Journal of General Virology, 70*(10), 2785-2789.

Riechmann, Laín, S., & García, J. A. (1992). Highlights and prospects of potyvirus molecular biology. *Journal of General Virology, 73*(1), 1-16.

Robaglia, C., Durand-Tardif, M., Tronchet, M., Boudazin, G., Astier-Manifacier, S., & Casse-Delbart, F. (1989). Nucleotide sequence of potato virus Y (N Strain) genomic RNA. *Journal of General Virology, 70 ( Pt 4)*, 935-947.

Robertson, N. L., French, R., & Gray, S. M. (1991). Use of group-specific primers and the polymerase chain reaction for the detection and identification of luteoviruses. *Journal of General Virology, 72*(6), 1473-1477.

Rodríguez-Cerezo, E., & Shaw, J. G. (1991). Two newly detected nonstructural viral proteins in potyvirus-infected cells. *Virology, 185*(2), 572-579. doi:10.1016/0042-6822(91)90527-i

Rohožková, J., & Navrátil, M. (2011). P1 peptidase - A mysterious protein of family Potyviridae. *Journal of Biosciences, 36*(1), 189-200.

Roossinck, M. J. (1997). Mechanisms of plant virus evolution. *Annual Review of Phytopathology, 35*(1), 191-209. doi:10.1146/annurev.phyto.35.1.191

Schaad, M. C., Haldeman-Cahill, R., Cronin, S., & Carrington, J. C. (1996). Analysis of the VPg-proteinase (NIa) encoded by tobacco etch potyvirus: effects of mutations on subcellular transport, proteolytic processing, and genome amplification. *Journal of Virology, 70*(10), 7039-7048.

Schaad, M. C., Jensen, P. E., & Carrington, J. C. (1997). Formation of plant RNA virus replication complexes on membranes: role of an endoplasmic reticulum-targeted viral protein. *The EMBO Journal, 16*(13), 4049-4059. doi:10.1093/emboj/16.13.4049

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics 11*(9), 647-657. doi:10.1038/nrg2857

Schneider, W. L., & Roossinck, M. J. (2000). Evolutionarily related Sindbis-like plant viruses maintain different levels of population diversity in a common host. *Journal of Virology, 74*(7), 3130-3134.

Schneider, W. L., & Roossinck, M. J. (2001). Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *Journal of Virology, 75*(14), 6566-6571. doi:10.1128/JVI.75.14.6566-6571.2001

Schwartz, S., Oren, R., & Ast, G. (2011). Detection and removal of biases in the analysis of Next-Generation Sequencing reads. *PLoS ONE, 6*(1), e16685.

Sharma, Singh, Nagpal, Virk, & Zaidi. (2009). Indexing Tools for Indian Citrus Ringspot Virus (ICRSV). *The Open Biology Journal, 2*, 27-31.

Shukla, Frenkel, & Ward. (1991). Structure and function of the potyvirus genome with special reference to the coat protein coding region. *Canadian Journal of Plant Pathology 13*, 178-191.

Shukla, McKern, Gough, Tracy, & Letho. (1988). Differentiation of potyviruses and their strains by high performance liquid chromatographic peptide profiling of coat proteins. *Journal of General Virology, 69*(3), 493-502.

Shukla, & Ward, C. W. (1988). Amino acid sequence homology of coat proteins as a basis for identification and classification of the potyvirus group. *Journal of General Virology, 69*(11), 2703-2710.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus – 15 years on. *Journal of General Virology, 85*(11), 3173-3188.

Steinhauer, D. A., Domingo, E., & Holland, J. J. (1992). Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene, 122*(2), 281-288. doi:10.1016/0378-1119(92)90216-c

Stewart, J. J., Watts, P., & Litwin, S. (2001). An algorithm for mapping positively selected members of quasispecies-type viruses. *BMC Bioinformatics, 2*, 1.

Sun, W., Jiao, K., Zhang, S., Zhang, C., & Zhang, Z. (2001). Electrochemical detection for horseradish peroxidase-based enzyme immunoassay using p-aminophenol as substrate and its application in detection of plant virus. *Analytica Chimica Acta, 434*(1), 43-50. doi:10.1016/s0003-2670(01)00803-0

Tatineni, S., Ziems, A. D., Wegulo, S. N., & French, R. (2009). Triticum mosaic virus: A distinct member of the family Potyviridae with an unusually long leader sequence. *Phytopathology, 99*(8), 943-950. doi:10.1094/phyto-99-8-0943

Teycheney, P. Y., Aaziz, R., Dinant, S., Salanki, K., Tourneur, C., Balazs, E., ... Tepfer, M. (2000). Synthesis of (-)-strand RNA from the 3' untranslated region of plant viral genomes expressed in transgenic plants upon infection with related viruses. *Journal of General Virology, 81*(4), 1121-1126.

Tordo, V. M.-J., Chachulska, A. M., Fakhfakh, H., Le Romancer, M., Robaglia, C., & Astier-Manifacier, S. (1995). Sequence polymorphism in the 5′NTR and in the P1 coding region of potato virus Y genomic RNA. *Journal of General Virology, 76*(4), 939-949.

Tucker, T., Marra, M., & Friedman, J. M. (2011). Massively Parallel Sequencing. In *Molecular Analysis and Genome Discovery* (pp. 114-134): John Wiley & Sons, Ltd. Retrieved from http://dx.doi.org/10.1002/9781119977438.ch6

Valli, A., López-Moya, J. J., & García, J. A. (2007). Recombination and gene duplication in the evolutionary diversification of P1 proteins in the family Potyviridae. *Journal of General Virology, 88*(3), 1016-1028.

Van Regenmortel, M. H., International, & Fauquet, C. M. (2000). *Virus Taxonomy: Classification and Nomenclature of Viruses: Seventh Report of the International Committee on Taxonomy of Viruses*: Academic Press. Retrieved from http://www.virustaxonomyonline.com/

Venter, J. C., Adams, M. D., Myers, E. W., & Li, P. W. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351.

Verchot, J., & Carrington, J. C. (1995a). Debilitation of plant potyvirus infectivity by P1 proteinase-inactivating mutations and restoration by second-site modifications. *Journal of Virology, 69*(3), 1582-1590.

Verchot, J., & Carrington, J. C. (1995b). Evidence that the potyvirus P1 proteinase functions in trans as an accessory factor for genome amplification. *Journal of Virology, 69*(6), 3668-3674.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., & Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature, 439*(19), 344-348.

*ViralZone*. (2012). [Swiss Institute of Bioinformatics]. Retrieved 2012, from www.expasy.org/viralzone

Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-Generation sequencing: From basic research to diagnostics. *Clinical Chemistry, 55*(4), 641-658.

Ward, Weiller, Shukla, & Gibbs. (1995). *Molecular systematics of the Potyviridae, the largest plant virus family Molecular Basis of Virus Evolution*: Cambridge University Press.

Wei, T., Pearson, M. N., & Fletcher, J. D. (2006). Molecular confirmation of New Zealand garlic yellow streak virus as Leek yellow stripe virus. *Australasian Plant Pathology, 35*(3), 341-346.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., ... Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature, 452*(7189), 872-876. doi:10.1038/nature06884

Wisler, G. C., Purcifull, D. E., & Hiebert, E. (1995). Characterization of the P1 protein and coding region of the zucchini yellow mosaic virus. *Journal of General Virology, 76*(1), 37-45.

Wittmann, S., Chatel, H., Fortin, M. G., & Laliberté, J.-F. (1997). Interaction of the viral protein genome linked of Turnip Mosaic potyvirus with the translational eukaryotic initiation factor (iso) 4E of *Arabidopsis thaliana* using the yeast two-hybrid system. *Virology, 234*(1), 84-92. doi:10.1006/viro.1997.8634

Wylie, & Jones, M. (2011). The complete genome sequence of a Passion fruit woodiness virus isolate from Australia determined using deep sequencing, and its relationship to other potyviruses. *Archives of Virology, 156*(3), 479-482. doi:10.1007/s00705-010-0845-3

Wylie, & Jones, M. G. (2011). Characterisation and quantitation of mutant and wild-type genomes of Hardenbergia mosaic virus isolates co-infecting a wild plant of Hardenbergia comptoniana. *Archives of Virology, 156*(7), 1251-1255. doi:10.1007/s00705-011-0965-4

Zeenko, V., & Gallie, D. R. (2005). Cap-independent Translation of Tobacco Etch Virus Is Conferred by an RNA Pseudoknot in the 5′-Leader. *Journal of Biological Chemistry, 280*(29), 26813-26824.

Zettler, F. W., & Hartman, R. D. (1987). Dasheen Mosaic Virus as a Pathogen of Cultivated Aroids and Control of the Virus by Tissue Culture. *Plant Disease, 71*(11), 958-963.

Zheng, Chen, J., Adams, M. J., & Hou, M. (2002). Bean common mosaic virus isolates causing different symptoms in asparagus bean in China differ greatly in the 5'-parts of their genomes. *Archives of Virology, 147*(6), 1257-1262. doi:10.1007/s00705-002-0805-7

Zheng, Rodoni, Gibbs, & Gibbs. (2010a). A novel pair of universal primers for the detection of potyviruses. *Plant Pathology, 59*(2), 211-220.

Zheng, Rodoni, B. C., Gibbs, M. J., & Gibbs, A. J. (2010b). A novel pair of universal primers for the detection of potyviruses. *Plant Pathology, 59*(2), 211-220.

Zheng, Wayper, P. J., Gibbs, A. J., Fourment, M., Rodoni, B. C., & Gibbs, M. J. (2008). Accumulating variation at conserved sites in Potyvirus genomes is driven by

species discovery and affects degenerate primer design. *PLoS ONE, 3*(2), e1586. doi:10.1371/journal.pone.0001586

Zheng, L., Wayper, P. J., Gibbs, A. J., Fourment, M., Rodoni, B. C., & Gibbs, M. J. (2008). Accumulating variation at conserved sites in potyvirus genomes is driven by species discovery and affects degenerate primer design. *PLoS ONE, 3*(2).