Development and optimization of a diagnostic system based on Illumina sequencing of genus wide PCR amplicons for the detection of viruses of grapevines

by

Jennifer Wayland

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae Microbiology (MSc.)

In the Faculty of Natural & Agricultural Sciences

University of Pretoria

Pretoria

Submitted October 2016

DECLARATION

I, Jennifer Wayland declare that this thesis, which I hereby submit for the degree *Magister Scientiae* Microbiology (MSc.) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Jennifer Wayland

Signature: ______________________

Date: ______________________

ACKNOWLEDGEMENTS

The success of this research project is based on the support of many individuals and entities. I wish to express my sincere gratitude to the following:

My mother, for both her emotional and financial support. For always being there to listen to my frustrations and stories of success in the lab, even when she had no idea what I was talking about. I am truly grateful for all your support and love.

My fiancé, for never doubting in my abilities even when I did, and for always making me see the best in every situation. Your love and support has carried me through many rough days and nights of lab work and writing, thank you.

The University of Pretoria and the Department of Microbiology and Plant Pathology for allowing me to complete this degree, as well as for their financial support.

Winetech, for their financial support throughout this project.

My supervisors, Professor Gerhard Pietersen and Doctor Elize Jooste, for all their time, advice and continuous support throughout.

My lab colleagues, David Read, Megan Harris, Jacolene Kleynhans, Kirsti Snyders, Ronel Roberts, Elrea Appelgryn and Azille Schulze, as well as my honorary lab colleagues, Marike Palmer, Gaby Carstens, Juanita Avontuur and Andele Conradie, for all of their advice and friendship.

SUMMARY

In this study we present a poly-specific PCR-based, high-throughput sequencing (HTS) diagnostic system together with an appropriate data analysis pipeline for the diagnosis of grapevine viruses. Poly-specific and virus-specific primers were established to be capable of detecting and identifying 37 grapevine infecting viruses from 11 genera. An analysis pipeline using CLC Genomics workbench was developed by utilising various defined artificial samples which were assembled and sequenced on the Illumina MiSeq platform. A threshold for percentage mapped reads of 0.4% during reference mapping was established to discriminate between presence or absence of viruses associated with reads. Various criteria for the evaluation of *de novo* assembled contigs and BLAST results were identified based on virus hits, E-value, percentage query overlap and percentage amplicon overlap. Various RT-PCR systems were used to screen 62 grapevine samples (field collected and candidate nuclear vines) for their virus populations. Seven samples were selected for Illumina MiSeq sequencing, and the data was analysed as per the optimized pipeline. The threshold established for reference mapping and the criteria for BLAST analysis was successfully implemented, proving the applicability of this PCR and HTS-based system in grapevine diagnostics. This system was compared to the standard ELISA system routinely utilised during certification. In our study, when samples evaluated by RT-PCR were tested using ELISA for the presence of GLRaV-1, -2 and -3, a false-negative rate in ELISA of 14.3% was observed, confirming that RT-PCR is the more sensitive test of the two. The capability of RT-PCR to readily detect viruses present in low concentrations in woody plants, the availability of primers for virus identification, the ease and rapidity of the technique, together with constant improvement of HTS platforms especially in the area of cost makes this an extremely useful method for grapevine virus diagnostics.

LIST OF ABBREVIATIONS

AILV	Artichoke Italian latent virus
ARC	Agricultural research council
ARC-PPRI	ARC-plant protection research institute
ArMV	Arabis mosaic virus
ATP	Adenosine triphosphate
BBLMV	Blueberry leaf mottle virus
BC	Buffer control
BLAST	Basic local alignment search tool
bp	Base pair
CCD	Charged coupled device
cDNA	Complimentary DNA
ChIP-Seq	Chromatin immunoprecipitation sequencing
C:I	Chloroform-Isomyl alcohol
CLRV	Cherry leafroll virus
CP	Coat protein
CTAB	Cetyltrimethylammonium bromide
CTV	Citrus tristeza virus
DNA	Deoxyribonucleic acid
dsRNA	Double stranded ribonucleic acid
E. coli	Escherichia coli
EDTA	Ethylene-diamine-tetra-acetic acid
emPCR	Emulsion polymerase chain reaction
ELISA	Enzyme-linked immunosorbent assay
GALV	Grapevine algerian latent virus
GAMaV	Grapevine asteroid mosaic associated virus
GARSV	Grapevine anatolian ringspot virus
Gb	Gigabase
GBLV	Grapevine bulgarian latent virus
GC	Guadine+Cytosine
GCMV	Grapevine chrome mosaic virus
GCSV	Grapevine cabernet sauvignon virus
GDefV	Grapevine deformation virus

GFkV	Grapevine fleck virus
GFLV	Grapevine fanleaf virus
GLD	Grapevine leafroll disease
GLRaV-1	Grapevine leafroll associated virus 1
GLRaV-2	Grapevine leafroll associated virus 2
GLRaV-3	Grapevine leafroll associated virus 3
GLRaV-4	Grapevine leafroll associated virus 4
GLRaV-5	Grapevine leafroll associated virus 4 strain 5
GLRaV-6	Grapevine leafroll associated virus 4 strain 6
GLRaV-9	Grapevine leafroll associated virus 4 strain 9
GLRaV-7	Grapevine leafroll associated virus 7
GTRSV	Grapevine tunisian ringspot virus
GVA	Grapevine virus A
GVB	Grapevine virus B
GVD	Grapevine virus D
GVE	Grapevine virus E
GVF	Grapevine virus F
GRBaV	Grapevine redblotch associated virus
GRGV	Grapevine redglobe virus
GRSPaV	Grapevine rupestris stem pitting associated virus
GRVFV	Grapevine rupestris vein feathering virus
GSyV-1	Grapvine syrah virus 1
h	Hour
HC	Healthy control
H_2O_2	Hydrogen peroxide
HTS	High-throughput sequencing
ICVG	International council for the study of viruses and virus diseases of the grapevine
Indel	Insertion-deletion
ISEM	Immuno-specific electron microscopy
LiCl	Lithium chloride
LF	Length fraction
m	Meter
M	Molar

ML	Maximum likelihood
mm	Millimetre
mM	Millimolar
MMLV	Moloney murine leukemia virus
mg	Milligram
min	Minutes
mRNA	Messenger ribonucleic acid
MTR	Methyltransferase
$MgCl_2$	Magnesium chloride
NaAc	Sodium acetate
NaCl	Sodium chloride
NCBI	National centre for biotechnology information
ng	Nanogram
nm	Nanometer
ORF	Open reading frame
PAMV	Petunia asteroid mosaic virus
PBS	Phosphate buffer saline
PC	Positive control
PCR	Polymerase chain reaction
PolyHiT-Seq	Poly-specific PCR and high-throughput sequencing
PPi	Pyrophosphate
PRMV	Peach rosetta mosaic virus
PVP	Polyvinyl-pyrrolidone
PVPP	Polyvinylpolypyrrolidone
QC	Quality score
rbcLa	Ribulose bisphosphate carboxylase large chain
RdRp	RNA dependant RNA polymerase
RM	Reference mapping/mapped
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RpRSV	Raspberry ringspot virus
RT-PCR	Reverse transcription polymerase chain reaction
RW	Rugose wood disease complex
s	Seconds

SBS	Sequencing by synthesis
SF	Similarity fraction
siRNA	Small/small interfering RNA
SIQ	Sepsis indicating quanitfier
SMS	Single molecule sequencing
SNP	Single nucleotide polymorphism
TBRV	Tomato blackring virus
TBSV-ch	Tomato bushy stunt virus cherry isolate
ToRSV	Tomato ringspot virus
TRSV	Tobacco ringspot virus
U	Units of enzyme
µl	Microliter
µM	Micromolar
UTR	Untranslated region
UV	Ultra-violet light
VIA	South African vine improvement association
V. vinifera	*Vitis vinifera*
ºC	Degree Celsius

LIST OF FIGURES

FIGURE **PAGE**

APPENDIX A

LIST OF TABLES

TABLE **PAGE**

APPENDIX A

TABLE OF CONTENTS

Chapter 1:

Literature Review

1.1. BRIEF BACKGROUND ON THE GRAPEVINE

Together with the civilization of man arose the domesticated grape, *Vitis vinifera* subspecies *vinifera* (McGovern, 2003, This *et al.*, 2006), derived from its wild progenitor, *Vitis. vinifera* subspecies *sylvestris* (Reisch *et al.*, 2012). The separation of the subspecies were based on their biochemical and morphological differences such as higher sugar content, better uniformity of berry maturity within clusters, broader selection in fruit colour, perfect flowers, greater yield and more regular production (This *et al.*, 2006, Reisch *et al.*, 2012). It is not certain if the events responsible for the changes that resulted in these differences occurred through human or natural selection over a long period of time, or quickly through mutations or vegetative propagation (This *et al.*, 2006). The wild form of *Vitis vinifera* is very rare and exclusively found in isolated areas of Western Europe, central Asia and Northern Africa (Reisch *et al.*, 2012), whereas the domesticated form [from here on referred to only as *Vitis vinifera* (*V. vinifera*)] is cultivated worldwide (Myles *et al.*, 2011).

Vitis is the only genus within the *Vitaceae* family that is considered to be of major agronomical importance (This *et al.*, 2006) since it is the only genus that produces edible fruits (Burger *et al.*, 2009). It consists of approximately 60 species with *V. vinifera* as the primary species used for fruit production (Reisch *et al.*, 2012). Through the years thousands of *V. vinifera* cultivars have been generated by crosses and vegetative propagation (Myles *et al.*, 2011). These cultivars are usually classified according to their production uses; table grapes, wine grapes and raisins (This *et al.*, 2006). Other uses of the grape includes the production of seed oil, vinegar, juices, jellies and jams (Naidu *et al.*, 2014), but the production of wine is considered the most important with regards to production area and tonnage (Reisch *et al.*, 2012). Although *V. vinifera* is the most widely cultivated grape species, other cultivated *Vitis* species include *V. berlandieri*, *V. riparia* and *V. rupestris*. These species are mainly used as rootstock material for the protection against various pathogens, their vectors and adverse soil conditions (Oliver and Fuchs, 2011).

The perennation of grapevine has enforced it’s propagation for thousands of years, allowing the spread of vines around the world from as early as the rise of the roman empire (This *et al.*, 2006). The exchange of source vines among nurseries worldwide together with insect and nematode transmission of infectious agents

(Martelli, 1993a, Zhang *et al.*, 2011) has resulted in grapevine harbouring the most intracellular pathogens known in a single crop (Martelli, 2014). These pathogens consist of approximately 65 viruses across 27 genera, various viroids and phytoplasmas (Martelli, 2014). Adverse effects on infected vines are associated with these pathogens, by causing a reduction in graft compatibility between scions and rootstocks, decline in plant vigour, serious crop losses, decreasing the productive life of vineyards and jeopardizing the survival of infected vines (Martelli and Boudon-Padiue, 2006, Martelli, 2014).

1.2. VIRUS DISEASES OF GRAPEVINE

In 2012 the International Council for the Study of Viruses and Virus Diseases of the Grapevine (ICVG) recognized 75 infectious agents of the grapevine, which include viruses, viroids, and phytoplasmas (Martelli, 2014). Of these 75 agents, 65 constitute viruses (*Table 1.1*) which are responsible for a wide range of economically important diseases worldwide (Martelli, 2014). The disease complexes considered of most economic importance include the grapevine leafroll disease complex, rugose wood complex, fleck complex, grapevine degeneration and grapevine decline (Martelli, 1993a).

Table 1.1: Viruses of grapevines and their taxonomic affiliation (Martelli, 2014, Al Rwahnih *et al.*, 2015b, Basso *et al.*, 2015).

FAMILY	GENUS	SPECIES
Viruses with single-stranded DNA genome		
Geminiviridae	Geminivirus	*Grapevine redblotch associated virus* (GRBaV)
Viruses with double-stranded DNA genome		
Caulimoviridae	Badnavirus	*Grapevine vein clearing virus* (GVCV) Unnamed virus
Viruses with double-stranded RNA genome		
Reoviridae	Oryzavirus Reovirus	Unnamed virus *Grapevine cabernet sauvignon virus* (GCSV)
Endornaviridae	Endornavirus	Two unnamed viruses
Partitiviridae	Alphacryptovirus	*Raphanus sativus cryptic virus 3* (RsCV-3) *Beet cryptic virus 3* (BCV-3)
Viruses with single-stranded RNA genome		
Closteroviridae	Closterovirus	*Grapevine leafroll associated virus 2* (GRLaV-2)
	Ampelovirus	*Grapevine leafroll associated virus 1* (GLRaV-1) *Grapevine leafroll associated virus 3* (GLRaV-3) *Grapevine leafroll associated virus 4* (GLRaV-4) GLRaV-4 strain 5 GLRaV-4 strain 6 GLRaV-4 strain 9 GLRaV-4 strain Pr

		GLRaV-4 strain Car
	Velarivirus	*Grapevine leafroll associated virus 7* (GLRaV-7)
Betaflexiviridae	Foveavirus	*Grapevine rupestris stem pitting associated virus* (GRSPaV)
	Vitivirus	*Grapevine virus A* (GVA) *Grapevine virus B* (GVB) *Grapevine virus D* (GVD) *Grapevine virus E* (GVE) *Grapevine virus F* (GVF)
	Trichovirus	*Grapevine berry inner necrosis virus* (GBINV) *Grapevine pinot gris virus* (GPGV)
Alphaflexviridae	Potexvirus	*Potato virus X* (PVX)
Potyviridae	Potyvirus	Unnamed virus
Virgaviridae	Tobamovirus	*Tobacco mosaic virus* (TMV) *Tomato mosaic virus* (ToMV)
Secoviridae	Nepovirus A	*Grapevine fanleaf virus* (GFLV) *Raspberry ringspot virus* (RpRSV) *Arabis mosaic virus* (ArMV) *Grapevine deformation virus* (GDefV) *Tobacco ringspot virus* (TRSV)
	Nepovirus B	*Grapevine anatolian ringspot virus* (GARSV) *Artichoke italian latent virus* (AILV) *Grapevine chrome mosaic virus* (GCMV) *Tomato blackring virus* (TBRV)
	Nepovirus C	*Blueberry leaf mottle virus* (BBLMV) *Cherry leafroll virus* (CLRV) *Grapevine bulgarian latent virus* (GBLV) *Grapevine tunisian ringspot virus* (GTRSV) *Peach rosetta mosaic virus* (PRMV) *Tomato ringspot virus* (ToRSV)
	Fabavirus	*Broadbean wilt virus* (BBWV)
	Sadwavirus	*Strawberry latent ringspot virus* (SLRSV)
Tymoviridae	Marafivirus	*Grapevine asteroid mosaic associated virus* (GAMaV) *Grapevine Syrah virus 1* (GSyV-1) *Grapevine rupestris vein feathering virus* (GRVFV)
	Maculavirus	*Grapevine fleck virus* (GFkV) *Grapevine redglobe virus* (GRGV)
Tombusviridae	Tombusvirus	*Grapevine algerian latent virus* (GALV) *Petunia asteroid mosaic virus* (PAMV)
	Carmovirus	*Carnation mottle virus* (CarMV)
	Necrovirus	*Tobacco necrosis virus D* (TNV-D)
Bromoviridae	Alfamovirus	*Alfalfa mosaic virus* (AMV)
	Cucumovirus	*Cucumber mosaic virus* (CMV)
	Ilarvirus	*Grapevine line pattern virus* (GLPV) *Grapevine angular mosaic virus* (GAMoV)
Unassigned viruses		
	Idaeovirus	*Raspberry bushy dwarf virus* (RBDV)
	Sobemovirus	*Sowbane mosaic virus* (SoMV)
		Grapevine ajinashika virus (GAgV)
		Grapevine stunt virus (GSV)
		Grapevine labile rod-shaped virus (GLRSV)
		Southern tomato virus (STV)
		Temperate fruit decay associated virus (TFDaV)
		Unnamed filamentous virus

The grapevine leafroll disease complex (GLD) is one of the most common diseases of grapevines worldwide (Abou Ghanem-Sabanadzovic *et al.*, 2010). It is induced by various individuals of a complex of viruses belonging in the *Closteroviridae* family (Almeida *et al.*, 2013) which are named accordingly as grapevine leafroll associated viruses (*Table 1.1*). Disease symptoms include the downward rolling of the leaf margins that is often associated with chlorotic mottling of the inter-veinal areas of the leaves. In red-berried cultivars purple-reddish discolouration and in white-berried cultivars a slight yellow discolouration is observed on the blades (Martelli, 1993a, Naidu *et al.*, 2014). GLD can result in the reduction of fruit size, without proper colouration of the berries (Martelli, 1993a), low sugar content, reduced fruit yield, plant vigour and longevity (Martelli, 2014). GLD is disseminated through infected propagation material, coccids and pseudococcids which are the natural vectors of the grapevine leafroll associated viruses (Martelli, 1993a, Naidu *et al.*, 2014).

The rugose wood complex (RW) consists of several diseases namely; grapevine rupestris stem pitting, grapevine kober stem grooving, grapevine corky bark and grapevine LN33 stem grooving (Martelli, 1993a), which can be differentially recognized by their symptom expression on their various indicators for graft transmission (Martelli, 2014). The causal agents for this complex are the viruses belonging to the genera *Vitivirus* and *Foveavirus* (*Table 1.1*) (Martelli, 2014). The characterization of this disease complex is based on the modification of the grapevine’s woody cylinder that is marked by pits and grooves (Martelli, 1993a). The symptoms induced consist of swelling above the bud union, delayed bud opening (Martelli, 1993a), graft incompatibility and severe decline which can be observed by most scion/rootstock combinations in the field (Fajardo *et al.*, 2012). The RW complex can be transmitted through infected propagation material and grafting (Fajardo *et al.*, 2012).

As with the RW the fleck complex also consists of several diseases; grapevine fleck, grapevine asteroid mosaic, grapevine rupestris necrosis and grapevine rupestris vein feathering (Martelli, 2014). The etiological agents for this complex belong to the genera *Maculavirus* and *Marafivirus* (*Table 1.1*). Disease symptoms include leaf wrinkling, twisting and stunting with reduced rooting ability, but latent infections are common for the viruses responsible (Oliver and Fuchs, 2011). There are no known vectors for any of the viruses involved in this complex but the complex is regularly spread through infected propagation material (Martelli, 2014).

Degenerate diseases of grapevine are caused by the viruses belonging to the genus *Nepovirus* (*Table 1.1*) (Oliver and Fuchs, 2011). Infectious degeneration is caused by European nepoviruses that cause symptoms similar to *Grapevine fanleaf* *virus* (GFLV). GFLV is the oldest known, most widespread and economically important *Nepovirus* (Martelli, 2014). Symptoms include chlorotic mottling with foliar deformations and stunting of the vines (Martelli, 1993a). Several *Nepoviruses* are naturally spread by nematode vectors (Martelli, 1993a).

1.3. PROPAGATION OF GRAPEVINES WITHIN SOUTH AFRICA

Grapevine is propagated vegetatively by cuttings to avoid the variability in progeny associated with sexual reproduction (Alley, 1980). Vegetative propagation introduces ample opportunity for the spread of viral diseases during large scale propagation for commercial use. The propagation of a virus infected source plant results in the carry-over of that virus into the progeny plants, which will upon the establishment in the field start to show signs of disease. Virus elimination is not feasible for vines established in the field. The only control is an integrated strategy of prophylactic measures; infected vine removal, vector exclusion and the production and establishment of virus-free vines (Almeida *et al.*, 2013). Certification schemes worldwide share the objective to produce plant material that have undergone specific procedures to ensure their health status and horticultural characteristics (Varveri *et al.*, 2015). Thus to prevent the spread of disease the propagated material is certified as free from specific infectious agents prior to large scale propagation. The principles of grapevine certification remain consistent throughout the globe; primary source selection, nuclear stock production (sanitation), nuclear stock propagation, foundation and mother block establishment, certification and labelling.

Plants selected as candidate material for propagation are mainly selected for their agronomical characteristics. The sanitary status of these plants is usually unknown (Varveri *et al.*, 2015), but they are selected due to the lack of evident viral symptoms (EPPO, 2008). The cuttings from these plants undergo thermotherapy, which consist of growing them at a temperature of 37°C from 4 weeks up to a year. Thermotherapy is employed for virus elimination from the tip of the vine, as the temperature is a compromise between virus degradation and plant survival. The meristem-tips (0.2 – 0.7 mm) of these plants are excised and grown *in vitro* for root and shoot development (Panattoni *et al.*, 2013). The combination of heat treatment and *in vitro* meristem-tip production has offered encouraging results but the mechanisms of virus exclusion from this part of the plant has not been fully determined (Panattoni *et al.*, 2013). Factors that might be responsible for the elimination of virus particles from the meristems at high temperatures, include the inactivation of virus particles, reduced movement of virus particles and thermally induced RNA silencing (Grout, 1999). The inactivation of virus particles might be ascribed to the change of the particles at heat treatments above 35°C, that include the rupture of hydrogen and disulphide bonds of the capsid protein and rupture of the nuclear phosphodiester covalent bonds that as a consequences can deteriorate viral infectivity (Panattoni *et al.*, 2013). Szittya *et al*., (2003) found that small interfering RNAs (siRNA) increased with temperature and with an increase of siRNAs, RNA silencing is triggered within the plant, which may explain the lack of virus particles within the meristem.

The plants generated by heat therapy and meristem-tip culture are considered candidate nuclear stock plants that are maintained in insect-free greenhouses (Almeida *et al.*, 2013). The sanitary status of these plants are evaluated by making use of biological indexing, serological and molecular assays (Maliogka *et al.*, 2015). Certification for the elimination of infectious agents differs among countries. In South Africa grapevines are certified under the authority of the South African Vine Improvement Association (VIA). The certification scheme of grapevine was licensed under the Plant Improvement Act No. 53 of 1976 of the South African Department of Agriculture, Forestry and Fisheries (http://www.gov.za). The phytosanitary requirements for plants under this Act require that both scion and rootstocks be free from the following virus diseases: grapevine fanleaf, grapevine fleck, grapevine leafroll, corky bark, grapevine stem pitting/grooving and Shiraz disease.

After the candidate nuclear plants have been screened and found negative for the specific diseases as required for certification, they are maintained in greenhouses to prevent infection by soil and air vectors. Planting material from these nuclear vines are propagated for commercial use and must be propagated in as few steps as possible under conditions that ensure freedom from infection (EPPO, 2008, Varveri *et al.*, 2015). Foundation blocks are established from the propagation of the nuclear vines either in greenhouses or open field plantings (Almeida *et al.*, 2013). Virgin soil (soil with no grapevine history) is preferred for the establishment of open field foundation blocks (Almeida *et al.*, 2013), or soil that has not hosted grapevine for at least 6 years (EPPO, 2008). The soil is required to be free from virus-transmitting nematodes and must be at least 26 m from other vineyards. These vines are tested for the presence of viruses associated with grapevine leafroll disease (GLRaV-1, -2 and -3), and when found free of them are used to establish mother blocks (commercial vineyards free from virus-infecting nematodes) (Almeida *et al.*, 2013).

A number of virus detection techniques are utilised to test candidate nuclear plants for their virus status. Biological indexing is a compulsory test during certification for the evaluation of disease, especially since many disease's etiological agents are unknown (EPPO, 2008). Testing is performed by grafting material from the candidate nuclear plants onto *Vitis* indicator vines. The indicator plants express disease symptoms in the presence of specific pathogens, but the test is based on the interpretation of expressed symptoms for disease diagnosis rather than that of a specific virus (Constable *et al.*, 2013). Once the candidate nuclear plants have been grafted onto indicator vines they are observed for symptom expression over 2 – 3 years (Rowhani *et al.*, 1997). Different indicator vines are used for different disease symptom expression, for example *V. rupestris* cv. rupestris St. George elicits symptoms for fanleaf, fleck and rupestris stem pitting disease, *V. vinifera* cv Pinot Nior and Cabernet Franc express symptoms for leafroll disease, *V. berlandieri x V. riparia* cv Kober 5BB for stem grooving and LN33 for corky bark (Rowhani *et al.*, 2005, EPPO, 2008, Constable *et al.*, 2013). There are several challenges associated with this test. It requires a lot of space for the plants to be kept over very long periods of time (Dovas and Katis, 2003b), the failure of the graft to take results in the failure of transmission of the virus from the bud to the indicator vine. The grafting of an uninfected bud due to uneven virus distribution in the candidate plant and the lack of experience for symptom interpretation has been reported to result in false negatives (Rowhani *et al.*, 1997, Constable *et al.*, 2013). The challenges associated with this technique make it unreliable when used without additional tests (Legrand, 2015).

The most widely applied serological assay used to monitor the sanitary status of candidate nuclear vines is the enzyme-linked immunosorbent assay (ELISA) (Boonham *et al.*, 2014, Varveri *et al.*, 2015). This technique is based on the affinity of antibodies to recognize the specific viral proteins that they have been raised against (Webster *et al.*, 2004). The advantages of this technique is that it is cost effective, simple to perform, results can be obtained within 2 days and large sample sizes can be tested easily (Boonham *et al.*, 2014). However, for ELISA to be performed successfully high-quality antiserum for every disease to be tested is required, the antiserum must also not react to plant proteins (Weber *et al.*, 2002, Webster *et al.*, 2004, Varveri *et al.*, 2015) and the virus titre within the sample must be high, thus sampling the appropriate tissue at the correct time is prudent (Weber *et al.*, 2002, Varveri *et al.*, 2015). Unfortunately antiserum has not yet been produced for all grapevine viruses or they do not react to all the strains of a specific virus, thus many viruses go undetected (Weber *et al.*, 2002).

Immuno-specific electron microscopy (ISEM) is another technique that makes use of antibodies for the detection of viruses. Virus specific antibodies are coated onto a carbon-treated plastic film of an electron microscope grid, allowing binding to the grid of homologous and related viruses present, for direct visualization by an electron microscope (Zimmermann *et al.*, 1990). With the addition of a second virus specific antibody coating step (decorating step) the virus particle can be identified to the species level, along with the additional viruses trapped by the coating antibody (Li *et al.*, 1993, Pietersen and Kasdorf, 1993), thus it is possible to detect viruses in mixed populations (Martin, 1998). However, to do this technique, an electron microscope is required as well as trained personnel for the operation of the microscope (Walter, 1993) and hence is very expensive. Therefore, this technique is also not suitable for large-scale routine testing because of its labour intensiveness and cost (Martelli, 1993b, Walter, 1993, Martin, 1998).

Molecular assays are based on the detection of genetic material (nucleic acid) of the viruses rather than their proteins (Weber *et al.*, 2002). The polymerase chain reaction (PCR) involves the amplification of a specific part of the genome, offering high levels of specificity. For RNA viruses, reverse-transcription PCR (RT-PCR) is used to first convert the RNA to complimentary DNA (cDNA) before amplifying to large quantities (Webster *et al.*, 2004). With the lack of antiserum for ELISA of many grapevine viruses this technique has proven to be very useful for the detection of those viruses (EPPO, 2008). This technique has proven successful in the presence of low virus titres, thus it’s sensitivity circumvents the need for sampling at specific times of the year where the virus titre is expected to be high (Maliogka *et al.*, 2015). Besides the sensitivity of RT-PCR, it is a rapid technique that generates results within a few hours (Maliogka *et al.*, 2015), and doesn’t require much skill to be performed (Webster *et al.*, 2004). Within sanitation programs this technique is very helpful in early screening of candidate nuclear vines as only a little material (200 mg) is required of the fresh *in vitro* grown tissue (EPPO, 2008). Methods based on nucleic acid are very generic and provide opportunities to obtain additional information to that of just virus presence or absence. Upon the sequencing of the amplified products produced by PCR, information can be obtained about virus strains and relatedness (Webster *et al.*, 2004, Boonham *et al.*, 2014). However, RT-PCR has not been used routinely in diagnostic laboratories because of the relatively high costs and problems associated with contamination. Since PCR is so sensitive post-PCR contamination by the liberation of small quantities of amplified products into the laboratory has resulted in diagnosis based on false positives (Boonham *et al.*, 2014). Practices are available to minimize the risk of cross contaminating samples by the introduction of mechanical barriers (creating master mixes in different areas/laboratories than where the template is added and the products are screened), sterilization of work spaces and apparatus with ultra-violet light irradiation (UV), 10 % bleach and ethanol, inactivation of nucleic acids with furocoumarins (base pair intercalating reagents), primer hydrolysis and many more (Aslanzadeh, 2004).

1.4. HIGH-THROUGHPUT SEQUENCING

High-throughput sequencing (HTS) has been used since the release of the first commercial HTS platform in 2005. The advent of this second generation of sequencing platforms has revolutionized our ability to sequence across all fields of biology. The reason for the wide adoption of these platforms is their ability to simultaneously sequence multiple samples in parallel without the need for *in vivo* amplification (bacterial cloning) as required with the first generation Sanger sequencing technology (Radford *et al.*, 2012), as well as the large amounts of data that they produced inexpensively (Kircher *et al.*, 2012).

The second generation of sequencing platforms share the requirement of *in vitro* amplification (by PCR) prior to sequencing, but the method of amplification among the platforms may vary. With the Roche 454 system; prior to sequencing, templates are amplified by emulsion PCR (emPCR). Adapters are ligated to sheared templates that allow their hybridization to microbeads via bound primers (*Figure 1.1 a*). Each individual bead is captured in a water-in-oil emulsion that contains PCR reagents creating an emulsion microreactor for amplification (*Figure 1.1 b*). Following thermal cycling emulsion-based amplification, each bead is coated by a single clonally amplified molecule (*Figure 1.1 c*) at which stage the emulsion is disrupted and the beads are washed over a picotitre plate that fits one bead per well. Templates are then sequenced by pyrosequencing which is based on sequencing by synthesis

(SBS), where pyrophosphate (PPi) is released with the incorporation of each nucleotide. The PPi is converted to ATP that in turn drives luciferase for light production proportional to base incorporation which is measured using a charged coupled device (CCD) (Kircher and Kelso, 2010, Radford *et al.*, 2012, Heather and Chain, 2016). A drawback of pyrosequencing is the difficulties associated with sequencing homopolymeric regions due to their attribution to insertions and deletions, increasing the error rate of a run (Morozova and Marra, 2008, Gilles *et al.*, 2011). Currently the GS FLX can generate 700 Mb/run at an approximate run time of 23 h, with a read length of 700 bp (Wu *et al.*, 2015).

Applied Biosystems SOLiD (Sequencing by Oligo Ligation Detection) is another second generation sequencing platform that makes uses of emPCR. As described above; templates are clonally amplified in a water-in-oil emulsion. Following the disruption of the emulsion, the clonally coated beads are attached to a glass slide for sequencing. Sequencing takes place by hybridization and ligation, making use of ligase rather than polymerase. A sequencing primer hybridizes to the template molecule followed by the ligation of a fluorescently labelled 8-mer probe (library of probes consist of 16 dinucleotide combinations). The fluorescence is read and the three 5’ universal bases including the label is cleaved from the probe, exposing a free 5’ phosphate on the remaining 5-mer, from which the process is repeated. The hybridization and ligation process is repeated with starting primers that are displaced by 1, 2, 3 and 4 nucleotides in the 5’ direction. Therefore, sequence information for different positions of the sequence is gathered for each starting primer e.g. the first primer gives information on positions 1 and 2, 6 and 7, etc., and the second primer gives information on positions 2 and 3, 7 and 8, etc. Accordingly each position in the sequence is investigated twice resulting in high accuracy base calls (Morozova and Marra, 2008, Kircher and Kelso, 2010, Radford *et al.*, 2012). To date SOLiD 5500xl can generate 95 Gb/run, 2 x 60 bp at a run time of approximately 60 days (Wu *et al.*, 2015).

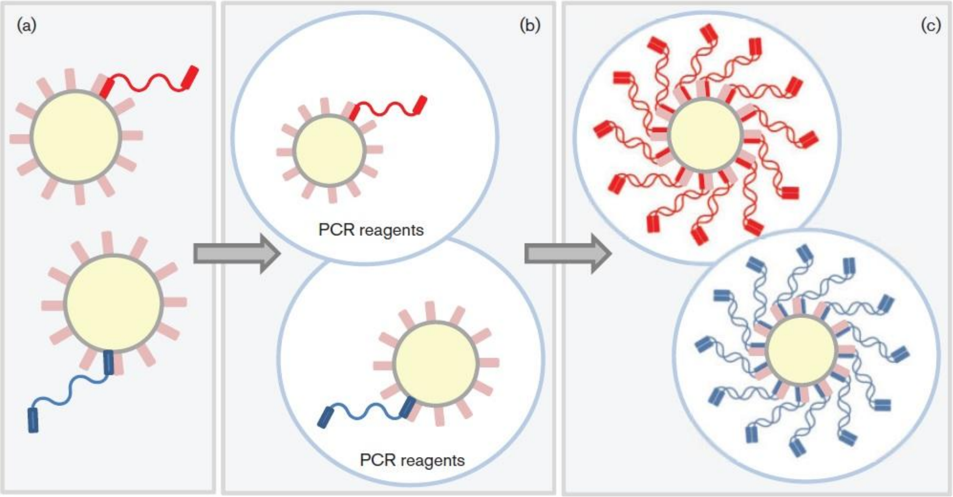

Figure 1.1: Basic principles of template amplification by emulsion PCR (Roche 454 and SOLiD). a) Single template molecules with ligated adapters are captured by microbeads via primer hybridization. b) Each microbead is incorporated into a water-in-oil emulsion that contains PCR reagents. c) Following amplification, each bead is coated with a single clonally amplified molecule (Radford *et al.*, 2012).

The method used for template preparation by Illumina platforms, is different from that of Roche 454 and SOLiD's emPCR. In the former, templates are clonally amplified by bridge amplification on a glass plate (flow cell) coated with oligonucleotides complimentary to the adapters ligated to the 3' and 5' ends of the template molecule. The adapters hybridize to their complimentary oligonucleotides on the flow cell which also acts to prime amplification (*Figure 1.2 a-b*). The template molecule and newly synthesized strand can bend and hybridize to an oligonucleotide complimentary to the second adapter to form a 'bridge' for amplification (*Figure 1.2 c-d*). Consecutive rounds of amplification results in clusters consisting of both forward and reverse strands across the flow cell of individual molecules (*Figure 1.2 e*). Prior to sequencing one of the strands is removed to prevent complimentary base pairing within the clusters, resulting in clonal single-stranded clusters. The clusters are sequenced using different reversible dye terminators that after incorporation, temporarily terminates the reaction so that the base can be read. After imaging, the terminator is chemically cleaved, allowing further extension of the molecule (Kircher and Kelso, 2010, Radford *et al.*, 2012). The HiSeq 2500 can sequence 2 x 125 bp with a throughput of 5 – 1000 Gb/run at a run time of 1 – 6 days , while the MiSeq can sequence 2 x 300 bp with a throughput of 0.3 - 15 Gb/run at a run time of 5 – 55 h (Wu *et al.*, 2015).

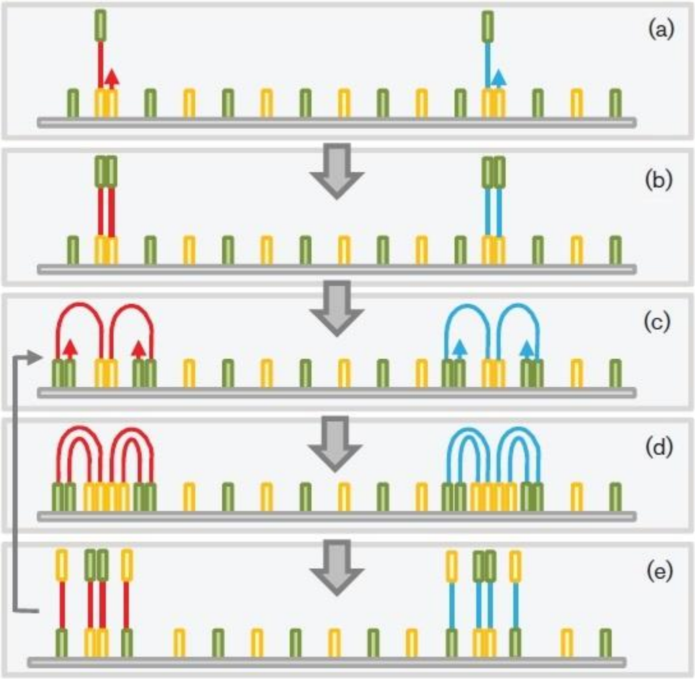

Figure 1.2: Basic principles of bridge amplification (Illumina). a) Single template molecules with ligated adapters anneal to glass plate (flow cell) via complimentary bound primer. b) The primer is used for extension. c) The free end of each single-stranded molecule anneals to a second primer bound to the flow cell, forming a 'bridge' that acts as template for (d) for amplification. e) Continuous amplification results in a cluster of clonal molecules (Radford *et al.*, 2012).

Single molecule sequencing (SMS) is considered to make up the third generation of sequencing technology. As the name implies no amplification step is required prior to sequencing, therefore eliminating the introduction of biases and errors, as introduced during clonal amplification (Egan *et al.*, 2012). The Helicos sequencer is considered the first single molecule sequencer. As with the second generation sequencers, adapters are ligated to sheared templates albeit in the form of poly(A)-adapters with a fluorescently labelled dATP. The templates are bound to a flow cell coated by oligo(dT) probes, and the position of the molecules are determined using fluorescence prior to the cleavage of the fluorescent label. DNA polymerase and fluorescently labelled, reversible terminator nucleotides are sequentially added and the incorporated base recorded (Shendure and Ji, 2008, Kircher and Kelso, 2010).

Together with the advancement of these technologies to reduce the run time and the cost per run, the applications of HTS technologies have diversified plant virus diagnostics and discovery. Due to the sequence independent nature of HTS technology, novel viruses have been discovered at an unprecedented rate (Zhang *et al.*, 2011, Poojari *et al.*, 2013). This feature has also proven useful in the detection of asymptomatic viruses and various viral strains that may go undetected by conventional serological and nucleic acid based methods such as ELISA and PCR (Al Rwahnih *et al.*, 2015a). Novel viruses of grapevines have been characterized through HTS from Italy, South Africa and the USA when making use of metagenomic approaches (Al Rwahnih *et al.*, 2009, Coetzee *et al.*, 2010, Giampetruzzi *et al.*, 2012). HTS has also been used as a diagnostic tool in comparison to the golden standard methodology used for plant virus diagnostics; biological indexing (Al Rwahnih *et al.*, 2015a). It was suggested that HTS could be used to replace biological indexing for grapevine virus diagnostics, given that results were obtained much quicker and viruses were detected amongst plants which tested negative with indexing (Al Rwahnih *et al.*, 2015a). Several studies have alluded to the potential for HTS in quarantine programs (Barba *et al.*, 2014, Candresse *et al.*, 2014), especially with the release of the more affordable bench top sequencers (Kircher and Kelso, 2010).

With the decrease in cost per run and the increase in affordability with bench top sequencers, there has been an influx in short read data which places a great demand on the availability of bioinformatic tools, computational resources and knowledge for their analysis and interpretation, as well as long-term storage options (Muir *et al.*, 2016, Raza and Ahmad, 2016). To address this need for bioinformatic tools, many programs have been developed, some with specific functions such as alignment (Bowtie, GenomeMapper, MAQ), *de novo* assembly (VCAKE, Velvet), SNP/Indel discovery (GATK, SAM tools), transcriptome analysis [RNA-Seq (RNA-star, TopHat)], genome annotation browsers (LookSeq, Sequence Assembly Manager), chromatin immunoprecipitation sequencing [ChIP-Seq (CisGenome, CNV-Seq)], and others as integrated tools that can perform several functions (BaseSpace, CLC Genomics Workbench, Galaxy, NextGENe) (Raza and Ahmad, 2016). Even though many analysis tools are being developed they are inaccessible to many potential users because of the expertise and computational resources required for their use (Rose *et al.*, 2016), therefore cloud computing is of great value as data can be stored and analysis scripts can be performed remotely (Muir *et al.*, 2016).

Besides the need for adequate bioinformatic tools for HTS data analysis, proper data interpretation is also required. Concomitant with the sensitivity of HTS comes the problem of persistent background noise or sample cross contamination (Capobianchi *et al.*, 2013, Massart *et al.*, 2014, Roossinck *et al.*, 2015). Contamination can be introduced at several steps during sample preparation for sequencing; sample amplification, purification, pooling, handling of barcodes, library preparation, during the actual sequencing process and demultiplexing (Kircher *et al.*, 2012, Capobianchi *et al.*, 2013, Quail *et al.*, 2014). Sample amplification requires DNA polymerase and in the instance of RNA viruses, reverse transcriptase. During the synthesis of cDNA there is a definite risk for incorrect base incorporations as reverse transcriptases lack 3’→5’ exonuclease activity (proofreading activity). Many DNA polymerases also lack proofreading activity and this has been proven to introduce base substitution errors at around 10^{-4} to 10^{-5}, but the presence of 3’→5’ exonuclease activity decreases this error rate approximately 100-fold to around 10^{-6} to 10^{-7} bases (Kunkel *et al.*, 1992). Sample cross contamination can occur during sample purification and pooling as well as during the handling of sample barcodes by physical contamination of one sample with another. In addition to physical contamination, sample barcoding can result in background noise due to reads being incorrectly assigned to their original samples during demultiplexing. During library preparation a risk exists that chimeric sequences can be constructed as a result of too close cluster formation on a flowcell (Kircher *et al.*, 2012). Furthermore, phasing and pre-phasing has been reported to cause noise in the cluster signal due to the difference in molecule length within the cluster (Schirmer *et al.*, 2015). As a result of the unavoidable presence of background noise or cross contamination in HTS data, thorough analysis pipelines with specific cut-off thresholds are required to provide a measure of relevance. Grumaz *et al.,* (2016) established a sepsis indicating quantifier (SIQ) during the analysis of bacteraemia-based HTS data from septic patients as a measure to discriminate significant reads from background noise to ensure confidence in diagnosis. In concert with the need for the establishment of appropriate thresholds, the reproducibility of HTS data analysis is very important

(Nekrutenko and Taylor, 2012). As a result; many HTS data analysis pipelines have been established to ensure accurate and reproducible data analysis. Some examples of these pipelines are; VirusHunter for the analysis of 454 and long read platform HTS data of viruses (Zhao *et al.*, 2013), QUASR a pipeline for the quality assessment of HTS short reads (Watson *et al.*, 2013) and PANGEA for the analysis of HTS amplicons (Giongo *et al.*, 2010), to name but a few. Generalized pipelines for the analysis of specific datasets such as metagenomic data, RNA-Seq and ChIP-Seq has also been established by a variety of research groups (Gogol-Döring and Chen, 2012, Grada and Weinbrecht, 2013, Raza and Ahmad, 2016). Based on the constant evolution of HTS platforms enhancing their throughput and sensitivity, together with their inherent and undesired background noise and cross contamination, efficient data analysis pipelines are required for the numerous applications of HTS data to ensure confidence in data interpretation as well as analysis reproducibility.

1.5. REFERENCES

Abou Ghanem-Sabanadzovic, N., Sabanadzovic, S., Uyemoto, J.K., Golino, D., Rowhani, A. (2010) A putative new ampelovirus associated with grapevine leafroll disease. *Archives of Virology* 155**:**1871-1876

Al Rwahnih, M., Daubert, S., Golino, D., islas, c.m., Rowhani, A. (2015a) Comparison of Next Generation Sequencing vs. Biological Indexing for the optimal detection of viral pathogens in Grapevine. *Phytopathology*

Al Rwahnih, M., Daubert, S., Golino, D., Rowhani, A. (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* 387**:**395-401

Al Rwahnih, M., Golino, D., Daubert, S., Rowhani, A. (2015b). Characterization of a novel reovirus species in Cabernet Grapevine in California. *Presented at the Proceedings of the 18th Congress of ICVG*, Ankara, Turkey, 7-11 September, p. 194-195

Alley, C.J. (1980) Propagation of Grapevine. *California Agriculture* 29-30

Almeida, R.P., Daane, K.M., Bell, V.A., Blaisdell, G.K., Cooper, M.L., Herrbach, E., Pietersen, G. (2013) Ecology and management of grapevine leafroll disease. *Frontiers in Microbiology* 4**:**94

Aslanzadeh, J. (2004) Preventing PCR Amplification Carryover Contamination in a Clinical Laboratory. *Annals of Clinical and Laboratory Science* 34**:**389-396

Barba, M., Czosnek, H., Hadidi, A. (2014) Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* 6**:**106-136

Basso, M.F., da Silva, J.C.F., Fajardo, T.V.M., Fontes, E.P.B., Zerbini, F.M. (2015) A novel, highly divergent ssDNA virus identified in Brazil infecting apple, pear and grapevine. *Virus Research* 210**:**27-33

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., Mumford, R. (2014) Methods in virus diagnostics: from ELISA to next generation sequencing. *Virus Research* 186**:**20-31

Burger, P., Bouquet, A., Striem, M.J. (2009). Grape breeding. *In: Breeding Plantation Tree Crops: Tropical Species*, Jain, S.M., Priyadarshan, P.M. (ed.) Springer Science, New York, p. 161-189

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.-H., Fernandez, E., Martin, D.P. (2014) Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS One* 9**:**e102945

Capobianchi, M., Giombini, E., Rozera, G. (2013) Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection* 19**:**15-22

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 a vineyard. *Virology* 400**:**157-163

Constable, F., Connellan, J., Nicholas, P., Rodoni, B. (2013) The reliability of woody indexing for detection of grapevine virus-associated diseases in three different climatic conditions in Australia. *Australian Journal of Grape and Wine Research* 19**:**74-80

Dovas, C., Katis, N. (2003) A spot nested RT-PCR method for the simultaneous detection of members of the *Vitivirus* and *Foveavirus* genera in grapevine. *Journal of Virological Methods* 107**:**99-106

Egan, A.N., Schlueter, J., Spooner, D.M. (2012) Applications of next-generation sequencing in plant biology. *American Journal of Botany* 99**:**175-185

EPPO (2008) Pathogen-tested material of grapevine varieties and rootstocks. *European and Mediterranean Plant Protection Organization Bulletin* 38**:**422–429

Fajardo, T.V.M., Eiras, M., Nickel, O., Dubiela, C.R., Souto, E.R.d. (2012) Detection and partial molecular characterization of Grapevine fleck virus, Grapevine virus D, Grapevine leafroll-associated virus-5 and-6 infecting grapevines in Brazil. *Ciência Rural* 42**:**2127-2130

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., Terlizzi, F., Credi, R., Saldarelli, P. (2012) A new grapevine virus discovered by deep sequencing of virus-and viroid-derived small RNAs in Cv *Pinot gris*. *Virus Research* 163**:**262-268

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., Martin, J.-F. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12**:**245

Giongo, A., Crabb, D.B., Davis-Richardson, A.G., Chauliac, D., Mobberley, J.M., Gano, K.A., Mukherjee, N., Casella, G., Roesch, L.F., Walts, B. (2010) PANGEA: pipeline for analysis of next generation amplicons. *The ISME Journal* 4**:**852-861

Gogol-Döring, A., Chen, W. (2012) An overview of the analysis of next generation sequencing data. *Next Generation Microarray Bioinformatics: Methods and Protocols***:**249-257

Grada, A., Weinbrecht, K. (2013) Next-generation sequencing: methodology and application. *Journal of Investigative Dermatology* 133**:**1-4

Grout, B.W.W. (1999). Meristem-Tip Culture for Propagation and Virus Elimination. *In: Methods in Molecular Biology: Plant Cell Culture Protocols*, Hall, R.D. (ed.) Human Press Inc., Totowa, NJ, vol. 111, p. 115-125.

Grumaz, S., Stevens, P., Grumaz, C., Decker, S.O., Weigand, M.A., Hofer, S., Brenner, T., von Haeseler, A., Sohn, K. (2016) Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Medicine* 8**:**1

Heather, J.M., Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* 107**:**1-8

Kircher, M., Kelso, J. (2010) High-throughput DNA sequencing–concepts and limitations. *Bioessays* 32**:**524-536

Kircher, M., Sawyer, S., Meyer, M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* 40**:**e3-e3

Kunkel, T., Hizi, A., Shaharabany, M., Tsygankov, A., Bröker, B., Fargnoli, J., Ledbetter, J., Bolen, J., De Vivo, M., Chen, J. (1992) DNA replication fidelity. *Journal of Biological Chemistry* 267

Legrand, P. (2015) Biological assays for plant viruses and other graft-transmissible pathogens diagnoses: a review. *EPPO Bulletin* 45**:**240-251

Li, Z., Guo, D.Y., Guo, Z.N., Feng, G.F., Kuai, C.H. (1993). Electron microscope observation of grapevine leafroll virus. *Presented at the Meeting of the International Concil for the Study of Viruses and Virus Diseases of Grapevine (ICVG)*, Montreux, Switzerland, 6-9 September, p. 50

Maliogka, V.I., Martelli, G.P., Fuchs, M., Katis, N.I. (2015) Chapter Six-Control of Viruses Infecting Grapevine. *Advances in Virus Research* 91**:**175-227

Martelli, G. (1993a). Graft-transmissible diseases of grapevines: handbook for detection and diagnosis. FAO

Martelli, G. (1993b) Immunosorbent electron microscopy (ISEM) and antibody coating. *Graft-transmissible diseases of grapevines. Handbook for detection and diagnosis***:**193-195

Martelli, G.P. (2014) Directory of Virus and Virus-like Diseases of the Grapevine and their Agents. *Journal of Plant Pathology* 96

Martelli, G.P., Boudon-Padiue, E. (2006) Directory of Infectious Diseases of Grapevines and Viroses and Virus-like Diseases of the Grapevine: Bibliographic Report 1998-2004. *CIHEAM; Options Méditerranéennes: Série B. Etudes et Recherches; n. 55*

Martin, R. (1998) Advanced diagnostic tools as an aid to controlling plant virus diseases. *Plant virus disease control. APS Press, St Paul, MN***:**381-391

Massart, S., Olmos, A., Jijakli, H., Candresse, T. (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research* 188**:**90-96

McGovern, P.E. (2003). Ancient Wine: The Search for the Origins of Viniculture. Princeton University Press, Princeton

Morozova, O., Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92**:**255-264

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J. (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* 17**:**1

Myles, S., Boyko, A.R., Owens, C.L., Brown, P.J., Grassi, F., Aradhya, M.K., Prins, B., Reynolds, A., Chia, J.-M., Ware, D. (2011) Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences* 108**:**3530-3535

Naidu, R., Rowhani, A., Fuchs, M., Golino, D., Martelli, G.P. (2014) Grapevine Leafroll: A Complex Viral Disease Affecting a High-Value Fruit Crop. *Plant Disease* 98**:**1172-1185

Nekrutenko, A., Taylor, J. (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* 13**:**667-672

Oliver, J.E., Fuchs, M. (2011) Tolerance and resistance to viruses and their vectors in vitis sp.: a virologist's perspective of the literature. *American Journal of Enology and Viticulture***:**ajev. 2011.11036

Panattoni, A., Luvisi, A., Triolo, E. (2013) Review. Elimination of viruses in plants: twenty years of progress. *Spanish Journal of Agricultural Research* 11**:**173-188

Pietersen, G., Kasdorf, G.G.F. (1993). Use of IEM for the detection of the viruses of the grapevine leafroll complex in South Africa. *Presented at the Meeting of the International Council for the Study of Viruses and Virus Disease of the Grapevine (ICVG)*, Montreux, Switzerland, 6-9 September, p. 140-141

Poojari, S., Alabi, O.J., Fofanov, V.Y., Naidu, R.A. (2013) A Leafhopper-Transmissible DNA Virus with Novel Evolutionary Lineage in the Family Geminiviridae Implicated in Grapevine Redleaf Disease by Next-Generation Sequencing. *PLoS One* 8**:**e64194

Quail, M.A., Smith, M., Jackson, D., Leonard, S., Skelly, T., Swerdlow, H.P., Gu, Y., Ellis, P. (2014) SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics* 15**:**110

Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C., Hall, N. (2012) Application of next-generation sequencing technologies in virology. *Journal of General Virology* 93**:**1853-1868

Raza, K., Ahmad, S. (2016) Principle, analysis, application and challenges of next-generation sequencing: a review. *arXiv preprint arXiv:1606.05254*

Reisch, B.I., Owens, C.L., Cousins, P.S. (2012). Grape. *In: Fruit breeding*, Badenes, M.L., Byrne, D.H. (ed.) Springer, New York, p. 225-262

Roossinck, M.J., Martin, D.P., Roumagnac, P. (2015) Plant Virus Metagenomics: Advances in Virus Discovery. *Phytopathology* 10.1094/phyto-12-14-0356-rvw

Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L., Prosperi, M. (2016) Challenges in the analysis of viral metagenomes. *Virus Evolution* 2**:**vew022

Rowhani, A., Uyemoto, J.K., Golino, D.A. (1997) A comparison between serological and biological assays in detecting grapevine leafroll associated viruses. *Plant Disease* 81**:**799-801

Rowhani, A., Uyemoto, J.K., Golino, D.A., Martelli, G.P. (2005) Pathogen Testing and Certification of Vitis and Prunus Species*. *Annual Review of Phytopathology* 43**:**261-278

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research***:**gku1341

Shendure, J., Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology* 26**:**1135-1145

Szittya, G., Silhavy, D., Molnár, A., Havelda, Z., Lovas, Á., Lakatos, L., Bánfalvi, Z., Burgyán, J. (2003) Low temperature inhibits RNA silencing-mediated defence by the control of siRNA generation. *The EMBO Journal* 22**:**633-640

This, P., Lacombe, T., Thomas, M.R. (2006) Historical origins and genetic diversity of wine grapes. *Trends in Genetics* 22**:**511-519

Varveri, C., Maliogka, V.I., Kapari-Isaia, T. (2015) Chapter One-Principles for Supplying Virus-Tested Material. *Advances in Virus Research* 91**:**1-32

Walter, B. (1993). Advances in grapevine virus disease diagnosis since 1990. *Presented at the Meeting of the International Council for the Study of Viruses and Virus Disease of the Grapevine (ICVG)*, Montreux, Switzerland, 6-9 September, p. 127-130

Watson, S.J., Welkers, M.R., Depledge, D.P., Coulter, E., Breuer, J.M., de Jong, M.D., Kellam, P. (2013) Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical*

Transactions of the Royal Society of London B: Biological Sciences 368**:**20120205

Weber, E., Golino, D., Rowhani, A. (2002) Laboratory testing for grapevine diseases. *Practical Winery and Vineyard*

Webster, C.G., Wylie, S.J., Jones, M.G. (2004) Diagnosis of plant viral pathogens. *CURRENT SCIENCE-BANGALORE-.* 86**:**1604-1607

Wu, Q., Ding, S.-W., Zhang, Y., Zhu, S. (2015) Identification of Viruses and Viroids by Next-Generation Sequencing and Homology-Dependent and Homology-Independent Algorithms. *Annual Review of Phytopathology*

Zhang, Y., Singh, K., Kaur, R., Qiu, W. (2011) Association of a novel DNA virus with the grapevine vein-clearing and vine decline syndrome. *Phytopathology* 101**:**1081-1090

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V.L., da Rosa, A.P.T., Guzman, H., Cao, S., Virgin, H.W., Tesh, R.B., Wang, D. (2013) Identification of novel viruses using VirusHunter--an automated data analysis pipeline. *PLoS One* 8**:**e78470

Zimmermann, D., Bass, P., Legin, R., Walter, B. (1990) Characterization and serological detection of 4 Closterovirus-like particles associated with leafroll disease on grapevine. *Journal of Phytopathology-Phytopathologische Zeitschrift* 130**:**205-218 10.1111/j.1439-0434.1990.tb01169.x

Chapter 2:

Optimization of poly-specific PCRs and high-throughput sequencing (PolyHiT-Seq) data analysis

2.1. INTRODUCTION

In recent years molecular methods have gained popularity as diagnostic tools for plant viral pathogens because of their low cost, sensitivity, specificity and speed of detection (Webster *et al.*, 2004, Rowhani *et al.*, 2005, James *et al.*, 2006). PCR based methods are the most commonly used techniques for the detection of RNA and DNA plant viruses, and have the advantage that additional information can be obtained by the sequencing of their products (Webster *et al.*, 2004).

The advent of HTS platforms has revolutionized viral discovery (Al Rwahnih *et al.*, 2009, Giampetruzzi *et al.*, 2012, Poojari *et al.*, 2013) with exciting prospects in virus diagnostics. HTS improved upon first generation Sanger sequencing by its ability to sequence multiple samples in parallel while generating large amounts of data inexpensively (Mamanova *et al.*, 2010, Kircher *et al.*, 2012). Reads generated from these platforms (25 to 400 bp) are shorter than that of Sanger sequencing (300 to 750 bp) (Barba *et al.*, 2014), but with the rapid advancement of these technologies they are expected to soon overcome this shortcoming.

One of the most sought after properties of the HTS platforms is their capacity to sequence multiple samples on a single lane of a flow cell (multiplexing) through the mechanism of barcoding. Barcoding is achieved by the addition of individual ‘barcode’ sequences to each sample so as to differentiate them from each other during data analysis; this not only reduces time but also the cost per sample when sequencing large numbers of samples (Illumina Inc, 2014). The risk associated with sample multiplexing is that of sample cross contamination. This may be the result of either the handling of sample barcodes resulting in their cross contamination prior to library preparation, reads being falsely assigned to their original samples, and chimera’s which can form during library preparation (Kircher *et al.*, 2012). In diagnostics, this is a big concern due to the high level of confidence required when evaluating the viral population of a sample. One of the approaches to monitor this is the introduction of additional unique barcodes into the samples, which is better suited for variant differentiation (Kircher *et al.*, 2012, Quail *et al.*, 2014). For diagnostic studies, we have developed a protocol whereby a number of viruses are detected through amplification by PCR, using primers to virus genus specific conserved regions (poly-specific) and virus species-specific regions followed by the HTS of a pool of the amplicons from these PCRs to identify the viruses present. We have optimized the data analysis procedure by establishing a threshold value to discriminate between virus positive samples and background contamination during data analysis. We have utilized artificially prepared preparations of known viral sequences to assess various parameters of targeted sequencing.

2.2. MATERIALS AND METHODS

Establishing poly-specific and virus-specific RT-PCR systems

Several of the poly-specific and virus-specific primer sets used within in this study was obtained from literature (*Appendix A: Table 1*). However, poly-specific primers for both the *Tombusvirus* and *Marafivirus* genera were designed using CLC Main workbench version 6.8.1. The *Tombusvirus* primers were designed based on the following viral nucleotide sequences obtained from GenBank; *Grapevine algerian latent virus* (GALV) accession NC_011535 and *Tomato bushy stunt virus* cherry isolate (TBSV-ch, synonymous to *Petunia asteroid mosaic virus*, PAMV) accession M21958. Primers for the *Marafivirus* genus were designed based on the nucleotide sequences for *Grapevine-Syrah virus 1* (GSyV-1) accession NC_012484, *Grapevine rupestris vein feathering virus* (GRVFV) accession AY706994 and *Grapevine asteroid mosaic associated virus* (GAMaV) accession AJ249357.

Positive control material in the form of infected plant material, non-infectious cDNA clones, non-infectious DNA clones as well as infected total RNA (*Table 2.1*) were used for the testing of the selected and designed primer pairs. Amplified positive control materials were cloned into the pGEM-T Easy® vector system to maintain stable controls throughout the study.

Cloning and verification of positive control viruses for various poly-specific and virus-specific RT-PCR systems

The various viral positive controls for each poly-specific and virus-specific RT-PCR system used within this study (*Table 2.1*) were ligated into the pGEM-T Easy® vector system and transformed into competent *E. coli* JM109 cells (Promega; Madison, WI, USA) according to the manufacturer's instructions. Potential recombinant colonies were selected based on the blue/white screening method and the plasmids were isolated by the alkaline lysis plasmid miniprep procedure

(Sambrook, 2001). Putative recombinant plasmids were evaluated for the presence of an insert by plasmid directed PCR, which consisted of: 1x PCR reaction buffer, 2.5 mM $MgCl_2$ (Bioline; London, UK), 0.14 mM dNTP mix (Promega; Madison, WI, USA), 2 µM forward (T7) and reverse (SP6) primers each, 2.5 U BIOTAQ DNA polymerase (Bioline; London, UK), 1 µl template plasmid DNA and molecular grade water (Sigma-Aldrich; St. Louis, MO, USA) to a final reaction volume of 50 µl. The PCR cycling conditions used were: 92°C for 2 min, followed by 35 cycles of 92°C for 30 s, 55°C for 45 s and 72°C for 1 min, with a final extension of 72°C for 10 min. Following amplification, the PCR products were run on a 1% agarose gel stained with ethidium bromide for UV visualization. Recombinant plasmids with products of the expected fragment length were selected for confirmatory sequencing of their insert.

Prior to sequencing, the selected PCR products were purified by the addition of 10 U Exonuclease and 2 U FastAP (Thermo Scientific; Wilmington, DE, USA). The reaction mixture was incubated at 37°C for 15 min followed by deactivation for 15 min at 85°C. Purified products were Sanger sequenced in the forward direction using the vector specific T7 primer. The sequencing reaction consisted of: 1 µl 2.5x BidDye® Terminator mix v3.1, 2.25 µl 5x BigDye® v3.1 sequencing buffer (Applied Biosystems; Foster City, CA, USA), 2 µM T7 primer, 1-6 µl PCR product (dependent on the intensity of the band seen on agarose gel) and molecular grade water (Sigma-Aldrich; St. Louis, MO, USA) to a final reaction volume of 10 µl. The sequencing cycling conditions used were: 94°C for 1 min, followed by 30 cycles of 94°C for 10 s, 50°C for 5 s and 60°C for 4 min. For precipitation of the cycle sequenced products the following was added; 1 µl 125 mM ethylene-diamine-tetra-acetic acid (EDTA), 1 µl 3 M sodium acetate (NaAc) and 25 µl 100% molecular grade ethanol (Sigma-Aldrich; St. Louis, MO, USA). The reaction mixture was incubated at room temperature for 15 min, followed by centrifugation at 14 000 rpm for 30 min at 4°C. The pellet was washed with 70% ethanol and centrifuged at 14 000 rpm for 15 min at 4°C. The supernatant was discarded and the pellet air-dried. The samples were submitted to the African Centre for Gene Technologies, Automated Sequencing Facility, Department of Genetics, University of Pretoria, South Africa and sequenced using an ABI Prism® 3130XL Genetic Analyser (Applied Biosystems, Foster City, CA, USA). The sequences obtained were identified using The National Centre for Biotechnology Information (NCBI) Basic Local Alignment Search Tool [BLAST; http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990)].

Table 2.1: Grapevine viral clones generated within this study (primers used in Appendix A, Table 1).

Genus	Virus Name	Source	Cloned Gene Region	Insert size (bp)
Ampelovirus	Grapevine leafroll associated virus 3 (GLRaV-3 isolate 623)	ARC-PPRI SA collection material	ORF4	226
	Grapevine leafroll associated virus 4 (GLRaV-4 isolate LR106; 93/0942 - Chasselas)		ORF6 and 3' coding region	485
	Grapevine leafroll associated virus 5 (GLRaV-5; 92/1027 - Barlinka)			
	Grapevine leafroll associated virus 6 (GLRaV-6 isolate Estellat; 92/1023 - Ohanez)			
Closterovirus	Citrus tristeza virus (CTV isolate 12-8)	Gerhard Pietersen, University of Pretoria	RdRp gene	500-536
	Grapevine leafroll associated virus 2 (GLRaV-2 isolate 93/955; 93/0933 - Barlinka)	ARC-PPRI SA collection material		
	Grapevine leafroll associated virus 2 (GLRaV-2 clone WC-HSP-15; 93/0936 - Barlinka)			
Foveavirus	Grapevine rupestris stem pitting associated virus (GRSPaV; 03/0225 - Black Spanish)			199
Geminivirus	Grapevine redblotch associated virus (GRBaV isolate NY147)	Mysore R Sudarshana, UC Davies	Partial IR and CP gene	557
Maculavirus	Grapevine fleck virus (GFkV)	Gerhard Pietersen, University of Pretoria	MTR gene	572
Marafivirus	Grapevine rupestris vein feathering virus (GRVFV; 14/8209, 14/8211, 14/8212)	L. Louw, Picardie, Paarl, South Africa	Marafibox and partial CP gene	436
Nepovirus A	Grapevine fanleaf virus (GFLV)	Joe Tang (Neogen 1147-11, UK)	RdRp gene	340
	Raspberry ringspot virus (RpRSV)	Joe Tang, New Zealand		
Nepovirus B	Tomato black ring virus (TBRV)			250
	Grapevine anatolian ringspot virus (GARSV)	Toufic El Beaino, Italy		
Nepovirus C	Grapevine bulgarian latent virus (GBLV)			
Reovirus	Grapevine cabernet sauvignon virus (GCSV)	Adib Rowhani, UC Davis	Genomic segment 4	368
Tombusvirus	Tomato bushy stunt virus (TBSV cherry isolate syn. Petunia asteroid mosaic virus; PAMV)	Herman Scholtof, Texas A&M University	Partial CP gene and 3' UTR	284
Velarivirus	Grapevine leafroll associated virus 7 (GLRaV-7)	Adib Rowhani, UC Davis	RdRp gene	500-536
Vitivirus	Grapevine virus A (GVA isolate P163-1; Black Spanish)	ARC-PPRI SA collection material		199
	Grapevine virus A (GVA; 92/0630)			
	Grapevine virus A (GVA isolate GTG11-1)			
	Grapevine virus B (GVB; 93/0953)			
'tag' sequence	Grapevine leafroll associated virus 3 (GLRaV-3 isolate 623 containing tag sequence; 14-8213)	Synthesized by IDT	ORF4	232

Preparation of clones for Illumina MiSeq sequencing

A clone of each virus within the respective genera included in this study (*Table 2.1*) was subject to amplification with either poly-specific or virus-specific primers (*Appendix A: Table 1*) to obtain viral amplicons for analysis by Illumina MiSeq sequencing.

The clones were amplified using the standard amplification protocol used throughout this study; which consisted of: 1x GoTaq® Flexi Buffer, 2 mM $MgCl_2$, 1.25 U GoTaqG2® DNA polymerase (Promega; Madison, WI, USA), 0.2 mM dNTPs (Kapa Biosystems; Cape Town, South Africa), 0.2 µM forward and reverse primers each (*Appendix A: Table 1*), 0.5 µl sample template and molecular grade water (Sigma; St. Louis, MO, USA) made up to a reaction volume of 25 µl. The PCR conditions for the initial denaturation step was 95°C for 5 min, followed by 40 cycles of 30 s at 95°C for denaturation, 30 s at the specific primer annealing temperature of the primer pair (*Appendix A: Table 1*), and 1 min at 72°C for elongation. For the amplification of the *Viti*-, *Fovea*-, *Velari*- and *Closteroviruses* a standard nested PCR protocol, with only the primers differing, was performed (Dovas and Katis, 2003a;2003b). The first round of the nested PCR consisted of: 1x Biotaq NH_4 buffer, 15 mM $MgCl_2$, 0.5 U BIOTAQ DNA polymerase (Bioline; London, UK), 3.5 mM dNTPs (Kapa Biosystems; Cape Town, South Africa), 10 mM Dithiothreitol (Thermo Scientific; Vilnius, Lithuania), 18 U Ribolock RNase Inhibitor (Thermo Scientific; Vilnius, Lithuania), 40 U Moloney-murine leukemia virus reverse transcriptase (Promega; Madison, WI, USA), 0.5 µM forward and reverse primers each, 1 µl RNA template and molecular grade water (Sigma: St. Louis, MO, USA) to a total volume of 25 µl. The PCR conditions used were 37°C for 45 min, 50°C for 2 min, 94°C for 4 min followed by 5 cycles of 30 s at 95°C, 10 s at 43°C, 5 s at 38°C and 15 s at 72°C. This was followed by 35 cycles of 30 s at 95°C, 30 s at 43°C and 20 s at 72°C. The final step was 72°C for 2 min. For the second round of amplification 1x Biotaq NH_4 buffer, 15 mM $MgCl_2$, 0.5 U BIOTAQ DNA polymerase (Bioline; London, UK), 3.5 mM dNTPs (Kapa Biosystems; Cape Town, South Africa), 0.5 µM forward and reverse primers each, 0.5 µl of the first round product and molecular grade water (Sigma: St. Louis, MO, USA) was set-up to a total volume of 25 µl. The cycling conditions were 95°C for 3 min, 48°C for 15 s, 72°C for 15 s followed by 39 cycles of 30 s at 95°C, 30 s at 54°C and 10 s at 72°C. The final elongation was at 72°C for 2 min. Following amplification, the reaction products were visualized on a 1% agarose gel stained with ethidium bromide for UV visualization.

The PCR products of the various viral clones were purified using the NucleoSpin® Gel and PCR clean up kit (Clonetech Laboratories, Inc., Mountain View, California) as per manufacturer's instructions and quantified using the Quant-iT™ dsDNA BR Assay Kit with the Qubit® 2.0 Fluorometer (Life Technologies, Grand Island, NY, United States). The products were pooled into various combinations of molar ratios and number templates, as treatments (*Table 2.2; Appendix A: Table 1 – 7, Figure 1*) and the final concentrations determined with the Nanodrop 2000 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Further preparation included the barcoding of each sample with the Illumina TruSeq adapters (Illumina; San Diego, CA, USA) for the sequencing of each on an eighth of a lane on the Illumina MiSeq platform (Illumina; San Diego, CA, USA). Samples were sequenced at the Agricultural Research Council Biotechnology Platform (ARC-BP, Pretoria; South Africa).

To limit sample contamination during preparation in the laboratory, aseptic techniques were employed. These included; the sterilization of work spaces prior to nucleic acid isolation with 12% bleach and 3% hydrogen peroxide and preparation of all PCR reagent mixes in an enclosed space pre-treated with ultraviolet light. Preparation of the master mix was made prior to thawing of the nucleic acid templates, and handling of positive controls only after all unknown samples were added to the PCR reagents. Amplified products were evaluated by agarose gel electrophoresis and stored at -20°C for future use. Once the concentrations of the products were determined and the molar pooling ratios calculated [number copies template = (ng x 0.022 x 10^{23})/(bp x 1 x 10^{9} x 650)], each sample was pooled individually in a UV box with only the required templates thawed with surface and apparatus sterilization between each sample pooling. The addition of barcodes and the library preparation of samples were performed by the sequencing facility. As an additional control a single sample containing a 144 bp amplified product of *Ralstonia solanacearum* and a 500 bp amplified product of *Citrus tristeza virus* (CTV), neither of which would not be expected to occur naturally in grapevine samples were included in the first (sample A) (*Table 2.2; Appendix A: Figure 1*) and second sequencing run (sample G) (*Table 2.2; Appendix A: Figure 1*). This allowed for the evaluation of the extent of sample contamination/leaching among the pooled samples as introduced by the sequencing facility.

Illumina MiSeq data analysis

Illumina paired-end data was analysed using CLC Genomics workbench version 6.5.1 (CLC Bio; Aarhus, Denmark). The raw sequence data was imported as paired-ends (distance 180-300 nucleotides), with the removal of failed reads. Quality scores for each data set was generated using the FastQC function according to default settings and the reads filtered by the removal of low quality sequences (quality limit of 0.05), ambiguous nucleotides (maximum of 2 nucleotides) and TruSeq adapter sequences (TruSeq1: TCT AGC CTT CTC GCC AAG TCG TCC; TruSeq2: CCT GCT GAA CCG CTC TTC CGA TCT).

For the first round of sequence analysis (samples A – D) (*Appendix A: Figure 1*); reference mapping against the cognate areas of the different virus clones were conducted using default settings with the exception of the length fraction (LF) and similarity fraction (SF) parameters, following which all data (samples A – R) was reference mapped using only the default settings for all parameters. The unmapped reads of each sample were subject to *de novo* assembly from which the contigs were generated with the default settings. Contigs were analysed with the multiBLASTn search against the viral nucleotide collection hosted by the NCBI portal via the CLC interface.

Table 2.2: Synthetic pooling of samples from established viral clones for Illumina MiSeq sequencing*

			Closterovirus			Ampelovirus					Velarivirus	Vitivirus				Foveavirus	Nepovirus A		Nepovirus B		Nepovirus C	Tombusvirus	Geminivirus	Maculavirus	Marafivirus
Sample	Molar pooling	Ralstonia	CTV	93/0933 GLRaV-2	93/0936 GLRaV-2	GLRaV-3 TAG	GLRaV-3	GLRaV-4	GLRaV-5	GLRaV-6	GLRaV-7	P163-1 GVA	92/0630 GVA	GTG11-1 GVA	GVB	GRSPaV	GFLV	RpRSV	GARSV	TBRV	GBLV	TBSV	GRBaV	GFkV	GRVFV
A	1:1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
B	1:1			+			+		+		+			+		+	+	+	+		+	+	+		
C	1:2			++			+		+		+			+		+	+	+	+		+	+	++		
D	1:2			+			++		+		+			++		+	+	+	+		+	+	+		
E	1:1			+			+		+		+			+		+		+	+		+	+	+		
F	1:1			+			+		+		+			+		+		+	+		+	+	+		
G	1:1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
H	1:1			+			+		+		+			+		+		+	+		+	+	+	+	+
I	1:2:3			+++			+		+++		+++			+		+		++	++		++	++	+++	+++	+++
J	1:3			+++			+		+++		+++			+		+		+	+		+	+	+++	+++	+++
K	1:4				+++		+		++++		++++			+		+		+	+		+	+	++++	++++	++++
L	1:1													+		+			+		+				
M	1:1													+		+			+		+				
N	1:1													+		+			+		+				
O	1:1													+		+			+		+				
P	1:1													+		+			+		+				
Q	1:1													+		+			+		+				
R	1:2:3			+++			++		+++		+++			+		+		++	++		++	++	+++	+++	+++
S	1:2:3			+++			++		+++		+++			+		+		++	++		++	++	+++	+++	+++
T	1:3			+++			+		+++		+++			+		+		+	+		+	+	+++	+++	+++

*+ = the pooling of a viral clone into the specific sample at a molar ratio of 1 times the other templates, ++ at a molar ratio of 2 times the least molar templates, +++ at a molar ratio of 3 times the least molar templates, and ++++ at a molar ratio of 4 times the least molar templates.

Statistical analysis of Illumina MiSeq data

The statistical significance of the difference in percentage mapped reads between comparable samples were based on the two-tailed t-test:

H_o: lacks a difference between the two groups (P≤0.05)

H_A: there is a difference between the two groups (P>0.05)

$$t_s = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{s_1^2 + s_2^2}{n_1 + n_2}}}$$

Where ȳ$_1$ is the mean reads for sample 1 and ȳ$_2$, the mean reads for sample 2, s_1^2 the standard error of difference for sample 1 and s_2^2, the standard error of difference for sample 2, n_1 the number of templates present within sample 1 and n_2, the number of templates present within sample 2.

The significance based on sample poolings as a quantitative measure were evaluated with the chi-square test:

H_o: data quantitative, reads mapped as expected based on pooling (P≤0.05)

H_A: data not quantitative, reads did not map as expected (P>0.05)

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Where O is the observed reads and E is the expected reads.

2.3. RESULTS

Optimizing reference mapping of reads

Results are discussed based on Figure 1 (*Appendix A*), and it is recommended that the reader makes use of this figure (provided as a fold out option) as a guide throughout.

The optimal LF and SF parameters for reference mapping were taken as the percentages at which they resulted in the highest percentage of mapped reads, since known, defined templates were sequenced. The highest percentage of mapped reads were observed when the data was reference mapped using the default settings with a LF of 0.5 and a SF of 0.8 (of the required 50% of the read that has to map to the reference sequence, 80% of the nucleotides must be exact)

(*Figure 2.1*). Previously observed sample cross contamination, necessitated the need to determine a threshold value in order to differentiate between positive samples and false positives (due to sequence contaminants) of viruses associated with reads during reference mapping. A threshold sensitivity value was determined by mapping the reads of samples with a mixture of several defined sequences of representative viruses from each genus (*Figure 2.2 – 2.6, Table 2.2; Appendix A: Table 1 – 7, Figure 1*) against the whole range of clonal cognate sequences (*Table 2.1*). Therefore, any reads that mapped to templates that were not included in these samples of known composition were contaminants (false positives). This process allowed the establishment of the threshold of a percentage mapped reads of 0.4%. Thus, any reads mapping to viruses with a read mapping percentage of less than 0.4% was considered to be false positives due to cross contamination and discarded.

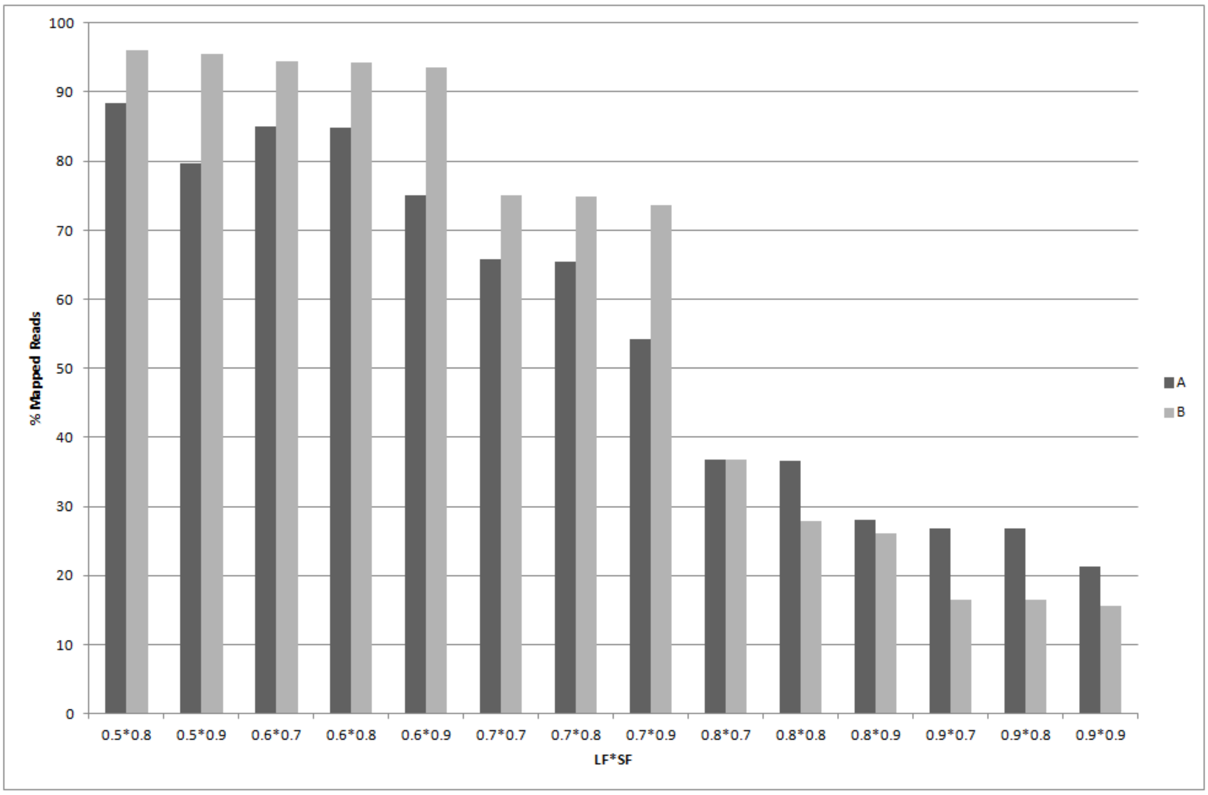

Figure 2.1: Reference mapping at various length fraction (LF) and similarity fraction (SF) parameters for sample A and B. Percentage mapped reads ranked from default settings to settings that are more stringent.

Ralstonia solanacearum and *Citrus tristeza virus* (CTV*)* were used as indicators for cross contamination/index leaching among samples. Using the positive/negative threshold determined above, *Ralstonia* and CTV would be interpreted as being present (true positive) only in samples A and G, the two samples in which they were included as templates (*Table 2.2*). However, *Ralstonia* and CTV reads were present at an average percentage mapped reads of 0.00015% and 0.0006% respectively, in the samples in which they were not included as templates, confirming the cross contamination problem inherent in the technique. However, this cross contamination rate falls far below the positive/negative threshold established at 0.4% of mapped reads. Samples from the second sequencing run (samples E – K) (*Table 2.2; Appendix A: Figure 1*) showed significant levels of contamination of *Grapevine fanleaf virus* (GFLV), a virus that had not been included as the representative template for the Nepovirus A genus PCR. The contamination however was tentatively traced back to the contamination of the preparations of dilutions for the clone which was used as the template for the Nepovirus A PCR. Because of this cross contamination, GFLV was present within all those samples containing the RpRSV template and was not due to contamination during the sequencing run. A fresh RpRSV clonal dilution was prepared for the third sequencing run (samples N – T) (*Appendix A: Figure 1*) whereupon the contamination was resolved.

To evaluate the level of experimental variation introduced at different stages of sample preparation in the laboratory, replicate samples were prepared from the same amplified templates after they were divided at specific stages during preparation (*Appendix A: Figure 1*). The specific stages in sample preparation chosen for evaluation were template purification (PCR cleanup, Exonuclease and FastAP), concentration determination and sample sequencing. Sample preparation as a whole was also investigated.

To measure the discrepancy from template purification by PCR cleanup (Exonuclease and FastAP); samples L and M were prepared from the same amplified products (consisting of four pooled templates; GRSPaV, GTG11-1 GVA, GARSV and GBLV) that were divided prior to purification (*Appendix A: Table 7, Figure 1*). The reference mapping data obtained from these two samples were compared (*Figure 2.2*). The greatest difference in percentage mapped reads between the two samples of a single template was 2.5% (for GRSPaV), however these differences are not statistically significant as the P-value equals 0.9986 (two-tailed t-test, P≤0.05). The templates for the samples of the final sequencing run (samples N – T, except for R) were not purified by PCR cleanup but by column purification prior to pooling. However, to compare these purification systems (*Figure 2.3*) the templates of sample R was purified by PCR clean up whilst those of sample S were column purified. Except for the purification step, the templates of sample R and S were prepared in the same way, 13 templates (GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV, GRVFV, GFkV and TBSV) were mixed at different molar ratios based on size (*Table 2.2; Appendix A: Table 4 and 5, Figure 1*). Upon the submission of the samples for sequencing, only sample R was column purified prior to library preparation (as a requirement by the sequencing facility when samples are not column purified prior to submission). Overall a higher percentage mapped reads was observed for the templates of sample R (PCR clean up) in comparison to S (column purified), but this difference was not considered significant (P=0.9965; two-tailed t-test, P≤0.05).

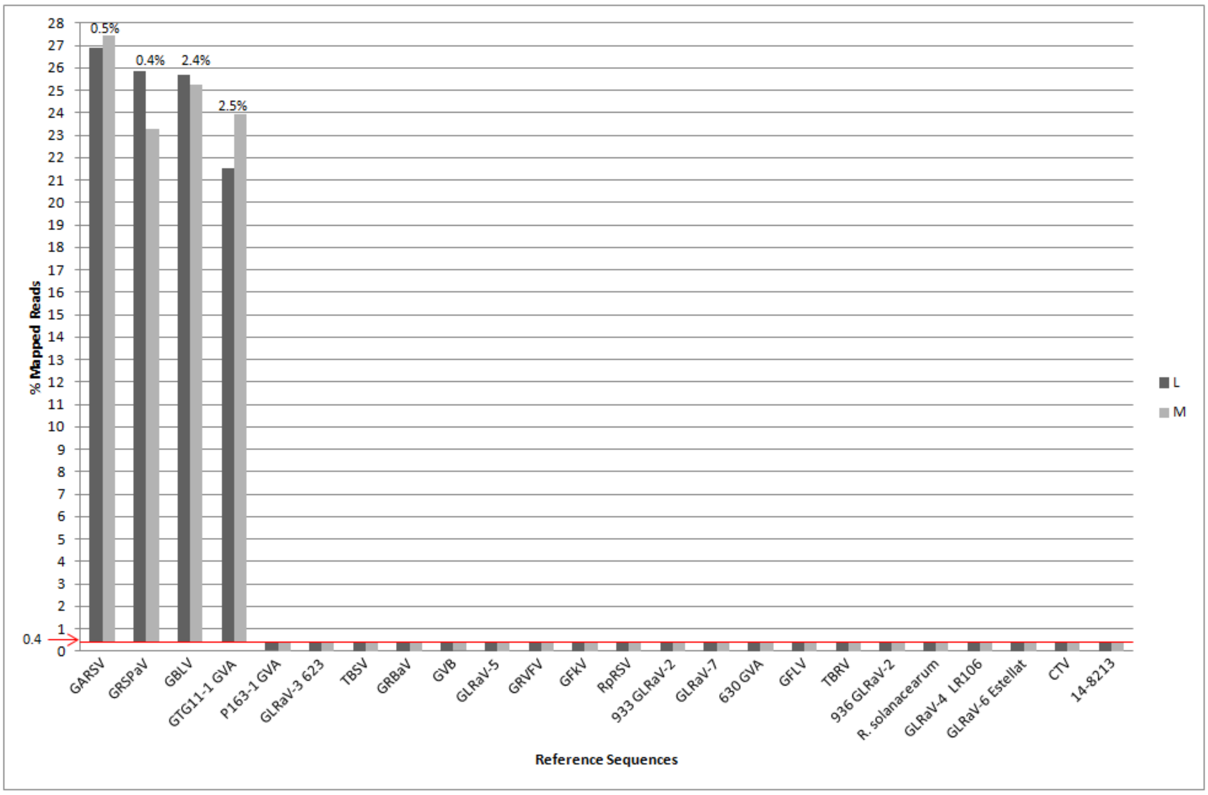

Figure 2.2: Graphic representation of the percentage mapped reads obtained for samples L and M (consisting of four templates pooled at equimolar ratios; GRSPaV, GTG11-1 GVA, GARSV and GBLV). Templates ranked by percentage mapped reads. The two samples are technical replicates, derived from single template amplification reactions, separated into two prior to PCR cleanup (Exonuclease and FastAP). To illustrate the small difference observed in mapped reads for the various templates in a sample the y-axis is set at 0.4% mapped reads (positive/negative threshold). Therefore, any template percentage mapped reads equal or above the y-axis are true positives and below is discarded as false positives. The difference in mapped reads as a percentage between the two samples for the various templates is displayed above the bars.

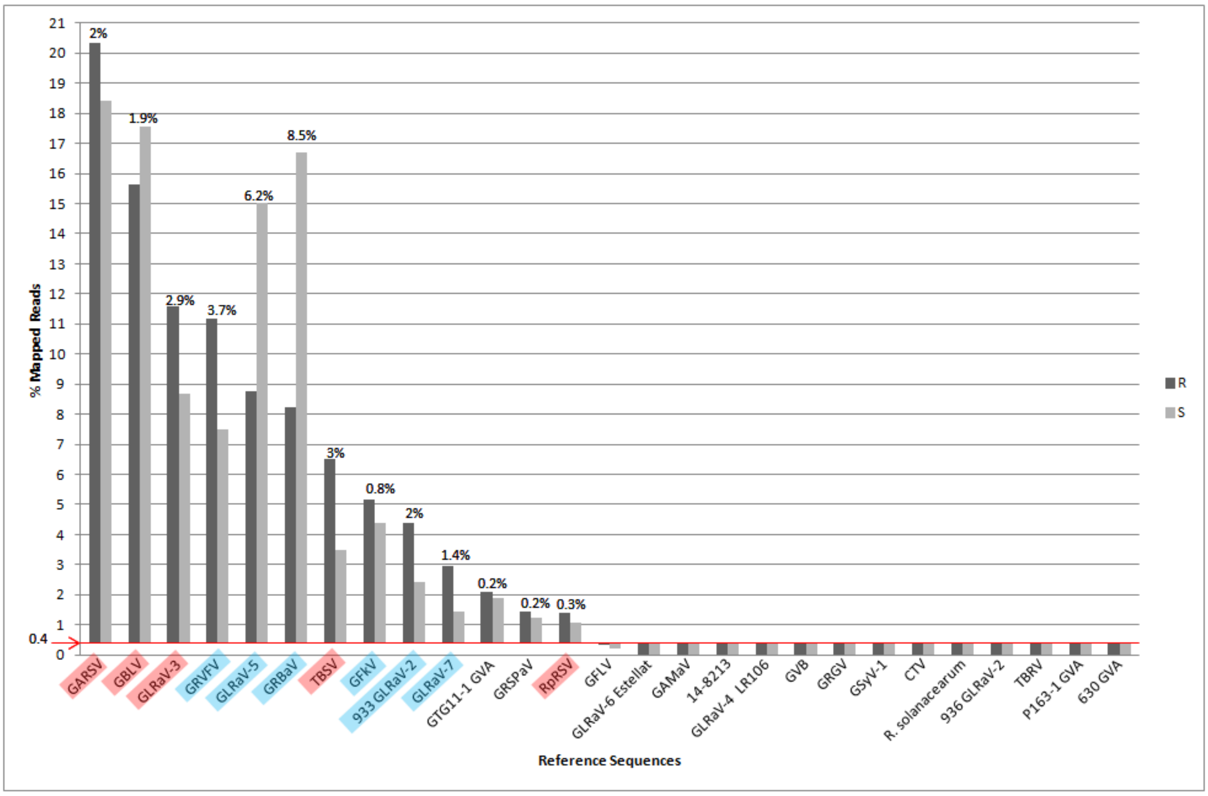

Figure 2.3: Graphic representation of the percentage mapped reads for samples R and S, consisting of 13 templates (GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV, GRVFV, GFkV and TBSV) pooled at various molar ratios based on size (white = one times, red = double, blue = three times). Templates ranked by percentage mapped reads. Templates for sample R were purified by PCR cleanup and upon submission for sequencing, the sample was column purified prior to sequencing preparation. Templates for sample S were column purified prior to pooling thus sample S was not required for purification upon submission for sequencing. The difference in mapped reads as a percentage between the two samples for the various templates is displayed above the bars.

The reference mapping data for samples P and Q (consisting of four pooled templates; GRSPaV, GTG11-1 GVA, GARSV and GBLV) were compared to investigate the variability introduced during the determination of the concentrations (Qubit) of the templates (*Figure 2.4; Appendix A: Figure 1*) prior to sample pooling. These samples were prepared as for L and M, except that the amplified templates were divided prior to concentration determination. The difference in percentage mapped reads between samples P and Q for the various templates were much larger than that observed for template purification (*Figure 2.2*). The greatest difference of percentage mapped reads for a template between samples P and Q was 8.9% (GBLV), closely followed by 8.5%, 7.6% and 6.9% for GTG11-1 GVA, GRSPaV and GARSV, respectively. Although the difference in percentage mapped reads between the two samples was much higher than for template purification, they are considered to be of no significance (P=0.993; two-tailed t-test, P≤0.05).

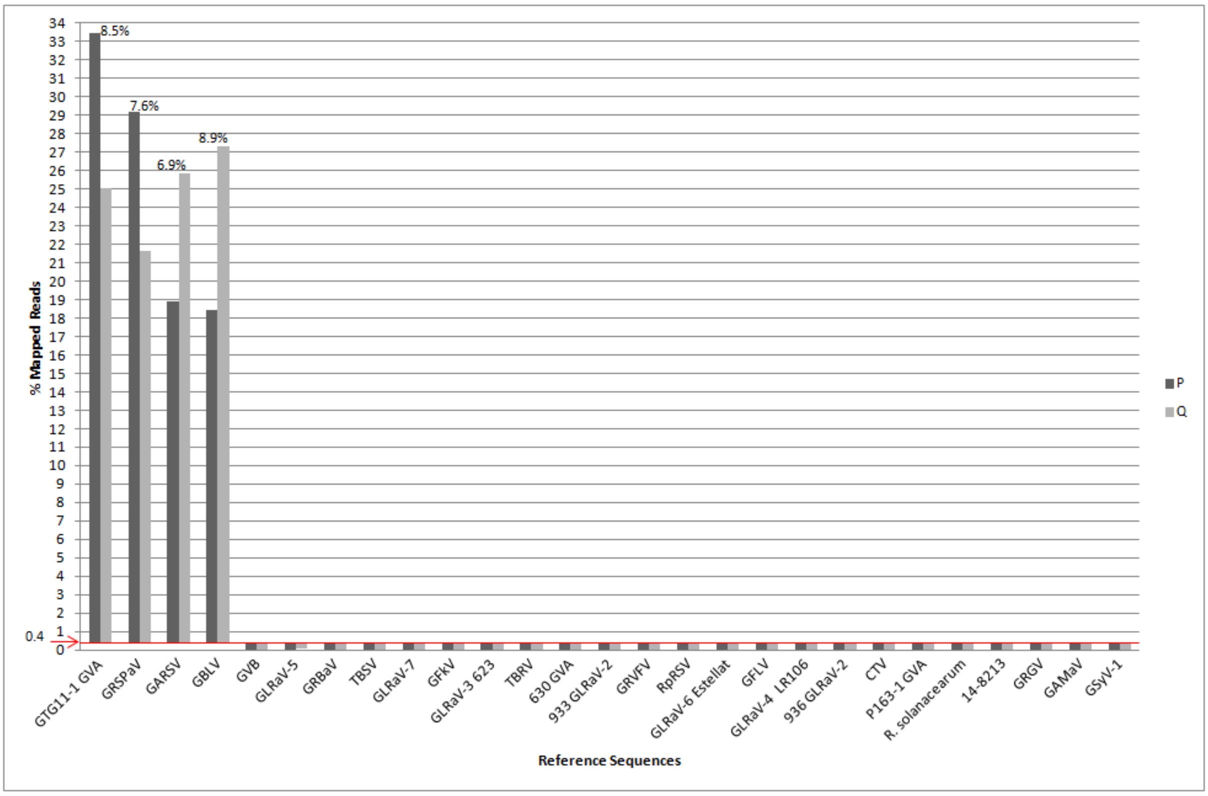

Figure 2.4: Graphic representation of percentage mapped reads for samples P and Q consisting of four templates (GRSPaV, GTG11-1 GVA, GARSV and GBLV) pooled at equimolar ratios. Templates ranked by percentage mapped reads. Templates were prepared as single reactions during amplification that were separated prior to concentration determination (Qubit). The difference in percentage mapped reads between the two samples for the various templates is displayed above the bars.

Samples (E – M) prepared from single library preparations were subject to two independent sequencing runs (*Appendix A: Figure 1*) to investigate the consistency of the sequencing process. Reference mapping of the two replicates of sequencing runs yielded very similar results suggesting that very little variation was introduced during the actual sequencing process (*Figure 2.5*). When considering the two sequencing runs of sample E (consisting of 11 pooled templates; GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV and TBSV, pooled equimolarly), the greatest difference in percentage mapped reads was less than 0.45% (GLRaV-5) between the technical replicate sample templates. This suggests that the actual sequencing process is very consistent, resulting in barely any variability due to experimental error (P=0.9874; two-tailed t-test, P≤0.05).

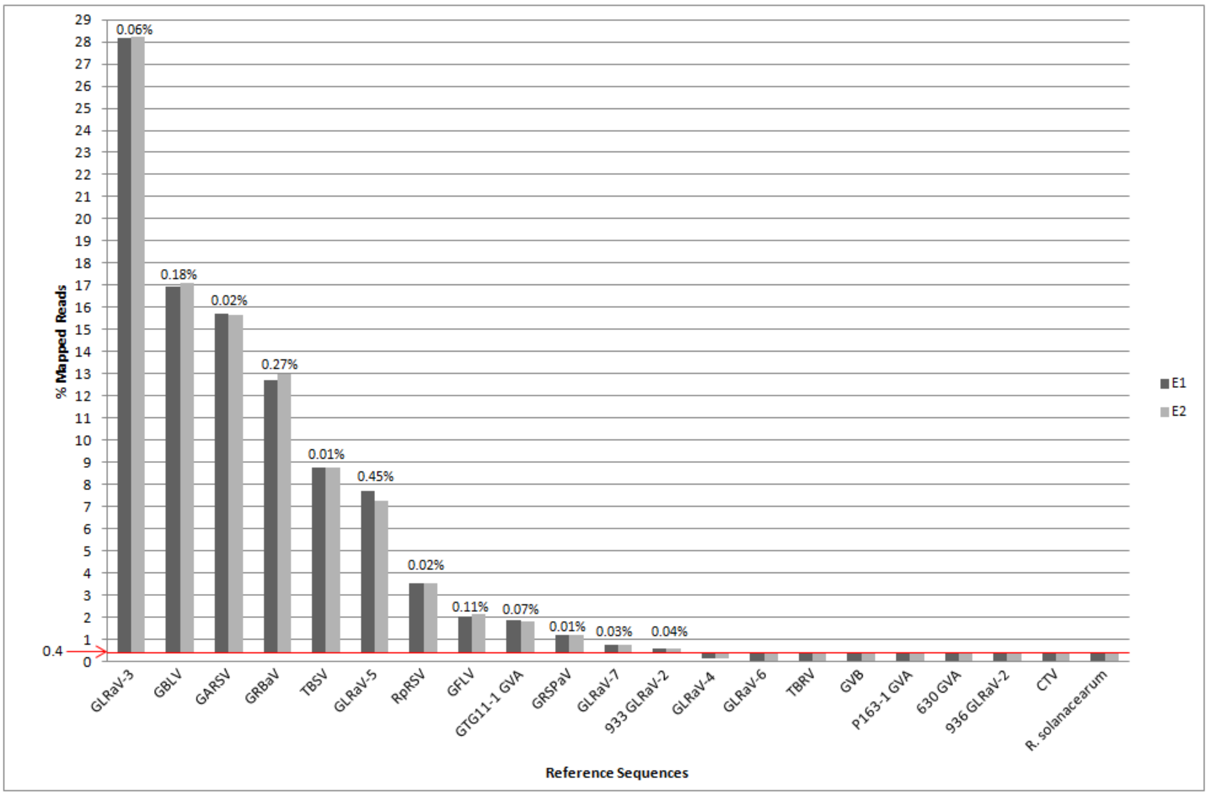

Figure 2.5: Graphic representation of the percentage mapped reads of the two sequencing run replicates of sample E consisting of 11 templates (GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV and TBSV) pooled at equimolar ratios, all derived from a single library preparation. Templates ranked by percentage mapped reads. The difference in percentage mapped reads between the two samples for the various templates is displayed above the bars.

The reproducibility of the entire HTS sample preparation was evaluated to determine if the process was robust enough for routine use. Replicate samples N and O consisting of four templates (GRSPaV, GTG11-1, GARSV and GBLV) pooled equimolarly, were subjected independently to the sample template preparation but placed in parallel on the same sequencing run for evaluation (*Figure 2.6; Appendix A: Figure 1*). Some differences in percentage mapped reads between the templates with the samples was observed, with the greatest difference being 5.5% for replicates of GVA GTG11-1 and the least being 0.8% for GARSV. The average difference in percentage mapped reads among the four templates was 2.75%. When compared to the 7.98% variance observed for replicates of concentration determination, it is clear that various components of the sample preparation protocol may have variances larger than the protocol in its entirety where the differences observed were not statistically significant (P=0.9988; two-tailed t-test, P≤0.05), and that these variances tend to cancel each other out to same extent. The overall sample preparation is quite robust and can be used routinely.

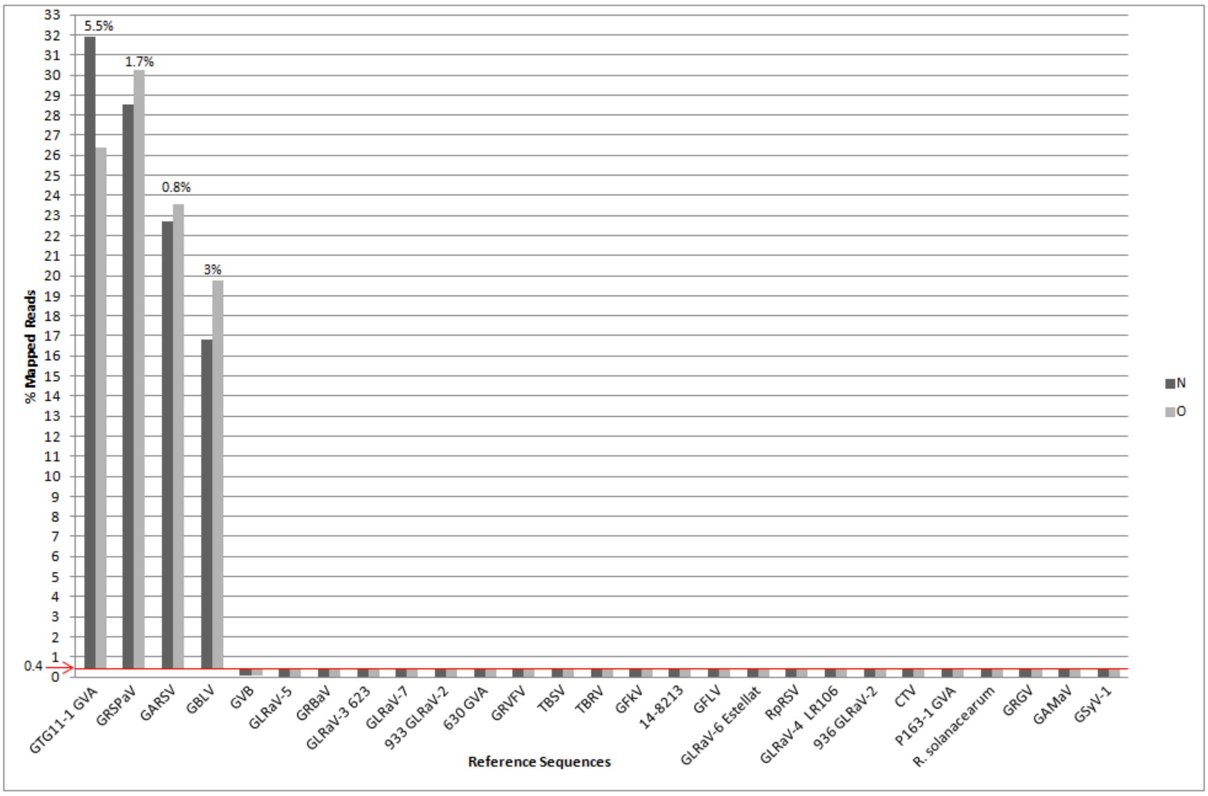

Figure 2.6: Graphic representation of percentage mapped reads of replicate samples N and O consisting of four templates (GRSPaV, GTG11-1, GARSV and GBLV) pooled equimolarly. Samples N and O were prepared separately but placed on the same sequencing run. Templates ranked by percentage mapped reads. The difference in percentage mapped reads between the two samples for the various templates is displayed above the bars.

The templates for samples A (21 templates), B (11 templates), E (11 templates), G (24 templates) and H (14 templates) were pooled at equimolar ratios (*Table 2.2*; Appendix *A: Table 2 and 3, Figure 1*) with the expectation that these ratios would be reflected in the number of reads obtained. However, upon mapping the reads to the reference sequences, the read distribution for each template were not equal (*Figure 2.7*). Within all the sample mixtures the variations in read distribution among templates differed significantly (P<0.0001; chi-square test, P≤0.05), consequently the protocol cannot be used quantitatively. The lack of relationship between template read mapping and template molarity could be due to differences introduced by varying template fragment length or GC content, and therefore the relationship of properties with the read distribution were assessed. Initial tests suggested that a bias against fragments might be due to fragment size, with fragments larger than approximately 400 bp having less reads. However, it was also observed that reads directed against the 199 bp templates obtained from the *Viti*- and *Foveavirus* PCR were also underrepresented (*Figure 2.8*). To assess whether the differences in GC content of the amplified templates were correlated with the observed unequal distribution of reads, these values were compared (*Figure 2.9*). The GC content of the various templates ranged primarily between 42 and 50%, with the exceptions of GLRaV-7 that was 35% and GFkV at 65.7%. No correlation was observed between the distribution of reads and GC content of the template sequences.

In view of the low reads obtained for templates from the *Viti, -Fovea, -Velari* and *Closterovirus* PCRs (*Figure 2.8*), the possibility that the low read mapping was due to the degree of primer degeneracy for the PCRs was also investigated (*Figure 2.10*). Many of the primer pairs used within this study are poly-specific and often contain degeneracies (*Appendix A: Table 1*) to allow the simultaneous amplification of multiple viruses. No discernible trend could be observed between the amount of primer degeneracy and template percentage mapped reads. Therefore, primer degeneracy is not considered responsible for the unequal distribution of reads.

To assess whether the bias in reads obtained from different templates could be compensated for by increasing the amount of template of larger fragments, templates were pooled according to their size at different molar ratio's (*Table 2.2, Figure 2.11 and 2.12; Appendix A: Table 3 – 6, Figure 1*). When templates larger than 400 bp (GBRaV 557 bp, GLRaV-; 485 bp, GLRaV-7 500 bp, 933 GLRaV-2 500 bp) were pooled at a molar ratio of 2:1, 3:1 and 4:1 of large template to smaller template (as determined based on the initial observation of bias against templates larger than approximately 400 bp, they were pooled at double, triple or quadruple the amount of template to templates smaller than 400 bp) an increase in percentage-mapped reads with increased amount of template could be observed. For specific templates (GLRaV-7 and GLRaV-2) when pooled at equimolar ratios with 22 other templates (*Figure 2.7*), they were considered false positives based on the determined positive/negative threshold, but when pooled at higher ratios of 2:1, 3:1 and 4:1, reads to these templates increased to levels higher than the positive/negative threshold (*Figure 2.11 and 2.12*). While it was observed that an increase in the amount of template increases the number of reads associated with that template, the percentage of reads remained unequally distributed.

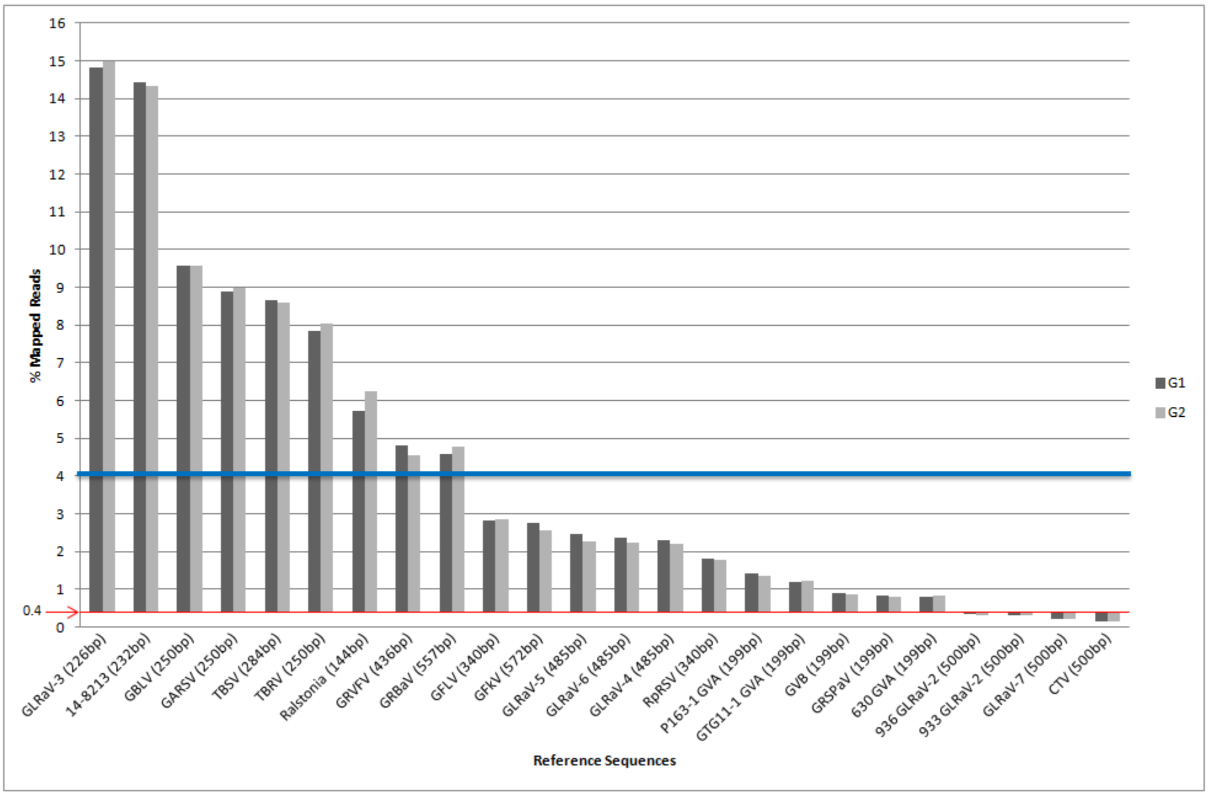

Figure 2.7: Graphic representation of percentage mapped reads of both sequencing runs from a single library preparation of sample G (24 templates pooled equimolarly). Templates ranked by mapped reads. The blue line represents the expected percentage mapped reads for each template, approximately 4.16%.

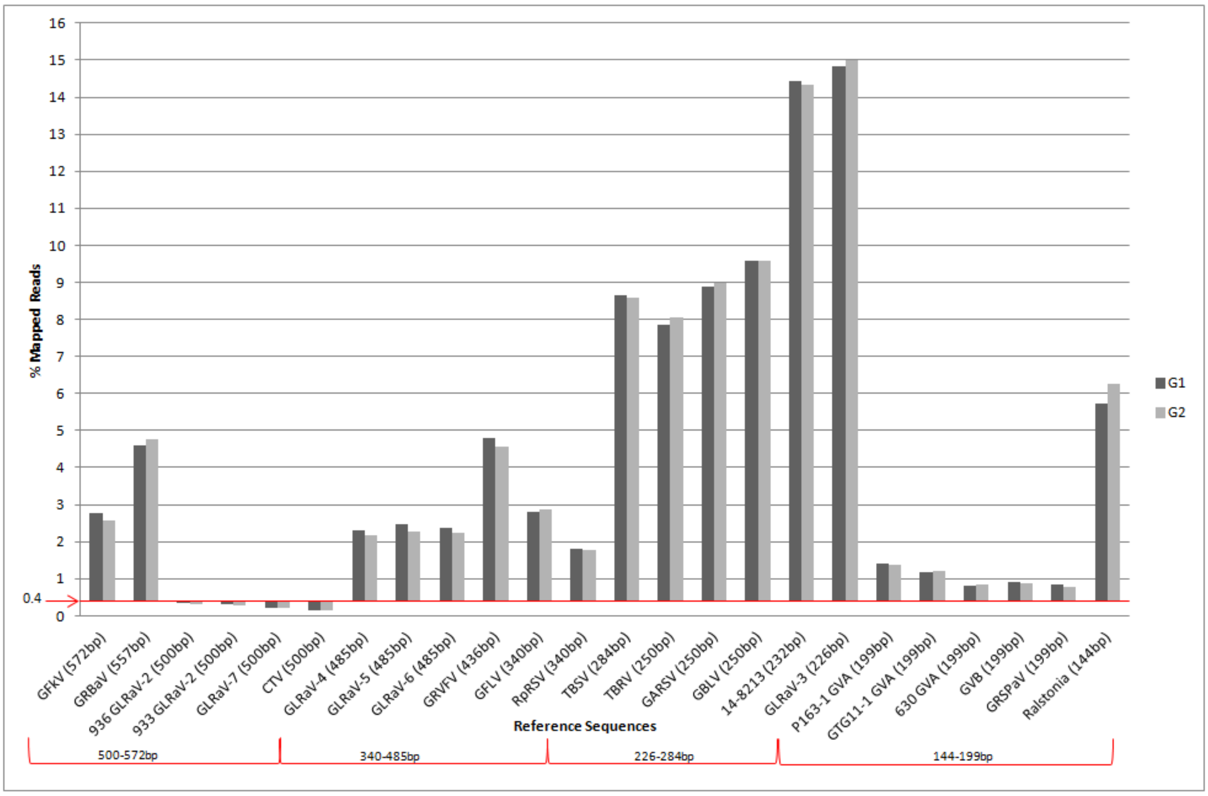

Figure 2.8: Graphic representation of percentage mapped reads of both sequencing runs from a single library preparation of sample G (24 templates pooled equimolarly). Templates ranked by amplified fragment size.

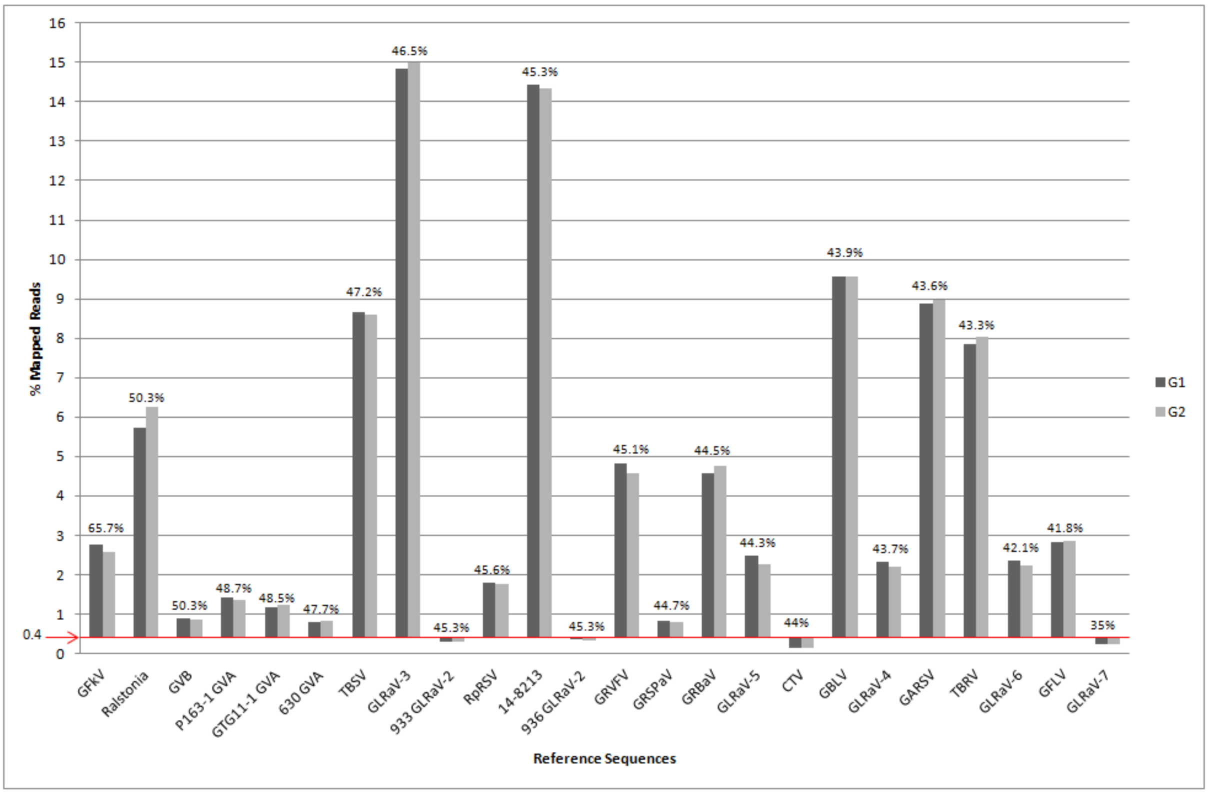

Figure 2.9: Graphic representation of percentage mapped reads of both sequencing runs from a single library preparation of sample G (24 templates pooled equimolarly). Templates ranked by GC content, as displayed above the bars.

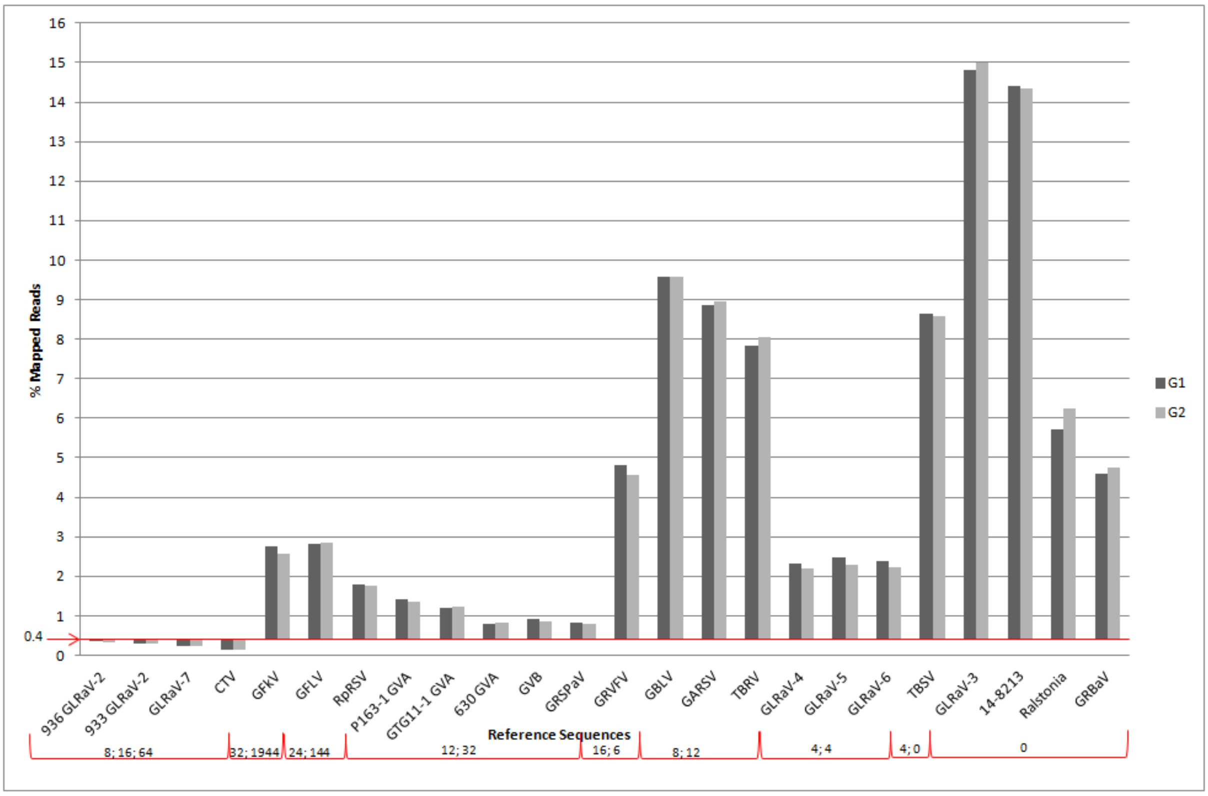

Figure 2.10: Graphic representation of percentage mapped reads of both sequencing runs from a single library preparation of sample G (24 templates pooled equimolarly). Templates ranked by primer degeneracy, in the order of forward followed by reverse primer e.g. GFkV; MTR1 forward primer degeneracy 32, MTR2 reverse primer degeneracy 1944. Degeneracy for the *Viti-, Fovea-, Clostero-* and *Velarivirus* nested PCRs (936 GLRaV-2, 933 GLRaV-2, GLRaV-7, CTV, P163-1 GVA, GTG11-1 GVA, 630 GVA, GVB and GRSPaV) are only indicated for the nested primers of the second amplification step.

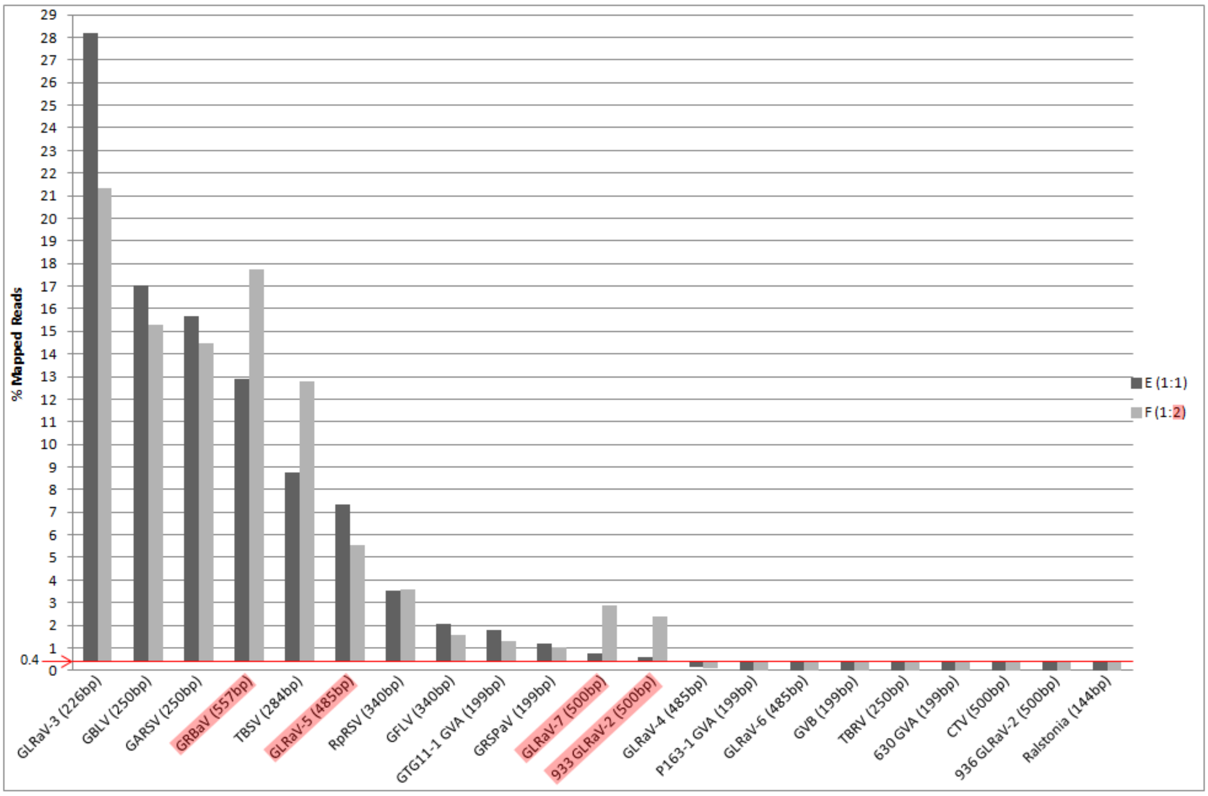

Figure 2.11: Graphic representation of percentage mapped reads for samples E and F, consisting of a mixture of 11 templates (GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV and TBSV); which in the case of sample E are pooled at equimolar ratios, while highlighted templates in sample F are present at twice the molar ratio of the other templates. Templates ranked by percentage mapped reads.

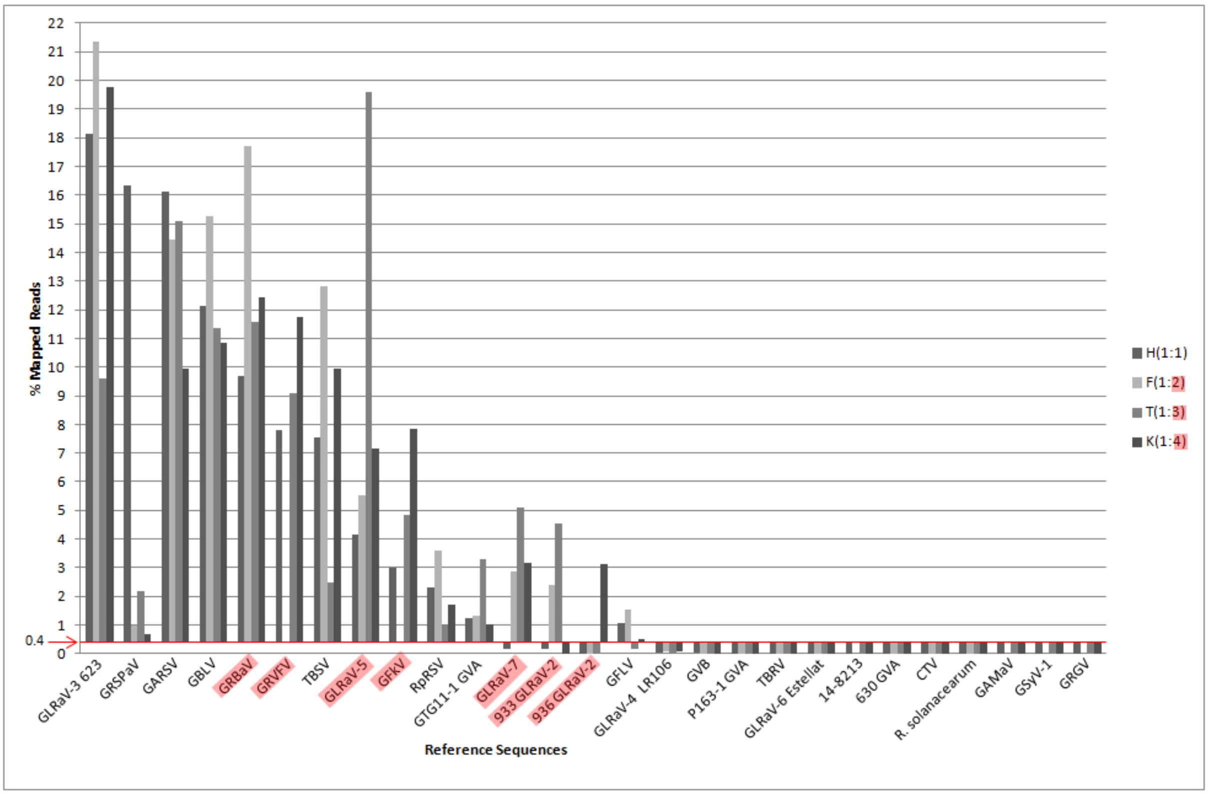

Figure 2.12: Graphic representation of percentage mapped reads for samples H (14 templates pooled equimolarly; GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV, TBSV, GFLV, GRVFV and GFkV), F (12 templates pooled at varying molar ratios of 2:1; GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV, TBSV and GFLV), T (13 templates pooled at varying molar ratios of 3:1; GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 933 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV, TBSV, GRVFV and GFkV)and K (13 templates pooled at varying molar ratios of 4:1; GRSPaV, GTG11-1 GVA, RpRSV, GARSV, GBLV, 936 GLRaV-2, GLRaV-3, GLRaV-5, GLRaV-7, GRBaV, TBSV, GRVFV and GFkV) each representing templates pooled at different molar ratios as indicated in brackets. The highlighted reference sequences were pooled at different molar ratios (1:1, 2:1, 3:1 and 4:1) for the various samples. Templates ranked by percentage mapped reads.

Optimizing contig BLAST analysis

During the reference mapping performed in this study, it was expected that all reads should map to the reference sequences as known, defined templates had been used. However, despite this, not all the reads mapped to the reference sequences under the parameters utilized. To understand this, the unmapped reads were collected and subjected to *de novo* assembly. Generated contigs were subjected to multiBLASTn against the NCBIs viral database using the CLC Genomics workbench interface, allowing all contigs to BLAST simultaneously. Due to the large number of possible BLAST hits available for each contig, various criteria (*Figure 2.13*) were assessed to reduce the data to relevant reads, potentially requiring confirmation by additional virus tests.

The first criteria utilised was the removal of all non-plant viral hits, as our interest is in plant viruses only. Within the remaining contigs, those with an E-value of more than 1E-10 were discarded. This was required given that the contigs were subjected to BLAST against a subset database of only the viruses in the NCBI database, and because of this, some contigs matched viruses based on very limited identity resulting in contigs with very high E-values. In these instances, it was observed that when these contigs were subject to BLAST against the entire nucleotide collection hosted by NCBI, many of the contigs would match sequences, primarily that of cloning vectors, at E-values much closer to 0.

The third criteria applied was the selection of only contigs with BLAST hits with a percentage query overlap (percentage of entire contig that aligns to the BLAST hit; i.e. hit length/contig length x 100) of more than 80%. However, to prevent discarding contigs which have percentage query overlaps of less than 80% due to them being assembly artifacts (chimeras of reads from various closely related viruses in the sample), an additional criteria was included for those contigs that did not adhere to the required 80% query overlap. In these instances they were evaluated for their percentage amplicon overlap (percentage of the entire amplicon size that is utilised for that BLAST hit; hit length/amplicon length for that BLAST hit x 100). Contigs with percentage amplicon overlaps of equal or more than 80% were retained. The inclusion of these ensured that no potentially relevant reads were discarded based on contig size alone. To illustrate the need of percentage amplicon overlap as a criterion contig 7 of sample R (*Table 2.3*) is discussed; its BLAST hit with the lowest E-value was against GBLV, a grapevine virus, with a hit length of 242 bp. As the length of contig 7 was 471 bp, the percentage query overlap was 51.4% (242 bp / 471 bp x 100). Based on this the contig would not be retained for further evaluation since the percentage query overlap is less than 80%. However, the amplicon length for GBLV (based on the PCR system used in this study, Nepovirus C, *Appendix A: Table 1*) is 250 bp, and actually represents the expected size contigs. Thus in this instance the percentage amplicon overlap was 96.8% (242 bp / 250 bp x 100). Therefore, the exclusion of this contig for further analysis based on the size of the contig alone would have resulted in the loss of potential valid reads.

Contigs that adhered to these four criteria were isolated and individually subjected to BLAST against the entire nucleotide collection hosted by NCBI. In initial assessments contigs had been subjected simultaneously to BLAST only against the virus sequences hosted by NCBI. While the inclusion of an individual BLAST against all the sequences hosted by NCBI, is primarily to confirm the original result, additional information is also obtained regarding the location of the contig alignment within the BLAST hit sequences, what other organisms the contig matched and whether the contig is a chimera of different template reads. To illustrate this, sample R (*Table 2.3*) is discussed. After the implementation of the four criteria, only three of the 30 contigs were retained (contig 5, 6 and 7). Each of these contigs were individually subjected to BLAST against NCBIs complete database. Both contig 5 and 6 when individually subjected to BLAST, matched the same virus (GRVFV) as observed when subject to multiBLASTn against only viruses. However, when subjected to an individual BLAST against NCBIs complete database, contig 7 was found to be a sequence chimera of different read templates. It aligned to three viruses, GLRaV-5 (1 – 118 bp), GBLV (167 – 405 bp) and GLRaV-3 (405 – 471 bp) in separate regions of the contig (*Figure 2.14*). As the viral composition of sample R was known (*Table 2.2; Appendix A: Table 4*), the viruses identified by *de novo* assembly and BLAST were all true positives, including those comprising the chimeric contig 7. However, when dealing with an unknown sample, if viruses identified by *de novo* assembly and BLAST were not identified during reference mapping, they should be analysed by phylogeny and confirmed by additional tests such as virus specific PCRs.

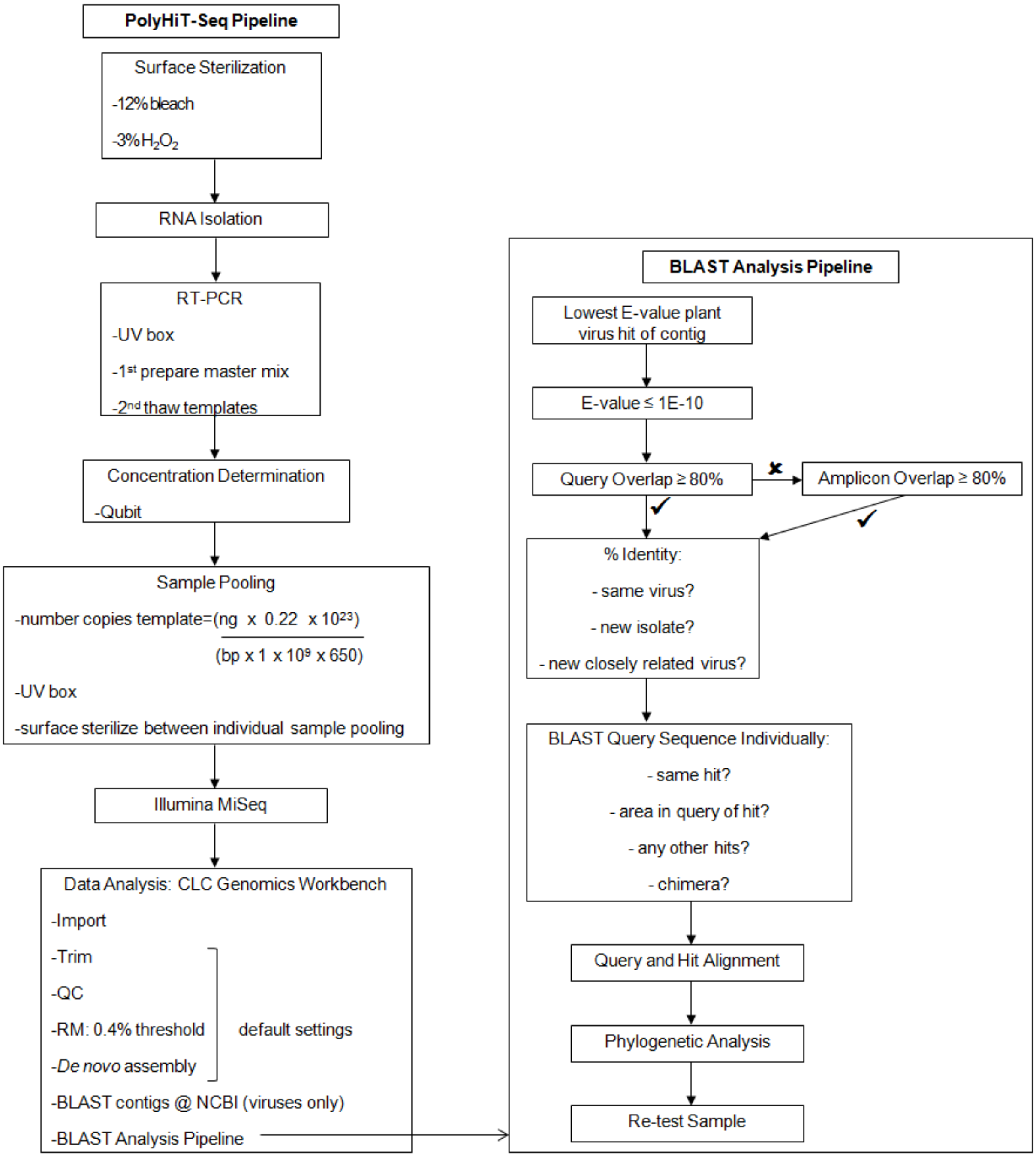

Figure 2.13: Schematic representation of sample preparation and data analysis for the PolyHiT-Seq system. During data analysis reads are evaluated for their quality (QC) and are trimmed for the removal of ambiguous nucleotides as well as adapter sequences. Trimmed reads are reference mapped (RM) with a length fraction (LF) of 0.5 and a similarity fraction (SF) of 0.8. Unmapped reads are *de novo* assembled and contigs subject to BLAST against NCBIs virus collection making use of the CLC Genomic workbenchs' default settings (materials and methods).

Table 2.3: Sample R BLAST results that adhered to the BLAST analysis criteria.

Query	Number of hits	Query length (bp)	Amplicon length (bp)	Lowest E-value	Hit lenght	% Query overlap	% Amplicon overlap	% Identity	Accession	Description
Contig 5	74	402	436	9.78E-102	322	80.1	73.9	85.71	AY706994	GRVFV
Contig 6	88	570	436	3.64E-145	432	75.8	99.1	86.57	AY706994	GRVFV
Contig 7	106	471	250	2.995E-89	242	51.4	96.8	90.5	FN691934	GBLV

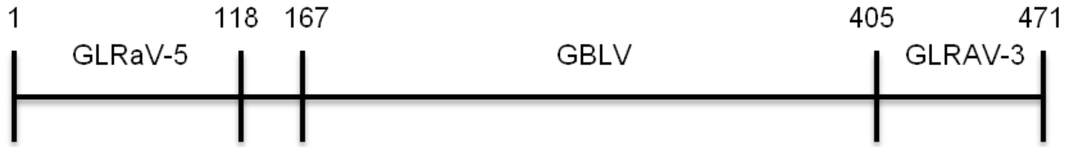

Figure 2.14: Schematic representation of sample R's contig 7 BLAST result when individually BLAST against NCBIs database. Contig 7 BLAST against GLRaV-5 (1-118 bp), GBLV (167-405 bp) and GLRaV-3 (405-471 bp).

2.4. DISCUSSION AND CONCLUSION

Here we describe the establishment of a novel approach to virus detection and identification that is based on poly-specific and virus-specific RT-PCRs in combination with HTS, so named the PolyHiT-Seq system. RT-PCR systems for the detection of 37 grapevine infecting viruses of 11 genera were established. These amplification systems together with HTS, in this case Illumina MiSeq sequencing, allows identification to the species level of the poly-specific PCR products. The establishment of an optimized HTS data analysis pipeline using CLC Genomics workbench, determined the successful analysis of HTS data.

Methods for the simultaneous detection of pathogens have been developed as a way to decrease the cost per reaction. Broad spectrum PCR is an example of such a technique; by focusing rather on higher-order taxonomic entities such as genera rather than species, short conserved regions can be amplified (James *et al.*, 2006), as described here. Although such a technique can lower reaction cost, it comes at a cost of itself, which usually is the need for degenerate bases within the primer pair, especially in the instance of viruses, as they do not have universal conserved regions similar to the 16S rRNA gene of bacteria. One of the disadvantages of having degenerate bases in a primer set is that there are lower concentrations of each primer variant within the reaction as degeneracy increases. Hence template priming is influenced by template primer specificity, which can result in specific primers within the reaction being depleted relatively fast, thus reducing amplification efficiency and sensitivity (James *et al.*, 2006, Ibarbalz *et al.*, 2014). Significant bias has been observed with degenerate bases within primer sets for the amplification of CTV genotypes, as evidenced by the sequencing of clones from an artificial template (Read, 2015). This bias appeared somewhat unpredictable with regards to degeneracy and the position of the degenerate nucleotide on the primer which suggests that primer degeneracy alone does not influence population representation, and that potentially other factors such as the amplicon size for sequencing is also involved (Read, 2015). A number of primer sets within the current study were selected for their ability to bind to a conserved region that amplifies a variable region in order to identify the virus species. This has resulted in many of these primers having relatively high degeneracy levels. As such, the degree of primer degeneracy in relation to template read distribution was investigated. We did not observe a trend between the level of primer degeneracy and template percentage mapped reads, which was expected, especially since templates were pooled based on molarity after amplification thus eliminating the influence of primer degeneracy on reads obtained.

A drawback of HTS sensitivity is the persistent problem of cross contamination or background noise. The cross contamination of reads amongst samples subjected to HTS were commonly observed during a pilot study (data not shown) and confirmed by previous reports in this regard (Capobianchi *et al.*, 2013, Massart *et al.*, 2014, Roossinck *et al.*, 2015). Cross contamination may be introduced in the laboratory by template contamination of one sample with another as a result of many steps in the sample preparation pipeline; sample amplification, purification, pooling, as well as during the synthesis and handling of sequencing barcodes, during library preparation at the sequencing facility, sequencing and at demultiplexing (Kircher *et al.*, 2012, Capobianchi *et al.*, 2013, Quail *et al.*, 2014). To address this, measures to prevent cross contamination in the laboratory were employed along with an optimized sequence data analysis protocol to account for any remaining contamination (*Figure 2.13*).

To limit cross contamination among samples during sample preparation the following measures were taken; 1) the sterilization of work spaces, in our case with 12% bleach and 3% hydrogen peroxide for the elimination of any contaminating amplicons and clones, RNases and DNases, and 2) preparation of sample amplification in either enclosed spaces that were sterilized by ultraviolet light, or spatially separated preparation of the master mix from the addition of templates and positive controls. Cross contamination among these is limited by the lack of, or reduced, exposure to one another. The risk of cross contaminating samples with each other during the pooling of their amplified templates is very high, but can also be limited by spatial separation (Aslanzadeh, 2004). Therefore, to minimize the risk in this step, templates for multiple samples should not be handled in parallel; collected together for pooling, but rather the templates for a single sample should be obtained, pooled and put away, followed by surface and apparatus sterilization prior to the pooling of the next sample.

In the sequencing reaction itself, phasing and pre-phasing has been reported to cause noise in the cluster signal due to difference in molecule length within the cluster. During phasing, problems with enzyme kinetics can cause molecules in the cluster to lag behind due to problems such as the incomplete removal of 3' terminators. In contrast, pre-phasing is where the synthesis of some molecules in the cluster is too fast, this can happen as a result of inadequate flushing of the flow cell, resulting in the incorporation of nucleotides lacking the 3' terminator (Schirmer *et al.*, 2015). As the sequencing process is provided as a service we cannot intervene during the actual sequencing process, therefore we have established a positive/negative threshold for reference mapping to account for any sample cross contamination or noise. The threshold to differentiate contamination from true template presence was determined by mapping samples with known sequence compositions to not only their homologous sequences, but to all the possible template reference sequences of prepared clones in this study. Samples containing only a single representative virus sequence of a specific genus gave the most useful results (samples pooled with either 11 or 13 templates); since mapping to viruses not present in those samples allowed the distinction between positive and negative results, which in turn assisted in the establishment of the positive/negative threshold.

These experiments were conducted based on an optimized reference mapping system of the Illumina reads in which various combinations of LF and SF were assessed with the optimal parameters being the default settings with a LF of 0.5 and a SF of 0.8, as these yielded the highest percentage of mapped reads. It was observed that with an increase in the stringency of these setting there was a decrease in efficiency in the mapping of reads.

Due to the short length of HTS reads, it is difficult to uniquely align reads to large reference sequences or complex genomes because of a higher probability of repetitive sequences (Voelkerding *et al.*, 2009). This was not expected to be a problem in this study, as defined amplified templates are used, thus short reads of known amplified sizes were aligned to only the cognate areas of the reference sequences. As known nucleic acid templates produced by specific PCR systems were sequenced, their identification to the species level was known, thus mapping the reads only to the amplicon cognate areas of the reference sequences was expected to result in the mapping of all the reads. For this reason, the LF and SF parameters that result in the greatest percentage of mapped reads were selected.

Since known, defined templates were sequenced, it was expected that all reads would map to the references sequences. However, this was not the case, and the unmapped reads were evaluated by *de novo* assembly and BLAST analysis to determine their origin. *De novo* assembly together with BLAST against NCBIs database are commonly used analysis techniques for metagenomic studies, especially in the absence of a reference genome (Casals *et al.*, 2012, Massart *et al.*, 2014, da Fonseca *et al.*, 2016, Rose *et al.*, 2016). During *de novo* assembly, contigs are generated through the connection of the HTS reads as a result of the mathematically based algorithms used; for CLC Genomics workbench, de Bruijn graphs (Compeau *et al.*, 2011). de Bruijn graph assembly is based on the formation of words of a specific length (*k*) from a given alphabet (sequencing reads) that occur only once consecutively as the shortest possible sequences (*k*-mers) within in a cyclic sequence. The de Bruijn graph algorithm assumes that all the *k*-mers are error free and that they appear only once in the genome (sequencing reads), however this is not true for HTS data (Compeau *et al.*, 2011). Mangul *et al*., (2014) reported that it was difficult for them to differentiate between rare viral variants and sequencing errors when using ultra-deep sequencing. Besides sequencing errors, mixed populations, sequence repeats and GC content biased regions are features of concern for accurate *de novo* assembly, such as the formation of chimeric sequences (Khalifa *et al.*, 2016). Chimeric sequences can also form during sample amplification in the laboratory and bridge amplification at the sequencing facility of the individual viruses and samples respectively (Zagordi *et al.*, 2012, Waugh *et al.*, 2015). During *de novo* assembly of sample R's reads that did not map during reference mapping we observed the formation of a chimeric sequence, contig 7. When subjected to BLAST against the nucleotide database at NCBI, it was observed that this contig consist of three virus sequences, GLRaV-3, GLRaV-5 and GBLV. As sample R was artificially assembled, the template composition is known, and all the virus templates comprising contig 7 were present within the sample. As these viruses were not amplified in a single reaction during sample preparation within the laboratory, it is expected that the chimera was formed during either library preparation, or *de novo* assembly.

There have been reservations in the acceptance of novel virus sequences obtained from HTS data, these reservations stem from the concern in the accuracy of *de novo* assemblies. For the formal recognition of new *de novo* based viral species, universally accepted criteria is required to ensure consistency (Khalifa *et al.*, 2016). As with metagenomic studies, amplicon-based HTS *de novo* assemblies also require evaluating criteria to ensure accurate and consistent analysis. This is the first study as far as we know where multiple virus specific amplicons have been subject to HTS for species identification. As with any novel approach, a guideline of what to do and how to do it are lacking, and these will likely be developed iteratively. We found that for the purpose of our study, when the result is not also supported by the reference mapping data, *de novo* assembly together with BLASTn, had to be evaluated with regards to E-value, percentage query overlap and percentage amplicon overlap. These parameters all present insight into the significance of the BLAST match, yet the data is mostly based on only a small part of the genome (< 600 bp). In the instance of the discovery of a novel virus or a novel virus to grapevine, the small amount of data does not lend confidence to the discovery. Therefore, confirmation of BLAST results are prudent, using both phylogenetic analyses and by additional laboratory tests.

Reproducibility of HTS results has been reported to be difficult, specifically based on data analysis (Nekrutenko and Taylor, 2012), but the use of HTS as a diagnostic system requires the reproducibility of the entire system from sample preparation to the actual sequencing process and the analysis of the data. As there are many steps involved throughout this system, it is necessary to evaluate the variability introduced at each step as well as the robustness of the entire process as a measure of reproducibility. Through the evaluation of these stages, interventions can be introduced to manage the variability. For the PolyHiT-Seq system we evaluated stages during sample preparation (template purification and concentration determination), the actual sequencing process, as well as the entire pipeline for variability and overall robustness as measures of reproducibility. Unfortunately, due to the costs involved, replicates conducted for all steps were limited.

The stage in sample preparation where the least variation was observed, was that of template purification, no significant variability was observed for our laboratory's standard template purification [Exonuclease and Alkaline Phosphatase (FastAP) treatment]. However, as column purification is part of the sample preparation protocol of the sequencing facility prior to sequencing, samples submitted for sequencing that were purified by Exonuclease-FastAP PCR cleanup (samples A – M and R) would also be column purified by the sequencing facility. As this was typically done after templates of the various viruses were pooled, the possibility exists that some size selection could occur during column purification (Nucleospin PCR clean-up and gel extraction user manual, Clonetech Laboratories) resulting in variation of the read numbers obtained. Therefore, the Exonuclease-FastAP treatment of the templates prior to pooling was replaced by column purification, and the two amplicon purification systems compared. No significant size selection during column purification of the already pooled samples was observed. This is probably due to the fact that we are utilizing amplicons of 144 to 572 bp and that size selection by column purification is mainly observed for fragments > 1000 bp which bind more stringently to the silica-membrane, and hence are more difficult to elute (Nucleospin PCR clean-up and gel extraction user manual, Clonetech Laboratories, Inc.,).

The stage in sample preparation where the largest amount of variability amongst replicates was observed, was during the determination of amplicon concentration. It is known that a variety of factors influence the Qubit® 2.0 fluorometers' readings of concentration, including the temperature of the solution, moisture on the outside of the tube, bubbles in the solution and the light sensitivity of the reagents (Qubit® 2.0 fluorometer user manual, Life technologies). In spite of determination of concentration being the step at which the most variability was observed, the differencse in percentage mapped reads among the replicates of the templates of the two samples were still not significant (P=0.993; two tailed t-test, P≤0.05).

The reproducibility of the actual sequencing process was evaluated and the consistency of the process was high (samples E – M; P=0.9874; two tailed t-test, P≤0.05). We did not investigate the variability which may be introduced during library preparation which is conducted at the sequencing facility and represents the costliest step in the process. However, Frey *et al.,* (2014) in their study on the comparison of metagenomic data of three HTS platforms (Roche 454, Ion Torrent PGM and Illumina MiSeq) from blood evaluated this step. Biological replicates derived from independent library preparations, and sequenced in a single run on the MiSeq platform yielded considerable variation in reads mapping to Influenza A H1N1. To determine if this variability was a result of the library preparation stage or the actual sequencing process, they performed replicated sequencing runs from a single library preparation, and observed reasonably consistent read mappings from these technical replicates (Frey *et al.*, 2014), similar to our results. Therefore, the actual sequencing process seems to be reproducible, specifically when sequencing from a given library preparation, suggesting that library preparation is the step in sample preparation besides concentration determination that may introduce high levels of variability. The selection of the Illumina MiSeq platform for this system does not only have the advantage of lower cost per sample, but its demonstrated reproducibility is a critical trait in diagnostic systems.

When evaluating the overall robustness of sample preparation, this was found to display non-significant variability (sample N and O; P=0.998; two tailed t-test, P≤0.05). Results presented here suggests that the sample preparation pipeline is quite robust and reproducible with the variability observed among sample replicates not being significant (P=0.9; two tailed t-test, P≤0.05). The reproducibility of the pipeline lends confidence to the established reference mapping threshold.

Although the PolyHiT-Seq system is a diagnostic system and its purpose is qualitative rather than quantitative, it was expected that when templates were pooled equimolarly it would be so represented in reads obtained. In previous studies it has been observed using Illumina MiSeq sequencing, that HTS is a better quantitative measure of sample population than conventional clonal sequencing (Read and Pietersen, 2016). However, amplicon sequencing-based detection has been reported to not be quantitative in metagenomic studies of soil (Zhou *et al.*, 2011), as is supported by the results we obtained (P<0.0001; chi-square, P≤0.05) based on equimolar template poolings after template amplification, proving that the PolyHiT-Seq system is only qualitative. It has also been found that the presence of unique sequences at equimolar ratios in the library production stage tend to result in various read depth coverage, and that this variation is only in part explained by molar pooling (Harismendy *et al.*, 2009). Other factors that may also attribute to bias in read coverage are features like fragment length and GC content (van Dijk *et al.*, 2014). In this study both fragment length and GC content was investigated as factors responsible for unequal read distribution in equimolarly pooled samples. GC bias can both be introduced during template amplification from RNA and DNA as well as during library preparation prior to sequencing as GC rich sequences are amplified with less efficiency than GC neutral sequences, which over a multitude of amplification cycles can result in bias (Quail *et al.*, 2012). We observed no clear trend between the distribution of reads and the GC content of the template sequences. Moreover, Aird *et al.,* (2011) reported that templates with a GC content between 13 and 56% have relatively similar read abundances when using the Illumina MiSeq platform. Even though the GC content of the templates in this study fall within this range, with the exception of GFkV (65.7%), the reads were not equally distributed among the templates. Hence it appears unlikely that GC content played a role in the uneven read distribution within this study. We observed that with an increase in the amount of template there was an increase in reads associated with that template, but this was not linear and that it could not compensate for the unequal distribution of reads amongst the different templates. However, it was also observed that the read bias was resolved when larger volumes of lower numbers of templates were pooled together.

Individual samples from the field are unlikely to contain as many viruses as the test is capable of detecting. Several of the viruses which can be detected by these systems have not been reported in South African grapevines but by including these, this will allow the detection of these viruses should they be introduced, or have gone undetected in previous tests. Additionally, their inclusion validates this method for use in other countries where the targeted viruses are present or are a biosecurity risk. Generally, relatively low numbers of viruses are expected from South African wine grapes as the vast majority of these are derived from nuclear plants that have undergone virus elimination procedures as part of the local certification scheme. In spite of this, mixed infections are prominent in grapevine viral communities (Jooste *et al.*, 2015) due to field spread of specific viruses to certified planting material. Field samples are likely to consist primarily of multiple strains of a single virus rather than a large variety of viruses, as quasi-species are prominent in RNA virus populations

(Zagordi *et al.*, 2012). During a survey of virus populations within South African vineyards it was observed that 79.68% of the evaluated plants consisted of mixed infections primarily comprising of GLRaV-3, *Vitiviruses* and GRSPaV (Jooste *et al.*, 2015). Based on these results virus populations within field samples tested with the PolyHiT-Seq system are expected to consist of GLRaV-3 (*Ampelovirus*), GLRaV-2 (*Closterovirus*) GVA (*Vitivirus*), GVE (*Vitivirus*), GVF (*Vitivirus*) and GRSPaV (*Foveavirus*). Therefore, equimolar pooling of templates is proposed, since read bias was resolved when samples consisted of only a few templates. However, a pooling ratio for when a sample does test positive for more than five viruses should be available. It is recommended that three times as much template from PCRs to *Clostero-, Fovea-, Velari-* and *Vitiviruses* be utilized as for PCRs to the other genera, as viruses belonging to the *Closterovirus, Foveavirus, Velarivirus* and *Vitivirus* genera had the lowest mapped reads percentages when reference mapped.

When applying the pipeline for data analysis established here on field samples the threshold values arrived at may underestimate the background sequence contamination, as these values were determined from plasmid DNA samples that did not undergo reverse transcription (Thys *et al.*, 2015) and are likely to have little sequence variation. This will require further evaluation when analyzing the MiSeq data from the field samples (Chapter 3). However, with the expectation of a small virus population in individual vine samples that will allow equimolar template pooling at large volumes; the established thresholds are considered to be accurate.

In this study we present a diagnostic system based on poly-specific and virus-specific PCRs together with HTS for the detection and identification of 37 grapevine infecting viruses. The PolyHiT-Seq system has been optimized from the start of sample preparation throughout to also include an optimized HTS data analysis pipeline for both reference mapping and *de novo* and BLAST data. This system has proved to be reliable and reproducible, properties which are sought after in diagnostics. The pipeline can also be expanded for the diagnosis of viruses of other crops, due to its robustness, versatility, relative rapidity and ease to perform.

2.5. REFERENCES

Abou Ghanem-Sabanadzovic, N., Sabanadzovic, S., Gugerli, P., Rowhani, A. (2012) Genome organization, serology and phylogeny of Grapevine leafroll-associated viruses 4 and 6: taxonomic implications. *Virus Research* 163**:**120-128

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12**:**R18

Al Rwahnih, M., Daubert, S., Golino, D., Rowhani, A. (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* 387**:**395-401

Al Rwahnih, M., Dave, A., Anderson, M., Uyemoto, J., Sudarshana, M. (2012) Association of a Circular DnA Virus in Grapevines Affected by Red blotch Disease in California. *Proceedings 17th ICVG, Davis, CA***:**104-105

Al Rwahnih, M., Golino, D., Daubert, S., Rowhani, A. (2015). Characterization of a novel reovirus species in Cabernet Grapevine in California. *Presented at the Proceedings of the 18th Congress of ICVG*, Ankara, Turkey, 7-11 September, p. 194-195

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215**:**403-410

Aslanzadeh, J. (2004) Preventing PCR Amplification Carryover Contamination in a Clinical Laboratory. *Annals of Clinical and Laboratory Science* 34**:**389-396

Barba, M., Czosnek, H., Hadidi, A. (2014) Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* 6**:**106-136

Bertazzon, N., Angelini, E. (2004) Advances in the detection of Grapevine leafroll-associated virus 2 variants. *Journal of Plant Pathology***:**283-290

Bester, R., Jooste, A.E., Maree, H.J., Burger, J.T. (2012) Real-time RT-PCR high-resolution melting curve analysis and multiplex RT-PCR to detect and differentiate grapevine leafroll-associated virus 3 variant groups I, II, III and VI. *Virology Journal* 9**:**219

Capobianchi, M., Giombini, E., Rozera, G. (2013) Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection* 19**:**15-22

Casals, F., Idaghdour, Y., Hussin, J., Awadalla, P. (2012) Next-generation sequencing approaches for genetic mapping of complex diseases. *Journal of Neuroimmunology* 248**:**10-22

Compeau, P.E., Pevzner, P.A., Tesler, G. (2011) How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* 29**:**987-991

da Fonseca, R.R., Albrechtsen, A., Themudo, G.E., Ramos-Madrigal, J., Sibbesen, J.A., Maretty, L., Zepeda-Mendoza, M.L., Campos, P.F., Heller, R., Pereira, R.J. (2016) Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*

Dovas, C., Katis, N. (2003a) A spot multiplex nested RT-PCR for the simultaneous and generic detection of viruses involved in the aetiology of grapevine leafroll and rugose wood of grapevine. *Journal of Virological Methods* 109**:**217-226

Dovas, C., Katis, N. (2003b) A spot nested RT-PCR method for the simultaneous detection of members of the *Vitivirus* and *Foveavirus* genera in grapevine. *Journal of Virological Methods* 107**:**99-106

Frey, K.G., Herrera-Galeano, J.E., Redden, C.L., Luu, T.V., Servetas, S.L., Mateczun, A.J., Mokashi, V.P., Bishop-Lilly, K.A. (2014) Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics* 15**:**1

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., Terlizzi, F., Credi, R., Saldarelli, P. (2012) A new grapevine virus discovered by deep sequencing of virus-and viroid-derived small RNAs in Cv *Pinot gris*. *Virus Research* 163**:**262-268

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* 10**:**R32

Ibarbalz, F.M., Pérez, M.V., Figuerola, E.L., Erijman, L. (2014) The bias associated with amplicon sequencing does not affect the quantitative assessment of bacterial community dynamics. *PLoS One* 9**:**e99722

James, D., Varga, A., Pallas, V., Candresse, T. (2006) Strategies for simultaneous detection of multiple plant viruses. *Canadian Journal of Plant Pathology* 28**:**16-29

Jooste, A.E., Molenaar, N., Maree, H.J., Bester, R., Morey, L., de Koker, W.C., Burger, J.T. (2015) Identification and distribution of multiple virus infections in Grapevine leafroll diseased vineyards. *European Journal of Plant Pathology***:**1-13

Khalifa, M.E., Varsani, A., Ganley, A.R.D., Pearson, M.N. (2016) Comparison of Illumina de novo assembled and Sanger sequenced viral genomes: A case study for RNA viruses recovered from the plant pathogenic fungus Sclerotinia sclerotiorum. *Virus Research* 219**:**51-57.

Kircher, M., Sawyer, S., Meyer, M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* 40**:**e3-e3

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7**:**111-118

Mangul, S., Wu, N.C., Mancuso, N., Zelikovsky, A., Sun, R., Eskin, E. (2014) Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics* 30**:**329-337

Massart, S., Olmos, A., Jijakli, H., Candresse, T. (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research* 188**:**90-96

Nekrutenko, A., Taylor, J. (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* 13**:**667-672

Osman, F., Leutenegger, C., Golino, D., Rowhani, A. (2007) Real-time RT-PCR (TaqMan®) assays for the detection of Grapevine Leafroll associated viruses 1–5 and 9. *Journal of Virological Methods* 141**:**22-29

Poojari, S., Alabi, O.J., Fofanov, V.Y., Naidu, R.A. (2013) A Leafhopper-Transmissible DNA Virus with Novel Evolutionary Lineage in the Family Geminiviridae Implicated in Grapevine Redleaf Disease by Next-Generation Sequencing. *PLoS One* 8**:**e64194

Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., Skelly, T.F., McQuillan, J.A., Swerdlow, H.P., Oyola, S.O. (2012) Optimal enzymes for amplifying sequencing libraries. *Nature Methods* 9**:**10-11

Quail, M.A., Smith, M., Jackson, D., Leonard, S., Skelly, T., Swerdlow, H.P., Gu, Y., Ellis, P. (2014) SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics* 15**:**110

Read, D. 2015. Overcoming bias in Citrus tristeza virus (CTV) genotype detection and a population study of CTV within Southern African Star Ruby grapefruit orchards. University of Pretoria, Pretoria, South Africa

Read, D.A., Pietersen, G. (2016) PCR bias associated with conserved primer binding sites, used to determine genotype diversity within Citrus tristeza virus populations. *Journal of Virological Methods* 237**:**107-113

Roossinck, M.J., Martin, D.P., Roumagnac, P. (2015) Plant Virus Metagenomics: Advances in Virus Discovery. *Phytopathology* 10.1094/phyto-12-14-0356-rvw

Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L., Prosperi, M. (2016) Challenges in the analysis of viral metagenomes. *Virus Evolution* 2**:**vew022

Rowhani, A., Uyemoto, J.K., Golino, D.A., Martelli, G.P. (2005) Pathogen Testing and Certification of Vitis and Prunus Species*. *Annual Review of Phytopathology* 43**:**261-278

Sabanadzovic, S., Abou-Ghanem, N., Castellano, M., Digiaro, M., Martelli, G. (2000) Grapevine fleck virus-like viruses in Vitis. *Archives of Virology* 145**:**553-565

Sambrook, J. (2001). Molecular Cloning: 'A laboratory manual'. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research***:**gku1341

Thys, K., Verhasselt, P., Reumers, J., Verbist, B.M., Maes, B., Aerssens, J. (2015) Performance assessment of the Illumina massively parallel sequencing platform for deep sequencing analysis of viral minority variants. *Journal of Virological Methods* 221**:**29-38

Turturo, C., Rott, M.E., Minafra, A., Saldarelli, P., Jelkmann, W., Martelli, G.P. (2000) Partial molecular characterization and RT-PCR detection of grapevine leafroll associated virus 7, Proceedings of the 13th Meeting of ICVG,, Adelaide, Australia

van Dijk, E.L., Jaszczyszyn, Y., Thermes, C. (2014) Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* 322**:**12-20

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

Waugh, C., Cromer, D., Grimm, A., Chopra, A., Mallal, S., Davenport, M., Mak, J. (2015) A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. *Virology Journal* 12**:**55

Webster, C.G., Wylie, S.J., Jones, M.G. (2004) Diagnosis of plant viral pathogens. *Current Science-Bangalore.* 86**:**1604-1607

Wei, T., Clover, G. (2008) Use of primers with 5′ non-complementary sequences in RT-PCR for the detection of nepovirus subgroups A and B. *Journal of Virological Methods* 153**:**16-21

Zagordi, O., Däumer, M., Beisel, C., Beerenwinkel, N. (2012) Read length versus depth of coverage for viral quasispecies reconstruction. *PLoS One* 7**:**e47046

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., Van Nostrand, J.D., He, Z., Yang, Y. (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME Journal* 5**:**1303-1313

Chapter 3:

Application of PolyHiT-Seq as a diagnostic system for the detection of grapevine viruses

3.1. INTRODUCTION

To date, 65 viruses have been identified as infectious agents of grapevine, and are responsible for a wide array of diseases worldwide (Martelli, 2014). The disease complexes considered of most economic importance include GLD, RW, fleck complex, grapevine degeneration and grapevine decline (Martelli, 1993a). The adverse effects of infected and diseased vines include the decline of plant vigour, the reduction in graft compatibility, the decrease in the vineyard productive lifespan and severe loss of crops (Martelli and Boudon-Padiue, 2006, Martelli, 2014). Owing to the importance of grapevine as an economic commodity in many countries worldwide, the sanitary status of commercial vines is of concern, and due to the unfeasibility of virus elimination from individual vines in established vineyards, integrated strategies for virus control is of great importance. Strategies for virus control include the removal of infected vines, the exclusion of viral vectors and the production of virus-free vines (Almeida *et al.*, 2013).

Certification schemes aim to produce propagative plant material that have undergone specific procedures to ensure their health status and horticultural characteristics (Varveri *et al.*, 2015). Within certification schemes, candidate nuclear plants' sanitary status is evaluated primarily by biological indexing and serology-based methods such as ELISA and ISEM, and more recently by molecular methods such as the PCR or RT-PCR (Rowhani *et al.*, 2005). Diagnostic tests used within certification schemes need to be highly reliable, sensitive and specific, especially when marketed commercial crops are termed 'virus tested' (Martin *et al.*, 2000, Mekuria *et al.*, 2003). Other desired characteristics of diagnostic techniques include their ability of large-scale testing, low running costs, speed of detection and ease of use (Martin *et al.*, 2000, Varveri *et al.*, 2015). Biological and serological tests are the primary diagnostic techniques used for virus detection during the certification process, although both these techniques have disadvantages that diminish the desired characteristics of diagnostic tests.

Biological indexing is based on the visualization of disease symptoms on indicator plants after graft inoculation. Typically symptoms appear after a long time, predominantly 2 – 3 years (Rowhani *et al.*, 1997), thus this process is extremely time consuming, requires a lot of space for completion and is not suitable for large scale testing (Dovas and Katis, 2003b, Lopez-Fabuel *et al.*, 2013, Varveri *et al.*, 2015).

ELISA on the other hand is suitable for large scale testing but high-quality antiserum is required for every virus to be tested (Varveri *et al.*, 2015), and the virus titre within the sample must be high. Sampling the appropriate tissue at the correct time is thus essential (Weber *et al.*, 2002, Dovas and Katis, 2003b, Varveri *et al.*, 2015). Furthermore, antiserum to many viruses are lacking and as such many viruses go undetected (Weber *et al.*, 2002). Disadvantages associated with ISEM, is the requirement of an electron microscope as well as trained personnel for the operation of the microscope (Walter, 1993). The cost and labour intensiveness of this technique makes it unsuitable for large scale routine testing (Martelli, 1993b, Walter, 1993, Martin, 1998).

Molecular based methods such as PCR and RT-PCR have proven very useful for the detection of viruses in the absence of antisera (EPPO, 2008) and when present at low titres (Maliogka *et al.*, 2015). Other valuable properties of these techniques is their rapidity, sensitivity and requirement of small quantities of testing material (EPPO, 2008, Maliogka *et al.*, 2015). An added advantage of molecular based methods is the additional information that can be obtained from the sequencing of the amplified products. This is especially so in instances when generic primers are used to only infer information to a higher taxonomic level such as genus or family, where by sequencing the products provides information about virus species, strains and relatedness can be obtained (Webster *et al.*, 2004, Boonham *et al.*, 2014).

In this study we describe the use of poly-specific and virus-specific RT-PCRs in combination with HTS as a diagnostic system, which we refer to here as the PolyHiT-Seq system for the detection and identification of grapevine viruses, in comparison to the serological technique ELISA.

3.2. MATERIALS AND METHODS

Application of poly-specific and virus-specific RT-PCRs on field-collected and candidate nuclear vine samples

A total of 56 wine grape samples were collected from the regions of Somerset West and Stellenbosch in the Western Cape, including samples of Cabernet Sauvignon, Affenthaler, Bacchus, Aleatico nero, Capes Donnes Seedling, Blaue Kadarka, Cabernet Franc, Grand noir de la Calmette, Majestic, Merlot, Malbec and

Shiraz. In addition, a further 49 table grape samples were collected from the Western Cape in the regions Hex river valley, Paarl and Wellington. Included were *Vitis vinifera* cv. Starlight, Crimson, La Rochelle, Red Globe, Prime, Melody, Autumn Royal, Sugra 19 and Waltham cross. All samples were selected based on various virus-like abnormalities or particularly severe leafroll symptoms. RNA was isolated from 24 wine and 11 table field collected grape samples respectively, selected from the initial collected samples as representatives of various symptoms. RNA was also isolated from 14 nuclear samples received from Vititec and 13 from Ernita nurseries according to White *et al*., 2008, with some modifications.

Two hundred milligram plant material was homogenised in liquid nitrogen with a mortar and pestle. To each homogenate 1.2 ml of pre-heated (65°C) CTAB extraction buffer [2% cetyltrimethylammonium bromide (CTAB), 2% polyvinyl-pyrrolidone k-40 (PVP), 25 mM ethylene-diamine-tetra-acetic acid (EDTA), 100 mM Tris-HCl (pH 8.0), 2 M sodium chloride (NaCl), 3% β-Mercaptoethanol] was added and incubated at 65°C for 30 min with occasional vortexing. Following this, samples were centrifuged at 20 000 x g for 10 min and the supernatant transferred to a new 1.5 ml Eppendorf tube. An equal volume of chloroform:isomyl-alcohol (C:I) (24:1) was added to the supernatant, vortexed for 30 s and centrifuged at 18 000 x g for 15 min. The upper aqueous phase was transferred to a new tube and the C:I step repeated. Thereafter, 2 volumes (v/v) 8 M lithium chloride (LiCl) was added and incubated at 4°C overnight. The samples were centrifuged at 18 000 x g for 60 min at 4°C and the supernatant discarded. The pellet was washed with 500 µl of 70% ethanol and centrifuged at 20 000 x g for 15 min at 4°C. The supernatant was discarded and the pellet centrifuged at 20 000 x g for 5 min at 4°C, and dried on ice for 15 min. To the pellet 50 µl of molecular grade water (Sigma; St. Louis, MO, USA) was added and left on ice for 2 h and briefly vortexed for resuspension. The concentration of the isolated RNA was determined with the NanoDrop 2000 Spectrophotometer (Thermo Scientific; Wilmington, DE, USA). The isolated nucleic acid was aliquoted into multiple tubes to ensure minimum freeze-thawing between reactions to maintain nucleic acid integrity. These were stored at -80°C.

Complimentary DNA (cDNA) was synthesized for each sample for each PCR system (*Appendix A: Table 1*) from the isolated RNA with only primer pairs differing (*Appendix A: Table 1*). The reaction volume was 5 µl and consisted of: 1 µl template RNA, 0.5 µM reverse primer and molecular grade water (Sigma; St. Louis, MO, USA), which was heated at 70°C for 5 min and incubated on ice for 5 min. Following the addition of 1x GoScript™ reaction buffer (Promega; Madison, WI, USA), 1.2 mM $MgCl_2$, 1 mM dNTPs (Kapa Biosystems; Cape Town, South Africa), 4 U Ribolock RNase Inhibitor (Thermo Scientific; Vilnius, Lithuania), 160 U GoScript™ Reverse Transcriptase and molecular grade water (Sigma; St. Louis, MO, USA) to a total volume of 10 µl. The cycling conditions were 5 min at 25°C, 60 min at 42°C and 15 min at 70°C. From the synthesized cDNA the samples were amplified using the standard amplification systems used throughout this study as previously described (Chapter 2), with the exception of the *Clostero-,* and *Velariviruses*.

Due to the inconsistency in performance of the nested *Clostero-* and *Velarivirus* RT-PCR established by Dovas and Katis (2003) virus specific primers for GLRaV-1 (Osman *et al.*, 2007), GLRaV-2 (Bertazzon and Angelini, 2004) and GLRaV-7 (Turturo *et al.*, 2000) (*Appendix A: Table 1*) were utilised. RNA positive controls were available for both GLRaV-1 and -2, but since no positive control was available for GLRaV-7 a DNA one was synthetically synthesized (IDT; Iowa; United States). To differentiate the positive control when contamination was suspected, some 70 nucleotides within the expected amplicon were not synthesized and therefore the amplicon was reduced from a 190 bp full product to one of 120 bp (*Figure 3.1*), allowing for the differentiation between a true positive and a contaminant. The sequence of GLRaV-7 from GenBank (JN383343) was used for the design of the synthetic DNA. The synthesized control had the sequence: 5' AAT GAC TGT GAT GTC GCT TTT ACA ACA AAT GAA TTA GTC GTT GGG TAT TCG AAC GAG AAG CTA AAG TTG ATG CTA TTG TCA AAA GTG CCT CCT TGG TGA ATA AAC CTC CTG GTA GTG GTA 3'. Multiple dilution series and temperature gradients were performed on the synthesized positive control to determine optimal dilution and temperature. This optima was obtained with a 1/10 000 dilution at an annealing temperature of 55ºC.

Similarly, an internal control for potential contamination with positive GLRaV-3, a synthetic 232 bp GLRaV-3 sequence based on the size of amplification with an additional internal 'tag' sequence (TAGTAG) was synthesized. The GLRaV-3 sequence was designed based on the isolate 623 (GQ352632) as follows: 5' TAA TCG GAG GTT TAG GTT CCG GAC CGG ACA CCT TAT TGA GGG TCG TTG ACG TGA TAT GTT TAT TCT TGA GAG CCT TGA TAC TGG AGT GCG AAA GGT ATA CGT CTA CTA CGG TTA CAG CCG CAG TTG TAA CGG TAC CGG CTG ATT

ATA ACT CCT TTA AAC G**TA GTA G**AA GCT TCG TTG TTG AGG CAC TGA AAG GTC TTG GTA TAC CGG TTA GAG GTG TTG TTA ACG AAC CGA C 3'. Based on preliminary Sanger sequencing results (data not shown) it was possible to differentiate between the 'tag' positive control sequence (sample 14-8213) and the original GLRaV-3 isolate 623 sequence. Therefore, 14-8213 was used as the positive control during sample evaluation by RT-PCR.

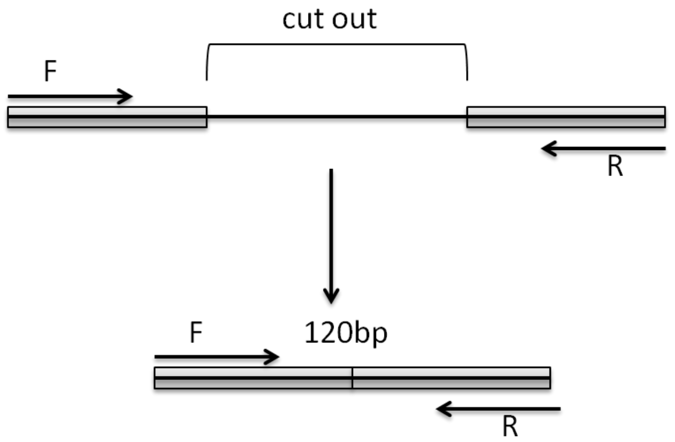

Figure 3.1: Schematic representation of the determinants of the GLRaV-7 DNA synthetic positive control that differed from the true GLRaV-7 sequence by a 70 bp deletion.

cDNA was synthesized for GLRaV-1, -2 and -7 from the isolated RNA as follows: a mixture of 1 µl template RNA, 0.5 µM reverse primer (*Appendix A: Table 1*) and molecular grade water (Sigma; St. Louis, MO, USA) to a final volume of 5 µl was heated at 70°C for 5 min and incubated on ice for 5 min. Following, the addition of 1x MMLV reaction buffer (Promega; Madison, WI, USA), 0.8 mM dNTPs (Kapa Biosystems; Cape Town, South Africa), 8 U Ribolock RNase Inhibitor (Thermo Scientific; Vilnius, Lithuania), 120 U MMLV Reverse Transcriptase and molecular grade water (Sigma; St. Louis, MO, USA) to a total volume of 15 µl. The reaction was incubated at 42ºC for 60 min. Amplification from cDNA was performed as follows: 1x Biotaq NH_4 buffer, 4 mM $MgCl_2$ (Bioline; London, UK), 0.4 mM dNTP mix (Promega; Madison, WI, USA), 0.2 µM forward and reverse primers each (*Appendix A: Table 1*), 1.25 U BIOTAQ DNA polymerase (Bioline; London, UK), 2 µl cDNA and molecular grade water (Sigma-Aldrich; St. Louis, MO, USA) to a final reaction volume of 25 µl. The PCR cycling conditions used were: 95°C for 1 min, followed by 40 cycles of 95°C for 15 s, 15 s at the specific primer annealing temperature of the primer pair (*Appendix A: Table 1*) and 72°C for 20 s, with a final extension of 72°C for 10 min. Following amplification the PCR products were run on a 1% agarose gel stained with ethidium bromide for UV visualization.

To determine which RNA extractions (lacking DNase treatment) contained sufficient DNA for amplification (since Geminiviruses have DNA genomes) the rbcLa plant host housekeeping gene was amplified. Samples that yielded amplicons for this gene, contained DNA as well as RNA and were utilised to test for the presence of the Geminivirus; GRBaV.

Sample preparation for Illumina MiSeq sequencing

Nine samples were selected for Illumina MiSeq sequencing based on their RT-PCR results. They were selected if they yielded amplicons in multiple RT-PCR systems or, in the case of nuclear material if they tested positive in any RT-PCR system. For Illumina MiSeq sequencing, multiple replicated RT-PCR reactions in larger reaction volumes were performed from original RNA extracts to obtain the required template amounts (*Table 3.1*). The products were evaluated by agarose gel electrophoresis, column purified using the NucleoSpin® Gel and PCR clean up kit (Clonetech Laboratories, Inc., Mountain View, California) as per manufacturer's instructions, and quantified using the Quant-iT™ dsDNA BR Assay Kit with the Qubit® 2.0 Fluorometer (Life Technologies, Grand Island, NY, United States). The various PCR products of each individual plant sample were pooled in equimolar amounts and the concentrations of the samples determined with the Nanodrop 2000 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Due to insufficient template for samples 14-8115 and 14-8165 after template pooling they were omitted from further study. Further preparation included the barcoding of each sample with the Illumina TruSeq adapters (Illumina; San Diego, CA, USA) for the sequencing of each on an eighth of a lane on the Illumina MiSeq platform (Illumina; San Diego, CA, USA). Samples were sequenced at the Agricultural Research Council Biotechnology Platform (ARC-BP, Pretoria; South Africa).

Table 3.1: RT-PCR sample preparation for Illumina MiSeq sequencing. The table represents the number of tubes per 50 µl reaction performed of each PCR system for each sample.

Sample	GLRaV-1	GLRaV-2	GLRaV-3	Viti/Fovea
14-8221				4x50
14-8222				4x50
14-8228				4x50
14-8115			2x50	2x50
14-8141		4x50	2x50	2x50
14-8155	4x50	4x50	2x50	2x50
14-8159		4x50	2x50	2x50
14-8160	4x50	4x50	2x50	2x50
14-8165		4x50	2x50	2x50

Illumina MiSeq data analysis

Illumina paired-end data was analysed using CLC Genomics workbench version 6.5.1 (CLC Bio; Aarhus, Denmark). The raw sequence data were imported as paired-ends (distance 180-300 nucleotides), with the removal of failed reads. Quality scores for each data set was generated using the FastQC function according to default settings and the reads filtered by the removal of low quality sequences (quality limit of 0.05), ambiguous nucleotides (maximum of 2 nucleotides) and TruSeq adapter sequences (TruSeq1: TCT AGC CTT CTC GCC AAG TCG TCC; TruSeq2: CCT GCT GAA CCG CTC TTC CGA TCT). Trimmed reads were reference mapped against the reference library using the default settings. The reference library consisted of the cognate areas of the viruses of the RT-PCR systems the samples tested positive for during amplification. The unmapped reads of each sample were subject to *de novo* assembly from which the contigs were generated with the default settings. Contigs were subject to multiBLASTn search against the viral nucleotide collection hosted by the NCBI portal via the CLC interface. Contigs' BLAST results were evaluated based on the BLAST analysis pipeline discussed in Chapter 2 (*Figure 2.13*). Contigs selected for phylogenetic analysis were aligned online with MAFFT (Katoh *et al.*, 2002) using default parameters and alignments trimmed in BioEdit (Hall, 1999). Nucleotide substitution models were selected and maximum likelihood trees were drawn in MEGA 6 (Tamura *et al.*, 2013).

Sanger sequencing of candidate nuclear material RT-PCR positives

Sanger sequencing was conducted on all amplicons obtained through RT-PCR of the candidate nuclear vine material. Samples 14-8221, 14-8222 and 14-8228 were amongst those selected and prepared for Illumina MiSeq sequencing, as they had tested positive for the *Viti-* and *Foveavirus* RT-PCR system. Further, sample 14-8220 had tested positive in the GLRaV-2 RT-PCR system. The PCR products were purified as previously described. Purified products were sequenced in the forward and reverse direction using the dRW-nest1 and dRW-nest2 primers for the *Viti-* and *Foveavirus* RT-PCR systems and the V2dCPf2 and V2CPr1 primers for the GLRaV-2 RT-PCR system, and precipitated as previously described. The samples were submitted to the African Centre for Gene Technologies, Automated Sequencing Facility, Department of Genetics, University of Pretoria, South Africa and sequenced using an ABI Prism® 3130XL Genetic Analyser (Applied Biosystems, Foster City, CA, USA). The sequences obtained were identified using The National Centre for Biotechnology Information (NCBI) Basic Local Alignment Search Tool [BLAST; http://www.ncbi.nlm.nih.gov (Altschul *et al.*, 1990)].

Indirect ELISA for the simultaneous detection of GLRaV-1, -2 and -3

The indirect ELISA was performed for the simultaneous detection of GLRaV-1, -2 and -3 using polyclonal antibodies according to manufacturer instructions (ARC-PPRI, South Africa). ELISA plates were coated as follows: to 0.2 M sodium carbonate buffer (pH 9.6) with 3 mM sodium azide, reagent 1 and 2 (IgG of goat antiserum) was added. Plates were incubated at 37ºC for 2 h. Following which the plates were washed five times with 3 min incubations with 0.2 M phosphate buffer saline (PBS; pH 7.4, 130 mM NaCl, 11 mM $Na_2HPO_4.2H_2O$, 1.5 mM KH_2PO_4, 2.7 mM KCl, 3 mM NaN_3) supplemented with Tween-20. One gram of plant material was homogenised in liquid nitrogen, followed by the addition of 5 ml extraction buffer [100 mM Tris-HCl (pH 7.6), 10 mM $MgSO_4.7H_2O$, 0.2% 2-Mercaptoethanol, 2% Trition X-100, 4% (w/v) Polyvinylpolypyrrolidone (PVPP)] and brief centrifugation. A hundred microliter of the supernatants was added to the plates and incubated at 4ºC overnight. Plant material was discharged and the plate washed as previously described. Hundred microliter of the antisera mixture [anti-GLRaV-1, -2 and -3 (pH

7.0), 0.2 M PBS-Tween-20, 2% (w/v) polyvinylpyrrolidone, 0.2% (w/v) egg albumen powder], was added to each well, and the plate incubated at 37ºC for 2 h. Plates were washed, and 100 µl of substrate was added [1 mg/ml ρ-nitrophenyl phosphate, 10 % diethanolamine (pH 9.8), 3 mM sodium azide].

Plates were read after a one hour incubation with substrates with a Multiscan™ GO Microplate Spectrophotometer (Thermo Scientific; Wilmington, DE, USA) ELISA reader at an optical density of 405 nm. A buffer control, healthy material control and virus-infected control was included in the test. The positive/negative threshold value utilised to interpret the ELISA result was taken to be the mean of the healthy control sample plus two standard deviations.

3.3. RESULTS

Application of poly-specific and virus-specific RT-PCRs on field-collected and candidate nuclear vine samples

None of the samples were infected with GLRaV-4 -like viruses, GLRaV-7, GRBaV, GCSV, any nepoviruses, tombusviruses, maculaviruses or marafiviruses. GLRaV-2, GLRaV-3 and *Viti-* and *Foveaviruses* however were prevalent amongst the samples (*Table 3.2*). GLRaV-1 was detected in two wine grape samples from ARC-Nietvoorbij. None of the samples yielded any amplicons in poly-specific RT-PCR systems other than the *Viti-* and *Foveavirus* nested RT-PCR system. The GLRaV-1, -2 and -3 systems did yield amplicons using virus-specific systems. Therefore very little further identification by HTS was needed. Only nine samples were selected for HTS, based on either their virus population complexity – either testing positive for multiple RT-PCR systems, or the potential of single source viruses (testing positive for a single RT-PCR system).

Table 3.2: RT-PCR or PCR (rbcLa and Geminivirus) results on various grapevine samples▲

Material	Accession	Sample	Farm	GLRaV-1	GLRaV-2	GLRaV-3	GLRaV-4 like viruses	GLRaV-7	Viti and- Foveavirus	Nepovirus A	Nepovirus B and C	Tombusvirus	Maculavirus	Marafivirus	Reovirus	rbcLa (DNA)	Geminivirus*
Candidate Nuclear	14-8231	01/2014	Ernita	-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8232	02/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8233	03/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8234	04/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8235	05/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8236	06/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8237	07/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8238	08/2014		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8239	09/2004		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8240	10/2014		-	-	-	-	-	-	-	-	-	-	-	-	-	
	16-1000	Ernita 1		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	16-1001	Ernita 2		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	16-1002	Ernita 3		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8219	*V. vinifera* cv Lombaardi (normal)	Vititec	-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8220	*V. vinifera* cv Lombaardi (red)		-	+	-	-	-	-	-	-	-	-	-	-	+	-
	14-8221	***V. vinifera* cv Verdelho**		**-**	**-**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8222	**R110 (rootstock for sample 14-8221)**		**-**	**-**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8223	*V. vinifera* cv Verdelho		-	-	-	-	-	+	-	-	-	-	-	-	+	-
	14-8224	R 110		-	-	-	-	-	+	-	-	-	-	-	-	+	-
	14-8225	*V. vinifera* cv Pinot Nior		-	-	-	-	-	+	-	-	-	-	-	-	+	-
	14-8226	*V. vinifera* cv Pinot Nior		-	-	-	-	-	-	-	-	-	-	-	-	-	
	14-8227	*V. vinifera* cv Pinot Nior glen elgin		-	-	-	-	-	+	-	-	-	-	-	-	+	-
	14-8228	**VB 1B**		**-**	**-**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8229	Pristine 2K		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	14-8230	HP 45 4K		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	15-6003	*V. vinifera* (NB230A)		-	-	-	-	-	-	-	-	-	-	-	-	+	-
	15-6004	*V. vinifera* (Cu1 D)		-	-	-	-	-	-	-	-	-	-	-	-	+	-
Table	**14-8115**	***V. vinifera* cv Crimson**	**De Hoop**	**-**	**-**	**+**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	
	14-8118	*V. vinifera* cv Crimson	Mooigesicht	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8125	*V. vinifera* cv La Rochelle	Werda	-	+	+	-	-	-	-	-	-	-	-	-	+	-
	14-8126	*V. vinifera* cv La Rochelle	Werda	-	+	+	-	-	-	-	-	-	-	-	-	+	-
	14-8127	*V. vinifera* cv La Rochelle	Werda	-	+	+	-	-	-	-	-	-	-	-	-	-	
	14-8131	*V. vinifera* cv La Rochelle	Werda	-	+	+	-	-	-	-	-	-	-	-	-	-	
	14-8132	*V. vinifera* cv La Rochelle	Werda	-	+	+	-	-	-	-	-	-	-	-	-	+	-
	14-8133	*V. vinifera* cv La Rochelle	Werda	-	+	+	-	-	-	-	-	-	-	-	-	+	-
	14-8136	*V. vinifera* cv Melody	Irene	-	-	+	-	-	-	-	-	-	-	-	-	+	-
	14-8141	***V. vinifera* cv Sugra19**	**St. Malo**	**-**	**+**	**+**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8142	*V. vinifera* cv Sugra19	St. Malo	-	+	+	-	-	-	-	-	-	-	-	-	-	
Wine	14-8154	*V. vinifera* cv Cabernet Saugvinon	Vergelegen	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8155	***V. vinifera* cv Affenthaler**	**ARC-Nietvoorbi**	**+**	**+**	**+**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8156	*V. vinifera* cv Bacchus	ARC-Nietvoorbij	-	-	+	-	-	-	-	-	-	-	-	-	-	
	14-8157	*V. vinifera* cv Aleatico nero	ARC-Nietvoorbij	-	+	+	-	-	-	-	-	-	-	-	-	-	
	14-8158	*V. vinifera* cv Capes donne seedling	ARC-Nietvoorbij	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8159	***V. vinifera***	**ARC-Nietvoorbi**	**-**	**+**	**+**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8160	***V. vinifera* cv Blaue kadarka**	**ARC-Nietvoorbi**	**+**	**+**	**+**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8161	*V. vinifera* cv Cabernet Franc	ARC-Nietvoorbij	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8162	*V. vinifera*	ARC-Nietvoorbij	-	+	+	-	-	+	-	-	-	-	-	-	+	-
	14-8163	*V. vinifera* cv Grand nior de la calmette	ARC-Nietvoorbij	-	+	+	-	-	-	-	-	-	-	-	-	-	
	14-8164	*V. vinifera*	ARC-Nietvoorbij	-	+	+	-	-	-	-	-	-	-	-	-	+	-
	14-8165	***V. vinifera* cv Majestic/Mureto**	**ARC-Nietvoorbi**	**-**	**+**	**+**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**
	14-8166	*V. vinifera* cv Merlot	ARC-Nietvoorbij	-	-	+	-	-	-	-	-	-	-	-	-	-	
	14-8167	*V. vinifera*	ARC-Nietvoorbij	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8168	*V. vinifera* cv Merlot	ARC-Nietvoorbij	-	-	+	-	-	-	-	-	-	-	-	-	+	-
	14-8169	*V. vinifera* cv Cabernet Saugvinon	Vergelegen	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8170	*V. vinifera* cv Malbeck	Vergelegen	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8174	*V. vinifera* cv Cabernet Franc	Vergelegen	-	-	+	-	-	-	-	-	-	-	-	-	+	-
	14-8176	*V. vinifera* cv Cabernet Saugvinon	Rust en Vrede	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8180	*V. vinifera* cv Shiraz	Rust en Vrede	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8184	*V. vinifera* cv Merlot	Rust en Vrede	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8187	*V. vinifera*	Simonsig	-	-	+	-	-	+	-	-	-	-	-	-	+	-
	14-8197	*V. vinifera*	Simonsig	-	-	+	-	-	+	-	-	-	-	-	-	-	
	14-8210	*V. vinifera* cv Pinot gris	Glen Elgin	-	-	+	-	-	+	-	-	-	-	-	-	+	-

▲ + Indicates viruses detected by RT-PCR. - Indicates viruses undetected by RT-PCR (rbcLa by PCR only). Samples prepared for Illumina MiSeq sequencing in bold. * Samples which lacked DNA, as tested for by PCR (lacking reverse transcription) for the rbcLa gene, were not tested for the Geminivirus, GRBaV.

Illumina MiSeq data analysis

From the seven samples subject to Illumina sequenceing approximately 200 000 reads were obtained per sample, except for samples 14-8221 and 14-8228 which had 532 148 and 32 948 reads, respectively (*Table 3.3*). On average 84% reads mapped during reference mapping (*Table 3.3*). As expected, based on the RT-PCR results, reads of samples 14-8221, 14-8222 and 14-8228 (candidate nuclear material) only mapped to viruses of the genera *Viti-* and *Foveavirus*. These reads were identified as being predominantly GRSPaV of the Foveavirus genus, with various strains of this virus occurring (*Table 3.4, Figure 3.2 a-c*). In both samples 14-8222 and 14-8228, significant, but low numbers of reads also mapped to GVB isolate 99B.Sdz5.2 (*Table 3.4*). In sample 14-8141 (a table grape) the identity of viruses found were confirmed by mapping of reads to reference sequences (*Figure 3.2 d*) and supported the RT-PCR results, with the majority of reads mapping to GVD (accessions Y15892 and AJ45782), followed by GLRaV-3 isolates 623, 621 and GP18, GVB isolate 99B.Sdz5.2, GLRaV-2 isolate 3138-07 and various strains of GRSPaV. Similarly, reads from sample 14-8155 (a wine grape) mapped to viruses within the same virus groups as detected by RT-PCR. The highest incidence of reads mapped to GVB isolate GVB-HI, while reads to GVB isolates 99B.Sdz5.2 and 94/971 were also identified albeit at much lower percentages. Other viruses identified by reference mapping included GLRaV-3 isolates 623, 621, GH30, GH11 and GP18, GLRaV-1 (accessions EF103901, EF195136 and AF195822), GLRaV-2 isolate 3138-07 and various strains of GRSPaV (*Figure 3.2 e*). Only 0.09% of reads mapped to GLRaV-2 in sample 14-8159 (a wine grape) in spite of the fact that this virus was detected by RT-PCR (*Figure 3.2 f*). Mapping of the remaining reads supported the RT-PCR results, identifying GVA isolates P163-M5 and I327-5, GLRaV-3 isolates 621, 623, GH30, GH11 and GP18, GVB isolate 99B.Sdz5.2 and various strains of GRSPaV. For sample 14-8160 (a wine grape), the reads to the same virus groups were detected as suggested by the occurrence of RT-PCR amplicons, the viruses involved were identified during reference mapping. The majority of reads mapped to GLRaV-3, specifically isolate 623, with a lower percentage of reads also mapping to isolates GH30, GH11 and GP18 (*Figure 3.2 g*). Other viruses identified include GVA isolates GTR ISD1, P163-M5, BMo32-1 and I327-5, GLRaV-1 (accessions EF103901, EF195136 and AF195822), GVB isolates 99B.Sdz5.2 and 94/971, GLRaV-2 isolate 31381-07 and various strains of GRSPaV.

Table 3.3: Illumina MiSeq summary statistics. Number of contigs for multiBLASTn results is only of contigs with plant virus BLASTn hits based on the lowest E-value.

	No. Reads			Reference Mapping		*De novo* Assembly			multiBLASTn virus	multiBLASTn all organisms
Sample	Raw	Trim	QC	% Mapped	% Unmapped	No. Assembled reads	No. Unassembled reads	No. Contigs	No. Contigs	No. Contigs
14-8221	532 148	529 574	37	98.43	3.57	1 304	17 612	20	2	2
14-8222	224 040	177 822	37	93.68	6.32	748	10 496	10	1	1
14-8228	32 948	27 866	37	94.25	5.75	227	1 374	5	0	0
14-8141	310 082	308 218	37	75.03	24.97	68 237	8 733	40	12	4
14-8155	209 946	208 486	37	93.68	6.32	1 171	12 009	41	9	5
14-8159	217 002	215 088	37	70.28	29.72	1 246	26 676	28	3	3
14-8160	284 682	241 612	37	61.8	38.2	10 962	81 328	82	24	8

Table 3.4: Percentage mapped reads for each sample. Percentages higher than the positive/negative threshold (0.4%) are indicated in bold.

Sample	GLRaV-1	GLRaV-2	GLRaV-3	GRSPaV	GVA	GVB	GVD	GVE	GVF	GLRaV-3 TAG
14-8221	0	0	0.004	**99.7**	0.03	0.25	0.02	0.0008	0	0.0004
14-8222	0	0	0.001	**98.85**	0.12	**0.95**	0.002	0.11	0	0
14-8228	0	0.02	0.09	**98.34**	0.29	**0.97**	0.19	0.11	0	0
14-8141	0	**3.85**	**14.37**	**2.78**	0.23	**6.64**	**72.05**	0.003	0.0004	0.06
14-8155	**6.5**	**3**	**17.95**	**1.39**	0.39	**70.71**	0.01	0.002	0.0005	0.04
14-8159	0	0.09	**27.43**	**9.28**	**62**	**0.97**	0.03	0.08	0.04	0.06
14-8160	**18.83**	**1.81**	**28.45**	**11.44**	**24.25**	**14.6**	0.11	0.05	0.29	0.18

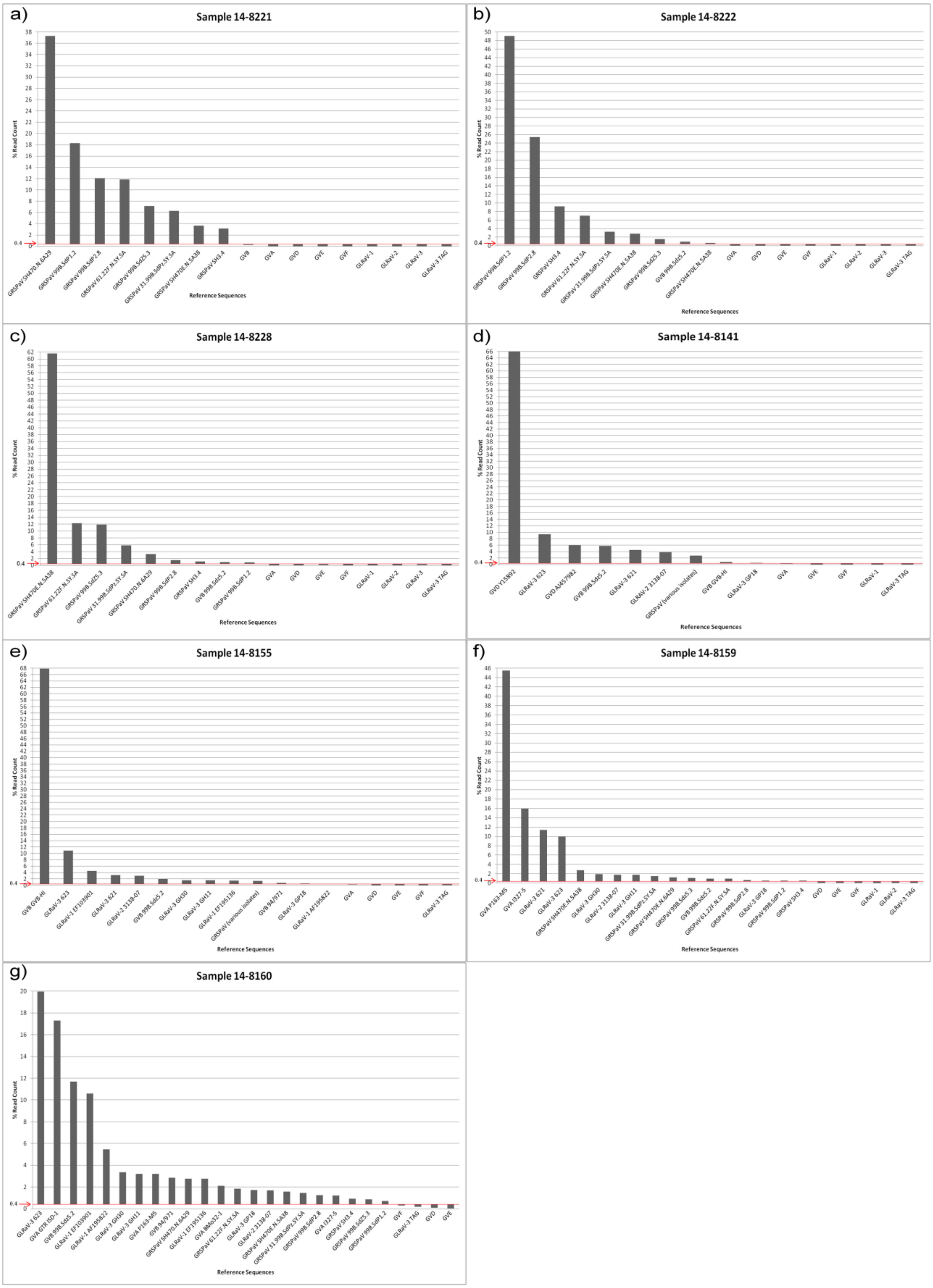

Figure 3.2: Graph of percentage of mapped reads against various virus templates, a) sample 14-8221, b) sample 14-8222, c) sample 14-8228, d) sample 14-8141, e) sample 14-

8155, f) sample 14-8159, g) sample 14-8160. Templates ranked by percentage mapped reads. The positive/negative threshold is 0.4% mapped reads. To better illustrate the difference observed in percentage mapped reads for the various templates in a sample the y-axis is set at 0.4% (positive/negative threshold). Therefore, any template percentage mapped reads equal or above the y-axis is considered true positives and below is discarded as false positives. The reference sequence accessions (NCBI) are FJ884335 GRSPaV SH470.N.6A29, FJ884327 GRSPaV 99.SdP1.2, FJ884328 GRSPaV 99B.SdP2.8, EU247951 GRSPaV 61.22F.N.SY.SA, FJ884333 GRSPaV 99B.SdZ5.3, EU247952 GRSPaV 31.99B.SdPz.SY.SA, FJ884334 GRSPaV SH470.N.5A38, DQ864489 GRSPaV SH3.4, DQ855082 GVA P163-M5, KC962564 GVA I327-5, EU247956 GVB 99B.Sdz5.2, DQ855081 GVA GTR ISD-1, DQ855087 GVA BMo32-1, GU733707 GVB GVB-HI, EF583906 GVB 94/971, Y15892 GVD, AJ457982 GVD, GQ352632 GLRaV-3 623, GQ352631 GLRaV-3 621, EU259806 GLRaV-3 GP18, JQ655296 GLRaV-3 GH30, JQ655295 GLRaV-3 GH11, JX559644 GLRaV-2 3138-07, EF103901 GLRaV-1, EF195136 GLRaV-1, AF195822 GLRaV-1

Reads that did not map to any of the reference sequences utilised were subject to *de novo* assembly and the resulting contigs were evaluated by multiBLASTn against both “all organisms”, as well as only “viruses”, in the NCBI database. Contigs that adhered to the optimized criteria (*Table 3.*5) established previously (*Chapter 2, Figure 2.13*) were individually subjected to BLAST against the complete NCBI nucleotide collection for hit confirmation. ML trees were constructed based on the contig individual BLAST hits to determine the relatedness of the contigs with their matched virus. Contigs with BLAST results supported by reference mapping to isolate level were not included in phylogenetic analysis. Samples 14-8221 and 14-8222, were re-evaluated for the presence of GLRaV-3 by RT-PCR due to the BLAST hit against GLRaV-3 isolate GH30 for contig 7 and 5 respectively. Both samples originally tested negative for the presence of GLRaV-3 by RT-PCR and upon re-evaluation were confirmed to be negative, therefore they were not subjected to phylogenetic analysis. Due to the unexpected occurrence of GVB in nuclear material (14-8221, 14-8222, 14-8228), albeit at extremely low percentages of reads (*Table 3.4*) or only a single contig (as in sample 14-8221) (*Table 3.5*), the samples were re-evaluated for the presence of GVB using GVB-specific primers which support that they did not contain it.

ML analysis of contig 4 of sample 14-8221 (*Figure 3.3*) revealed a clustering of contig 4 with GVA, GVB, GVF and Actinidia virus B, but due to low bootstrap values (60%), that could not be resolved individually. Contigs 2 and 9 of sample 14-8141 (*Figure 3.4*), contig 3 of sample 14-8159 (*Figure 3.6*) as well as contigs 3 and 40 of sample 14-8160 (*Figure 3.7*) all clustered with GLRaV-3. Contigs 6, 8 and 40 of sample 14-8155 (*Figure 3.5*) have a strong bootstrap support (100%) that they are not identical to their respective BLAST hits. For sample 14-8160, both contigs 23 (*Figure 3.7 b*) and 46 (*Figure 3.7 e*) clustered with GLRaV-2 and contig 25 (*Figure 3.7 c*) clustered with GLRaV-1. Contig 38 (*Figure 3.7 d*) clustered with GVA, GVB, and GVE, but due to low bootstrap values (57%), they could not be resolved individually.

Table 3.5: MultiBLASTn-“virus” and multiBALSTn-“all organisms” results on *de novo* assembled contigs. Only plant virus hits selected with the lowest E-value, while “all organisms” selected only for the description of hits with the lowest E-value.

Sample	Query	Number of hits	Query length (bp)	Amplicon length (bp)	Lowest E-value	Hit length	% Query overlap	% Amplicon overlap	% Identity	Accession	Description (virus only)	Description (all organisms)
14-8221	Contig 4	100	204	199	3.24E-25	199	97.55	100	73.63	EU247957	GVB clone 99B.SdP2.10	GVB clone 99B.SdP2.10
	Contig 7	20	577	226	0	549	95.15	242.92	99.64	JQ655296	GLRaV-3 isolate GH30	GLRaV-3 isolate GH30
14-8222	Contig 5	19	212	226	3.873E-90	189	89.15	83.63	99.47	JQ655296	GLRaV-3 isolate GH30	GLRaV-3 isolate GH30
14-8141	Contig 2	100	215	226	1.73E-101	215	100	95.13	98.6	KP867003	GLRaV-3 isolate MO-84	GLRaV-3 isolate MO-84
	Contig 5	7	256	515	1.17E-122	254	99.22	49.32	98.82	JX559644	GLRaV-2 isolate 3138-07	GLRaV-2 isolate 3138-07
	Contig 9	22	628	226	0	549	87.42	242.92	99.82	JQ655296	GLRaV-3 isolate GH30	GLRaV-3 isolate GH30
	Contig 24	101	484	199	1.44E-56	204	42.15	102.51	84.31	Y15892	GVD	Vitis
14-8155	Contig 6	85	219	199	1.813E-69	198	90.41	99.5	89.9	JX105428	GVF isolate AUD46129	GVF isolate AUD46129
	Contig 8	67	209	515	8.78E-99	205	98.09	39.81	99.51	KF007292	GLRaV-2 isolate 6693_LR2CP_PN3	GLRaV-2 isolate 6693_LR2CP_PN3
	Contig 10	94	374	226	1.39E-102	272	72.73	120.35	90.81	JQ655295	GLRaV-3 isolate GH11	GLRaV-3 isolate GH11
	Contig 11	19	292	226	5.18E-146	292	100	129.2	99.66	JQ655296	GLRaV-3 isolate GH30	GLRaV-3 isolate GH30
	Contig 40	4	320	320	2.63E-138	298	93.13	93.13	96.98	JQ023131	GLRaV-1 isolate 1050	GLRaV-1 isolate 1050
14-8159	Contig 3	100	283	226	7.182E-88	182	64.31	80.53	100	KM058745	GLRaV-3 isolate GH24	GLRaV-3 isolate GH24
	Contig 14	101	183	226	5.906E-51	207	113.11	91.59	80.68	JQ655296	GLRaV-3 isolate GH30	GLRaV-3 isolate GH30
14-8160	Contig 3	100	276	226	8.278E-80	171	61.96	75.66	99.42	KM058745	GLRaV-3 isolate GH24	GLRaV-3 isolate GH24
	Contig 23	27	328	515	0	558	170.12	108.35	98.75	KF220376	GLRaV-2 isolate GLRaV-2-SG	GLRaV-2 isolate GLRaV-2-SG
	Contig 25	76	258	320	9.438E-98	228	88.37	71.25	95.18	KF029725	GLRaV-1 isolate 82PL1	GLRaV-1 isolate 82PL1
	Contig 38	100	283	199	4.027E-20	279	98.59	140.2	68.68	AB432911	GVE strain TvP15	GVE strain TvP15
	Contig 46	100	233	515	1.83E-106	232	99.57	45.05	97.42	KF220376	GLRaV-2 isolate GLRaV-2-SG	GLRaV-2 isolate GLRaV-2-SG
	Contig 48	54	269	226	2.237E-86	182	67.66	80.53	99.45	KM058745	GLRaV-3 isolate GH24	GLRaV-3 isolate GH24
	Contig 63	12	259	226	1.3E-120	252	97.3	111.5	98.41	JQ655296	GLRaV-3 isolate GH30	GLRaV-3 isolate GH30

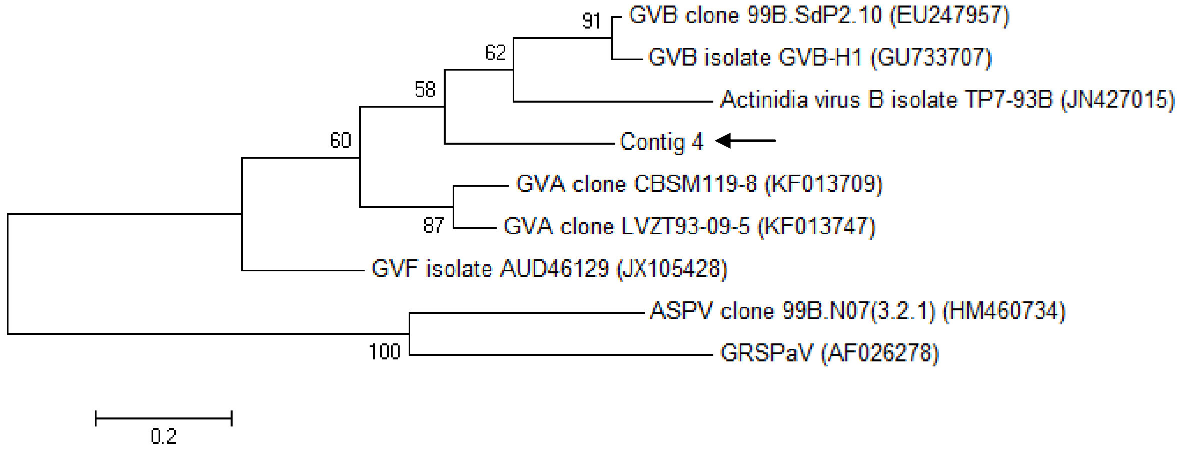
91
GVB clone 99B.SdP2.10 (EU247957)
62
GVB isolate GVB-H1 (GU733707)
58
Actinidia virus B isolate TP7-93B (JN427015)
60
Contig 4
GVA clone CBSM119-8 (KF013709)
87
GVA clone LVZT93-09-5 (KF013747)
GVF isolate AUD46129 (JX105428)
ASPV clone 99B.N07(3.2.1) (HM460734)
100
GRSPaV (AF026278)
0.2

Figure 3.3: Maximum likelihood phylogenetic tree for sample 14-8221 contig 4, with bootstrap values indicated on the branches as a percentage of a 1000 replicates. Tree drawn with cognate regions of viruses to contig 4.

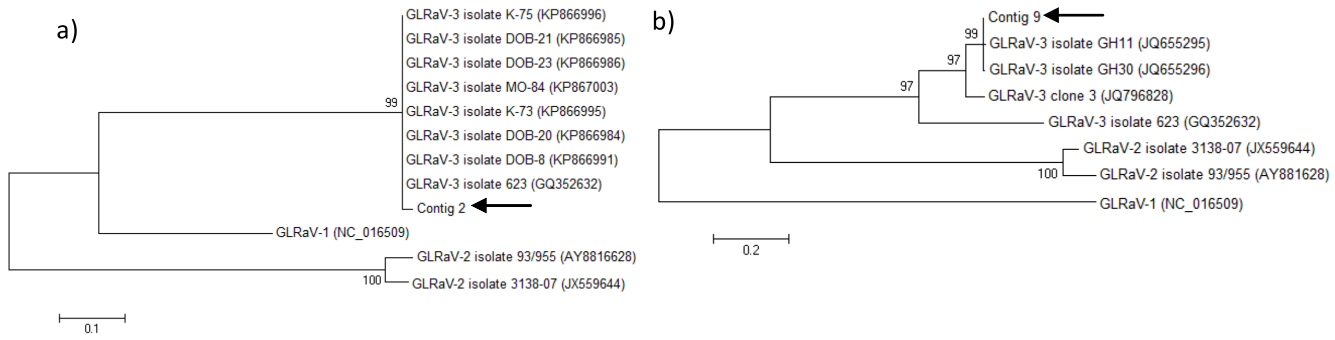
a)
GLRaV-3 isolate K-75 (KP866996)
GLRaV-3 isolate DOB-21 (KP866985)
GLRaV-3 isolate DOB-23 (KP866986)
GLRaV-3 isolate MO-84 (KP867003)
99
GLRaV-3 isolate K-73 (KP866995)
GLRaV-3 isolate DOB-20 (KP866984)
GLRaV-3 isolate DOB-8 (KP866991)
GLRaV-3 isolate 623 (GQ352632)
Contig 2
GLRaV-1 (NC_016509)
GLRaV-2 isolate 93/955 (AY8816628)
100
GLRaV-2 isolate 3138-07 (JX559644)
0.1
b)
Contig 9
99
GLRaV-3 isolate GH11 (JQ655295)
97
GLRaV-3 isolate GH30 (JQ655296)
97
GLRaV-3 clone 3 (JQ796828)
GLRaV-3 isolate 623 (GQ352632)
GLRaV-2 isolate 3138-07 (JX559644)
100
GLRaV-2 isolate 93/955 (AY881628)
GLRaV-1 (NC_016509)
0.2

Figure 3.4: Maximum likelihood tree for sample 14-8141, with bootstrap values indicated on the branches as a percentage of a 1000 replicates. a) contig 2, b) contig 9. Trees drawn with cognate regions of viruses to contig 2 and 9.

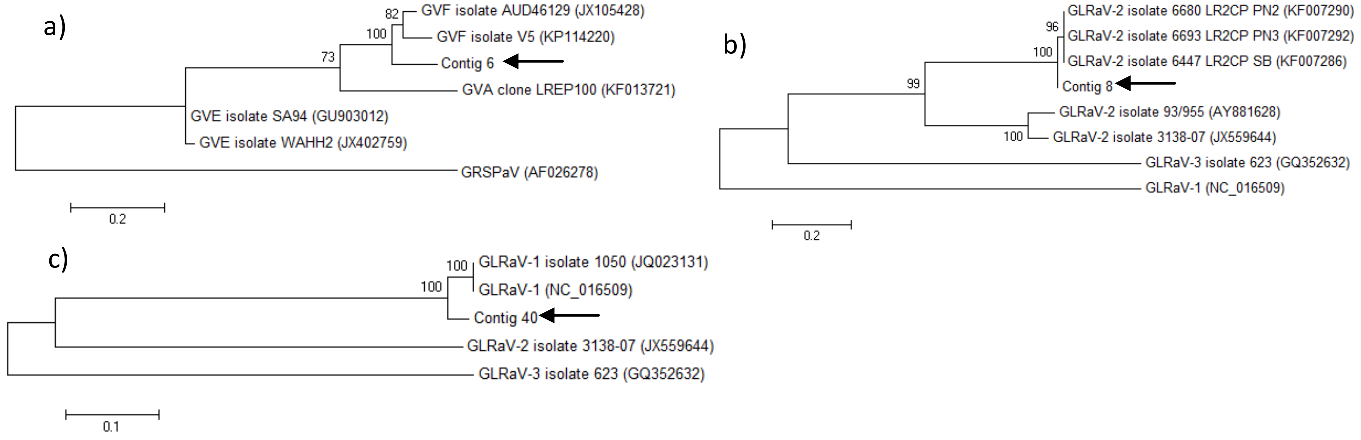
a)
82
GVF isolate AUD46129 (JX105428)
100
GVF isolate V5 (KP114220)
73
Contig 6
GVA clone LREP100 (KF013721)
GVE isolate SA94 (GU903012)
GVE isolate WAHH2 (JX402759)
GRSPaV (AF026278)
0.2
b)
GLRaV-2 isolate 6680 LR2CP PN2 (KF007290)
96
GLRaV-2 isolate 6693 LR2CP PN3 (KF007292)
100
GLRaV-2 isolate 6447 LR2CP SB (KF007286)
99
Contig 8
GLRaV-2 isolate 93/955 (AY881628)
100
GLRaV-2 isolate 3138-07 (JX559644)
GLRaV-3 isolate 623 (GQ352632)
GLRaV-1 (NC_016509)
0.2
c)
100
GLRaV-1 isolate 1050 (JQ023131)
100
GLRaV-1 (NC_016509)
Contig 40
GLRaV-2 isolate 3138-07 (JX559644)
GLRaV-3 isolate 623 (GQ352632)
0.1

Figure 3.5: Maximum likelihood tree for sample 14-8155, with bootstrap values indicated on the branches as a percentage of a 1000 replicates. a) contig 6, b) contig 8, c) contig 40. Trees drawn with cognate regions of viruses to contig 6, 8 and 40.

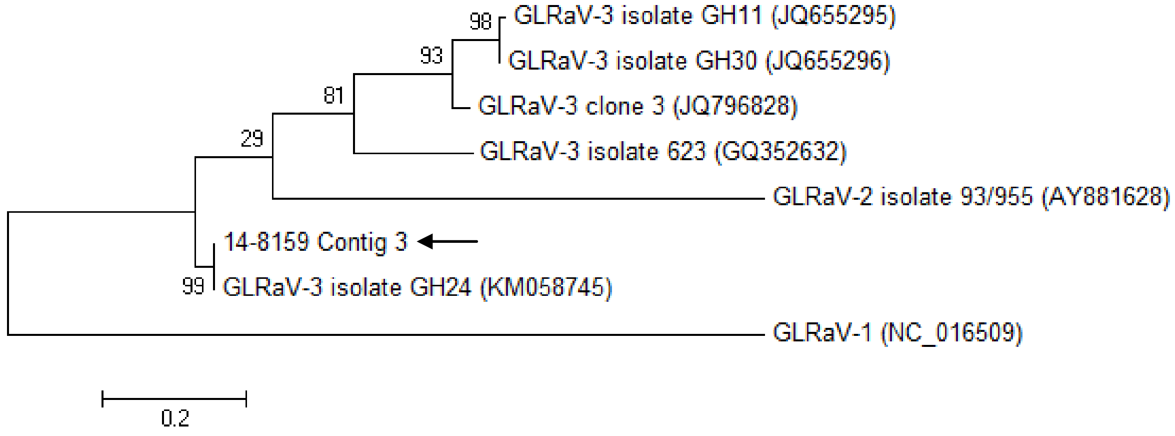
98
GLRaV-3 isolate GH11 (JQ655295)
93
GLRaV-3 isolate GH30 (JQ655296)
81
GLRaV-3 clone 3 (JQ796828)
29
GLRaV-3 isolate 623 (GQ352632)
GLRaV-2 isolate 93/955 (AY881628)
14-8159 Contig 3
99
GLRaV-3 isolate GH24 (KM058745)
GLRaV-1 (NC_016509)
0.2

Figure 3.6: Maximum likelihood tree for sample 14-8159 contig 3, with bootstrap values indicated on the branches as a percentage of a 1000 replicates. Tree drawn with cognate regions of viruses to contig 3.

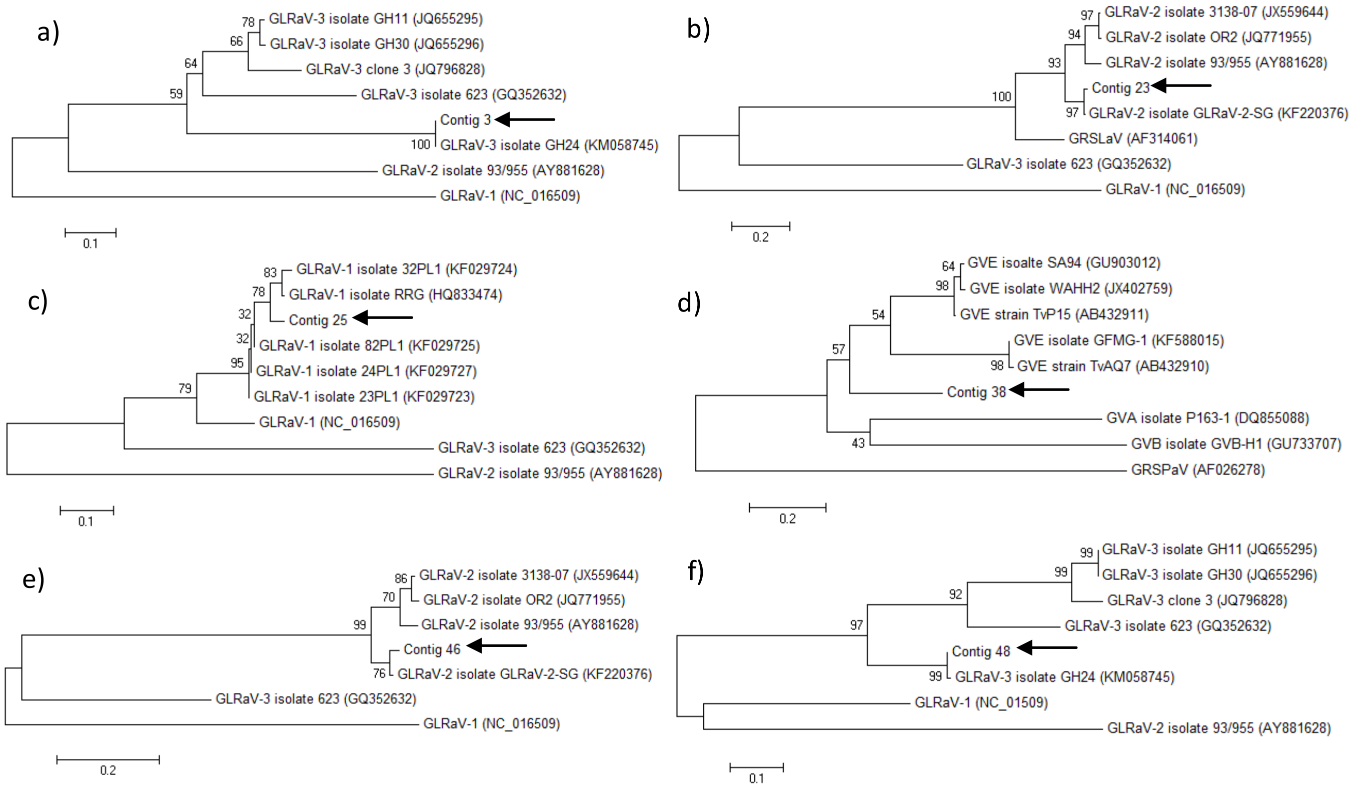
a)
78
GLRaV-3 isolate GH11 (JQ655295)
66
GLRaV-3 isolate GH30 (JQ655296)
64
GLRaV-3 clone 3 (JQ796828)
59
GLRaV-3 isolate 623 (GQ352632)
Contig 3
100
GLRaV-3 isolate GH24 (KM058745)
GLRaV-2 isolate 93/955 (AY881628)
GLRaV-1 (NC_016509)
0.1
b)
97
GLRaV-2 isolate 3138-07 (JX559644)
94
GLRaV-2 isolate OR2 (JQ771955)
93
GLRaV-2 isolate 93/955 (AY881628)
Contig 23
100
97
GLRaV-2 isolate GLRaV-2-SG (KF220376)
GRSLaV (AF314061)
GLRaV-3 isolate 623 (GQ352632)
GLRaV-1 (NC_016509)
0.2
c)
83
GLRaV-1 isolate 32PL1 (KF029724)
78
GLRaV-1 isolate RRG (HQ833474)
32
Contig 25
32
GLRaV-1 isolate 82PL1 (KF029725)
95
GLRaV-1 isolate 24PL1 (KF029727)
79
GLRaV-1 isolate 23PL1 (KF029723)
GLRaV-1 (NC_016509)
GLRaV-3 isolate 623 (GQ352632)
GLRaV-2 isolate 93/955 (AY881628)
0.1
d)
64
GVE isoalte SA94 (GU903012)
98
GVE isolate WAHH2 (JX402759)
54
GVE strain TvP15 (AB432911)
57
GVE isolate GFMG-1 (KF588015)
98
GVE strain TvAQ7 (AB432910)
Contig 38
GVA isolate P163-1 (DQ855088)
43
GVB isolate GVB-H1 (GU733707)
GRSPaV (AF026278)
0.2
e)
86
GLRaV-2 isolate 3138-07 (JX559644)
70
GLRaV-2 isolate OR2 (JQ771955)
99
GLRaV-2 isolate 93/955 (AY881628)
Contig 46
76
GLRaV-2 isolate GLRaV-2-SG (KF220376)
GLRaV-3 isolate 623 (GQ352632)
GLRaV-1 (NC_016509)
0.2
f)
99
GLRaV-3 isolate GH11 (JQ655295)
99
GLRaV-3 isolate GH30 (JQ655296)
92
GLRaV-3 clone 3 (JQ796828)
97
GLRaV-3 isolate 623 (GQ352632)
Contig 48
99
GLRaV-3 isolate GH24 (KM058745)
GLRaV-1 (NC_01509)
GLRaV-2 isolate 93/955 (AY881628)
0.1

Figure 3.7: Maximum likelihood tree for sample 14-8160, with bootstrap values indicated on the branches as a percentage of a 1000 replicates. a) contig 3, b) contig 2, c) contig 25, d) contig 38, e) contig 46, f) contig 48. Trees drawn with cognate regions of viruses to contig 2, 3, 25 38, 46 and 48.

Sanger sequencing of candidate nuclear material RT-PCR positives

Several candidate nuclear vines tested positive for the *Viti*- and *Foveavirus* RT-PCR system (*Table 3.2*). Amplicons to these samples were subjected to Sanger sequencing, except for samples 14-8221, 14-8222 and 14-8228 which had been subject to Illumina MiSeq sequencing, along with sample 14-8220 which had tested positive for GLRaV-2 (*Table 3.6*) a *Clsoterovirus*. Since these samples were sanitized by thermotherapy and meristem tip culture, they were expected to be virus-free. Therefore, the positive RT-PCR results were not expected and the sequences analyzed. Each sample contained the *Foveavirus*, GRSPaV except for sample 14-8220 that contained GLRaV-2 a *Closterovirus*.

Table 3.6: Sanger sequencing BLAST results for candidate nuclear vines that tested positive by RT-PCR.

Sample	BLAST result	Accession	Query Cover	E-value	Identities
14-8220	GLRaV-2 isolate Pinot Nior	EF118033	99%	0	96%
14-8223	GRSPaV strain GR1	JN683371	91%	3.00E-65	88%
14-8224	GRSPaV strain GR2	JN683372	93%	8.00E-77	93%
14-8225	GRSPaV stran GR6	JN683376	100%	3.00E-21	82%
14-8227	GRSPaV clone 61.22F.N.SY.SA	EU247951	98%	9.00E-83	95%

Indirect ELISA for the simultaneous detection of GLRaV-1, -2 and -3

To compare the results obtained by RT-PCR for GLRaV-1, -2 and -3 to the standard serological technique used within the certification scheme – ELISA – a total of 35 field samples were tested by ELISA for the simultaneous detection of GLRaV-1, -2 and -3. Eight of the samples tested negative (14-8136, 14-8154, 14-8169, 14-8170, 14-8174, 14-8219, 14-8231, 14-8232), of which 62.5% were observed to be positive for GLRaV-3 when evaluated by RT-PCR (*Figure 3.8, Table 3.2*). In this survey ELISA had a 14.3% false negative rate in comparison to RT-PCR.

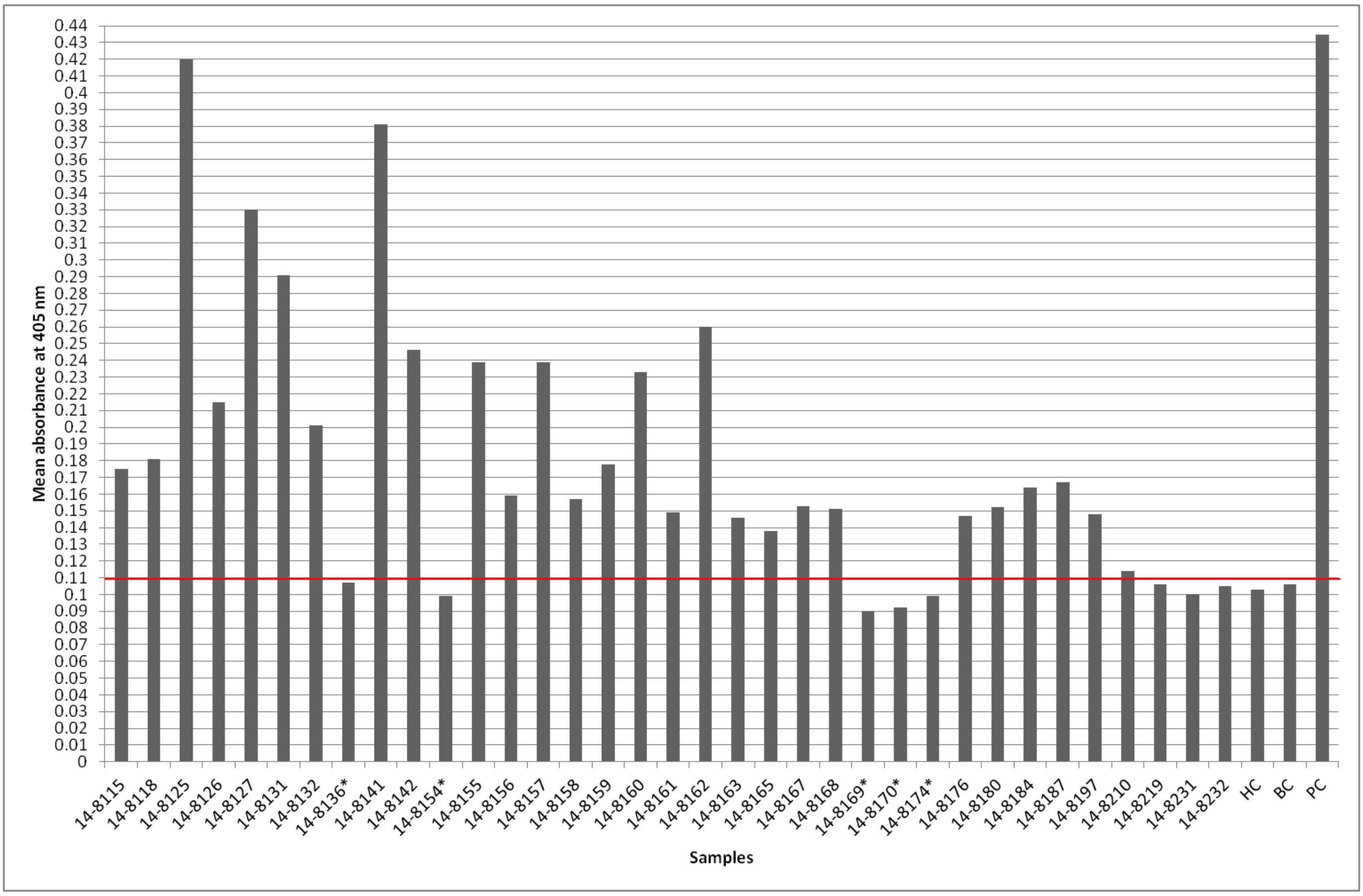

Figure 3.8: ELISA results of field samples evaluated for GLRaV-1, -2 and -3. Horizontal red line represents the mean of the healthy control (HC) plus two standard deviations. BC = buffer control. PC = positive control. (*samples tested negative by ELISA but positive by RT-PCR).

3.4. DISCUSSION AND CONCLUSION

In this study we make use of a novel approach to virus detection and identification, that we call the PolyHiT-Seq system, which involves using HTS sequencing of PCR products generated from poly-specific and virus-specific primers along with an optimized data analysis pipeline. While 62 samples were tested by PCR, the entire approach was only applied to seven grapevine samples. We compared the RT-PCR results obtained for 35 grapevine samples to that of the standard ELISA used within the sanitation program. We found that making use of RT-PCR is superior to that of the standard ELISA, and that the HTS of their products together with the implementation of the optimized pipeline leads to a very useful diagnostic system. The capability of RT-PCR to readily detect viruses present in low concentrations in woody plants, the availability of primers for virus identification, the ease and rapidity of the technique, together with constant improvement of HTS platforms especially in the area of cost makes this a useful method for virus diagnostics.

Many studies have been published on the detection of multiple viruses through either multiplex PCRs, poly-specific PCRs or the use of multiple PCR systems (Saldarelli *et al.*, 1998, Sabanadzovic *et al.*, 2000, Gambino and Gribaudo, 2006, Nakaune and Nakano, 2006, Digiaro *et al.*, 2007). In this study we report the use of poly-specific or virus-specific primers for the detection of 37 grapevine infecting viruses from 11 genera. The PCR products generated by poly-specific primers can detect viruses to the genus level while HTS of the products allows identification to species or strain level, in this case using the Illumina MiSeq platform. The PCR component of this system was compared to the standard ELISA performed during certification.

The application of various HTS approaches and the successive bioinformatic analysis of the generated reads, hold the potential for routine, generic detection of viruses. The primary reasons for this is the ability of HTS platforms to produce unprecedented amounts of data per single run as well as the fact that no prior information about the infectious population of the sample is required (Barba *et al.*, 2014). HTS has been applied to several areas in plant virology, including whole genome sequencing, discovery and detection, replication and transcription; by the sequencing of total RNA (Poojari *et al.*, 2013), messenger RNA (mRNA), double stranded RNA (dsRNA) (Coetzee *et al.*, 2010, Al Rwahnih *et al.*, 2013) and small RNAs (siRNA)(Zhang *et al.*, 2011, Giampetruzzi *et al.*, 2012). Several studies on grapevines have made use of sequencing cDNA libraries generated from dsRNA since endogenous plant RNAs do not form extensive double stranded structures as replicative RNA viruses do, thus this approach enriches for viral nucleic acid. Coetzee *et al*., (2010) sequenced cDNA from dsRNA of 44 pooled grapevines and they discovered the presence of grapevine virus E, a virus previously unreported in South African vineyards. Al Rwahnih *et al*., (2013) making use of the Illumina Genome Analyzer platform sequenced cDNA libraries prepared from dsRNA and identified GRBaV. This virus had been identified previously by Poojari *et al*., (2013) as *Grapevine cabernet franc associated virus* when they sequenced total RNA. Viruses associated with decline symptoms in Syrah grapevines were also investigated using HTS. Both total RNA and cDNA from dsRNA were sequenced for the identification of multiple viruses and viroids, confirming that decline symptoms are the result of a mixed infection (Al Rwahnih *et al.*, 2009). siRNA sequencing is based on the natural anti-viral defence system of RNA silencing, where small 21-24 nucleotide siRNAs are generated that corresponds to the invading viruses or viroids. The use of siRNA sequencing has proven very efficient in the discovery of novel viruses. Both Zhang *et al*., (2011) and Giampetruzzi *et al*., (2012), discovered novel grapevine infecting viruses when they used this approach for HTS. While these HTS approaches have many advantages, there are also some disadvantages associated with them. For the sequencing of sample RNAs relatively large amounts of nucleic acid templates are required and may be difficult to obtain and for many of these approaches, the presence of low virus titres will result in a large number of the reads obtained being directed at the host (Boonham *et al.*, 2014).

To our knowledge, no standard pipeline for data analysis has been established for plant virus diagnosis, and this poses its own set of challenges with regards to new viruses and virus variants. Only once the need for an automated and universally applicable bioinformatics pipeline have been met, will these technologies be used for mainstream virus diagnostics (Boonham *et al.*, 2014). In this study we make use of novel approach of HTS by sequencing PCR products generated from multiple poly-specific and virus-specific primers. We were able to successfully apply the, here named, PolyHiT-Seq system to seven grapevine samples selected from an initial screening of 62 samples. Viruses from the *Viti-* and *Foveavirus* poly-specific nested RT-PCR were identified by Illumina MiSeq sequencing to the species level. The threshold established for reference mapping as well the criteria for BLAST analysis (established in Chapter 2) could successfully be implemented, proving the applicability of the PolyHiT-Seq system in grapevine diagnostics.

Seven nuclear samples obtained from Vititec tested positive for the presence of either a *Vitivirus* or a *Foveavirus*. Upon the sequencing of the PCR products, the samples were found to be infected by GRSPaV. This is not an unexpected occurrence since GRSPaV is known to be rather recalcitrant to elimination by any of the standard virus elimination techniques employed for grapevines; thermotherapy, meristem, shoot tip or auxiliary bud culture, somatic embryogenesis as well as chemotherapy. Many studies have found that viruses from the Closteroviridae family are much easier to eliminate than GRSPaV by thermotherapy together with meristem and shoot tip culture (Maliogka *et al.*, 2009, Skiada *et al.*, 2009). Maliogka *et al.*, (2009) concluded from their findings that success of thermotherapy is dependent not only on the virus species to be eliminated but that the specific interaction between the pathogen and the grapevine genotype also plays a role. Somatic embryogenesis is technically more difficult, time consuming, and cultivar dependant, than thermotherapy and meristem and shoot tip culture. However when used as a sanitation technique in comparison to thermotherapy and meristem tip culture, more plants were found to be free from GRSPaV infection (Gribaudo *et al.*, 2006). The technique however may result in somaclonal variation in plantlets, an undesirable trait for vegetatively propagated plants (Skiada *et al.*, 2013). More recently chemotherapy has been proven to be quite successful in the elimination of GRSPaV from plantlets after treatment. In a study by Skiada *et al.*, (2013), they tested the efficacy of three antiviral compounds for the elimination of GRSPaV. They observed that, with a higher concentration of the antiviral compound, there was a higher percentage of virus eradication. However, they also observed a very high percentage of phytotoxicity. Even though they observed a higher efficacy of chemotherapy in comparison to thermotherapy and meristem tip culture, they had a much lower plantlet survival rate (Skiada *et al.*, 2009, Skiada *et al.*, 2013). The reason for GRSPaV recalcitrance in grapevine is however still not known, therefore the presence of this virus in nuclear samples that were derived from thermotherapy and meristem tip culture, as is used by Vititec, is of no surprise, as these sanitation techniques have been found to be the least efficient in GRSPaV elimination.

ELISA has been the foremost tool used for virus testing in fields such as breeding, quarantine and certification. Due to the relative ease of implementation and interpretation of results and its speed of application this technique is well suited for high throughput testing (Boonham *et al.*, 2014). Even though nucleic acid-based methods such as RT-PCRs can be a more generic-type test than ELISA, and their flexibility makes it possible to address a broader range of diagnostic questions, there have been reservations for its use as a routine diagnostic system (Boonham *et al.*, 2014). In our study, when samples evaluated by RT-PCR were tested using ELISA a false-negative rate of 14.3% was observed, confirming the well-known fact that RT-PCR is the more sensitive test of the two. This confirms a number of previous studies. Specifically on grapevine, Constable *et al.*, (2012) evaluated vines over three growing seasons for the presence of GLRaV-2, GLRaV-3, GVA and GFkV. In their study viruses were detected in a higher proportion of samples by RT-PCR compared to ELISA, showing that RT-PCR is a more reliable detection system than ELISA. Fiore *et al*., (2009) made a similar observation regarding the higher sensitivity of RT-PCR in comparison to ELISA. They also observed that there is a difference in virus detection of different viruses, plant tissues and time of the year, therefore samples tested negative by ELISA may be as a result of the temporarily low virus concentrations in the sampled tissue (Fiore *et al.*, 2009). The same was observed in a study by Kominek *et al*., (2009), who observed a difference in virus presence between individual shoots of the same plant.

Making use of the PolyHiT-Seq system, we could successfully determine the virus population of seven samples to the species and in many instances the isolate level. This system is relatively time efficient and easy, which are desired characteristics in diagnostics, therefore it is suitable for routine use. The establishment of various poly-specific and virus-specific primer sets into standard RT-PCR systems (GoTaqG2 and BioTaq) makes them easy to apply routinely, and in the instance of large sample sizes to be tested, poly-specific primers can easily be used as a pre-screening method. With the comparison of the RT-PCR results with the standard ELISA used within the sanitation program, the higher sensitivity of the RT-PCRs also lends to its superiority as a diagnostic test. The sequencing of the poly-specific RT-PCR products, together with proper data analysis of the reads, allows the identification of the viruses present and for some systems identification to isolate level. We expect that the data analysis pipeline is robust enough for use in other amplicon HTS projects that target different plant viruses. The availability of virus sequences enables the design of many more poly-specific primer sets for the detection of viruses across several genera and crops not included in this study, thereby expanding the PolyHiT-Seq system for the detection of not only grapevine viruses but viruses of multiple crops. The ease of expansion of this system for the diagnosis of other virus infected crops, together with its sensitivity, illustrates the robustness and applicability of this system as a diagnostic system.

3.5. REFERENCES

Al Rwahnih, M., Dave, A., Anderson, M., Uyemoto, J., Sudarshana, M. (2013) Association of a Circular DNA Virus in Grapevines Affected by Red blotch Disease in California. *Phytopathology* 103**:**1069-1076

Almeida, R.P., Daane, K.M., Bell, V.A., Blaisdell, G.K., Cooper, M.L., Herrbach, E., Pietersen, G. (2013) Ecology and management of grapevine leafroll disease. *Frontiers in microbiology* 4**:**94

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *Journal of molecular biology* 215**:**403-410

Barba, M., Czosnek, H., Hadidi, A. (2014) Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* 6**:**106-136

Bertazzon, N., Angelini, E. (2004) Advances in the detection of Grapevine leafroll-associated virus 2 variants. *Journal of Plant Pathology***:**283-290

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., Mumford, R. (2014) Methods in virus diagnostics: from ELISA to next generation sequencing. *Virus research* 186**:**20-31

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 a vineyard. *Virology* 400**:**157-163

Constable, F.E., Connellan, J., Nicholas, P., Rodoni, B.C. (2012) Comparison of enzyme-linked immunosorbent assays and reverse transcription-polymerase chain reaction for the reliable detection of Australian grapevine viruses in two climates during three growing seasons. *Australian Journal of Grape and Wine Research* 18**:**239-244 10.1111/j.1755-0238.2012.00188.x.

Digiaro, M., Elbeaino, T., Martelli, G.P. (2007) Development of degenerate and species-specific primers for the differential and simultaneous RT-PCR detection of grapevine-infecting nepoviruses of subgroups A, B and C. *Journal of virological methods* 141**:**34-40

Dovas, C., Katis, N. (2003b) A spot nested RT-PCR method for the simultaneous detection of members of the< i> Vitivirus</i> and< i> Foveavirus</i> genera in grapevine. *Journal of virological methods* 107**:**99-106

EPPO (2008) Pathogen-tested material of grapevine varieties and rootstocks. *European and Mediterranean Plant Protection Organization Bulletin* 38**:**422–429 10.1111/j.1365-2338.2008.01258.x.

Fiore, N., Prodan, S., Pino, A. (2009) Monitoring grapevine viruses by ELISA and RT-PCR throughout the year. *Journal of Plant Pathology* 91**:**489-493

Gambino, G., Gribaudo, I. (2006) Simultaneous detection of nine grapevine viruses by multiplex reverse transcription-polymerase chain reaction with coamplification of a plant RNA as internal control. *Phytopathology* 96**:**1223-1229

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., Terlizzi, F., Credi, R., Saldarelli, P. (2012) A new grapevine virus discovered by deep sequencing of virus-and viroid-derived small RNAs in Cv< i> Pinot gris</i>. *Virus research* 163**:**262-268

Gribaudo, I., Gambino, G., Cuozzo, D., Mannini, F. (2006) Attempts to eliminate Grapevine rupestris stem pitting-associated virus from grapevine clones. *Journal of Plant Pathology***:**293-298

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Presented at the Nucleic acids symposium series*, p. 95-98

Jooste, A.E., Molenaar, N., Maree, H.J., Bester, R., Morey, L., de Koker, W.C., Burger, J.T. (2015) Identification and distribution of multiple virus infections in Grapevine leafroll diseased vineyards. *European journal of plant pathology***:**1-13

Katoh, K., Misawa, K., Kuma, K.i., Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30**:**3059-3066

Komínek, P., Glasa, M., Komínková, M. (2009) Analysis of multiple virus-infected grapevine plant reveals persistence but uneven virus distribution. *Acta virologica* 53**:**281

Lopez-Fabuel, I., Wetzel, T., Bertolini, E., Bassler, A., Vidal, E., Torres, L.B., Yuste, A., Olmos, A. (2013) Real-time multiplex RT-PCR for the simultaneous detection of the five main grapevine viruses. *Journal of virological methods* 188**:**21-24 10.1016/j.jviromet.2012.11.034.

Maliogka, V., Skiada, F., Eleftheriou, E., Katis, N. (2009) Elimination of a new ampelovirus (GLRaV-Pr) and Grapevine rupestris stem pitting associated virus (GRSPaV) from two Vitis vinifera cultivars combining in vitro thermotherapy with shoot tip culture. *Scientia Horticulturae* 123**:**280-282

Maliogka, V.I., Martelli, G.P., Fuchs, M., Katis, N.I. (2015) Chapter Six-Control of Viruses Infecting Grapevine. *Advances in virus research* 91**:**175-227

Martelli, G. (1993a). Graft-transmissible diseases of grapevines: handbook for detection and diagnosis. FAO.

Martelli, G. (1993b) Immunosorbent electron microscopy (ISEM) and antibody coating. *Graft-transmissible diseases of grapevines. Handbook for detection and diagnosis***:**193-195

Martelli, G.P. (2014) Directory of Virus and Virus-like Diseases of the Grapevine and their Agents. *Journal of Plant Pathology* 96

Martelli, G.P., Boudon-Padiue, E. (2006) Directory of Infectious Diseases of Grapevines and Viroses and Virus-like Diseases of the Grapevine: Bibliographic Report 1998-2004. *CIHEAM; Options Méditerranéennes: Série B. Etudes et Recherches; n. 55*

Martin, R. (1998) Advanced diagnostic tools as an aid to controlling plant virus diseases. *Plant virus disease control. APS Press, St Paul, MN***:**381-391

Martin, R.R., James, D., Lévesque, C.A. (2000) Impacts of molecular diagnostic technologies on plant disease management. *Annual Review of Phytopathology* 38**:**207-239

Mekuria, G., Ramesh, S.A., Alberts, E., Bertozzi, T., Wirthensohn, M., Collins, G., Sedgley, M. (2003) Comparison of ELISA and RT-PCR for the detection of Prunus necrotic ring spot virus and prune dwarf virus in almond (Prunus dulcis). *Journal of virological methods* 114**:**65-69

Nakaune, R., Nakano, M. (2006) Efficient methods for sample processing and cDNA synthesis by RT-PCR for the detection of grapevine viruses and viroids. *Journal of virological methods* 134**:**244-249

Osman, F., Leutenegger, C., Golino, D., Rowhani, A. (2007) Real-time RT-PCR (TaqMan®) assays for the detection of Grapevine Leafroll associated viruses 1–5 and 9. *Journal of virological methods* 141**:**22-29

Poojari, S., Alabi, O.J., Fofanov, V.Y., Naidu, R.A. (2013) A Leafhopper-Transmissible DNA Virus with Novel Evolutionary Lineage in the Family

Geminiviridae Implicated in Grapevine Redleaf Disease by Next-Generation Sequencing. *PLoS One* 8**:**e64194

Rowhani, A., Uyemoto, J.K., Golino, D.A. (1997) A comparison between serological and biological assays in detecting grapevine leafroll associated viruses. *Plant Disease* 81**:**799-801

Rowhani, A., Uyemoto, J.K., Golino, D.A., Martelli, G.P. (2005) Pathogen Testing and Certification of Vitis and Prunus Species*. *Annual Review of Phytopathology* 43**:**261-278

Sabanadzovic, S., Abou-Ghanem, N., Castellano, M., Digiaro, M., Martelli, G. (2000) Grapevine fleck virus-like viruses in Vitis. *Archives of virology* 145**:**553-565

Saldarelli, P., Rowhani, A., Routh, G., Minafra, A., Digiaro, M. (1998) Use of degenerate primers in a RT-PCR assay for the identification and analysis of some filamentous viruses, with special reference to clostero-and vitiviruses of the grapevine. *European journal of plant pathology* 104**:**945-950

Skiada, F., Grigoriadou, K., Maliogka, V., Katis, N., Eleftheriou, E. (2009) Elimination of Grapevine leafroll-associated virus 1 and Grapevine rupestris stem pitting-associated virus from grapevine cv. Agiorgitiko, and a micropropagation protocol for mass production of virus-free plantlets. *Journal of Plant Pathology***:**177-184

Skiada, F., Maliogka, V., Katis, N., Eleftheriou, E. (2013) Elimination of Grapevine rupestris stem pitting-associated virus (GRSPaV) from two Vitis vinifera cultivars by in vitro chemotherapy. *European journal of plant pathology* 135**:**407-414

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution* 30**:**2725-2729

Turturo, C., Rott, M.E., Minafra, A., Saldarelli, P., Jelkmann, W., Martelli, G.P. (2000) Partial molecular characterization and RT-PCR detection of grapevine leafroll associated virus 7, Proceedings of the 13th Meeting of ICVG,, Adelaide, Australia.

Varveri, C., Maliogka, V.I., Kapari-Isaia, T. (2015) Chapter One-Principles for Supplying Virus-Tested Material. *Advances in virus research* 91**:**1-32

Walter, B. (1993). Advances in grapevine virus disease diagnosis since 1990. *Presented at the Meeting of the International Council for the Study of Viruses and Virus Disease of the Grapevine (ICVG)*, Montreux, Switzerland, 6-9 September, p. 127-130

Weber, E., Golino, D., Rowhani, A. (2002) Laboratory testing for grapevine diseases. *Practical Winery and Vineyard*

Webster, C.G., Wylie, S.J., Jones, M.G. (2004) Diagnosis of plant viral pathogens. *CURRENT SCIENCE-BANGALORE-*. 86**:**1604-1607

Zhang, Y., Singh, K., Kaur, R., Qiu, W. (2011) Association of a novel DNA virus with the grapevine vein-clearing and vine decline syndrome. *Phytopathology* 101**:**1081-1090

APPENDIX A

Table 1: List of viruses, primer specificities, expected amplicon sizes and annealing temperatures.

Virus	Genus	Primer (degeneracy)	Nucleotide sequence* (5' to 3')	Amplified product (bp)	Annealing Temperature (^{o}C)	Product	Literature
GALV, PAMV/TBSV	Tombusvirus	tombusF (4) tombusR (0)	GGC AMG TTT GTC ATA TYC GG TAT CCA TGA ACT GGT CTT GTT CAA G	284	54	Partial CP gene and 3' UTR	This study
GFkV, GRGV	Maculavirus	MTR1 (32) MTR2 (1944)	TTC ATG CAY GAY GCY MTS ATG T TCC GAV GCN BHB GVR GTG ACC CA	572	50	MTR gene	(Sabanadzovic *et al.*, 2000)
GAMaV, GRVFV, GSyV-1	Marafivirus	marafiF (16) marafiR (6)	CRM STT CTG CGG STA CTA HTG AAG CAA TTC ACC YTG	436	51	Marafibox and partial CP gene	This study
GRBaV	Geminivirus	GVG f1 (0) GVG r1 (0)	CTC GTC GCA TTT GTA AGA ACT GAC AAG GCC TAC TAC G	557	50	Partial IR and CP gene	(Al Rwahnih *et al.*, 2012)
GCSV	Reovirus	Ctg468F (0) Ctg468R (0)	ACG TTG GAT CAA CTA GCC GAA G TAT TCA CGA GGC TCA GAC GAC T	368	56	Genomic segment 4	(Al Rwahnih *et al.*, 2015b)
GLRaV-4 like viruses (5,6,9)	Ampelovirus	LRAmp-F (4) LRAmp-R (4)	ATT TAG GTA ATG TWG TRG CTA C TAT CCT CAG WGA GGA ARC GG	485	46	ORF6 + 3' coding region	(Abou Ghanem-Sabanadzovic *et al.*, 2012)
GLRaV-3		LR3.HRM4.F (0) LR3.HRM4.R (0)	TAA TCG GAG GTT TAG GTT CC GTC GGT TCG TTA ACA ACA C	226	53	ORF4	(Bester *et al.*, 2012)
GFLV, GDefV, RpRSV, TRSV, ArMV	Nepovirus A	NepoA_F (24) NepoA_R (144)	ACD TCW GAR GGI TAY CC RAT DCC YAC YTG RCW IGG CA	340	45	RdRp gene	(Wei and Clover, 2008)
GARSV, GCMV, GTRSV, TBRV **GBLV, CLRV, ToRSV**	Nepovirus B Nepovirus C	NepoB_F (8) NepoB_R (12)	TCT GGI TTT GCY TTR ACR GT CTT RTC ACT VCC ATC RGT AA	250	45	RdRp gene	(Wei and Clover, 2008)
GLRaV-1	Closterovirus	HSP70-417F (0) HSP70-737R (0)	GAG CGA CTT GCG ACT TAT CGA GGT AAA CGG GTG TTC TTC AAT TCT	320	63	HSP70	(Osman *et al.*, 2007)
GLRaV-2		V2dCPf2 (0) V2CPr1 (0)	ACG GTG TGC TAT AGT GCG GCA GCT AAG TAC GAA TCT	515	63	HSP70	(Bertazzon and Angelini, 2004)
GLRaV-7	Velarivirus	LR7 G23metF (0) LR7 G23metR (0)	ATT GAC TGT GAT GTC GCT TTT AC TAC CAC TAC CAG GAG GTT TAT TCA	190	55	RdRp gene	(Turturo *et al.*, 2000)
GRSPaV, GVA,B,D,E,F	Foveavirus, Vitivirus	dRW_up1 (8) dRW_do2 (128) dRW_nest1 (12) dRW_nest2 (32)	WGC IAA RDC IGG ICA RAC RMY TCI CCI SWR AAI CKC AT GGG GCA RAC IHT IGC ITG YTT AAI GCY TCR TAR TCI GAI TCN GT	199	Nested PCR	RdRp gene	(Dovas and Katis, 2003b)
GLRaV-1, -2 **GLRaV-7**	Closterovirus Velarivirus	dHSPup1 (6) dHSPup1G (2) dHSPdo2 (8) dHSPdo2C (8) dHSPnest1 (8) dHSPnest2 (16) dHSPnest3 (64)	GGI HTI GAI TTY GGI ACI ACI TT AGT TYG GGA CGA CGT T GTI CCI CCI CCN AAR TC GTI CCI CCC CCN AAR TC TTY GGG ACG ACG TTY AGY AC TYG GGA CGA CGT TYT CAN C SCI GCI GMI SWI GGY TCR TT	500 - 536		RdRp gene	(Dovas and Katis, 2003a)

* B = T+C+G; H = A+T+C; M = A+C; N = I = A+C+G+T; R = A+G; S = C+G; V = A+G+C; Y = T+C; W = A+T

Table 2: Viral clones with their concentration (ng/µl), number molecules/µl and treatments in which they were pooled for Illumina MiSeq sequencing; first sequencing run.

Genus	Viral Clones[a]	ng/µl	Molecules/µl	A (1:1)	B (1:1)	C 1:2	D 1:2
Nepovirus B	GARSV	16.4	6.08 x 10^{10}	0.25 µl	0.25 µl	0.25 µl	0.25 µl
Ralstonia	RCS	8.27	5.32 x 10^{10}	0.286 µl			
Nepovirus C	GBLV	7.2	2.67 x 10^{10}	0.569 µl	0.569 µl	0.569 µl	0.569 µl
Nepovirus A	RpRSV	9.03	2.46 x 10^{10}	0.618 µl	0.618 µl	0.618 µl	0.618 µl
Tombusvirus	TBSV	6.9	2.25 x 10^{10}	0.675 µl	0.675 µl	0.675 µl	0.675 µl
Ampelovirus	GLRaV-5	10.5	2.01 x 10^{10}	0.756 µl	0.756 µl	0.756 µl	0.756 µl
Geminivirus	GRBaV	11.9	1.98 x 10^{10}	0.768 µl	0.768 µl	1.536 µl	0.768 µl
Ampelovirus	GLRaV-4	9.79	1.87 x 10^{10}	0.813 µl			
Verlarivirus	GLRaV-7	7.11	1.32 x 10^{10}	1.152 µl	1.152 µl	1.152 µl	1.152 µl
Closterovirus	933 GLRaV-2	6.67	1.24 x 10^{10}	1.225 µl	1.225 µl	2.45 µl	1.225 µl
Vitivirus	GTG 11-1 GVA	2.57	1.20 x 10^{10}	1.268 µl	1.268 µl	1.268 µl	2.536 µl
Closterovirus	CTV	6.12	1.13 x 10^{10}	1.345 µl			
Vitivirus	630 GVA	2.23	1.04 x 10^{10}	1.463 µl			
Vitivirus	P163-1 GVA	1.97	9.17 x 10^{9}	1.658 µl			
Closterovirus	936 GLRaV-2	4.97	8.65 x 10^{9}	1.758 µl			
Nepovirus B	TBRV	2.09	7.75 x 10^{9}	1.96 µl			
Ampelovirus	GLRaV-6	3.44	6.57 x 10^{9}	2.313 µl			
Vitivirus	GVB	1.29	6.01 x 10^{9}	2.528 µl			
Foveavirus	GRSPaV	1.22	5.68 x 10^{9}	2.675 µl	2.675 µl	2.675 µl	2.675 µl
Ampelovirus	GLRaV-3	1.28	5.25 x 10^{9}	2.895 µl	2.895 µl	2.895 µl	5.79 µl
Nepovirus A	GFLV	1.7	4.63 x 10^{9}	3.283 µl			
Final Volume (µl):				35.40	12.85	14.84	17.0
Average concentration (ng/µl)				566.45	685.61	624.95	702.38

[a] Viral clones sorted form highest number molecules/µl to lowest. 🟦 Amplicon ≥ 230bp 🟧 Amplicon ≥ 400bp

Table 3: Viral clones with their concentration (ng/µl), number molecules/µl and treatments in which they were pooled for Illumina MiSeq sequencing; second sequencing run.

Genus	Viral Clones[a]	ng/µl	Molecules/µl	E (1:1)	F (1:2)	G (1:1)	H (1:1)	I (1:2:3)	J (1:3)	K (1:4)
Ampelovirus	GLRaV-3	18	7.38×10^{10}	0.5 µl	0.5 µl	0.5 µl	0.5 µl	1.0 µl	0.5 µl	0.5 µl
Ampelovirus	14-8213	16.6	6.63×10^{10}			0.557 µl				
Ralstonia	Solanacearum	7.88	5.07×10^{10}			0.728 µl				
Marafivirus	GRVFV	21.7	4.61×10^{10}			0.801 µl	0.801 µl	2.403 µl	2.403 µl	3.204 µl
Tombusvirus	TBSV	10.13	3.3×10^{10}	1.118 µl	1.118 µl	1.118 µl	1.118 µl	2.236 µl	1.118 µl	1.118 µl
Vitivirus	630 GVA	7.06	3.29×10^{10}			1.122 µl				
Geminivirus	GRBaV	19.17	3.19×10^{10}	1.157 µl	2.313 µl	1.157 µl	1.157 µl	3.471 µl	3.471 µl	4.628 µl
Maculavirus	GFkV	18.57	3.01×10^{10}			1.226 µl	1.226 µl	3.678 µl	3.678 µl	4.904 µl
Nepovirus C	GBLV	7.67	2.88×10^{10}	1.282 µl	1.282 µl	1.282 µl	1.282 µl	2.563 µl	1.282 µl	1.282 µl
Vitivirus	GTG11-1 GVA	5.65	2.63×10^{10}	1.403 µl	1.403 µl	1.403 µl	1.403 µl	1.403 µl	1.403 µl	1.403 µl
Nepovirus B	GARSV	7.08	2.62×10^{10}	1.409 µl	1.409 µl	1.409 µl	1.409 µl	2.817 µl	1.409 µl	1.409 µl
Ampelovirus	GLRaV-4	12.2	2.33×10^{10}			1.584 µl				
Nepovirus B	TBRV	5.31	1.97×10^{10}			1.873 µl				
Ampelovirus	GLRaV-6	10.2	1.95×10^{10}			1.893 µl				
Foveavirus	GRSPaV	4.17	1.94×10^{10}	1.902 µl	1.902 µl	1.902 µl	1.902 µl	1.902 µl	1.902 µl	1.902 µl
Vitivirus	P163-1 GVA	4	1.86×10^{10}			1.984 µl				
Nepovirus A	RpRSV	6.4	1.74×10^{10}	2.121 µl	2.121 µl	2.121 µl	2.121 µl	4.241 µl	2.121 µl	2.121 µl
Vitivirus	GVB	3.68	1.71×10^{10}			2.158 µl				
Ampelovirus	GLRaV-5	7.75	1.48×10^{10}	2.493 µl	2.493 µl	2.493 µl	2.493 µl	7.479 µl	7.479 µl	9.972 µl
Closterovirus	CTV	4.34	8.04×10^{9}			4.59 µl				
Nepovirus A	GFLV	2.75	7.49×10^{9}			4.927 µl				
Velarivirus	GLRaV-7	2.94	5.45×10^{9}	6.771 µl	13.541 µl	6.771 µl	6.771 µl	20.313 µl	20.313 µl	27.084 µl
Closterovirus	933 GLRaV-2	2.67	4.95×10^{9}	7.455 µl	14.909 µl	7.455 µl	7.455 µl	22.365 µl	22.365 µl	
Closterovirus	936 GLRaV-2	2.35	4.35×10^{9}			8.483 µl				33.932 µl
Final Volume (µl):				27.611	42.991	59.537	29.638	75.871	69.444	93.459
Average concentration (ng/µl)[b]:				344 325	325 304	344 331	376 448	361 356	355 354	359 275

[a] Viral clones sorted form highest number molecules/µl to lowest. □ Amplicon < 200bp 🟦 Amplicon > 200bp 🟧 Amplicon ≥ 400bp

[b] Each pooling was completed in replicate. The average ng/µl is representative of each replicate of each pool.

Table 4: Viral clones with their concentration (ng/µl), number molecules/µl and treatments in which they were pooled for Illumina MiSeq sequencing; third sequencing run.

Genus	Viral Clones[a]	ng/µl	Molecules/µl	R (1:2:3)
Vitivirus	GTG11-1 GVA	21.2	9.87×10^{10}	0.75 µl
Nepovirus C	GBLV	25.0	9.26×10^{10}	1.60 µl
Ampelovirus	GLRaV-3	21.0	8.61×10^{10}	1.72 µl
Geminivirus	GRBaV	45.6	7.58×10^{10}	2.93 µl
Nepovirus B	GARSV	19.5	7.23×10^{10}	2.05 µl
Marafivirus	GRVFV	32.6	6.93×10^{10}	3.20 µl
Tombusvirus	TBSV	16.1	5.25×10^{10}	2.82 µl
Maculavirus	GFkV	31.5	5.10×10^{10}	4.35 µl
Ampelovirus	GLRaV-5	23.7	4.53×10^{9}	4.90 µl
Foveavirus	GRSPaV	9.16	4.26×10^{9}	1.74 µl
Velarivirus	GLRaV-7	9.75	1.81×10^{9}	12.27 µl
Nepovirus A	RpRSV	2.63	7.17×10^{9}	20.65µl
Closterovirus	933 GLRaV-2	3.77	6.99×10^{9}	31.77 µl
			Final Volume (µl):	90.75
			Average concentration (ng/µl)[b]:	646.43 633.07

[a] Viral clones sorted form highest number molecules/µl to lowest. □ Amplicon < 200bp 🟦 Amplicon > 200bp 🟧 Amplicon ≥ 400bp

[b] Each pooling was completed in replicate. The average ng/µl is representative of each replicate of each pool

Table 5: Viral clones with their concentration (ng/µl), number molecules/µl and treatments in which they were pooled for Illumina MiSeq sequencing; third sequencing run.

Genus	Viral Clones[a]	ng/µl	Molecules/µl	S (1:2:3)
Geminivirus	GRBaV	122.3	2.03×10^{11}	2.25 µl
Marafivirus	GRVFV	80.9	1.72×10^{11}	2.66 µl
Ampelovirus	GLRaV-3	38.8	1.59×10^{11}	1.92 µl
Tombusvirus	TBSV	39.0	1.27×10^{11}	2.40 µl
Maculavirus	GFkV	59.7	9.67×10^{10}	4.73 µl
Ampelovirus	GLRaV-5	41.2	7.87×10^{10}	5.82 µl
Nepovirus C	GBLV	17.7	6.56×10^{10}	4.65 µl
Foveavirus	GRSPaV	12.8	5.96×10^{10}	2.56 µl
Nepovirus B	GARSV	15.8	5.86×10^{10}	5.21 µl
Vitivirus	GTG11-1 GVA	12.5	5.82×10^{10}	2.62 µl
Closterovirus	933 GLRaV-2	12.3	2.28×10^{10}	20.07 µl
Velarivirus	GLRaV-7	11.3	2.09×10^{10}	21.90 µl
Nepovirus A	RpRSV	3.23	8.80×10^{9}	34.67 µl
			Final Volume (µl):	111.46
			Average concentration (ng/µl)[b]:	26.9 27.3

[a] Viral clones sorted form highest number molecules/µl to lowest. □ Amplicon < 200bp 🟦 Amplicon > 200bp 🟧 Amplicon ≥ 400bp

[b] Each pooling was completed in replicate. The average ng/µl is representative of each replicate of each pool

Table 6: Viral clones with their concentration (ng/µl), number molecules/µl and treatments in which they were pooled for Illumina MiSeq sequencing; third sequencing run.

Genus	Viral Clones[a]	ng/µl	Molecules/µl	T (1:3)
Ampelovirus	GLRaV-3	37.4	1.53×10^{11}	1.0 µl
Geminivirus	GRBaV	89.3	1.49×10^{11}	3.08 µl
Tombusvirus	TBSV	34.2	1.12×10^{11}	1.37 µl
Marafivirus	GRVFV	51.8	1.10×10^{11}	4.17 µl
Maculavirus	GFkV	61.4	9.94×10^{10}	4.62 µl
Nepovirus C	GBLV	20.8	7.71×10^{10}	1.98 µl
Vitivirus	GTG11-1 GVA	15.5	7.22×10^{10}	2.12 µl
Nepovirus B	GARSV	17.2	6.37×10^{10}	2.40 µl
Foveavirus	GRSPaV	12.3	5.73×10^{10}	2.67 µl
Ampelovirus	GLRaV-5	20.6	3.94×10^{10}	11.65 µl
Closterovirus	933 GLRaV-2	12.6	2.33×10^{10}	19.70 µl
Velarivirus	GLRaV-7	12.0	2.22×10^{10}	20.68 µl
Nepovirus A	RpRSV	4.21	1.15×10^{10}	13.30 µl
			Final Volume (µl):	88.74
			Average concentration (ng/µl)[b]:	31.8 31.7

[a] Viral clones sorted form highest number molecules/µl to lowest. □ Amplicon <400bp ▨ Amplicon ≥ 400bp

[b] Each pooling was completed in replicate. The average ng/µl is representative of each replicate of each pool

Table 7: Viral clones with their concentration (ng/µl), number molecules/µl and treatments in which they were pooled for Illumina MiSeq sequencing; second and third sequencing runs.

Pool	Viral Clones[a]	ng/µl	Molecules/µl	µl Pooled (1:1)	Total Volume (µl)	Average ng/µl[b]
L	GBLV	15.73	5.83×10^{10}	8.0	50.08	404.87
	GARSV	14.0	5.19×10^{10}	8.96		
	GTG 11-1 GVA	7.12	3.31×10^{10}	14.08		414.1
	GRSPaV	5.27	2.45×10^{10}	19.04		
M	GARSV	17.03	6.31×10^{10}	8.0	53.12	403.8
	GBLV	16.4	6.08×10^{10}	8.32		
	GTG 11-1 GVA	6.87	3.2×10^{10}	15.76		416.26
	GRSPaV	5.15	2.4×10^{10}	21.04		
N	GTG 11-1 GVA	18.8	8.75×10^{10}	9.0	47.71	25.2
	GRSPaV	14.7	6.84×10^{10}	11.51		
	GARSV	16.3	6.04×10^{10}	13.04		25.2
	GBLV	15.0	5.56×10^{9}	14.16		
O	GTG 11-1 GVA	26.8	1.25×10^{11}	7.0	51.96	23.2
	GARSV	18.6	6.89×10^{10}	12.70		
	GBLV	15.1	5.60×10^{10}	15.63		23.2
	GRSPaV	11.3	5.26×10^{10}	16.63		
P	GRSPaV	15.0	6.98×10^{10}	10.0	42.18	29.4
	GTG 11-1 GVA	14.9	6.94×10^{10}	10.06		
	GARSV	17.7	6.56×10^{10}	10.64		27.5
	GBLV	16.4	6.08×10^{10}	11.48		
Q	GRSPaV	14.9	6.94×10^{10}	10.0	42.52	18.3
	GARSV	18.3	6.78×10^{10}	10.24		
	GTG 11-1 GVA	14.4	6.70×10^{10}	10.36		18.8
	GBLV	15.7	5.82×10^{10}	11.92		

[a] Viral clones sorted form highest number molecules/µl to lowest for each pool. Viruses included in these pools belong to the Nepovirus B (GARSV), Nepovirus C (GBLV), Vitivirus (GTG 11-1 GVA) and Foveavirus (GRSPaV) genera.

[b] Each pooling was completed in replicate. The average ng/µl is representative of each replicate of each pool.