Unbiased, next-generation sequencing for the characterization of *Citrus tristeza virus* populations

By

Olivier Zablocki

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae Microbiology (MSc.)

In the Faculty of Natural & Agricultural Sciences

University of Pretoria

Pretoria

Submitted February 2013

DECLARATION

I, Olivier Zablocki 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.

Olivier Zablocki

Signature: ______________________

Date: ______________________

ACKNOWLEDGMENTS

This research project would not have been possible without the support of many individuals and entities. I wish to express my sincere gratitude to the following:

My family, which although from a great distance, have never ceased to give me their love and encouragements.

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.

Citrus Research International (CRI) for their financial support throughout this project.

My supervisor, Professor Gerhard Pietersen, for his continuous support, advices and opportunities.

My lab colleagues, David Read, Helen Walsh, Ronel Viljoen and Katherine Scott for their advices and friendship.

Professor Fourie Joubert and Dr. Jasper Rees for their valuable inputs and suggestions.

Dr. Stephanus van Vuuren, Kobus Breytenbach and Glennis Cook for their help and support in collecting samples and bud-grafting experiments.

Graham Pienaar for allowing me to sample his farm.

Peet Wolmarans, for always being there for me

- *This work is dedicated to my parents, Nadine and Roman.*

SUMMARY

A high-throughput sequencing pipeline to characterize *Citrus tristeza virus* isolates was developed. Three alternative viral templates (total RNA, double-stranded RNA and virus particles) were first tested on a single, previously characterized GFMS12 sub-isolate for their enrichment qualities, and combined with random RT-PCR amplification were subjected to Illumina paired-end sequencing. Double-stranded RNA was found to be most useful and was selected for further characterization of additional isolates (glasshouse-kept and field-derived). A novel South African genotype, named CT-ZA3 was assembled *de novo* and shown to be the dominant component in all GFMS12 sub-isolates tested. Genotype distributions within field-derived isolates collected from commercial orange (*Citrus sinensis*) orchards revealed a mixed infection status, dominated by a resistance breaking (RB)-like component (Tai-SP) coupled with a minor, VT-like (mild) (Kpg3) component. Based on read mapping patterns from field isolates, it is further suggested that two previously unknown recombinants may be present: a SP/Kpg3 and HA16-5/Kpg3 combination. This study underlined the effectiveness of next-generation sequencing for genotype discovery as well as whole-genome characterization of CTV isolates to a level of detail previously unreachable with classical methods such as SSCP and Sanger sequencing of multiple clones.

LIST OF ABBREVIATIONS

AMV	**Avian Myeloblastosis virus**
ARC	**Agricultural Research Council**
A-T	**Adenosine- Tyrosine**
BCA	**Brown citrus aphid**
BLAST	**Basic local alignment search tool**
bp	**base pair**
BSA	**Bovine serum albumin**
cDNA	**complementary DNA**
CP	**major coat protein**
CPm	**minor coat protein**
CTV	**Citrus tristeza virus**
DNA	**Deoxyribonucleic acid**
dNTP	**Deoxyribonucleotide triphosphate**
dsRNA- RI	**Double-stranded RNA replicative intermediate**
EDTA	**Ethylenediaminetetraacetic acid**
ELISA	**Enzyme-linked immunosorbent assay**
G	**g-force**
GFMS12	**Grapefruit mild strain 12**
HSP	**Heat shock protein**
IDT	**Integrated DNA technologies**
Kb	**kilo base**
kDa	**kilo Dalton**
MEGA	**Molecular Evolutionary Genetics Analysis**

ml	**millilitre**
mM	**millimolar**
mRNA	**messenger ribonucleic acid**
NCBI	**National Center for Biotechnology Information**
nm	**nanometer**
nt	**nucleotide**
ORF	**Open reading frame**
PBS	**Phosphate buffered saline**
PCR	**Polymerase chain reaction**
pH	**potential of hydrogen**
RDOD comp	**Complementary random dodecamer**
RDOD	**random dodecamer**
RdRp	**RNA-dependent RNA polymerase**
RNA	**Ribonucleic acid**
rRNA	**Ribosomal ribonucleic acid**
RT-PCR	**Reverse-transcription polymerase chain reaction**
SDS	**Sodium dodecyl sulphate**
siRNA	**Short interfering RNA**
sgRNA	**Sub-genomic RNA**
SNP	**Single nucleotide polymorphisms**
SSCP	**Single strand conformation polymorphisms**
STE	**Sodium-Tris-EDTA**
TRIS	**Tris- (hydroxymethyl) aminomethane**
U	**Units of enzyme**
µl	**microliter**

UTR	**Untranslated region**
vol	**volume**
WGA	**Whole-genome amplification**
ZMW	**Zero-mode waveguide**

LIST OF FIGURES

FIGURE NUMBER **PAGE**

LIST OF TABLES

TABLE NUMBER **PAGE**

TABLE OF CONTENTS

CHAPTER 1:

LITERATURE REVIEW: CITRUS TRISTEZA VIRUS EPIDEMIOLOGY AND ITS CURRENT ANALYTICAL TOOLS

1.1 SUMMARY

One of the most serious diseases of citrus worldwide is caused by viruses in the species *Citrus tristeza virus* (CTV; family *Closteroviridae*, genus *Closterovrius*). Infection with this RNA virus can cause several economically important syndromes, such as “quick decline”, “stem pitting” and “seedling yellows”. Many factors can contribute to the severity of infection, including climatic conditions, host/scion combination, as well as which CTV genotype predominates within a mixed infection. Indeed, many CTV genotypes exist, each with varying degree of virulence, and as of 2012, twenty-eight genomes have been fully sequenced. It has been observed that typically, a host is infected with a mixture of strains, which can complicate biological indexing and characterization studies.

This pathogen is endemic to South Africa, and in efforts to control the disease, a cross-protection scheme has been put into place. This procedure entails creating CTV free plants via meristem tip culture and pre-immunization plants derived from these of commercially grown citrus cultivars infected with a mild strain mixture of the virus, which by virtue of the superinfection exclusion principle prevents infection from more virulent strains. However, no system being perfect, severe infections may still occur on pre-immunized hosts. One hypothesis to explain the phenomenon has been proposed by Folimonova *et al.,* (2010), where they observed cross-protection occurring only against specific genotypes of which a mild form was present in the pre-immunizing mixture. Therefore, in order to improve future cross-protecting sources, accurate genotypes of CTV occurring in a given location must be obtained.

Determining which genotypes are present within a host can be a challenging task. Many techniques have been developed to not only detect the virus, but to also identifying individual genotypes. The most commonly used methods, as well as their uses, advantages and limitations will be discussed in this chapter. However, characterization of isolates based on selected genomic regions has been shown to produce incongruences in phylogenetic studies (Scott *et al.,* 2012). To tackle this issue, a high-throughput sequencing approach was suggested, which, combined with alternative reverse-transcription methods, would yield full genome sequences to compare. The advent of “next-generation” sequencing, its uses and associated challenges are discussed towards the end of this chapter.

1.2 INTRODUCTION

Viruses of the species *Citrus tristeza virus* (CTV) have been recognized as one of the most devastating diseases of citrus (Bar-Joseph *et al.,* 1989). This aphid-borne Closterovirus can induce severe syndromes on its host, such as “quick decline” and “stem pitting”. This translates into massive economic losses for the producers, which have to live with the disease due to its endemic status in many parts of the world, including South Africa. The most promising control measure used is cross-protection, but unfortunately severe symptoms may still occur despite pre-immunization with a mild strain (van Vuuren *et al.,* 1993).

Over the last decades, the development of many molecular tools has enabled the detection and characterization of this complex virus, which, besides its quasi-species status typically occurs as a mixture of strains. Due to high mutation and recombination rates resulting in incongruent phylogenies (Scott *et al.,* 2012) when several gene regions are sequenced, characterization of CTV isolates requires whole-genome comparisons. This goal is now within reach as a new generation of sequencing technologies has become available.

1.3 TRISTEZA DISEASE AND CONTROL MEASURES

1.3.1 ETIOLOGICAL AGENT & MOLECULAR BIOLOGY ASPECTS

Tristeza disease is caused by a virus called *Citrus tristeza virus* (CTV). Its host range encompasses members of the *Citrus* genus (family *Rutaceae*; subfamily *Aurantoidea*) (Bar-Joseph *et al.*, 1979). CTV belongs to the *Closteroviridae* family, genus *Closterovirus* and has been identified as the largest plant RNA virus. Its genome is monopartite, composed of a positive sense single-stranded RNA molecule approximately 19.3 kilobases (kb) in length. Virions occur as long (2000 nm by 12 nm) flexuous threads encapsidated by two coat proteins (Karasev *et al.*, 1995). The major coat protein (CP) covers 97% of the virion’s surface, while the remaining 3% is encapsidated by the minor coat protein (CP_m). The genome is organized into 12 open reading frames (ORFs), potentially encoding 19 protein products (Karasev *et al.*, 1995) (Figure 1). Short untranslated regions (UTRs) are located at both the 5' and 3' ends of the genome, with no 5' methylated cap and poly (A) tail. With a tropism for phloem cells, CTV achieves successful infection and replication by relying on several strategies, such as the formation of double-stranded RNA intermediates, polyprotein cleavage, translational frameshifting and formation of subgenomic RNAs (sgRNAs) (figure 2).

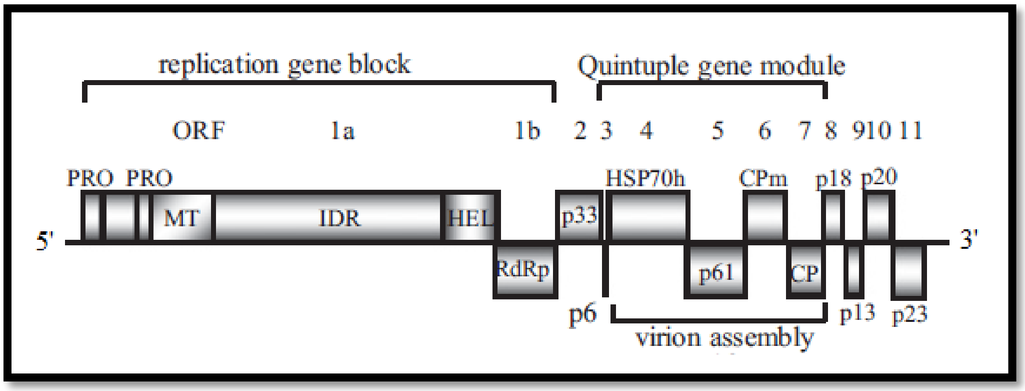


Figure 1: Genomic organization of Citrus Tristeza virus (Satyanarayana *et al.*, 2011)

The genome can be divided into 2 parts, namely, the replication gene block (often referred to as the 5' half) and the quintuple gene module, encompassing most of the other ORFs (also referred to as the 3' half). The 5'half comprises genes that mainly encode proteins and enzymes for replication purposes. Upon entering a compatible phloem cell, the virus starts uncoating, thereby releasing a naked ssRNA molecule, capable of being directly translated, due to its positive polarity. Host cell's ribosomes are redirected to translate ORF1 a and b, a large 400-kDa polyprotein which subsequently undergoes proteolytic cleavage resulting into nine protein products. These products include two papain-like proteases, a type I methyltransferase and a helicase. Most importantly, the 1b portion encodes a 54 kDa RNA-dependent-RNA polymerase (RdRp) which is translated by means of a +1 ribosomal frameshift (Karasev *et al.,* 1995). Upon RNA polymerase production, the genome can start replicating by acting as a template for the polymerase, by the synthesis of a complementary minus (-) strand, thus forming a double-stranded RNA replicative intermediate (dsRNA- RI). The replicative intermediate serves as a template for replication (Figure 2). The 3' half of the genome encodes the remainder of the ORFs (2-11), expressed as a collection of ~30 sub-genomics RNAs (sgRNAs), allowing both expression and regulation of genes, including controller elements, capsid and HSP70 proteins (Gowda *et al.*, 2003).

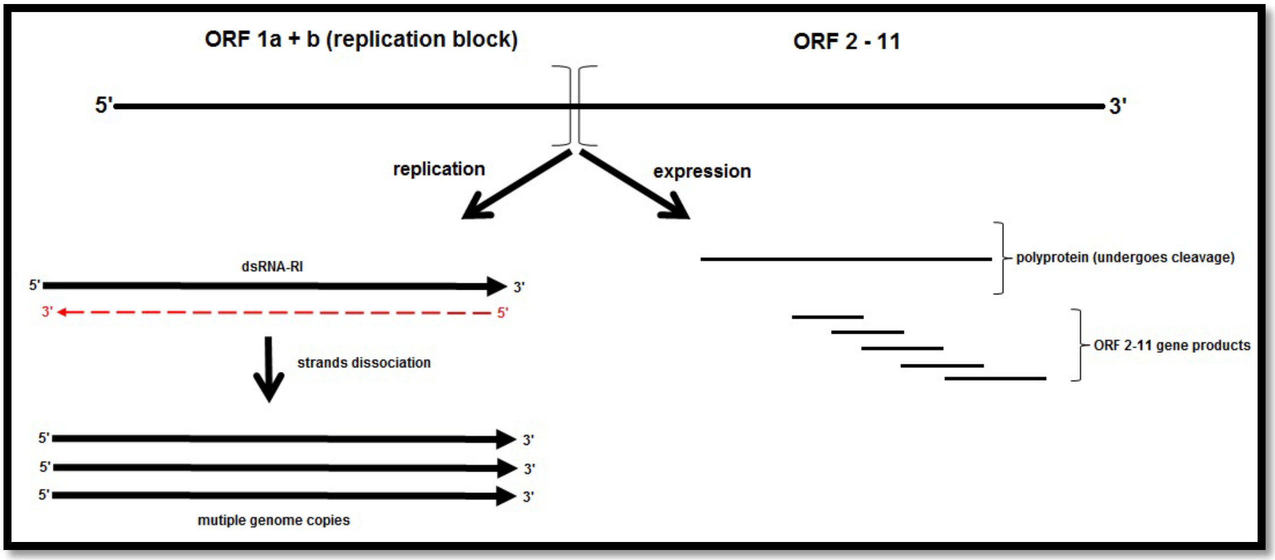


Figure 2: Generalized expression and replication strategies of *Citrus tristeza virus*.

1.3.2 TRANSMISSION, STRAIN DIVERSITY, QUASI-SPECIES & EVOLUTION

The primary mode of CTV transmission is by insect vectors. Specifically, aphids (Order *Homoptera*; Family *Aphididae*) seemed to have co-evolved selectivity for CTV particles (Karasev, 2000). Several aphid species are able to transmit CTV, such as *Toxoptera citricida*, *Aphis gossypi*, *Aphis spiraecola* and *Toxoptera aurantii*. Within these species, it has been determined that *T. citricida* was the most efficient at transmitting CTV, in a semi-persistent manner (Marroquin *et al.*, 2004). Aphids feed on the phloem tissues of citrus species and in doing so pierce and suck on plant sap. In this way, a viruliferous aphid physically injects its viral passenger directly into sieve tube cells. Conversely, a CTV-free aphid may acquire the virus and enhance its spread. CTV transmission can also occur by grafting virus-infected material onto healthy susceptible citrus material (rootstocks or scions).

In its host, CTV exists as a mixture of strains (Moreno *et al.*, 1991). Moreover, it is well documented (Rubio *et al.*, 2000; van Vuuren *et al.*, 2000) that the strain composition, in terms of ratio and dominance of a particular strain over others, can be influenced by factors such as the rootstock/scion combination and environmental conditions. Many distinct genotypes have been fully sequenced and are available from the National Institute for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). These include: VT (EU937519) from Israel (Mawassi *et al.,* 1996), T30 (AF260651) from the USA (Albiach-Marti *et al.,* 2000), B165 (EU076703) from India (Roy and Brlansky, 2010), CTV-Mexico (Quiroz et al., unpublished), Nuaga (AB046398) from Japan (Suastika et al., 2001), Quaha (AY340974) from Egypt (Abdelmaksoud, unpublished) and the resistance breaking (RB) group (FJ525431-35) from New Zealand (Harper *et al.,* 2010). Along with others, these isolates form the basis of current phylogenetic analyses, being the most important strains recognized today (Roy *et al*, 2009). Additional CTV whole genomes are constantly being sequenced and the number has grown to 28 currently, from all over the world. From whole-genome sequence comparisons, isolates have been associated into six genotypes (Melzer *et al.*, 2010), namely VT, B165, HA16-5, T30, T36 and RB (Figure 3).

Sequence divergence amongst CTV genotypes has been shown to be region-specific. In the 3'half, sequence similarity amongst genotypes was >90%. However, the 5'half shows high dissimilarity, with the 5'UTR being the most divergent region of the genome. Without the 5'UTR the 5' half has 70% similarity amongst genotypes (Ayllon *et al.,* 2001). Sequence variation amongst different CTV genotypes is a very prominent feature of this virus. The main reason for the variability observed amongst CTV isolates derives from the quasi-species nature of RNA viruses. A good definition of the concept has been put forward by Domingo *et al.,* (1998): "closely related (but not identical) mutant and recombinant genomes subjected to continuous genetic variation, competition and selection". Still, for the continuous genetic variation to operate, very frequent mutation events must occur. Spontaneous mutations are rather rare events, and they would not be sufficient enough to account for the rapid mutation rates that are observed. However, RNA viruses are notoriously more prone to mutations when compared to double-stranded RNA viruses or DNA viruses. There are two major mechanisms that allow rapid generation of sequence variants. The first source of mutations comes from the virus-encoded RNA-dependent RNA polymerase (RdRp) or RNA replicase. The RNA replicase, used to make copies of the viral genome, has little or no ability to proofread newly synthesized RNA strands, which leads to an accumulation of genomes differing from the original “master” sequence (Levy *et al.*, 1994). A second source of sequence diversity is by reassortment, where multiple strains of a virus with multiple components co-infect the same cell (superinfection). Recombination, which in the case of CTV is now known to be common occurrence (Weng *et al.,* 2007), has been shown to occur following the isolation of defective RNAs and sequence analysis that revealed the features of a chimeric genome (Mawassi *et al.*, 1995). In a recombination study, Hilf *et al.,* (1999) hypothesized that the severe Florida T36 strain could have been the result of a recombination event between the ORF1a of an unknown isolate and the 3’half of a “normal” CTV isolate.

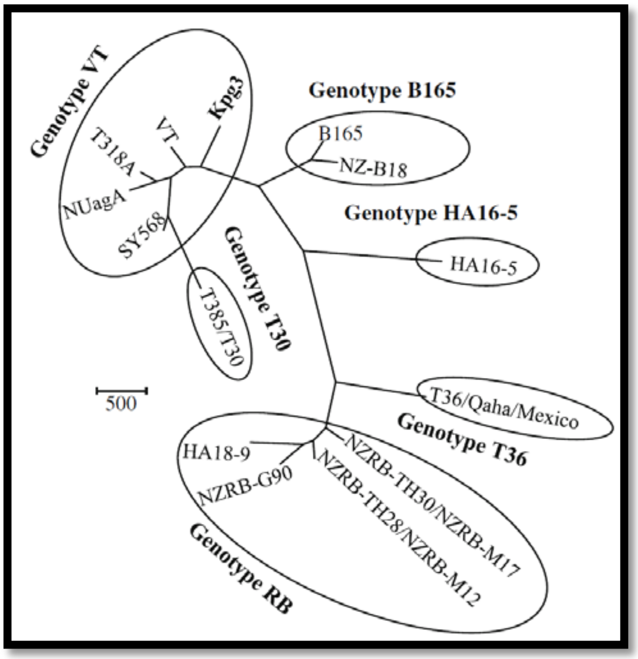


Figure 3: Rooted dendrogram depicting the genetic divergence and the 6 genogroups amongst CTV reference genomes (Biswas *et al.*, 2012).

1.3.3 SYMPTOMS, DISEASE CONTROL AND CROSS-PROTECTION BREAKDOWN

CTV-induced symptoms are very variable, from very mild with regards to general tree health and fruit yield, to complete tree death. In terms of economic loss, the three most important syndromes of CTV infection observed in the field are quick decline (**QD**), stem pitting (**SP**) and seedling yellows (**SY**) (Figure 4). Quick decline (also called *Tristeza* disease) is the worst syndrome, resulting in tree death when grafted on sour orange rootstocks. In this syndrome, an excessive amount of non-functional phloem cells induce a loss of root mass, stunting, reduced fruit size, poor growth, dieback, wilting, which culminates in tree death (Shneider, 1957). The second most important syndrome, SP, may arise on many citrus crops such as lime, grapefruit, and sweet orange when grafted onto *any* rootstock. Identified visually as pits in the wood, it does not induce tree death, but because it still affects general tree health, such as growth stunting and very low yield of fruits, it makes this syndrome important in terms of economic losses.

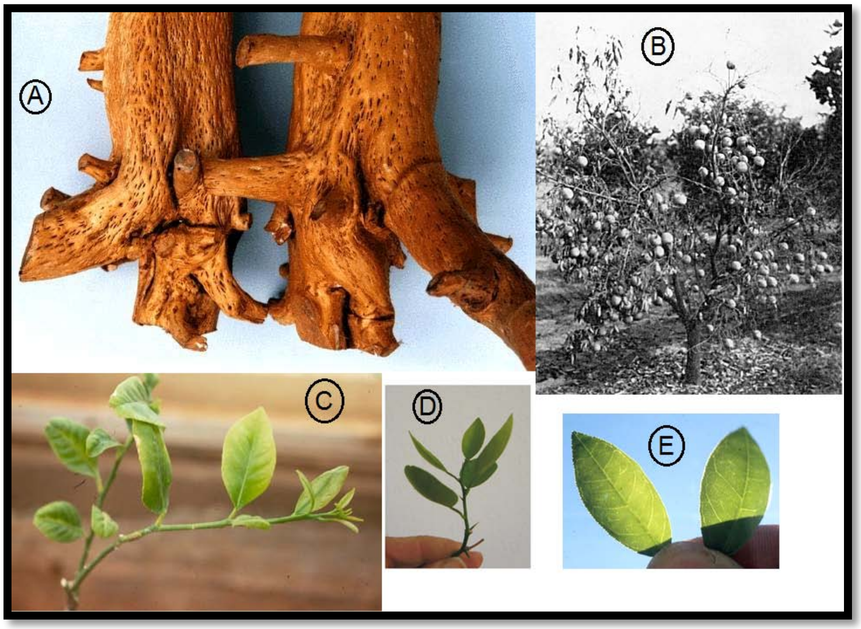


Figure 4. The three main syndromes caused by CTV infections. (a) Stem pitting; (b) Quick decline and (c) Seedling yellows; (d) leaf cupping and (e) vein clearing. Photos: USDA website.

The last syndrome, termed seedling yellows, occurs mainly under glasshouse conditions but is occasionally observed in the field. Its effects include stunting and leaf chlorosis (Fraser, 1952). In addition to the three main syndromes, two symptoms have been observed exclusively under glasshouse conditions, vein clearing and leaf cupping.

Control of plant viruses is a continuous process, and may require multiple methods that act in concert for effective protection of major food crops. Generally, these include (Hull, 2002): 1) the production of virus-free propagation material, 2) controlling the insect vector, 3) conventional breeding of virus-resistant varieties, 4) generation of virus-resistant transgenic lines, 5) the examination of resistance-inducing factors, 6) removal of infected material and 7) cross-protection. Among these, the most effective control method for CTV is cross-protection, which has been developed more than eighty years ago (McKinney, 1929). This process has been defined as: “a phenomenon in which infection with mild or attenuated strains protects plants against subsequent (or “challenge”) inoculation or infection with more severe strains of the same virus” (Ziebell, 2008). One of the earliest and most successful application of cross-protection on commercial crops was in Brazil, for CTV control (Grant and Costa, 1951) and since then has been applied throughout the world on several other important crops for a great diversity of plant viruses (Ziebell, 2010). `

Initially, control measures for CTV were not uniformly widespread. At first, removal of severely infected trees and propagation of sour orange as a tristeza-resistant rootstock (Bar-Joseph *et al.,* 1974) were used. However, these practices were deemed insufficient, as many symptomless infections were not detected and severe infections still occurred on other scions/rootstock combinations. This led to a wider adoption of cross-protection as the most efficient mean to control severe CTV isolates (Costa and Muller, 1980).

The molecular mechanism(s) underlying the concept of cross-protection is/are still elusive. There have been many theories to explain the phenomenon, reviewed by Ziebell and Carr (2010). Amongst them, Palukaitis and Zaitlin (1984) have suggested a RNA homology-mediated mechanism, in which excessive presence of mild strain single-stranded RNA genomes would hybridize in a homologous manner with the negative-sense part of the challenge virus, thereby inhibiting its replication. This view prompted work that eventually led to the discovery of RNA silencing in plants (Hamilton and Baulcombe, 1999), which is currently thought to be the main drive behind cross-protection.

Although efficient, cross-protection is not fully capable at preventing the onset of severe symptoms, and in some instances breakdown occurs. The reasons for this are not fully understood, but several factors have been identified as playing a role in the process (Bar-Joseph, 1978; Wang *et al.,* 1987, Powell *et al.,* 2003) including; 1) introduction of new virus vectors, 2) physiological changes in the host over time, 3) mutations of mild strains into severe ones, 4) climatic conditions, 5) natural infection of severe strains, 6) insufficient spread of the mild strain throughout the host tissues and 7) infection with a severe strain too soon after pre-immunization for protection to occur. It has been observed that successful protection is hugely host-dependent (Pelosi *et al.,* 2000). More recently, it was also demonstrated that cross-protection only protects between isolates of the same strain and not between isolates of different strains (Folimonova *et al.,* 2010). Is therefore crucial to continuously search for potential mild strains, and be able to combine as many as possible in future pre-inoculation mixtures.

In South Africa, CTV is a widespread, endemic pathogen mainly due to the presence of *T. citricida* which introduces natural, potentially severe CTV strains (Schwarz, 1965). Cross-protection is the most effective way to prevent CTV-related losses, without which a lucrative citrus industry would be impossible (van Vuuren *et al.,* 1993). In 1973, the South African Citrus Improvement Scheme (SACIS) was initiated (von Broembsen and Lee, 1988), which entailed the selection of high quality cultivars, the production of virus-free budwood by shoot-tip grafting and the pre-inoculation of mild CTV strains. Currently, GFMS35 (grapefruit mild strain 35) and LMS6 (lime mild strain 6) are used to pre-immunize grapefruits, oranges, mandarins and other relatives. In 2000, van Vuuren *et al.* determined by means of single aphid transmissions (SATs) the strain composition of LMS6 and a predecessor to GFMS 35, known as GFMS 12). Aphid-transmitted sub-isolates were evaluated on Mexican Lime and Marsh Grapefruit, and their results showed the separation of mild and more severe isolates as compared to the original isolate. They characterized these individual isolates by SSCP based on the coat protein gene and found that for GFMS12, different SSCP profiles were found. In the case of LMS6 sub-isolates, additional bands were also observed that were not present in the original source. These results demonstrated the mixed strain nature of these two cross-protecting sources.

As an alternative to cross-protection, transgenic resistance to CTV has also been investigated. Gutierrez *et al.,* (1997) was the first group to successfully transform foreign genes in sour orange (*C. aurantium* L.) and Key lime (*C. aurantifolia* (Christm.) Swing.). This was accomplished via *Agrobacterium*-mediated transformation of the coat protein (CP) gene of CTV. However, the transformation efficiency was very low (0.12%) and pathogen-derived resistance against challenge strains was not assessed. Later, this same process was optimized by Dominguez *et al.,* (2000) which produced higher transformation efficiencies (3.5%) and recorded variable levels of gene expression amongst transgenic plants. Following the optimization of the transformation process, the same group introduced the *p25* CP gene of CTV isolate T-305 and T-317 in an attempt to confer resistance in Mexican Lime, the model system for CTV symptom expression. Transgenic lines were inoculated with isolate T300 and were examined for symptom expression. The majority of plants showed symptoms, although with delay as compared to non-transgenic controls. The remaining transgenic lines (10-33%) were shown to protect against the challenge virus. In another study by Fagoaga *et al.,* (2005), a *p23* construct was used as a potential candidate for immunity. They observed high levels of *p23*-specific siRNAs with concomitant low levels of *p23* mRNA suggesting that these are signs of post-transcriptional gene silencing (PTGS). Once challenged, transgenic lines responded similarly to the study of Dominguez *et al.,* (2000). A low percentage of plants were immune, others showed attenuated symptoms and the remainder were completely susceptible, suggesting that other factors besides the transgenic constructs are at work in this kind of resistance. Another highly susceptible *Citrus* member, grapefruit (*Citrus paradisi*), was also transformed with CTV-derived sequences (Febres *et al.*, 2003; Febres *et al.*, 2008). Many parts of several CTV isolates were transformed separately in grapefruit, which included: the CP gene from stem-pitting isolate B249 and T30, NTCP (non-translatable region of CP) from T36, the CP$_m$ from T36, the 3'end from a stem-pitting isolate in Florida (DPI38000) and the RdRp gene from T36. Upon challenge with a severe CTV isolate (T66-E) results indicated that the majority of transgenic lines were still susceptible, some were slightly resistant and only one line, expressing the 3'end of CTV, was resistant. Although the transgenic route for CTV control appears promising, the low efficiency of the process and the mechanism by which it operates still being unclear, prevent this method from becoming popular. Additionally, public acceptance as well as food regulations in some countries further halts this technology from becoming mainstream.

1.4 CLASSICAL METHODS FOR THE ANALYSIS OF RNA PLANT VIRUSES

1.4.1 INTRODUCTION

This section describes the most commonly used techniques to study RNA plant viruses, along with their strengths and weaknesses. They are not, of course, only used for the analysis of viruses, but due to their versatility, have become common in a vast array of different fields of biology. A clear trend that has emerged over the last two decades is a movement towards the analysis of organisms (viruses included) at the nucleotide sequence level, mainly due to the advent of sequencing technologies, such as the Sanger and Maxam-Gilbert method. Indeed, the very first gene and genome which was published belonged that of a RNA virus, bacteriophage MS2, which was completed in Belgium by Fiers *et al*., in 1976 and thus marked the humble beginnings of the genomics era. Another paradigm shift occurred with the development of the polymerase chain reaction (PCR) in 1983 by Kary Mullis, which was later modified into many forms, including reverse transcription-PCR. Twenty years later, a new generation of genomics tools emerged: affordable, high-throughput sequencing**.** This technology is already revolutionizing a broad array of fields, such as diagnostics, epidemiology and infection control (Studholme *et al.*, 2011). The analysis of life forms at the nucleotide level can be considered the ultimate comparative tool as there are no known smaller units of life-defining boundaries than at sequence level. Moreover, as computer power and sequencing technologies continue to evolve, research focus is rapidly shifting from the analysis of single genes to whole genomes. However, one of the greatest challenges biologists now have to face is the ability to transform masses of sequence data into information of biological significance.

1.4.2 ELISA

One of the gold standards for screening large numbers of plants remains the enzyme-linked immunosorbent assay (ELISA) (Clark *et al.*, 1977). This serological method relies on the interaction between antibodies raised against a particular virus or part of the virus (e.g. coat protein) and the homologous virus in plants. By coupling an enzyme to the antibody the binding process can be monitored by adding substrate to produce a color reaction. Many variants of ELISA exist, but for CTV detection, the triple antibody sandwich method is mostly used (Bar-Joseph *et al.*, 1979). The multiple advantages of ELISA are reflective of its wide use and acceptance. Firstly, it is easy to perform and inexpensive in terms of reagents and apparatus. Secondly, a large amount of samples can be tested at the same time, further reducing time constraints and costs. Thirdly, ELISA is sensitive and can be very specific, due to the mechanism of detection itself, but also because monoclonal antibodies are available (Vela *et al.*, 1986), permitting the binding of certain CTV strains only. However, development of other techniques are increasingly making ELISA less useful for diagnostics and detection, especially in cases where viral titers are low and accurate genotype composition is required. Although relatively sensitive, it has been demonstrated by Mathews *et al.*, (1997) that at very low titer periods of the virus, ELISA could not achieve enough resolution in terms of absorbance readings. In contrast, when RT-PCR was used, CTV could be detected. In addition, for the specificity of ELISA to operate, adequate anti-CTV IgGs must be produced, which requires prior knowledge and purification of the pathogen. For plant viruses such as CTV, which encompasses several strains, ELISA might not be able to detect all known CTV variants, yet alone uncharacterized ones. Overall, ELISA remains a powerful tool for pre-screening large amounts of plant material. If discrepancies occur, other methods may be used to assess not only the infection status, but also which genotype is present within a host plant.

1.4.3 DsRNA PROFILING

A technique that allows not only detection, but also genotype identification is based on the analysis of viral dsRNA banding patterns. Originally developed by Morris and Dodds (1979), it enables the purification of viral dsRNA from plant and fungal tissues and has been optimized over time for different viruses and hosts (Moreno *et al.*, 1990; Balijja *et al.*, 2008). Its use is based on the replication strategy of plant RNA viruses. Virtually all ssRNA(+) viruses produce at some point, in their replication cycle, a replicative intermediate (RI) of dsRNA. The method usually employs phenol-chloroform for total nucleic acid extraction, from which dsRNA is purified from using specific ethanol concentrations (~15-16%) in the presence of cellulose. Purified dsRNA can be visualized either on agarose or polyacrylamide gels (figure 5), where separation of the molecules produces a specific banding pattern. This banding pattern can be specific to screen a particular virus, and in the case of CTV, can serve to differentiate between different strains of the same virus (Moreno *et al.*, 1991).

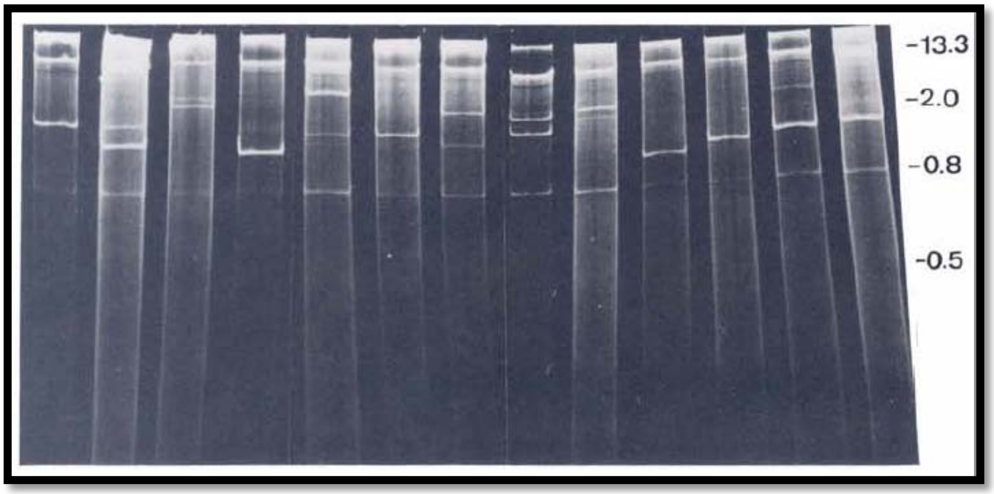


Figure 5: DsRNA banding patterns of CTV. Each lane in the polyacrylamide gel represents a different variant of the virus, identified as a distinct banding pattern of dsRNAs (Moreno *et al.*, 1991)

This method has several advantages and disadvantage and is mostly used for general screening purposes. It can provide a means to characterize co-infecting CTV strains in a single host without relying on sequence data or prior knowledge of the pathogen and so can potentially identify novel strains if an unusual banding pattern is observed. However, a major drawback of this technique lies in what it selects for dsRNA, or just the RI of ssRNA viruses. Using only this technique, the full spectrum of CTV strain diversity cannot be assessed. This is partly due to host selection, which can dictate which strains will or will not replicate or at least influence rates of replication amongst isolates. Additionally, even if several strains replicate at equal rates, how would one interpret and differentiate banding patterns? It would be easy to misinterpret a pattern as belonging to a single strain, but it could also be the result of two strains whose banding patterns have been mixed. In this case, relying on band visualization alone would not be accurate enough in characterization studies, where it is of prime importance to know exactly which strains are present. Additionally, by relying on dsRNA alone, other forms of the virus are missed, such as encapsidated forms and naked single-stranded RNAs, which could represent strains that are not replicating but could potentially do so in an alternative host. Therefore, as it was the case of ELISA, additional methods are required for a more accurate discrimination between strains. For the purpose of selecting only virus-specific nucleic acids as part of an enrichment strategy it works well.

1.4.4 SSCP

First introduced by Orita *et al.,* in 1989, single-strand conformation polymorphisms (SSCP) allows for the detection of minor changes in DNA strands without the need of sequencing. It is based on the ability of single-stranded nucleic acids to separate when subjected to electrophoresis due to differences in secondary structures that result from minor alterations in their nucleotide sequences. Typically, an asymmetric PCR is first performed on the gene of interest. This generates a very high concentration of single-stranded DNA fragments, which are subsequently run through a non-denaturing polyacrylamide gel. The differences in band migration patterns are interpreted as the level of nucleotide polymorphisms between identical gene regions.

SSCP has been extensively used for the analysis of CTV sequence variant present as a mixed population in several citrus hosts (Kong *et al.*, 1999; Roy *et al.*, 2009; van Vuuren *et al.*, 2000) and has been the cornerstone for determining the mixed nature of CTV infections. A big advantage of SSCP is its ability to observe quasi-species. Rubio *et al.,* (2000) demonstrated this by performing this technique on dsRNA extracts, observing effects of host selection on identical isolates as well as their relative dominance by using RT-PCR on the coat protein gene. They recorded different SSCP profiles before and after grafting from citron to sweet orange for example, which clearly indicated a shift in strain dominance from host to host. In addition, SSCP is an ideal tool if sequencing data does not need to be obtained.

1.4.5 RT-PCR AND SEQUENCING OF CLONES

One of the most popular approaches to detect and/or assess the CTV genetic diversity is by reverse-transcription, followed by PCR, and cloning of the amplicons thereafter (Mehta *et al.*, 1997). Generally, the cloning step is omitted for screening-only purposes. For both methods, this system first synthesizes cDNA from dsRNA (Melzer *et al.*, 2010) as well as virus purification through immunocapture (Hilf and Garnsey, 2000). Following this step, cDNA is mixed with a set of gene-specific primers and PCR amplification of cDNA products is performed. The resulting amplicons represents a single amplified gene but which potentially belongs to different strains.

In 2000, sequence-specific primers for a PCR-based identification of CTV isolates were developed, termed "Hilf markers", in reference to its author (Hilf and Garnsey, 2000). These markers (primers), derived from several unique regions of VT, T3, T30 and T36, were applied on uncharacterized CTV sources and formed for each isolate a marker profile, also called an isolate "genotype". This created an efficient method for rapid identification of a given CTV isolate. At the time, this differentiation method was the best around, and was proposed as an alternative to full genome comparisons (which the author claimed were most optimal), but was judged too difficult and impractical as a routine characterization tool for multiple isolates.

To simplify routine screenings, Roy *et al.,* (2010) have developed a multiplex RT-PCR assay which could detect five genotypes: T3, T30, T36, VT and B165 based on their ORF1a gene region. Genotype-specific primers were designed for each of the afore-mentioned genotype, and were used to characterize a global CTV isolate collection. However, personally I believe this system is flawed, because it assumes the presence of a whole genotype-specific genome based on a single gene analysis, which is not not sufficiently representative. As Hilf and Garnsey (2000) noted: "[…] comparison of the sequence of a single gene or region may reflect differences for just that region, but this region may not be reflective of the entire genome."

For more in depth characterization, cDNA is generally cloned into a suitable plasmid vector, which after extraction may either be used for capillary sequencing and SSCP analysis (Lopez *et al.*, 1998; Kong *et al.*, 1999). One of the most comprehensive studies using a cloning-based characterization method was published by Rubio *et al.* (2001), where four genomic regions were used to assess the genetic variation within natural CTV populations in Spain. They observed that most isolates contained a population of sequence variants, with one being dominant. However, incongruences arose between genotype classifications between the different genomic regions used, which prompted the suggestion that recombination events between diverged sequence variants had occurred.

The assessment of the genetic diversity of a given CTV isolate based on the analysis of several genes has so far been the best method to do so, but has been shown to contain some flaws. The amplification step for example, based on gene-specific primers, might not be able to amplify all components of a population if a particular genotype is sufficiently divergent. Additionally, it is possible that if a minor CTV component has too low titers, the probability for their amplification is reduced compared to more abundant templates. Another problem lies in the number of clones analyzed. There are no current standards pertaining to the amount of clones one should select to reflect the actual CTV population and therefore components may be missed by a lack of clones sequenced. Lastly, Hilf's comments on preferentially comparing full genomes as opposed to gene regions (which show incongruences) has now come to a stage where it may become a reality to do so, since the recent commercial availability of next-generation sequencing technologies.

1.4.6 MICROARRAYS

A DNA microarray or oligonucleotide microarray is a device which can be used to compare DNA sequence similarities based on nucleic acid hybridization and can be applied to do parallel processing of many different sequences (Maskos, 1992). The method entails the binding of DNA sequences on a glass surface or "chip", onto which query sequences are deposited with the nature and amount of binding to the fixed DNA being determined with dyes such as fluorophores. Because of their small size, whole virus genomes can be deposited on such chips, which enables characterization of more than one gene as well as the monitoring of gene expression levels. Weng *et al.,* (2007) used this technology to assess CTV diversity and to determine where recombination between genotypes takes place. Accepting that a genome-wide comparison surpasses looking at selected genes, they designed universal CTV primers that insured full genome amplification by long-range RT-PCR without genotype-specific sequence bias. They fitted their DNA chip with four full length CTV genomes (T3, T30, T36 and VT), as well as several unique sequences from other isolates, which were mixed with total RNA derived from a naturally occurring isolate that caused stem-pitting in sweet orange. Their results suggested that this isolate contained multiple genotypes, mostly VT, T30 and T36. In addition, by looking at the most divergent part of the genome, the 5' termini, they could assess recombination events and the origin of these recombinant sequences. This study was at the time the closest one could get to the current high-throughput sequencing technologies available today. One of the main drawbacks of this study was the lack of sufficient reference genomes present on the chip, which does introduce a bias. Additionally, as with many CTV analyses, the results could differ if this same isolate was grafted onto another species or if the same infected tree was experiencing different climatic conditions. In addition they used total RNA as a template, which does not always reflect virus replication. Nonetheless, these challenges can be overcome, but even if the genotype distribution would change overtime, this technology shed light on CTV recombination mechanisms, its hotspots and origins.

1.5 NGS: A BIOLOGY REVOLUTION

1.5.1 INTRODUCTION

Up to 2005, researchers around the world had to rely on the Sanger method using capillary sequencing if they wanted to accomplish any sequencing projects (Harismendy *et al.*, 2009). This trend changed however, when a range of three next-generation sequencing (NGS) platforms were made commercially available: 454 (Roche), SOLiD (Applied Biosystems) and Solexa (Illumina), each with their own proprietary sequencing chemistries (Mardis, 2007). These instruments started a revolution in the sequencing world by allowing parallel processing of millions of sequence reads instead of the 96 previously achieved by capillaries. In addition, the number of generated sequence reads is huge, with accuracy claimed to be as high as over 99.99% (Mir, 2009). A marked difference lies in sequence read lengths, which ranges from 35bp to 250bp, in contrast to 650-800bp typically achieved by "classical" capillary-based instruments. A significant advantage from high-throughput sequencing allows bypassing cloning procedures, which may introduce sequence bias in terms of ligated DNA and the amount of clones sequenced (Mardis, 2007). In addition, the process can be lengthy and only a handful of cloning vectors can accommodate insert sizes as large as a 20kb- long virus genome. In the near future, as more instruments are made available, such as single-molecule sequencing (PacBio, Starlight), template preparations and PCR steps may also become optional, as with single molecule sequencing DNA *and* RNA can be sequenced directly, in real-time, (Stuholme *et al.*, 2011). It has also opened the doors to metagenomics, or in this context *metaviromics*, which enables the determination of the total viral diversity within a given sample, which includes the discovery of unknown viruses.

1.5.2 RANGE OF NGS TECHNOLOGIES

Over the last few years, as sequencing technologies evolved at a rapid rate, they can now be classified into two broad categories (Glenn, 2011): 2nd generation (454, Illumina, SOLiD) and 3rd generation (PacBio, Heliscope). They are not solely classified this way due their time of release, but mainly because of a shift from having a pre-amplification step prior to sequencing itself (2nd generation) to direct sequencing of nucleic acids (3rd generation). Each sequencing platform was originally best suited for only certain studies, but due to constant improvement in their characteristics, are starting to stand on equal grounds to suit every need. In this section, the most commonly used platforms are described, along with their sequencing chemistries, range of applications, advantages and disadvantages are provided.

The 454 sequencing instrument was the very first next-generation sequencer made available (www.454.com). It is characterized by its "pyrosequencing mechanism" and production of fewer (700 Mb) but much longer reads (up to 1000 bp) as compared to other platforms such as Illumina. These characteristics provide in effect, Sanger-like sequences on a massive scale, which makes this technology very popular, albeit being one the most expensive (Glenn, 2011). The concept of 454 sequencing lies in three key components: 1) Adapter ligation to sheared DNA fragment; 2) clonal amplification of DNA on beads via emulsion PCR and 3) pyrosequencing in individual wells of a PicoTiter™ plate. In the first step, DNA is sheared into fragments, ligated with specific adapters which will serve to bind to magnetic beads as well as primer-binding sites for the emulsion PCR reaction that follows. This results in beads coated with DNA clones which are subsequently deposited into wells of a PicoTiter™ plate designed to only hold one bead per well. The pyrosequencing reaction take place simultaneously in each of the wells. Briefly, a DNA polymerase is added and extends the 454 adapter sequence that serves as a primer-binding site. Next, fluorescently-labelled bases are added, which upon binding release a pyrophosphate molecule. This molecule reacts with a Sulphurylase and Luciferase enzyme, inducing light and oxyluciferin production. These light signatures are then interpreted by a laser and recorded for each base. Due to the long read length, *de novo* assemblies are easier, more accurate and less computationally demanding.

Illumina technology, originally called Solexa (www.illumina.com) was the second next- generation sequencer available. It is characterized by its short read lengths (~100 bp) and its “sequencing by synthesis” mechanism using a glass plate called a flow cell. There are three major steps in the Illumina workflow: 1) Library preparation; 2) Cluster generation and 3) Sequencing. In the library step, DNA is sheared into small pieces, followed by the ligation to these fragment of Illumina-specific adaptors. Cluster generation is then performed on a flow cell, consisting of a glass plate densely coated by oligos that hybridizes to the adapter. Every bound DNA molecule subsequently undergoes bridge amplification, which produces clonal copies of the original molecule. This produces millions of sequences clusters over the flow cell, which are now ready to be sequenced simultaneously. Each base is determined individually, by flowing over the glass four fluorescently-labelled, reversibly terminated nucleotides. At each cycle, a laser excites the nucleotides, which produces a specific color allowing the identification of the added base. The competition for binding to the template between the four nucleotides insures high base-call accuracy in general, but also through homopolymer regions and repetitive sequences. One of the biggest drawbacks of Illumina is its short read length, but because it produces millions of sequences, is still a very valuable tool. Read length is continuously improving, being 35 bp when the instrument was first released, to 100 bp for paired-end and 500 bp in mate-paired libraries. Still, short read lengths are harder to assemble for *de novo* purposes, and therefore require more intense computational resources as compared to longer reads (Glenn, 2011). Therefore, although read lengths have been greatly improved, Illumina can be mostly used for targeted re-sequencing projects, SNP detection (for small genomes) and metagenomics. Lastly, multiplexing is permitted (as on all NGS platforms), which brings high-throughput even further by sequencing multiple samples in a single run.

The SOLiD platform, owing its name from "Sequencing by Oligonucleotide *Li*gation and *D*etection", was the third 2nd generation sequencing technology that was released (www.appliedbiosystems.com). As its name suggests, the sequencing chemistry for this instrument employs sequential rounds of ligation reactions of fluorescently-labelled dinucleotide probes to determine a base. The steps involved in SOLiD sequencing have much in common with 454, in terms of DNA fragmentation, magnetic bead hybridization and emulsion PCR. The main differences between the two are the glass support (similar to Illumina) referred to as a microfluidic flowchip and the base detection mechanism, called 2-base encoding. In a SOLiD run, each base is interrogated twice, by means of a probe pool consisting of 4 defined dinucleotides per pool, each with 4 different dyes. This "double-checking" scheme inherent to the sequencing mechanism alone, allows the best base-call accuracy, especially for differentiating between sequencing errors and true SNPs. However, this technology allows the generation of short reads (~60-75 bp, in amount similar to Illumina), but compensates with its unparalleled accuracy (especially for homopolymer regions), which can slightly facilitate contig assemblies.

The PacBio *RS* instrument is one the latest, third generation instruments made available (www.pacificbiosciences.com). This technology is capable of achieving single molecule sequencing in real time, which was termed "SMRT" sequencing (*S*ingle *M*olecule *R*eal *T*ime). Its ability to sequence and generate very long reads (up to 10kb in some cases, but 3kb is the norm) is based on an optical device called a "zero-mode waveguide", or ZMW (Levene *et al.*, 2003). The exact physical principles at work in this sequencing mechanism are beyond the scope of the present description, and only the basic functioning is described. To produce a base call, a DNA polymerase is fixed at the bottom of the ZMW, along with the target DNA molecule. Every four nucleotides used by the polymerase are uniquely fluorescently labelled, and every time a new nucleotide is incorporated to a growing DNA strand, the fluorophore is cleaved, emitting light that can be detected and translated into a base call based on the dye color. The biggest advantage from this technology are length of the reads generated, which make it ideal for *de novo* genome sequencing, re-sequencing and notably DNA methylation (Tyson *et al.*, 2011).

Another recent addition to the 3rd generation machines is the Ion Torrent platform (www.iontorrent.com). This technology also uses the sequencing by synthesis principle, but is unique due to its ability to base call not by an optical device, but by a semi-conductor circuit (an "ion chip") that detects cleaved hydrogen ions during DNA polymerization. The ion chip is where the actual sequencing happens. Merely the size of a typical desktop processor chip, it is a multilayered device composed of microscopic wells layer under which the detection mechanism is positioned. Briefly, the sequencing woks as follows: the target DNA to be sequenced is deposited into wells along with a DNA polymerase. Each of the four un-modified dNTPs are sequentially added, and upon incorporation of a particular dNTP to the growing strand, a pyrophosphate, as well as a hydrogen ion is released. The detection layer, composed of an ion sensitive layer (or ISFET ion sensor) records the change in pH induced by the released ion and can record which base was incorporated. The rapidity of the sequencing process (~4 sec/base), as well the lower cost of the machine and its reagents (e.g. unmodified dNTPs, no pre-amplification, no magnetic beads, re-use of the ion chip) make this new technology very promising for lower budget sequencing projects. However, the short read lengths generated (~200 bp) may become problematic for *de novo* assemblies of large genomes (not so for viral and bacterial genomes). Another limitation of this technology is its inability to accurately determine the base composition of homopolymer repeats (e.g. AAAAAA), a problem shared by other platforms such as 454.

1.5.3 ADVANTAGES, LIMITATIONS & CHALLENGES

The application of high-throughput sequencing for viral populations studies surpasses (and sometimes renders obsolete) classical methods including those described in section 1.4, in many ways. Although accurate to some extent, sequencing of multiple clones (as previously done in our laboratory) allows a focus on only a few genes. Various steps of that technique may introduce biases, including the RT-PCR step, ligation, repeated PCR reactions and also the limited number of clones sequenced. For CTV characterization specifically, incongruent phylogenies have been observed when different genomic regions (e.g. A-fragment, *p23*) where analyzed of the same isolate (Scott *et al.,* 2012), with whole genome analysis recommended to remedy this problem. This is achievable with NGS. Combined with massive sequence data output, the small genome size of viruses are likely to yielding high coverage values, as opposed to fungal, bacterial or other large genomes. Coverage, also called “read redundancy”, which may exceed a thousand, can offer accelerated genotype discovery, as well as providing enough resolution to detect rare mutations, which can be significant when dealing with viral quasi-species. Additionally, high read redundancy may also help determining base call accuracies when dealing with low PHRED scores (base quality value) (Beerenwinkel & Zagordi, 2011).

Every new technology is by definition, not fully optimized. Some of the current problems associated with NGS will be improved over time, but there are some which remain problematic. One of the biggest drawbacks of NGS is the relatively short read lengths obtained, which can have two major impacts for downstream analysis: *de novo* assembly and SNP detection (Morazova, 2008; Beerenwinkel & Zagordi, 2011). In contig assemblies, the longer the read, the easier it is for them to contain regions of overlap. Moreover, computational power required for assembly will become less intensive as the length of the reads become greater. The exact opposite is true for short reads. In terms of SNP detection, if insufficient areas of overlap are obtained, there are virtually no ways to determine if a nucleotide change (e.g. indel) is the result of a sequencing error or an actual biological artifact. However, if sufficient coverage is obtained, multiple reads can be compared and a distinction between error and fact can be made. Progress regarding read lengths have greatly improved, from 35 bp in 2006 to

450 bp in 2012 (Illumina, Inc) without compromising the amounts of reads generated. But there are some instances where read length is not a major concern. In resequencing projects, whole-genome data for the target organism is available, which is then used in read assemblies to create contigs. When reference genomes are available, reads are “guided” to where they belong and therefore their length becomes irrelevant (there is minimal limit of course). This methodology also requires less computer power, which is another major limitation of NGS. Handling NGS data is notoriously computationally demanding, especially for contig generation yielded by *de novo* algorithms using short reads. In addition, there currently remains a lack of standard in terms methodologies, the versions of software used, parameters etc… which hinders peer-reviewed processes and reproducibility (Nekrutenko and Taylor, 2012).

When it comes to RNA viruses, one of the most important considerations is the choice of template that is used for the sequencing process (Studholme *et al,* 2011). Template choice *and* preparation are both critical for future data interpretation and depend largely on the research questions. Due to the current inability to sequence RNA directly, a reverse-transcription step must be included. This is where the first potential bias may arise, not only in terms of the enzymes used to synthesize DNA, but also which regions will be amplified. To minimize this issue, several alternatives are available, including the use of random/degenerate primers or long-range RT-PCR (section 1.6). This represents one of the biggest problems associated with NGS analysis of viral populations, in which the determination of genuine mutations cannot be distinguished by those generated by technical errors (Beerenwinkel & Zagordi, 2011). In addition, enrichment methods need to be robust enough so as to minimize the amount of unwanted sequences, such as those derived from the host.

The use of next-generation sequencing in plant virology is already well underway and is set to become the standard in the near future. Combined with unbiased amplification methods, it has accelerated the discovery of novel haplotypes and previously unseen viruses. Furthermore, targeted re-sequencing may become the method of choice to characterize viral populations/quasi-species due to high coverage values and resolution one can achieve with such massive amounts of sequence data. In current literature, the application of NGS technology is mostly used in metaviromics studies, which requires unbiased amplification methods as well as a means to discard host-derived sequences. Adams *et al.,* (2009) were amongst the first to publish an analysis pipeline for the identification of unknown viruses in plant. Based on a RNA extract, two types of primers, random (universal) and oligo- (d)T , enabled enrichment for viruses having a poly-A tail as well as those who do not. Of course, relying on random priming causes sequencing also of contaminants, such as mRNA, rRNA, etc…Their solution was to use subtractive hybridization, which removed contamination *prior* to sequencing. This pre-sequencing removal allowed higher coverage of viral sequences and facilitated filtering *in silico*. In another study, Roossinck *et al.,* (2010) tackled the host contamination problem in a different way. Instead of using total RNA as a template, double-stranded RNA was used as an enrichment method. Also using a set of random primers, it was expected that dsRNA species would belong only to viruses within the host, thereby eliminating contamination.

1.6 WHOLE GENOME AMPLIFICATION (WGA) SYSTEMS

At present, there are two ways in which one can achieve whole virus genome amplification: long-range PCR and random priming. In the first method, a high fidelity polymerase is used to synthesize large DNA stretches, and uses sequence-specific primers. In order to minimize biases, universal species-specific primers must be designed carefully. Random amplification of DNA was first developed by Reyes and Kim (1991), commonly known as SISPA or “sequence-independent single primer amplification”. One year later, Froussard (1992) adapted the technique for the amplification of viral RNA using random hexameric primers in reverse-transcription, followed by PCR. Several other methods allow for random amplification of all RNA species in a sample, including mRNA, single-stranded and double-stranded RNA viruses and ribosomal RNA. Based on the same random priming principle, each of these methods has been modified to select which RNA species should mostly be amplified. This allows for amplification of whole-virus genomes as well as transcriptome analyses for a particular tissue.

One of the best ways to generate full-length whole cDNA libraries is by means of random-PCR (rPCR) (Froussard, 1992). This methodology employs a 26 -nucleotide long primer, containing a random stretch (6-12) of nucleotides at its 3’ end, commonly called “universal primer- dN_6”. The random hexamer end is able to bind virtually everywhere along an RNA template, and primes reverse transcription. Careful consideration must be given to determining primer concentration, as this parameter dictates the size of the synthesized cDNAs. Single-stranded cDNAs are then converted to a double-stranded form by means of the Klenow fragment of DNA polymerase I, which results in double-stranded cDNAs flanked by the original anchor domain of universal-dN_6. After column purification to remove excess primers, PCR amplification of the template using primers with complementary sequences of the dN_6 anchor domain is initiated. This results in a DNA smear, ranging from 300 bp to several kilobases in size (Figure 6).

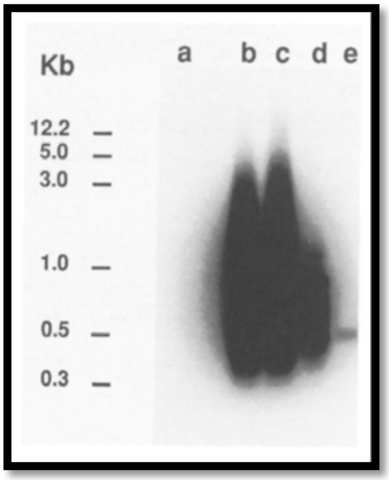


Figure 6: rPCR agarose gel. Randomly amplified cDNAs produce DNA fragments ranging from 300 bp to several kilobases in size (Froussard, 1992).

The main advantage of this method is its complete lack of discrimination towards any RNA species, regardless of initial amount of RNA template. It has been applied on double-stranded RNA viruses (Roossinck *et al*, 2010; Rwahnih *et al.*, 2011), single-stranded RNA viruses (Djikeng *et al.,* 2008; Wylie *et al.*, 2012), both poly- and non-polyadenylated RNA, as well as in a vast array of research fields such as neurobiology and developmental biology. However, this inherent unparalleled non-selectivity is a double-edged sword, as it introduces masses of undesirable background sequences, which have to be removed by either using an alternative template to total RNA (e.g.

dsRNA), physically removing undesirable DNA (e.g. subtractive hybridization), or by removal of sequences *in silico*.

A variant to rPCR, which selectively amplifies polyadenylated RNAs (host mRNA, ssRNA viruses, etc…) is homopolymer tailing, also referred to as oligo-d (T) priming (Primrose and Twyman, 2009). Due to repeated adenosine monophosphates at their 3' ends, these RNAs can selectively be reverse-transcribed using a primer composed of only thymine residues linked with an anchor sequence for subsequent PCR amplification. This methodology has been applied for in-depth analyses of plant viruses (Adams *et al.*, 2009; Wylie *et al.*, 2012). In these studies, pooled double-stranded cDNAs were subjected to high-throughput sequencing, which showed more than 15 species of known viruses as well as novel ones. The main disadvantage of homopolymer tailing is its "random spectrum", because it only identifies polyadenylated viruses, therefore omitting greater potential for virus discovery. Additionally, because of the template used (total RNA), a large amount of host contamination is observed, which complicates *de novo* assemblies and lowers virus-derived reads. This effect could be reduced to a certain extent by using subtractive hybridization prior to sequencing (Pradel *et al.*, 2002).

1.7 REFERENCES

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackevicienne, E., Navalinskiene, M., Samuitiene, M. and Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology* 10(4): 537-545.

Albiach-Marty, M.R., Mawassi, M., Gowda, S., Satyanarayana, T., Hilf, M.E., Shanker, S., Almira, E.C., Vives, M.C., Lopez, C., Guerri, J., Flores, R., Moreno, P., Garsney, S.M., and Dawson, W.O., 2000. Sequences of Citrus tristeza virus separated in time and space are essentially identical. *Journal of Virology* 74:6856-6865.

Ayllon, M.A., Lopez, C., Navas-Castillo, J., Garsney, S.M., Guerri, J., Flores, R., and Moreno, P., 2001. Polymorphism of the 5' terminal region of citrus tristeza virus (CTV) RNA: Incidence of three sequence types in isolates of different origin and pathogenicity. *Archives of Virology* 146:27-40.

Bar-Joseph, M., 1978. Cross protection incompleteness: A possible cause for natural spread by mild mutants of papaya ringspot virus for control of ringspot disease of papaya in citrus tristeza virus cross-protection trial in Florida. *In Proceedings of the14th Conference IOCV*, pp. 111–114.

Bar-Joseph, M., Garsney, S.M., Gonsalves, D., Moscovitz, M., Purcifull, D.E., Clark, M.F., and Loebenstein, G., 1979. The use of enzyme-linked immunosorbent assay for the detection of Citrus tristeza virus. *Phytopathology* 69(2): 190-194.

Bar-Joseph. M., Marcus, R., and Lee, R.F., 1989. The continuous challenge of citrus tristeza virus control. *Annual Reviews of Phytopathology* 27:291-316.

Costa, A.S., and Muller, G.W., 1980. Tristeza control by cross-protection: A U.S-Brazil cooperative success. *Plant Disease* 64: 538-541.

Djikeng, A., Halpin, R., Kuzmickas, R., DePasse, J., Feldblyum, J., Sengamalay, N., Afonso, C., Zhang, X., Anderson, N.G., Ghedin, E., and Spiro, D.J., 2008. Viral genome sequencing by random priming methods. *BMC Genomics* 9:5.

Fagoaga, C., Lopez, C., Hermoso de Mendoza, A., Moren, P., Navarro, L., Flores, R., and Pena, L., 2006. Post-transcriptional gene silencing of the *p23* silencing suppressor of Citrus tristeza virus confers resistance to the virus in transgenic Mexican lime. *Plant Molecular Biology* 60:153-165.

Febres, V.J., Ashoulin, L., Mawassi, M., Frank, A., Lee, R.F., Bar-Joseph, M., Manjunath, K.L., Lee, R.F., and Niblett, C.L., 2003. The *p27* protein is present at one end of *Citrus tristeza virus* particles. *Phytopathology* 86:1331-1335.

Febres, V.J., Lee, R.F., and Moore, G.A., 2008. Transgenic resistance to *Citrus tristeza virus* in grapefruit. *Plant Cell Reports* 27:93-104.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D, Merregaert, J., Min Jou, W., Molemans, F., Raeymackers, A., Van der Berghe, A., Volckaert, G., and Ysebaert, M., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* 260 (5551): 500–7.

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11:759-769.

Gowda, S., Ayllon, M.A., Satyanarayana, T., Bar-Joseph, M. and Dawson, W.O., 2003. Transcription stategy in a Closterovirus: a novel 5'-proximal controller element of Citrus Tristeza Virus produces 5'-and 3'- terminal subgenomic RNAs and differs from 3 open reading frame controller elements. *Journal of Virology* 77(1): 340-352.

Grant, T. J., and Costa, A. S.,1951. A mild strain of the tristeza virus of citrus. *Phytopathology* 41:114–122.

Gutierez-E, M.A., Luth, D., and Moore, G.A., 1997. Factors affecting Agrobacterium-mediated transformation in Citrus and production of sour orange (*Citrus aurantium* L.) plants expressing the coat protein gene of citrus tristeza virus. *Plant cell reports* 16: 745-753.

Hamilton, A.J., and Baulcombe, D.C., 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Nature* 286 (5441):950-952.

Harper, S.J., Dawson, T.E., and Pearson, M.N., 2010. Isolates of Citrus tristeza virus that overcome Poncirus trifoliata resistance comprise a novel strain. *Archives of Virology* 155(4): 471-480.

Hilf, M.E. and Garnsey, S.M., 2000. Characterization and classification of Citrus Tristeza virus isolates by amplification of multiple molecular markers. *In proceedings of the fourteenth IOCV conference.*

Hilf, M.E., Karasev, A.V., Albiach-Marti, M.R., Dawson, W.O. and Garnsey, S.M., 1999. Two paths of sequence divergence in the citrus tristeza virus complex. *Phytopathology* 89:336–42

Hull, R., 2002. Matthews' Plant Virology. 4th Edition. *Academic Press, London and San Diego.*

Innis, M. A., Gelfand, D. H., Sninsky, J. J., White, T. J., 1990. PCR protocols: a guide to methods and applications: 21-27. *Academic Press*, New York.

Karasev, A.V., Boyko, V.P., Gowda, S., Nikolaeva, O.V., Hilf, M.E., Koonin, E.V., Niblett, C.L., Cline, K., Gumpf, D.J., Lee, R.F., Garsney, S.M., Lewandowski, D.J. and Dawson, W.O., 1995. Complete sequence of the Citrus Tristeza Virus RNA genome. *Virology* 208: 511-520.

Karasev, A.V., 2000. Genetic diversity and evolution of Closteroviruses. *Annual Review of Phytopathology* 38: 293-324.

Kong, P., Rubio, L., Polek, M., and Falk, B.W., 2000. Population structure and genetic diversity within California *Citrus tristeza virus* (CTV) isolates. *Virus genes* 21:139-145.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., Webb, W.W., 2003. Zero-Mode Waveguides for single-molecule analysis at high concentrations. *Science* 299: 682-686.

Levy, J.A., Fraenkel-Conrat, H. and Owens, R.A., 1994. Virology, 2nd Edition. *Prentise Hall publishers*: 26-31.

Marroquin, C., Olmos, A., Gorris, M.T., Bertolini, E., Martinez, M.C., Carbonell, E.A., De Mendoza, A.H. and Cambra, M., 2004. Estimation of the number of aphids carrying Citrus tristeza virus that visit adult citrus trees. *Virus Research* 100: 101-108.

Maskos, U., and Southern, E.M., 1992. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. Nucleic Acid Research 20(7): 1679-1684.

Mawassi, M., Karasev, A.V., Mitkiewska, E., Gafny, R., and Lee, R.F., 1995. Defective RNA molecules associated with citrus tristeza virus. *Virology* 208:383–87.

Mawassi, M., Mitkiewska, E., Gofman, R., Yang, G and Bar-Joseph, M., 1996. Unusual sequence relationships between two isolates of Citrus tristeza virus. Journal of General Virology 77(9):2359-2364.

McKinney, H. H. (1929). Mosaic diseases in the Canary Islands, West Africa and Gibraltar. Journal of Agricultural Research 39:557–578.

Melzer, M.J., Borth, W.B., Sether, D.M., Ferreira, S., Gonsalves, D. and Hu, J.S., 2010. Genetic diversity and evidence for recent modular recombination in Hawaiian Citrus Tristeza virus. *Virus Genes* 40: 111-118.

Moreno, P., Guerri, J., Ballester, J.F., and Martinez, M.E., 1991. Segregation of Citrus Tristeza strains evidenced by double stranded RNA (dsRNA) analysis. *In Proceedings of the11th Conference IOCV.*

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

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., and Sekiya, T., 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci U S A* :86(8):2766–2770

Powell, C. A., Pelosi, R. R., Rundell, P. A., and Cohen, M., 2003. Breakdown of cross-protection of grapefruit from decline-inducing isolates of citrus tristeza virus following introduction of the brown citrus aphid. *Plant Disease* 87:1116–1118.

Pradel, N., Leroi-Setrin, S., Joly, B., and Livrelli, V., 2002. Genomic subtraction to identify and characterize sequences of Shiga toxin-producing Escherichia coli 091:H21. *Applied Environmental Microbiology* 68:2316-2325.

Primrose, S.B. and Twyman, R.M., 2009. Principles of gene manipulation and genomics, 7th edition. Blackwell Publishing, USA.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G. and Roe, B.A. 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology* 19: 81-88.

Roy, A., Ananthakrishnan, G., Hartung, J. and Brlansky, R.H., 2010. Development and application of a multiplex reverse-transcription polymerase chain reaction assay for screening a global collection of Citrus tristeza virus isolates. *Phytopathology* 100(10): 1077-1088.

Roy, A., and Brlansky, R.H., 2009. Population dynamics of a Florida Citrus tristeza virus isolate and aphid transmitted subisolates: identification of three genotypic groups and recombinants after aphid transmission. *Phytopathology* 99, 1297–1306.

Roy, A., and Brlansky, R.H., 2010. Genome analysis of an orange stem pitting citrus tristeza virus isolate reveals a novel recombinant genotype. *Virus Research* 151(2): 118-130.

Rubio, L., Ayllon, M.A., Kong, P., Fernandez, A., Polek, M., Guerri, J., Moreno, P. and Falk, B.W., 2001. Genetic variation of Citrus Tristeza Virus isolates from California and Spain: evidence for mixed infection and recombination. *Journal of Virology* 75(17): 8054-8062.

Rwahhnih, M.A., Daubert, S., Urbez-Torres, J.R., Cordero, F., and Rowhani, A., 2011. Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. *Archives of Virology* 156: 397-403.

Schneider, H., 1959. The anatomy of Tristeza virus-infected citrus. Citrus virus diseases, *J.M. Wallace Editions*, California.

Schwarz, R.E., 1965. Aphid-borne virus diseases of citrus and their vectors in South Africa. An investigation into the epidemiology of aphid transmissible diseases of citrus by means of trap plants. *South African Journal of Agricultural Science* 8:839-852.

Scott, K.A., Hlela, Q., Zablocki, O., Read, D., van Vuuren, S., and Pietersen, G., 2012. Genotype composition of populations of grapefruit-cross-protecting citrus tristeza virus strain GFMS12 in different host plants and aphid-transmitted sub-isolates. *Archives of Virology* DOI: 10.1007/s00705-012-1450-4.

Suastika, G., Natsuaki, T., Terui, H., Kano, T., Ieki, H., and Okuda, S., 2001. Nucleotide sequence of Citrus tristeza virus seedling yellows isolate. *Journal of General Plant Pathology* 67:73-77.

Palukaitis, P., and Zaitlin, M., 1984. A model to explain the "cross-protection" phenomenon shown by plant viruses and viroids. *Plant-Microbe interactions.*

Studholme, D.J., Glover, R.H., and Boonham, N., 2011. Application of high-throughput DNA sequencing in Phytopathology. *Annual Reviews of Phytopathology* 49:87-105.

Tyson, A.C., Murray, I.A., Morgan, R.D., Kislyuk, A.O., Spittle, K.E., Boitano, M., Fomenkov, A., Roberts, R.J., and Korlach, J., 2011. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acid Research* doi: 10.1093/nar/gkr1146.

Van Vuuren, S.P., Collins, R.P., and Da Graca, J.V., 1993. Growth and production of lime trees pre-immunized with different mild citrus tristeza virus isolates in the presence of natural disease conditions. *Phytophylactica* 25:49-52.

Van Vuuren, S.P., van der Vyver, J.B., and Luttig, M., 2000. Diversity among sub-isolates of cross-protecting citrus tristeza virus isolates in South Africa. *Fourteenth IOCV Conference- Citrus Tristeza Virus*: 103-110.

Vela, C., Cambra, M., Cortez, E., Moreno, P., Miguet, J.G., Perez De San Roman, C., and Sanz, A., 1986. Production and characterization of monoclonal antibodies specific for Citrus tristeza virus and their use for diagnosis. *Journal of General Virology* 67(1): 91-96.

Von Broembsen, L., and Lee, A.T.C., 1988. South Africa's Citrus Improvement Programme. *In Proceedings of the 10th IOCV Conference*, Riverside.

Wang, H. L., Yeh, S. D., Chiu, R. J., and Gonsalves, D.,1987. Effectiveness of cross-protection by mild mutants of papaya ringspot virus for control of ringspot disease of papaya in Taiwan. *Plant disease* 71(6): 491-497.

Weng, Z., Barthelson, R., Gowda, S., Hilf, M.E., Dawson, W.O., Galbraith, D.W., and Xiong, Z., 2007. Persistent infection and promiscuous recombination of multiple genotypes of an RNA virus within a single host generate extensive diversity. *PLoS One* 2(9): e917.

Wylie, S.J., Luo, H., Li, H., and Jones, M.G.K., 2012. Multiple polyadenylated RNA viruses detected in pooled cultivated and wild plant samples. *Archives of Virology* 157:271-284.

Ziebell, H., 2008. Mechanisms of cross-protection. *Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources* 3: 49-62.

Ziebell, H., and Carr, J.P., 2010. Cross-protection: a century of mystery. *Advances in Virus Research* 76: 211-264

CHAPTER 2:

APPLICATION OF DODECAMER-PRIMED RT-PCR ON DIFFERENT CTV TEMPLATES SUBJECTED TO HIGH THROUGHPUT SEQUENCING & COMPARISON WITH CLONING RESULTS

2.1 INTRODUCTION

The objective of this part of the work was to develop and evaluate an alternative workflow for characterizing CTV isolates as it is critical to assess which genotypes are present in pre-immunization sources for efficient cross-protection. This workflow aimed to address three fundamental steps usually employed to characterize CTV isolates: the enrichment template, target genomic region and sequencing method. Typically, several gene regions, such as *p23* and the A-fragment, are used to differentiate between strains (Roy *et al.,* 2010). For survey purposes, these gene regions may be directly sequenced (by the Sanger method) and are often used as a preliminary identification system. In in-depth characterization studies, sequencing of multiple clones spanning several gene regions are used.

However, depending on the gene region used, it has been observed that phylogenies may appear incongruent. For example, in a recent study by Scott *et al.,* (2012), while characterizing the same isolate, cloning of the ORF1a region inferred a CTV population dominated by B165. *A contrario*, when p23 was used, it indicated dominance for the VT genotype. To explain this, it was argued that a recombinant possessing similar regions as these two parental genotypes may be present. This observation could also be attributed in part to variability differences between the two halves of the CTV genome, where genes located towards the 3' end tend to be highly conserved (~90%) and towards the 5' end, highly variable (<70%) (Ayllon *et al.*, 2001). However with CTV being highly prone to recombination events (Kong *et al.*, 2000; Rubio *et al*., 2001), the original conclusion may have so credence. To illustrate: If only one or several portions of the genome is/are sequenced (e.g. A-fragment), analysis might attribute it to the wrong genotype. For example, if a portion from the 1a fragment was used, it may be attributed to Kpg3. Therefore, it would be inferred that Kpg3 was the genotype present. However, this is not necessarily the case, as it is possible that instead of Kpg3, VT was present, but because the 1a region was exchanged, it was interpreted as Kpg3. This example illustrates the danger of using selected genomic regions and underlines the need to characterize isolates at a whole-genome level. Another example for this phenomenon can also occur when phylogenies are inferred using selected genes of *only* reference strains. Indeed, if identical evolutionary models are used, differing tree topologies are observed depending on the gene region used. Moreover, if whole genome dendrograms are constructed, another topology arises. Therefore, the best estimate for inferring relationships between CTV genotypes may be by analyzing CTV isolates at a whole-genome level.

To remediate these problems, a non-specific whole genome amplification technique combined with high throughput sequencing (Illumina) was implemented. This approach was based on the assumption that whole genome comparisons would offer the most optimal comparative tool as opposed to the comparison of a few genes, and therefore reduce incongruences.

In planta, CTV may be present in three distinct forms: as an encapsidated particle, as a naked single-stranded RNA and as a double-stranded RNA intermediate. As each form represents a distinct stage in the virus's replication cycle, it is of importance to choose a template which reduces bias to a minimum. Bias in this regard may be due to host selection**,** abiotic factors such as temperature and possibly inter-strain competition, in a scenario where multiple strains co-infect the same host, but only one or few achieve a full replication cycle. Therefore, the choice of template can play a critical role in data interpretation. Furthermore, the ideal template should sufficiently enrich so as to obtain the highest coverage and minimize host contamination (Studholme *et al*., 2011).

In order to evaluate and compare results generated by the new characterization technique, analysis was initially performed on a single sub-isolate (12-7) for which cloning data based on the ORF1a and *p23* gene regions is available.

2.2 MATERIALS & METHODS

2.2.1 Virus sub-isolates

The CTV sub-isolate used in this study (12-7, accession number 08-0010) was derived from a GFMS12- inoculated Mexican lime (*Citrus aurantifolia*) (L.) Swingle) tree onto which single aphid transfers (*Toxoptera citricida*) (Kirkaldy) have been performed on Mexican lime trees (van Vuuren *et al.*, 2000). The 12-7 sub-isolate tree was kept in insect-free glasshouse conditions for a minimum of 7 years.

2.2.2 Total RNA extraction

CTV-infected leaf petioles were cut and 0.3 grams were pulverized with liquid nitrogen and ground with a mortar and pestle until a fine powder was obtained. Extraction was performed with the SV Total RNA Isolation System (Promega, Madison, WI, USA) according to the manufacturer's instructions. Extracted RNA was stored at -80ºC until further use.

2.2.3 Double-stranded RNA purification

Purification of double-stranded RNA was performed according to Morris and Dodds (1983) with minor modifications. A total of 0.5 grams of leaf petioles were pulverized with liquid nitrogen and ground to a fine powder with a mortar and pestle. The powder was transferred to a 2 ml microcentrifuge tube and 1 vol of 2X STE buffer (0.1 M NaCl, 0.05 Tris, 0.001 M EDTA, pH 7.0, supplemented with 1% SDS and 0.1% 2- mercaptoethanol), 1 vol phenol and 0.5 vol of chloroform: pentanol (25:1) was added. The tube was centrifuged at 8,000 g for 10 minutes. The aqueous phase was decanted and transferred to a new tube and its volume measured. A 100% ethanol solution was added to a concentration of 15 % (v/v). The solution was then added to a tube containing CF-11 cellulose pre-equilibrated with 1X STE-E 15% buffer. Double-stranded RNA was allowed to adsorb to the cellulose for 15 min with intermittent vortexing for a total of 15 min followed by centrifugation for 3 min to pack the cellulose. The cellulose was then washed 3 times by adding 1 ml STE-E 15%, vortexing, centrifuging briefly and decanting the supernatant. After the final wash, 0.5 ml of 1X STE was added. The tube was vortexed and centrifuged, and the supernatant transferred to a 1.5 ml eppendorf tube (Eppendorf, Germany). The dsRNA was precipitated with 100% ethanol and stored at -20ºC overnight. The tube was centrifuged at 14,000 *g* for 20 min, the supernatant was discarded and the pellet was left to dry. Lastly, dsRNA was resuspended in 50µl of molecular-grade water.

2.2.4 Virus particle enrichment through Immunocapture

Immunocapture was performed according to Nolasco *et al*., (1993), with slight modifications from Le Provost *et al*., (2006) in order to obtain viral particles specifically.

Briefly, 0.2 ml thin-walled polypropylene microfuge tubes (Molecular BioProducts, San Diego, California) were coated with a 100 µl of CTV polyclonal antiserum CREC 29 diluted (1:5000) in coating buffer (1.59 g/l Na_2CO_3, 2.93 g/l $NaHCO_3$, pH 9.6) and incubated for 2 hours at 30ºC. The tubes were washed 3 times for 3 min with PBS-Tween buffer (8g/l NaCl, 0.2g/l $KH_2PO_{4,}$ 1.15 g/l Na_2HPO_4, 0.2 g/l KCl, pH 7.4 and supplemented with 0.1 % Tween 20). One gram of CTV-infected leaf petioles was macerated in coating buffer (1:10 v/v) using a HOMEX 6 homogenizer (Bioreba, Reinach, Switzerland). The resulting liquid was transferred to a 2ml tube and briefly centrifuged. A 100µl of recovered supernatant was applied to the antibody-coated tubes (in triplicates) and incubated at 4ºC overnight. The tubes where washed 3 times as previously described and washed one final time with sterile water followed by brief drying at room temperature.

2.2.5 Random- primed, two-step reverse-transcription PCR

The random, reverse transcription PCR reaction was conducted according to Roosinck *et al*., (2010) with minor modifications. Twelve microlitres of template (total RNA, dsRNA or virus particles from immunocapture tests) was mixed with 5µl universal RDOD primer (Table 1) (IDT, USA) at a concentration of 20 µM in a 0.2 ml microcentrifuge tube. The tubes were incubated for 10 min at different temperatures, according to which template they contained. Denaturation temperatures for total RNA, dsRNA and viral particles were 70ºC, 95ºC and 80ºC, respectively. Following denaturation, tubes were transferred on ice for 5 min. A 7µl cDNA synthesis reaction mix containing 1X Avian Myeloblastosis virus (AMV) buffer (Roche, Mannheim, Germany), 200 µM of each dNTP (Promega, Madison, WI, USA), 5 units of RNAse inhibitor (Roche, Mannheim, Germany) and 10 units of AMV reverse transcriptase (Roche, Mannheim, Germany) was added, for a total volume of 24 µl. The tubes were incubated for 10 min at 25ºC, followed by 60 min at 50ºC. A 10 µl cDNA aliquot was used as template in PCR. The reaction mix consisted of 5 µl 1X PCR buffer, 2.5 µl $MgCl_2$(50mM), 2 µl dNTPs (3.5 mM), 10 µl 2 µM $RDOD_{comp}$ complimentary primer (Table 1) (IDT, USA) , 2.5 units of BioTaq DNA polymerase (Bioline, UK), 5 µl BSA (20 µg/ µl) to a final volume of 50 µl with molecular grade water. The samples were loaded into a

Bio-Rad T100 thermal cycler (Bio-Rad, CA, USA) under the following cycling conditions: 94ºC for 2 min, followed by 40 cycles of 94 ºC for 30 sec, 42 ºC for 45 sec,72 ºC for 1 min and a final extension at 72 ºC for 10 min.

Table 1: Primers used for the rRT-PCR assay and the cloning of the CTV A-fragment / *p23* gene.

Primer name	**Sequence (5'-3')**	**Genomic region**	**Author**
A-F (forward)	ACGTGTTCGTGAAACGCGG	A-fragment	Rubio *et al*.
A-R (reverse)	GTCGATAACTCGACAAACGAGC	A-fragment	
PM50 (forward)	ACTAACTTTAATTCGAACA	p23	Sambade *et al*.
PM51 (reverse)	AACTTATTCCGTCCACTTC	p23	
SP6	ATTTAGGTGACACTATAGAA	pGEM-t Easy polylinker	Promega, USA
T7	TAATACGACTCACTATAGGG	pGEM-t Easy polylinker	
RDOD	CCTTCGGATCCTCCN_{12}	Random	Roossink *et al*.
RDOD$_{comp}$	CCTTCGGATCCTCC	Random	

2.2.6 PCR clean-up & Illumina sequencing

Randomly amplified cDNAs from each template were visualized on a 1% agarose gel to confirm the reaction had taken place. The amplicons were cleaned prior to sequencing with the Qiagen MinElute PCR purification kit according to the manufacturer's instruction. For the preparation of Illumina compatible libraries, the Nextera DNA sample kit (Adey *et al.*, 2010) was used, and the resulting libraries were sequenced on 1/8th of a lane in a Illumina HiScanSQ located at the ARC Biotechnology platform in Pretoria, South Africa.

2.2.7 Illumina data analysis

Illumina paired-end datasets were analyzed using CLC Genomics version 5.1. Raw reads were imported as paired-end (distance of 180-250 nucleotides) and quality scores were assessed with the Illumina pipeline 1.8. Reads were filtered by removal of low quality sequences (limit of 0.05), removal of ambiguous nucleotides (maximum 2 nucleotides allowed) and removal of adapter sequences (Nextera v2 transposase 1 forward/reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ; Nextera v2 transposase 2 forward/reverse: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG) Quality control for each read dataset was performed using the FastQC function imbedded into the software. A set of 29 CTV full genome sequences (NZRB-G90 (FJ525432), NZRB-M12 (FJ525431), NZRB-TH30 (FJ525434), NZRB-M17 (FJ525435), NZRB-TH28 (FJ525433), Qaha (AY340974), T385 (Y18420), B165 (EU076703), CT11A (JQ911664), CT14A (JQ911663), VT (EU937519), T30 (EU937520), T36 (EU937521), A18 (JQ798289), Mexico (DQ272579), B301 (JF957196), Kpg 3 (HM573451), CTV (AF001623), NUaGa (AB046398), T68-1 (JQ965169), SP (EU857538), NZ-B18 (FJ525436), AT-1 (JQ061137), Taiwan-Pum/SP/T1 (JX266712), Taiwan-Pum/M/T5 (JX266713), T318A (DQ151548), VT defective RNA CTV-L5' complete sequence (AY206452), HA18-9 (GQ454869) and HA16-5 (GQ454870)) obtained from the NCBI database imbedded within CLC was imported and used in reference assembly, performed with the following settings: mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 1, similarity of 1, conflict resolution by voting (A, T, C, G) and non-specific matches ignored. *De novo* assemblies were also made with the same long-read settings described for reference mapping, except for the length fraction (0.5) and similarity (0.8). In addition, a minimum contig length of 200 nucleotides along with the creation of “full contigs” option was selected. *De novo*-generated contigs were subjected to a multiBLASTn search against the nucleotide collection hosted by the NCBI portal (http://www.ncbi.nlm.nih.gov/) via the CLC interface. Multiple sequence alignments were made with the online version of MAFFT (version 6.952) (http://mafft.cbrc.jp/alignment/software/), applying the following parameters: non-use of structural alignments, “same as input” for the direction of sequences, “aligned” as output order, “automatic” strategy, amino acid scoring matrix to BLOSUMM62, nucleotide matrix scoring of “1PAM / k=2”, a gap opening penalty of 1.53 and an offset value of 0.0. Alignments were imported into BioEdit v7.1.3.0 (Tom Hall, Isis pharmaceuticals, Inc. 1997-2004). Phylogenetic analysis were performed using MEGA version 4 (http://www.megasoftware.net/). Unrooted dendrograms were inferred using the

Neighbor-joining method with a bootstrap test of a thousand pseudo-replicates. Evolutionary distances were determined by applying the Jukes-Cantor base substitution model. Recombination analysis was performed by RDP (version beta 4.16) by selecting the "X-over" function (accessible from http://darwin.uvigo.es/rdp/rdp.html).

2.2.8 PCR cloning, Sanger sequencing & phylogenetic analysis of the p23 gene

Total RNA from sub-isolate 12-7 was submitted to PCR amplification according to Rubio *et al*. (2001) using primer pair PM50/PM51 and A-F/A-R (Table 1) (IDT, USA). Amplicons were cleaned with the Wizard SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA) and cloned into the pGEM-T Easy vector (Promega) according to the manufacturer's instructions. Thirty recombinant plasmids were extracted and PCR amplified with vector-specific primers SP6 and T7 (IDT, USA). Amplicons were then cleaned and sequenced in one direction in an ABI Prism 3130XL Genetic Analyzer (Applied Biosystems, Foster City, California, USA). BioEdit and the MAFFT algorithm (see Section 2.2.7) were used to trim and perform a multiple sequence alignment, respectively. Phylogenetic relationships were determined as previously described in Section 2.2.7.

2.3 RESULTS

2.3.1 Illumina sequencing and data analysis

A total of 33,733,904 raw 100bp-long paired-end reads were obtained, over all three preparation methods (Table 2) (raw data available in Appendix 5 on supplementary DVD). All templates yielded datasets which showed high sequence duplication levels (>80%) and PHRED scores between 37 and 38, which was indicative of very good read quality.

Table 2: Illumina data statistics, encompassing dataset quality scores (PHRED) and raw reads processing.

Template	***Average PHRED score (/40)***	***Total reads (unfiltered)***	***Total reads (post-trim)***	***% reads trimmed***	***Average length after trim***
Double-stranded RNA	37/40	10, 804,032	10,796,479	99.92	79.1
Total RNA	37/40	11,897,404	11,887,387	99.93	78
Virus particles	38/40	11,032,468	11,022,613	99.92	76.6

Parameter optimization for mapping reads against CTV reference genotypes

Calibration of read mapping parameters was first determined prior to further analysis. In each trial run, reads were mapped against a set of 29 reference CTV genomes. Two critical assembly parameter values, “read fraction length” (FL) and “similarity” (S), which determine the stringency at which reads match any given reference sequences were first optimized for this study. The default values of the software were first utilized, where FL=0.5 and S= 0.8, to assess if it was sufficient to yield acceptable mapping counts. These parameters (FL=0.5; S= 0.8), required that a minimum of 50% from a given read length had to map at least 80% of a given reference sequence. Various FL and S values were tried and it was found that a value of 0.99 (a value of 1 is not allowed by the software) for both was most suitable, as it would ensure that 99% of the read length maps 99% of a given reference sequence, thus mapping regions in a specific manner. Two additional parameters were also tested, which dealt with how non-specific reads should be handled by the software while it mapped against reference sequences. These handling options were: “mapped randomly” and “ignore”. The first option dictated that if a read obtained more than one match (and therefore maps to multiple references), it should be mapped randomly across all references. The second option (“ignore”), dictated that if a read obtained more than one match, it should be ignored or, in effect, be discarded from the input read dataset during the mapping process.

A total of four parameter combinations were tested on a single randomly chosen template (virus particles) obtained by immunocapture:

#1. Mapped randomly / (FL=0.5; S= 0.8)
#2. Mapped randomly / (FL=0.99; S=0.99)
#3. Ignored / (FL=0.5; S= 0.8)
#4. Ignored / (FL=0.99; S=0.99)

Figure 7 shows the different combinations of these parameter values and their associated effects on mapping counts and consensus lengths obtained for each reference genome. In combination 1, read counts and consensuses were the highest obtained for all reference genomes when compared to the other parameter combinations. This was expected due to the allowance of reads to map to more than one reference sequence combined with low mapping stringency for each read. These results are not helpful in trying to determine which genotypes are actually present in the sample, as all reference genomes yielded similar mapping values, due to the presence of conserved regions. In combination 2, a higher stringency for read mappings was applied, but random read mapping redundancy was allowed. This affected both consensus lengths and reads counts, which were reduced by about half compared to combination 1. The top ranking genotypes remained identical (T68-1, SP, B165, VT, NZ-B18) to those of parameters combination 1, but there still remained too many read counts and lengthy consensus sequences for any distinctions to be made. In combination 3, which forbade read mapping redundancies, almost identical read counts were obtained as in combination 2, but shorter consensus lengths were obtained, since mapped reads were specific to any given reference. Finally, in combination 4, the strictest parameters were applied, where high stringency in read mapping and only specific reads were utilized. For 11 of the CTV genotypes, read counts and consensus lengths were essentially zero. Only two genotypes (SP and VT) had relatively high values remained in significantly higher values as compared to the others. However, the filtering in unique reads induced a dramatic decrease in mapping metrics, which should not be interpreted as an indication of poor template quality, as these values depend hugely on the mapping parameters (Table 3). Based on this data, all further analyses on other samples was applied by using parameter set 4, since it has proved to best filter "background noise" genotypes.

Read mapping for the remaining templates was made using optimal parameters obtained. For each dataset, three key metrics were recorded: read count (RC), average coverage (AC, not shown in graphs throughout as its presence would not be clearly visible due to the scale) and consensus length (CL). The results are shown in Figure 8.

For dsRNA, the Tai-SP reference genome yielded the highest CL (7281), while VT had the highest RC (47340) and AC (188.5), but a lower CL (4034). Other genomes with relatively higher metrics than the whole was SP, T68-1 and CT14A. The rest of the genotypes averaged a CL of 344, RC of 106 and AC of 0.35. With total RNA as template for NGS, the Tai-SP was almost not represented (CL=70; RC=13; AC=0.2). With this template VT (CL=4057; RC=5083), SP (CL=3763; RC=5194) and T68-1 (CL=3191; RC=2202) constituted the top three rankings, but not in terms of AC which was 0.0 for all three. However, T318A and Tai-M rose to the top in terms of AC which was 21.7 and 21.5, respectively. Lastly for virus particles as template, a similar pattern as seen in total RNA was observed. VT (CL=3099; RC=28772; AC=122.4), SP (CL=2936; RC=28686; AC=120.1) and T68-1 (CL=1940; RC=1318857; AC=79.4) constituted the top three hits, except this time they also obtained the highest AC as well for this template as compared to total RNA. The remaining of the genotypes were very poorly represented, as also observed for the other two templates.

Table 3: The effects of parameter combinations on read mapping metrics (consensus length, CTV read count and average coverage) against a single template dataset (virus particles).

Parameter combination #	Average consensus length in nucleotides (min/max)[1]	Total CTV read count	% CTV reads from dataset	Average coverage (min/max)[1]
1	7767.62 (4405/12165)	724483	6.57 %	94 (2.5/4089)
2	3159.72 (1631/6812)	294357	2.67%	39.95 (0.4/229.8)
3	2134.62 (472/6767)	221230	2.007%	30.54 (0.09/312.2)
4	543.68 (8/3027)	77343	0.701%	11.13 (0.00/121.3)

[1]= minimum and maximal value for a given metric (min/max).

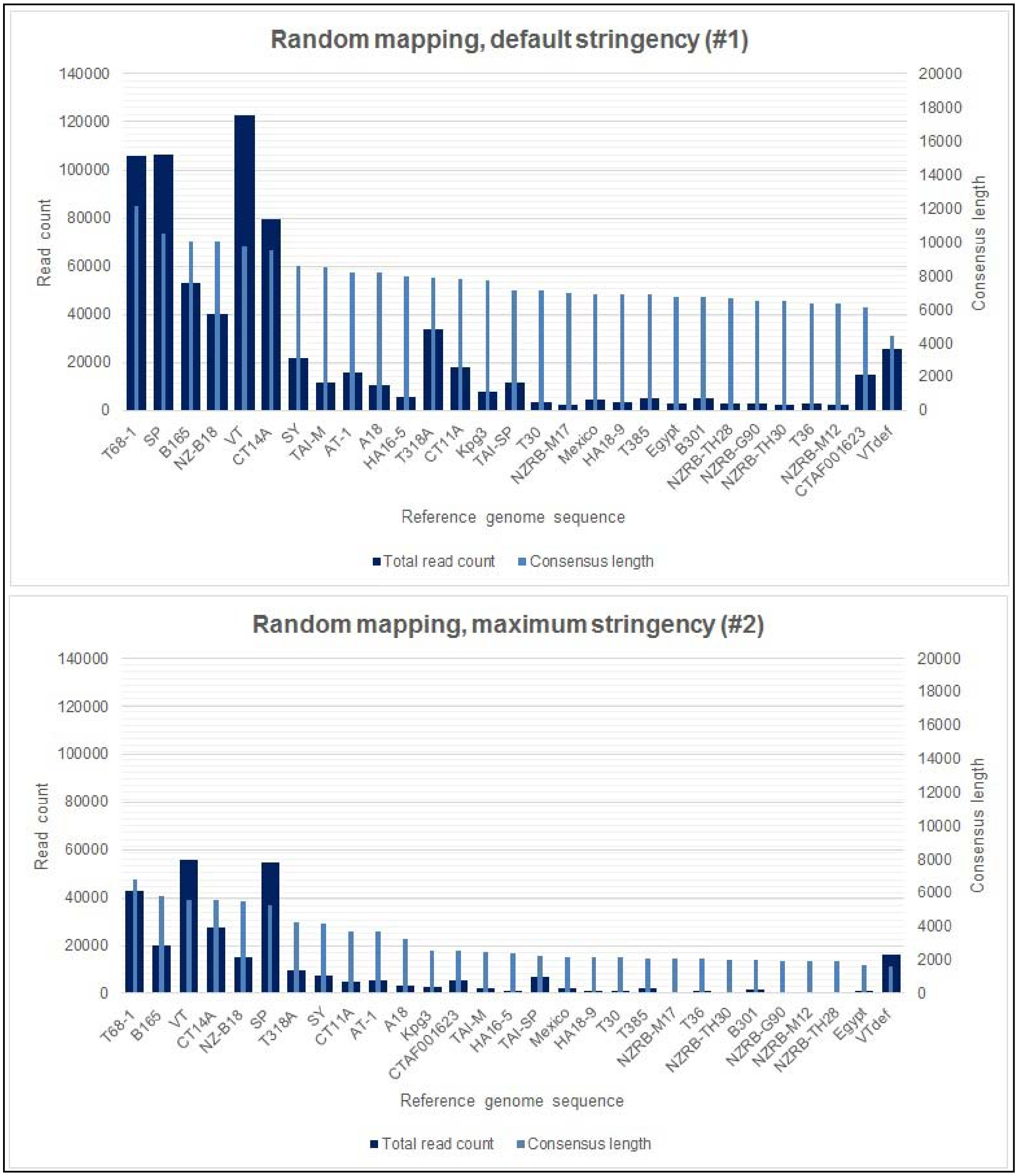


Figure 7: Read mapping results against 29 CTV reference sequences according to several parameter combinations (**#1:** reads mapped randomly/ default mapping stringency (FL=0.5/S=0.8); **#2:** reads mapped randomly/ maximum mapping stringency (FL=0.99/S=0.99); **#3:** redundant reads ignored/ default mapping stringency FL=0.5/S=0.8); **#4:** redundant reads ignored/ maximum mapping stringency FL=0.99/S=0.99)). FL= read fraction length %; S= read similarity %. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained from reads. The primary vertical axis (left) scale (amount of reads) is kept constant to 140000 throughout the graphs. The secondary vertical axis (consensus length, in nucleotides) is kept constant to 20000, the average length of a typical CTV genome.

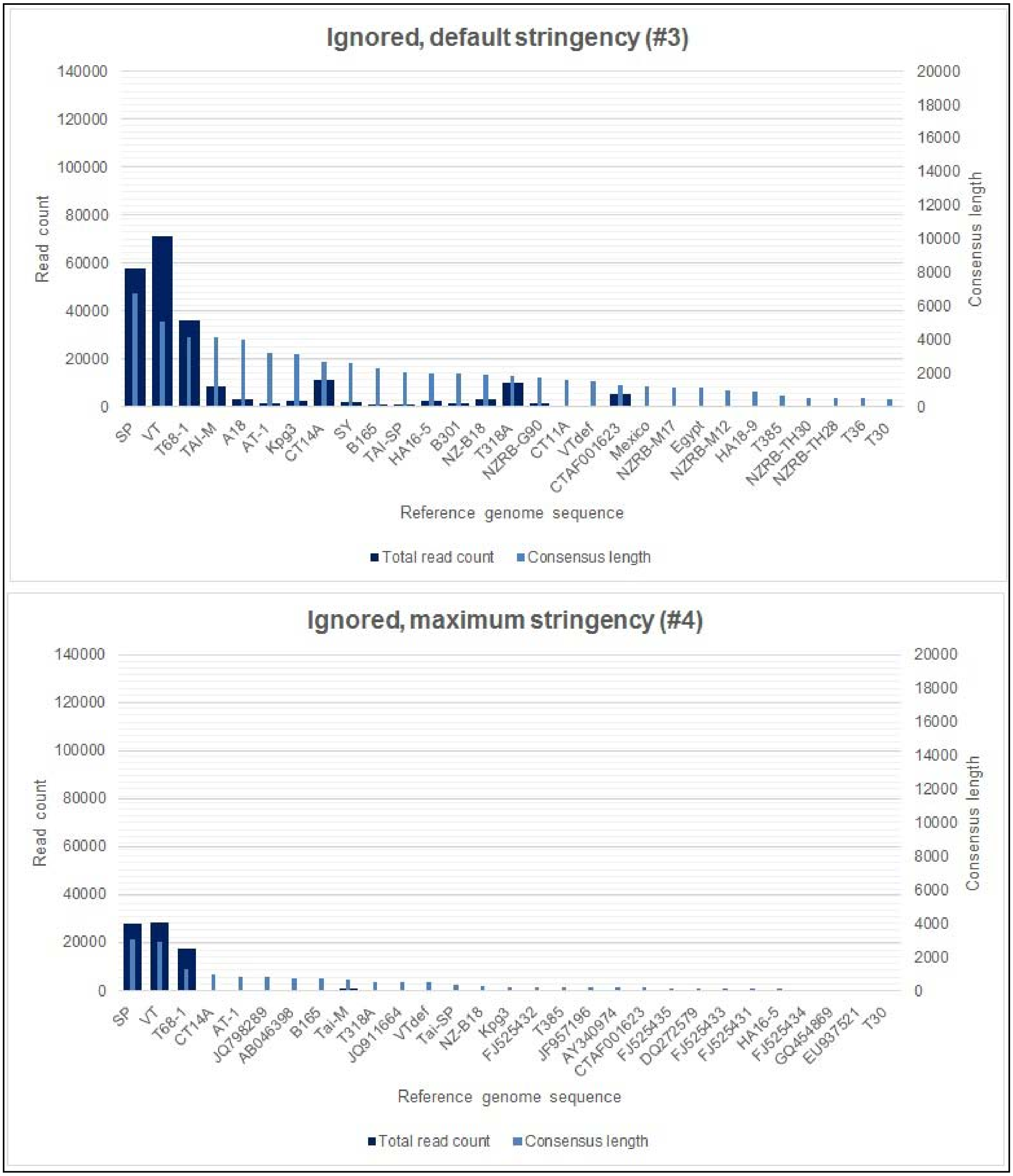


Figure 7: Continued

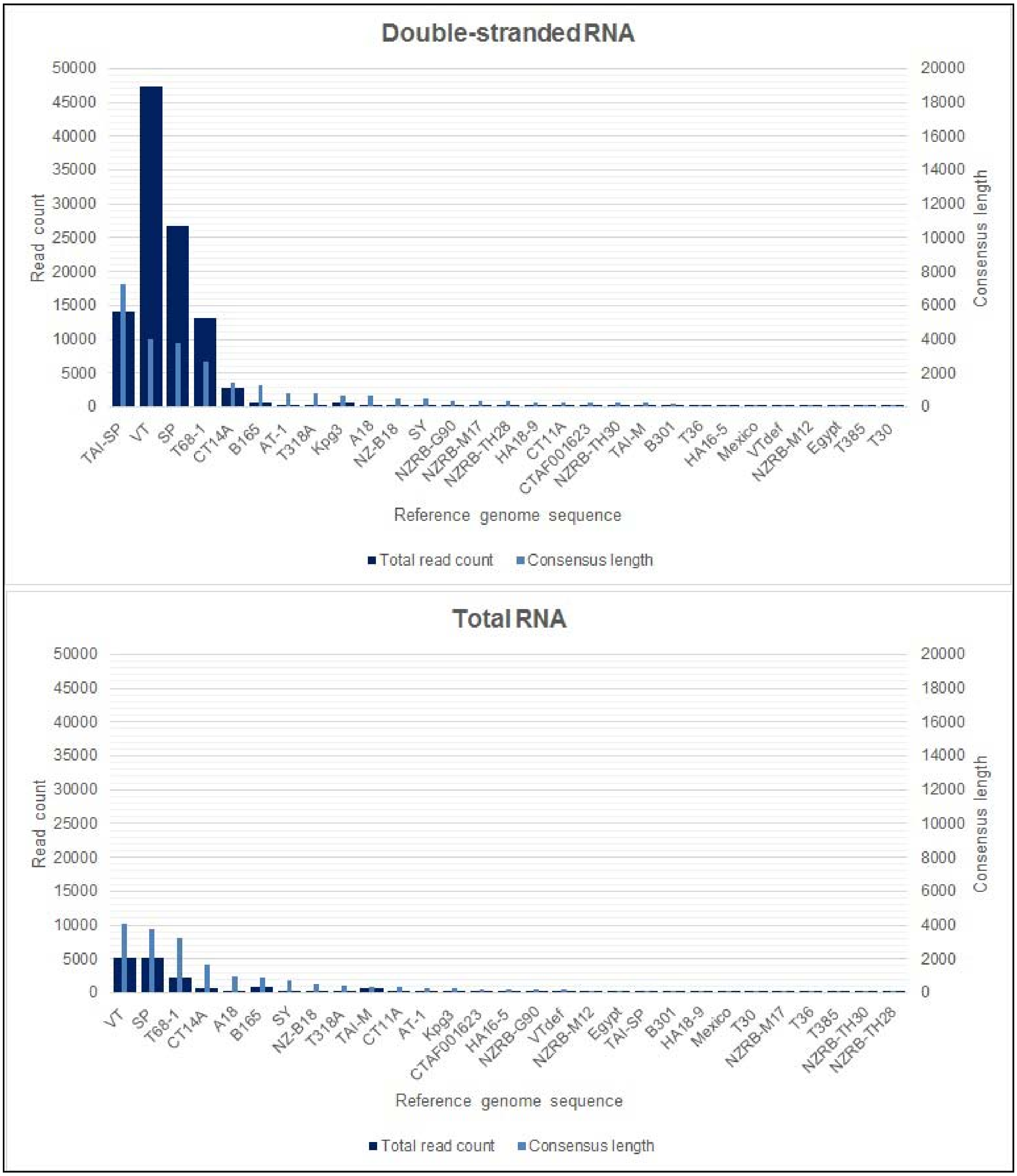


Figure 8: Read mapping results spanning 29 CTV reference genomes for each type of template of GFMS12 12-7. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale (amount of reads) is kept constant to 50000 throughout the graphs. The secondary vertical axis (consensus length, in nucleotides) is kept constant to 20000, the average length of a typical CTV genome.

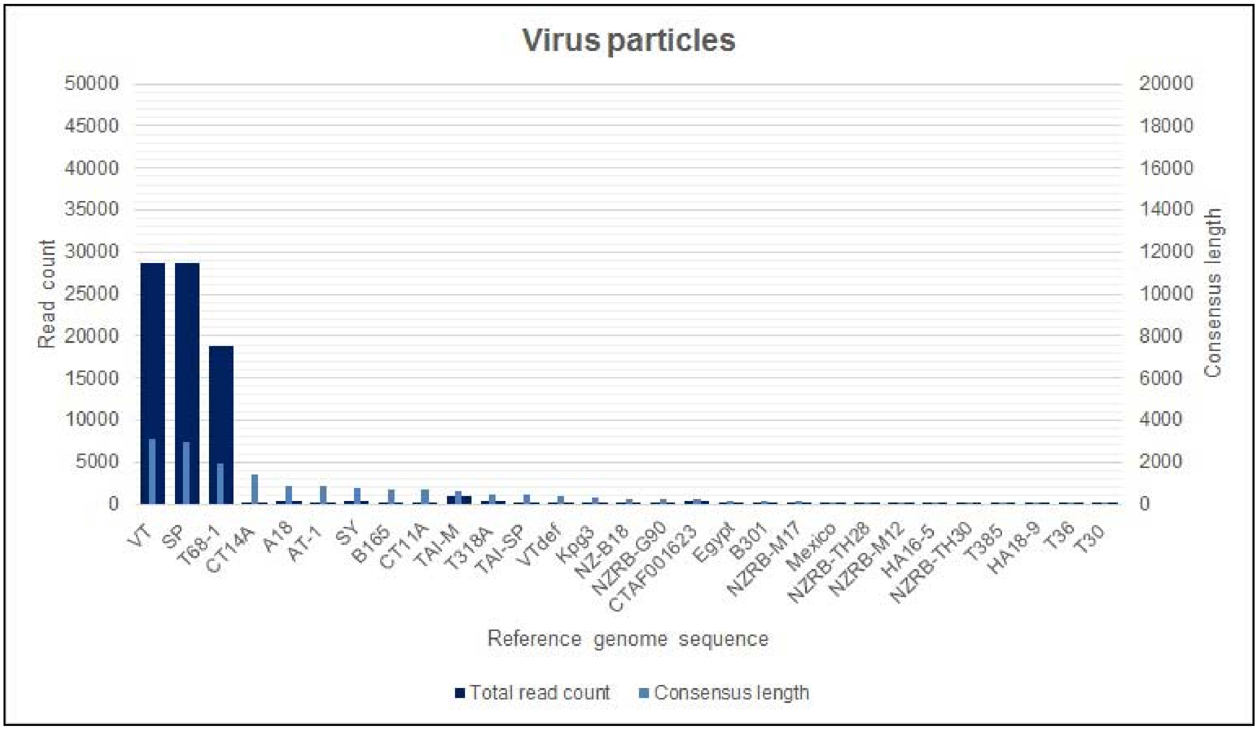


Figure 8: continued.

De novo assemblies

During mapping reads using the most stringent parameters (parameter set #4), none of the reference genome came close to having consensus length close to that of the whole genome (~19200 bp). This suggested that the genotype(s) present in the 12-7 may not be present in the reference set. Therefore, a *de novo* read assembly approach was attempted on each template dataset. Contig statistics are summarized in Table 4. To identify and isolate CTV-only contigs, a MultiBLASTn was performed. CTV-specific contigs were subsequently aligned with a single reference genome (T68-1, randomly chosen) to identify where these contigs mapped relative to the ORF organization of a typical CTV genome. The contigs were subsequently used to build longer contigs based on their overlapping regions (CLC could not do this in some instances, since some contigs required reverse complementation for correct alignment). This resulted in "master" contigs, one from total RNA (18801 nt) and one for dsRNA (4715 nt). In the case of virus particles, a single contiguous sequence could not be assembled, but instead four separate contigs of short lengths (1007 nt, 784 nt, 550 nt, 519 nt) were obtained.

For clarity, the "master contig" originating from total RNA will be referred to as "CT-ZA1" (Citrus Tristeza-Zuid Afrika1). Since it spanned the total, typical CTV genome length (~19200 nt), further analysis with this "partially" complete genome was made (next section).

Table 4: Summary statistics for *de novo* assemblies performed on each template dataset.

	DsRNA	***Total RNA***	***Virus particles***			
Input reads	10,796,479	11,887,387	11,022,613			
Matched reads	7,629,376	9,335,657	8,767,670			
Contigs generated	742	1067	951			
Min. length	195	199	201			
Max. length	5386	5469	5386			
Avg. length	450	408	404			
CTV master contig lengths	4715	18801	1007	550	519	784

Analysis of CT-ZA1

The total length of the CT-ZA1 construct was 18,801 nucleotides. A multiple alignment with the set of 29 reference genomes and the new construct was performed and used to construct a dendrogram showing its relationship with other genotypes (Figure 9). CT-ZA1 was found clustering within the same node as T68-1, a Floridian isolate. Inspection of the alignment showed several regions were missing in CT-ZA1. Regions 4691-4745, 4759-4930, 9485-9321 were missing when compared to cognate regions of the other reference genomes. The construct ended at position 19264, which represented the end of the p23 gene, which indicated a lack of the 3' UTR region. The whole genome alignment was also used for an RDP analysis, in order to assess any recombinant properties (Figure 10). It was found that the majority of the 3' half of the genome appeared to be VT in origin. Smaller fragments, overlapping with the VT fragment were also recombinations attributed to SP or of unknown origin. The 5' was relatively unique, except for the 1b portion of ORF1, which showed regions of similarities with HA16-5 and AT-1 genotypes. A 95.04% nucleotide identity was found

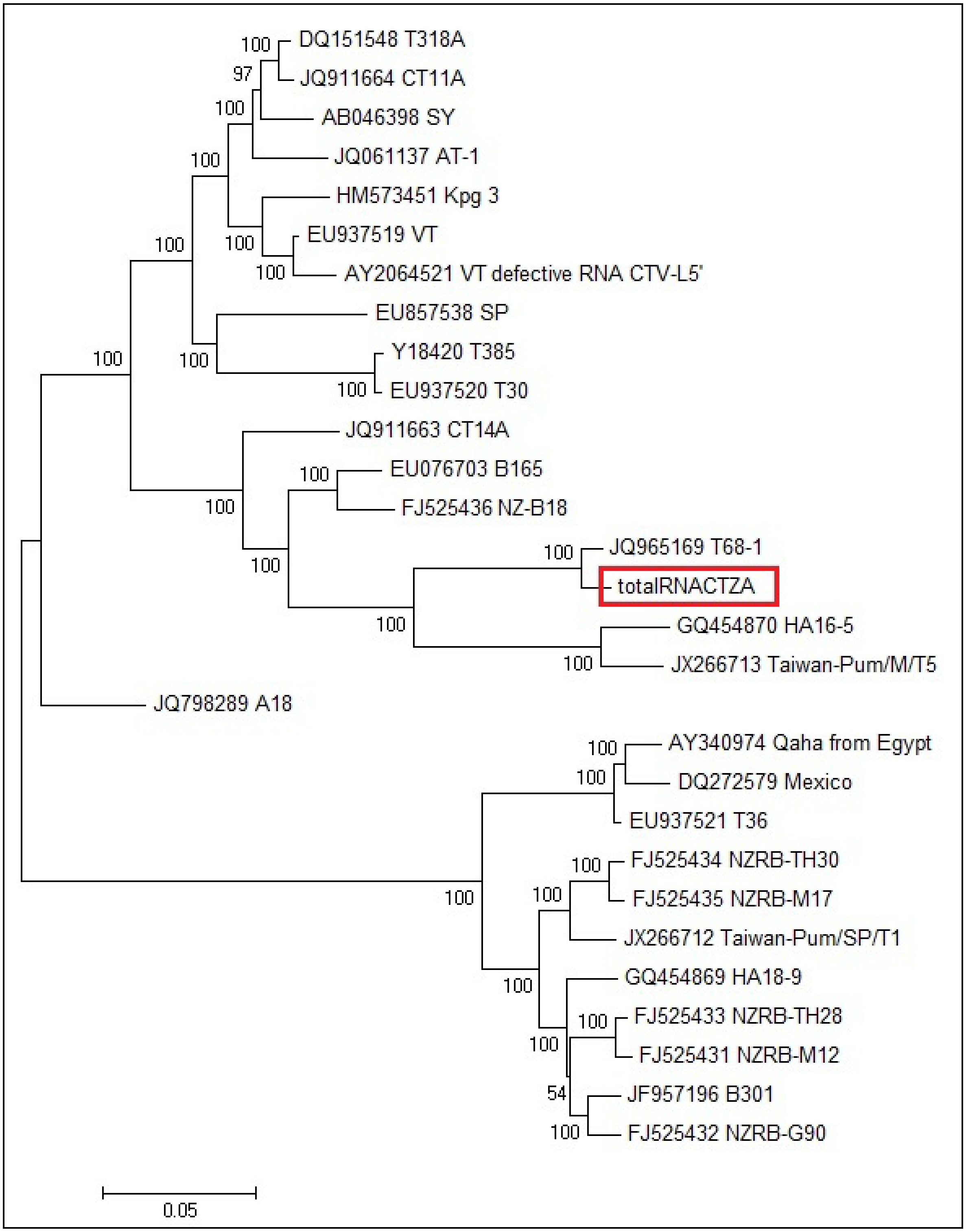


Figure 9: Neighbor-joining dendrogram showing the phylogenetic relationship between CT-ZA1 and other CTV genomes. Confidence levels are shown as bootstrap values at each node. In the dendrogram, “totalRNACTZA” represents the CT-ZA1 construct (within the red rectangle).

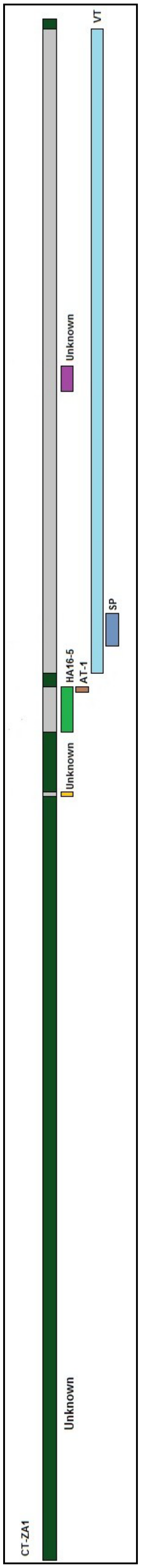


Figure 10: Visualization of the CT-ZA1 genome after an RDP analysis showing the regions of recombination originating from other CTV genomes. Colored blocks depict the full genome sequence starting from the far left (5') to the far right (3'). Block were color-coded to facilitate differentiation. Light green= HA16-5; dark blue= SP; light blue= VT, brown= AT-1, unknown= dark green or yellow or purple, and grey= serves to provide continuity to the full genome length on top while highlighting were multiple recombinations have occurred.

between T68-1 and CT-ZA1 genome sequences, specifically.

Read mapping using CTV reference genotypes including the novel construct, CT-ZA1

It was found in the previous section that CT-ZA1 represented a distinct genotype partially assembled from the total RNA dataset. Since all of the reference genomes used so far in mappings failed to yield consensus lengths approaching the average full CTV genome length (~19,200 nt), CT-ZA1 was included to determine whether other genotypes are present. With identical mapping parameters as previously used, Figure 11 shows the results of these mappings with the new construct.

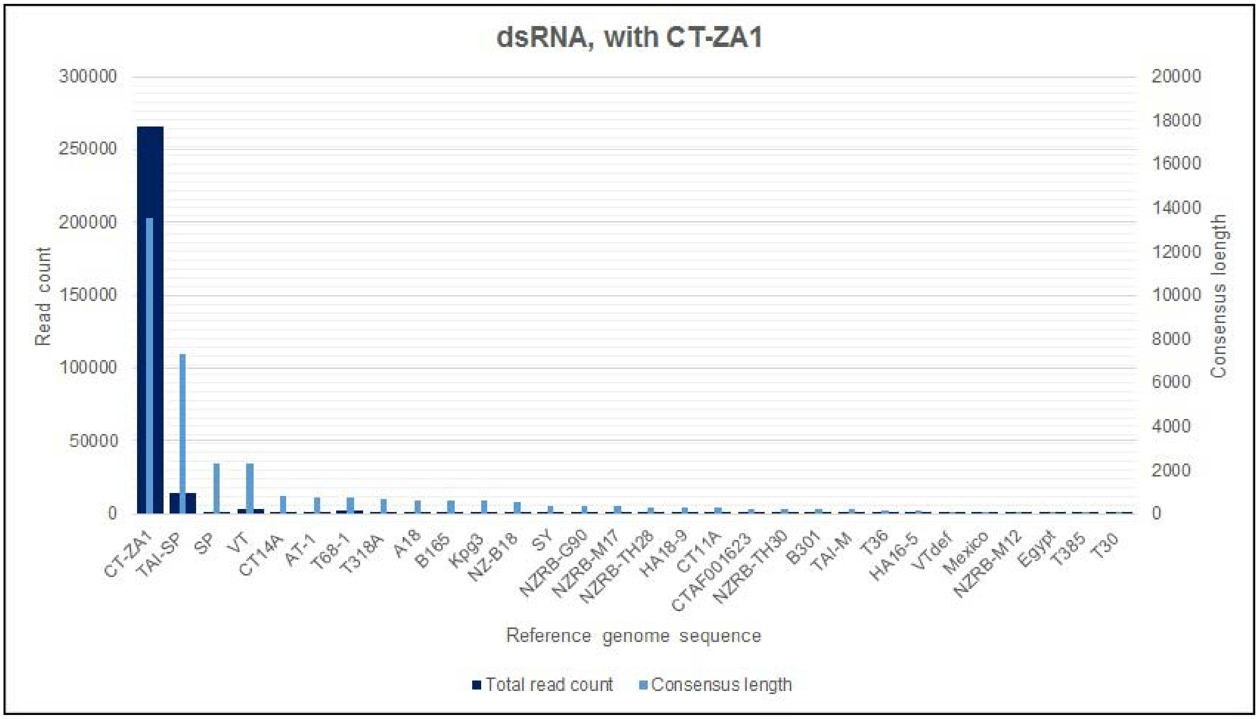


Figure 11: Read mapping results spanning 30 CTV reference genomes (including CT-ZA1) for each type of template of GFMS12 12-7. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale (amount of reads) is kept constant to 300000 throughout the graphs. The secondary vertical axis (consensus length, in nucleotides) is kept constant to 20000, the average length of a typical CTV genome.

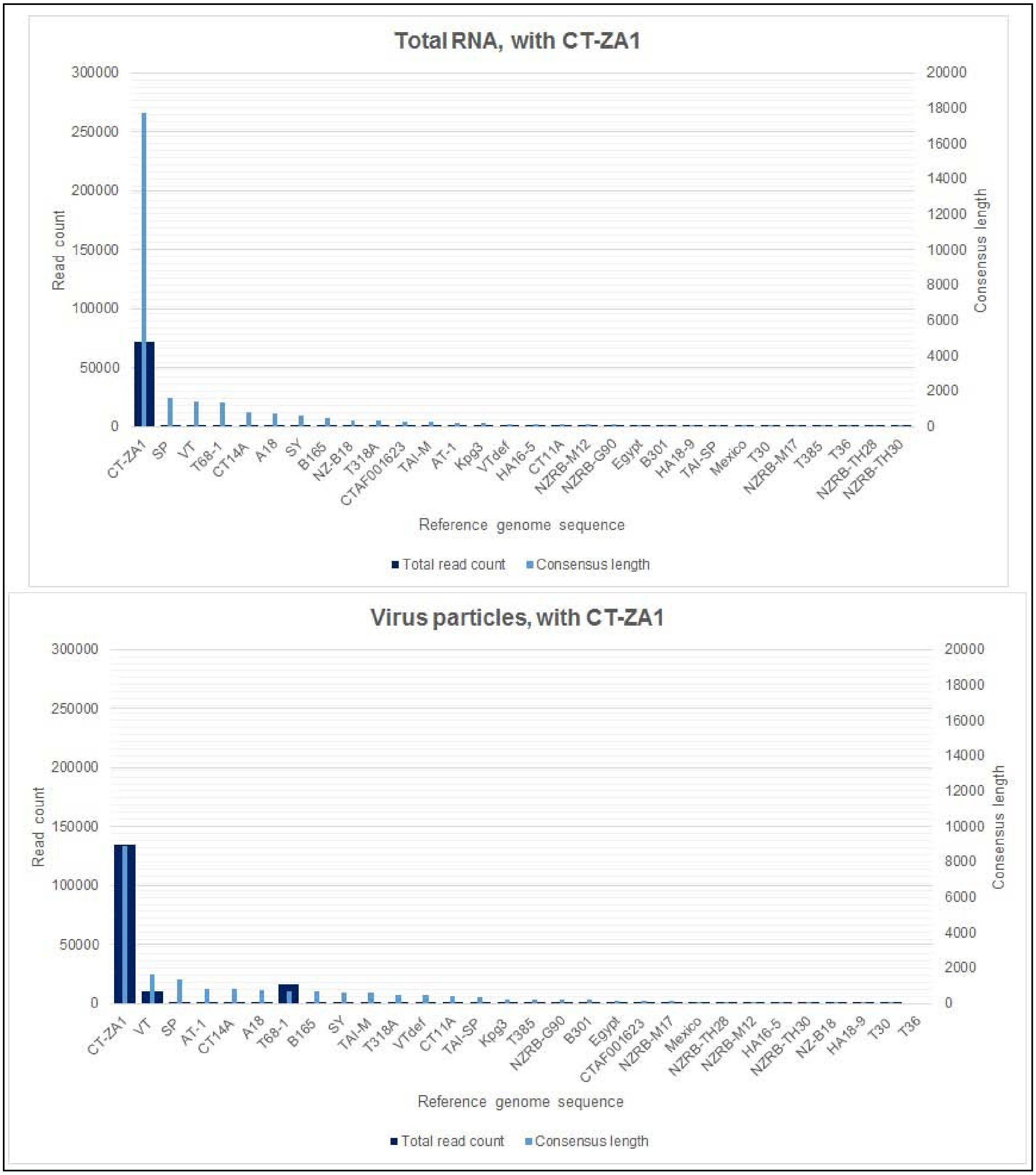


Figure 11: continued

In all templates, CT-ZA1 was the reference sequence that obtained the longest consensus lengths, read counts and coverage compared to other genotypes (SP, VT, and T68-1) that previously achieved the higher rankings. Expectedly, CT-ZA1 performed best pertaining to total RNA reads in terms of all mapping metrics (CL=17727; RC=72258; AC=325.8) since it was derived from the same dataset. Besides CT-ZA1 specific reads, 1045 CTV-like reads in total, which mapped to other genotypes, mostly distributed between SP (84 reads), VT (337 reads) and T68-1 (151 reads). These read mappings were visualized vis-à-vis their respective reference sequences and displayed in Figure 12. CT-ZA1 was almost fully mapped, with some scattered gapped regions. Read mappings for the next best hits (SP, VT, and T68-1) were extremely scattered across their respective genomes, with slight clustering within the 3' half for SP and VT genomes.

When mapped against the dsRNA read dataset, CT-ZA1 also achieved top metrics (CL=13491; RC=265892; AC=1190.2), followed by Tai-SP (CL=7317; RC=14619; AC=62.6) previously identified as the most represented genotype prior to including CT-ZA1. Discounting reads from these two references, there remained 6176 reads distributed between the remaining reference, but mostly shared between SP (386 reads), VT (2752 reads) and T68-1 (1917 reads). The read mappings derived from these genotypes, as well as from CT-ZA1 and Tai-SP are shown in Figure 13. VT and SP were mapped almost only within their 3' half, while Tai-SP was mapped mainly within its 5' half with slight mapping in the 3' half as well.

Lastly, mapping against virus particle reads showed a similar pattern as in the two previous templates, but with lower values for CT-ZA1 (CL=8910; RC=134341; AC=602.1). Despite this, CT-ZA1 appeared to be the most represented sequence. Additional genotypes remained in much lower read numbers (28045), but were mostly originating from VT (10455 reads) and T68-1 (15925 reads). Figure 14 shows the mapping patterns for the virus particle template. CT-ZA1 was mapped poorly compared to the other two templates, but still remained in top position in terms of CL value. VT and SP were mostly mapped against their 3' half, with the remaining reads mapped in a very scattered manner across their respective genotypes.

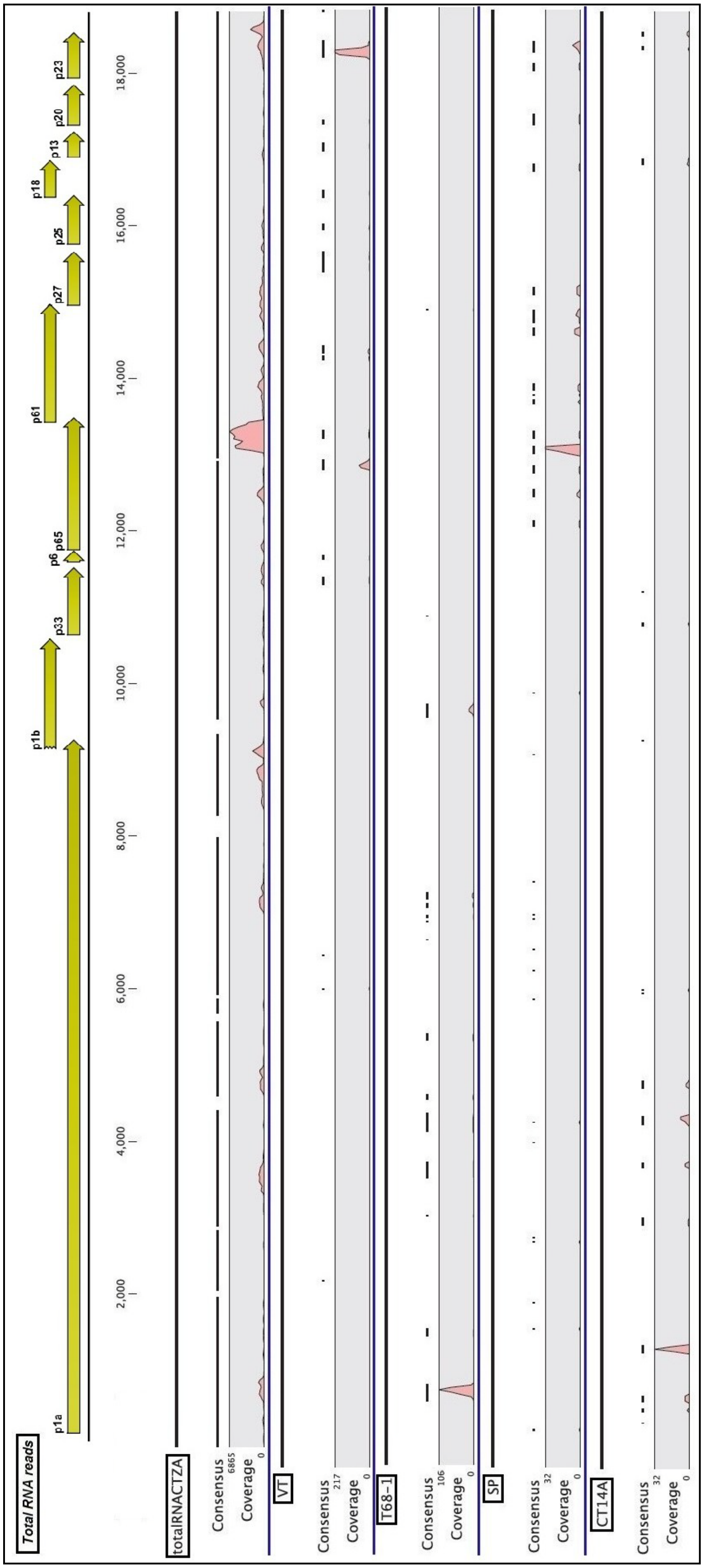


Figure 12: Visual mappings of total RNA-derived reads against the five mostly represented genotypes (CT-ZA1 or “totalRNACTZA”, VT, T68-1, SP and CT14A). The scale on top of the image represents the full-genome length in nucleotides. The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read counts. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

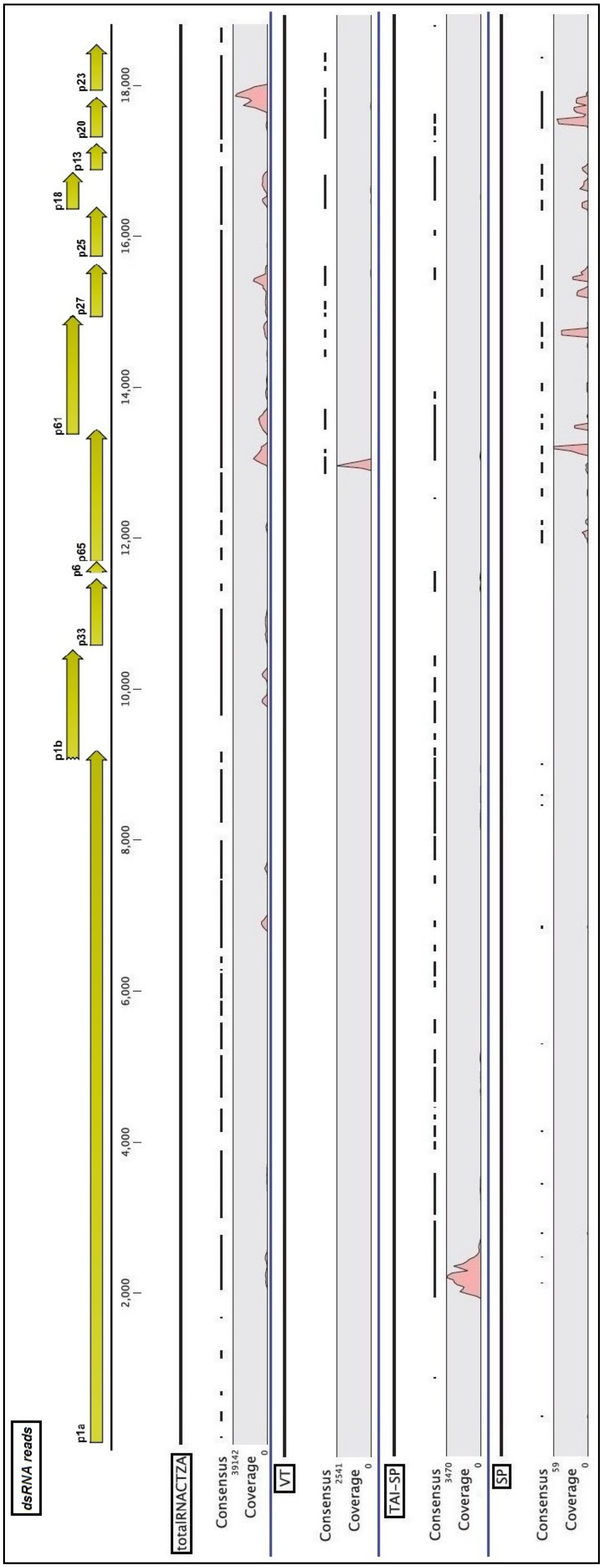


Figure 13: Visual mappings of dsRNA-derived reads against the four mostly represented genotypes (CT-ZA1 or "totalRNACTZA", VT, Tai-SP and SP). The scale on top of the image represents the full-genome length in nucleotides. The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read counts. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

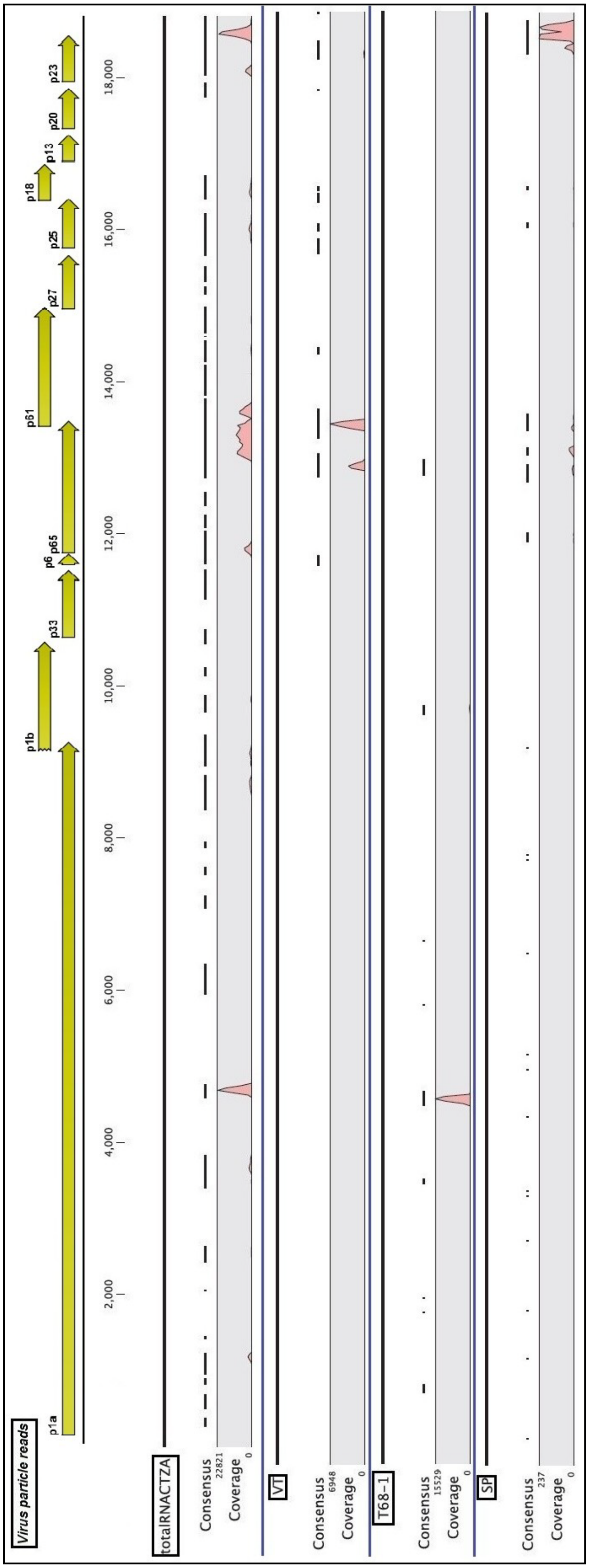


Figure 14: Visual mappings of virus particles-derived reads against the four mostly represented genotypes (CT-ZA1 or "totalRNACTZA", T68-1, SP and VT). The scale on top of the image represents the full-genome length in nucleotides. The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read count. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

2.3.2 A comparison with cloning-based phylogenetic analysis of sub-isolate 12-7 using specific genes

Clones generated during this study (of the p23 gene from 12-7), as well as previously obtained A-fragment clones (courtesy of K. Scott), were used as a means of comparison to validate NGS data. For this analysis data from total RNA extracts alone were utilized. Two neighbor-joining dendrograms were created for each genomic region, which both used 29 reference CTV genomes, but only one set containing CT-ZA1 additionally. Dendrograms depicting the A-fragment clones (Figure 15) indicated that when CT-ZA1 was not included, 12-7 clones grouped closely to a clade consisting of CT14A, NZ-B18, T68-1 and B165 (A). With CT-ZA1 included (B), no significant changes in terms of tree topology or clade groupings was observed. For p23 dendrograms (Figure 16), all clones clearly clustered at a common node with SP, a VT-like genotype (A). However, once the CT-ZA1 sequence was included, clones were almost identical to it (B).

Additionally, an "*in silico* cloning" experiment was tested. This entailed mapping read datasets against reference sequences that only spanned specific gene regions. In this case these were the same regions used in the cloning strategy described above (A-region and p23). This simulation provided the opportunity to not only serve as a comparison, but also to observe the effects of a large-scale analysis of clones (except here the clones are shorter, and are actually sequencing reads or "read clones") as well as testing multiple templates.

With dsRNA as the NGS template, both gene regions confirmed that CT-ZA1 was the dominant component (based on read counts). However, a second dominant component was found, but whose identity differed according to the gene region used. Tai-SP was found in high counts according to the 1a region and VT according to the p23 gene region. Additionally, a series of other genotypes were also identified, but at very low read counts (Figure 17).

In the total RNA 1a region, only two genotypes obtained hits, albeit at very low read counts: CT-ZA1 (n=54 and Tai-SP (n=2). For the p23 regions, a wider diversity of genotypes obtained hits, but still showed a dominance of CT-ZA1 (4334 reads) and VT (228 reads) compared to other genotypes.

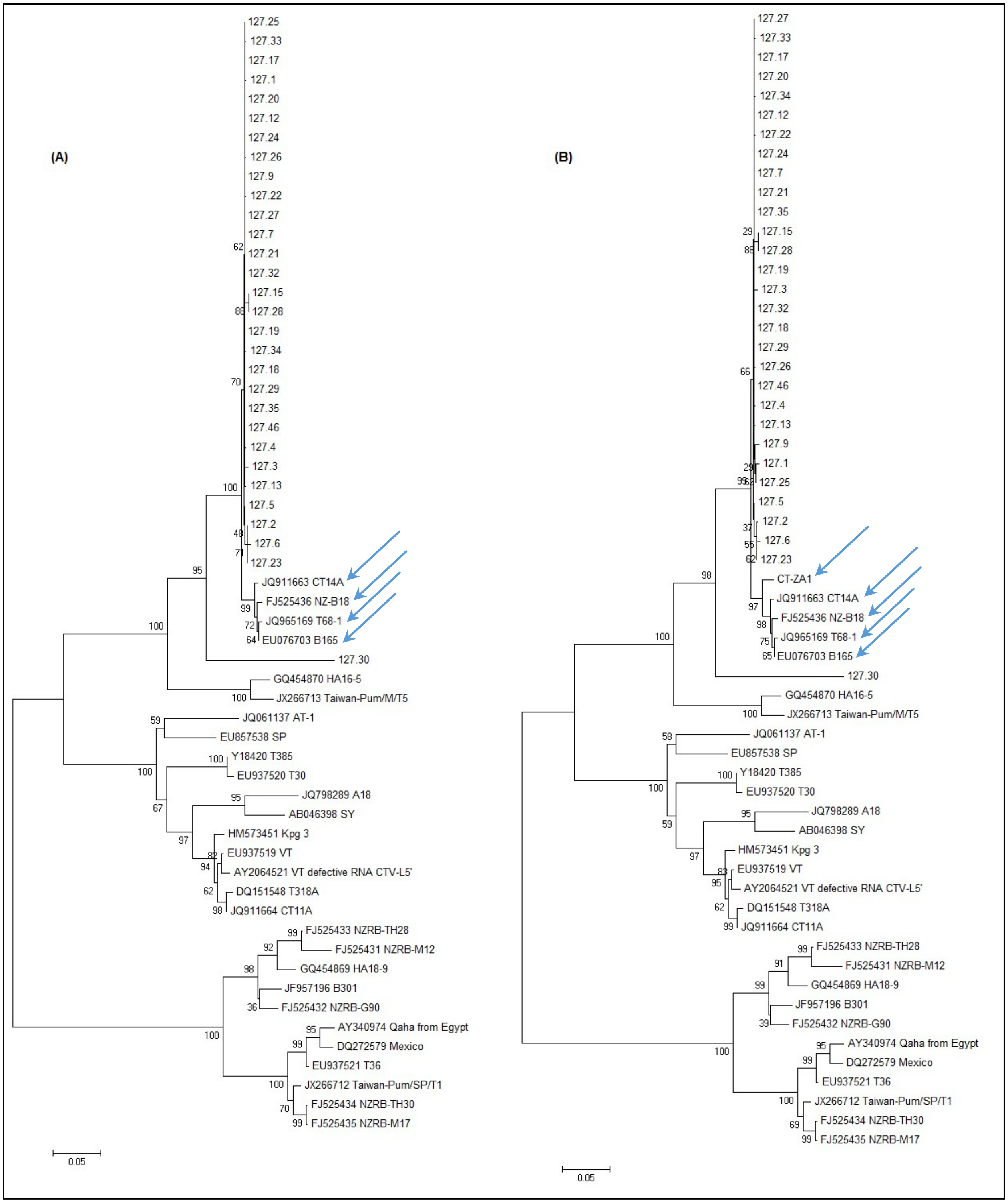


Figure 15: Neighbor-joining dendrograms showing the phylogenetic relationship between GFMS12 sub-isolate 127 A-region clones and cognate regions of other genomes. (A) *excluding* CT-ZA1 and (B) *including* CT-ZA1. Confidence levels are shown as bootstrap values at each node. Arrows indicate genotypes with closest homology to the clones.

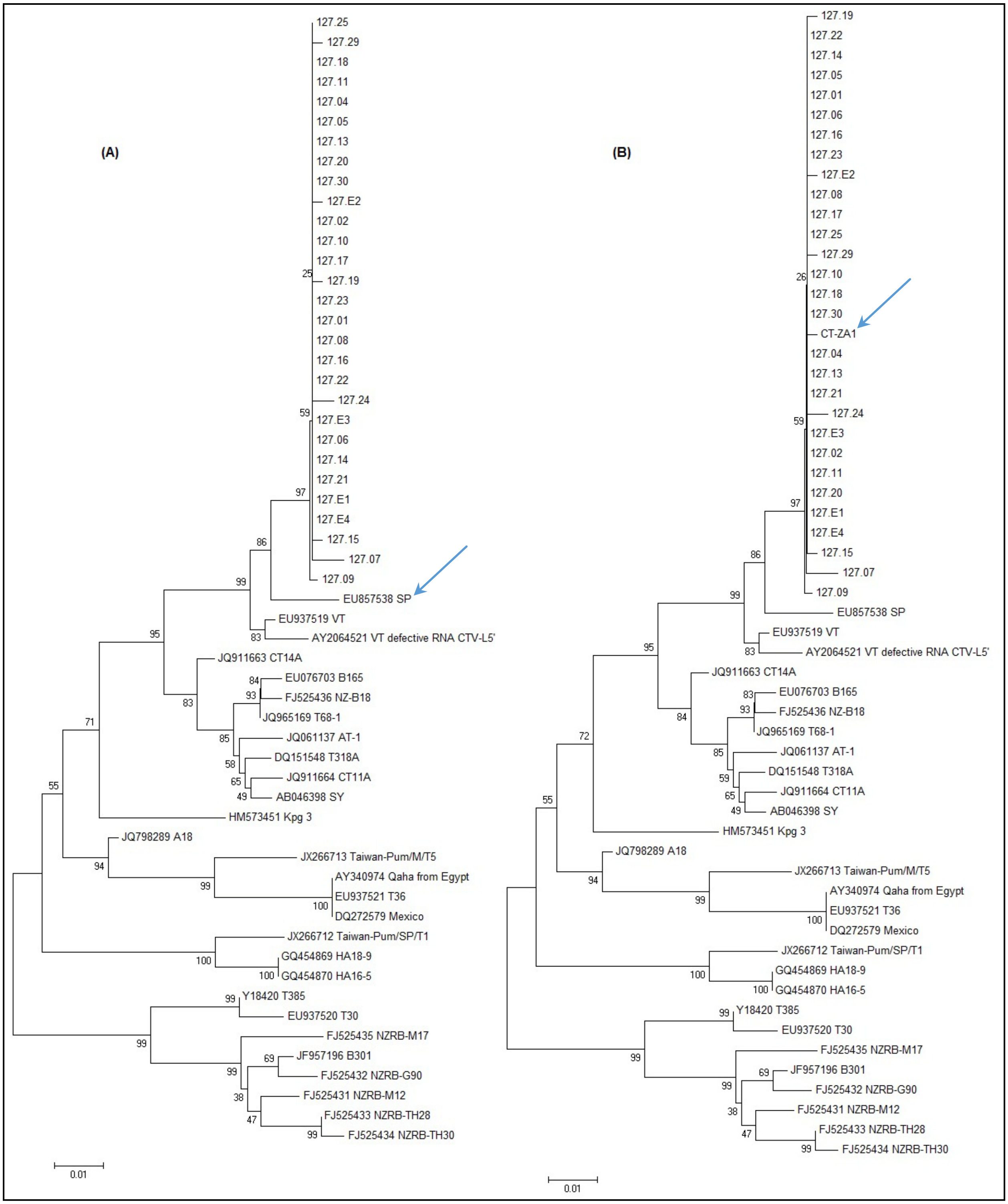


Figure 16: Neighbor-joining dendrograms showing the phylogenetic relationship between GFMS12 sub-isolate 12-7 p23 gene clones and cognate regions from other genomes. (A) *excluding* CT-ZA1 and (B) *including* CT-ZA1. Confidence levels are shown as bootstrap values at each node. Arrows indicate genotypes with closest homology to the clones.

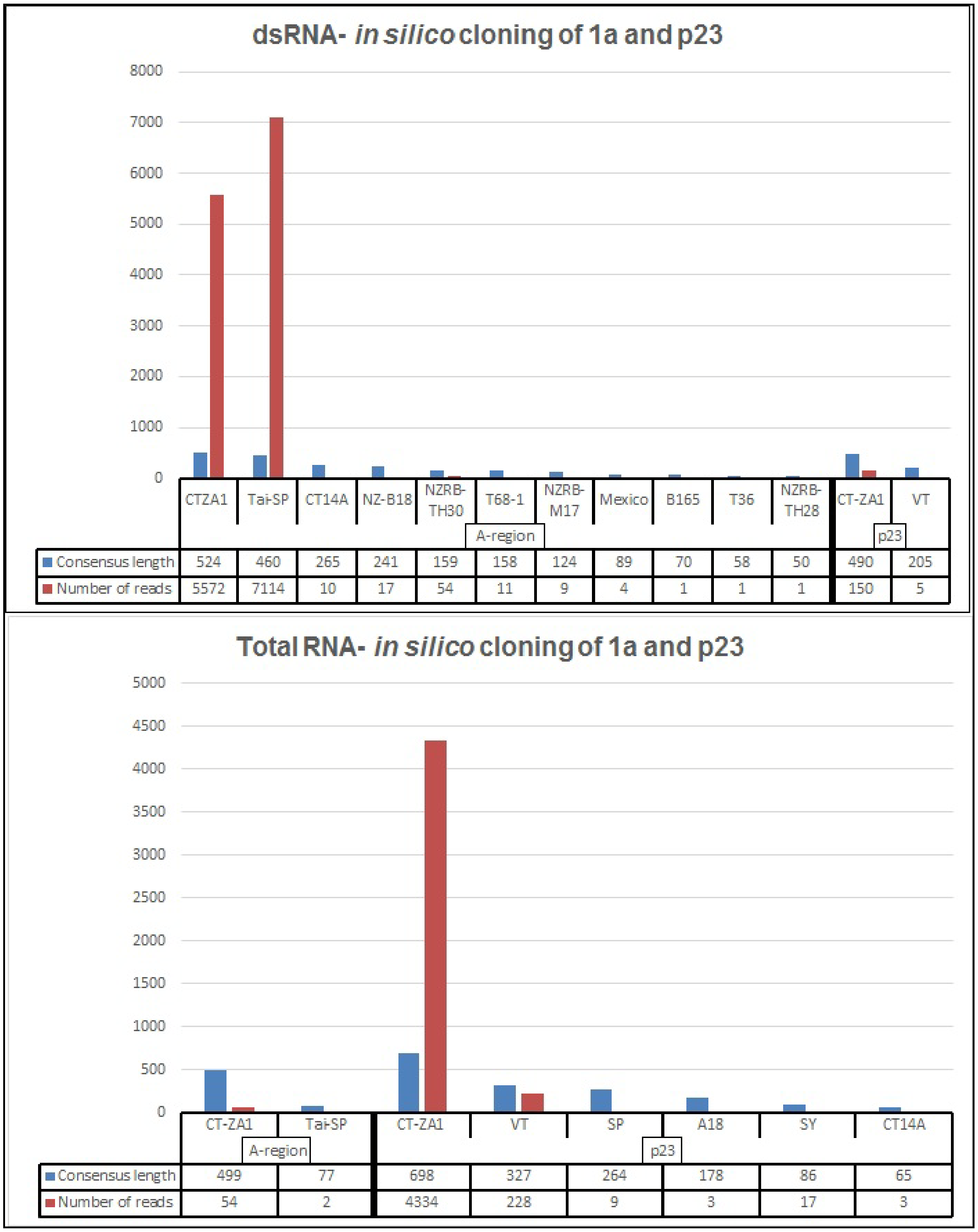
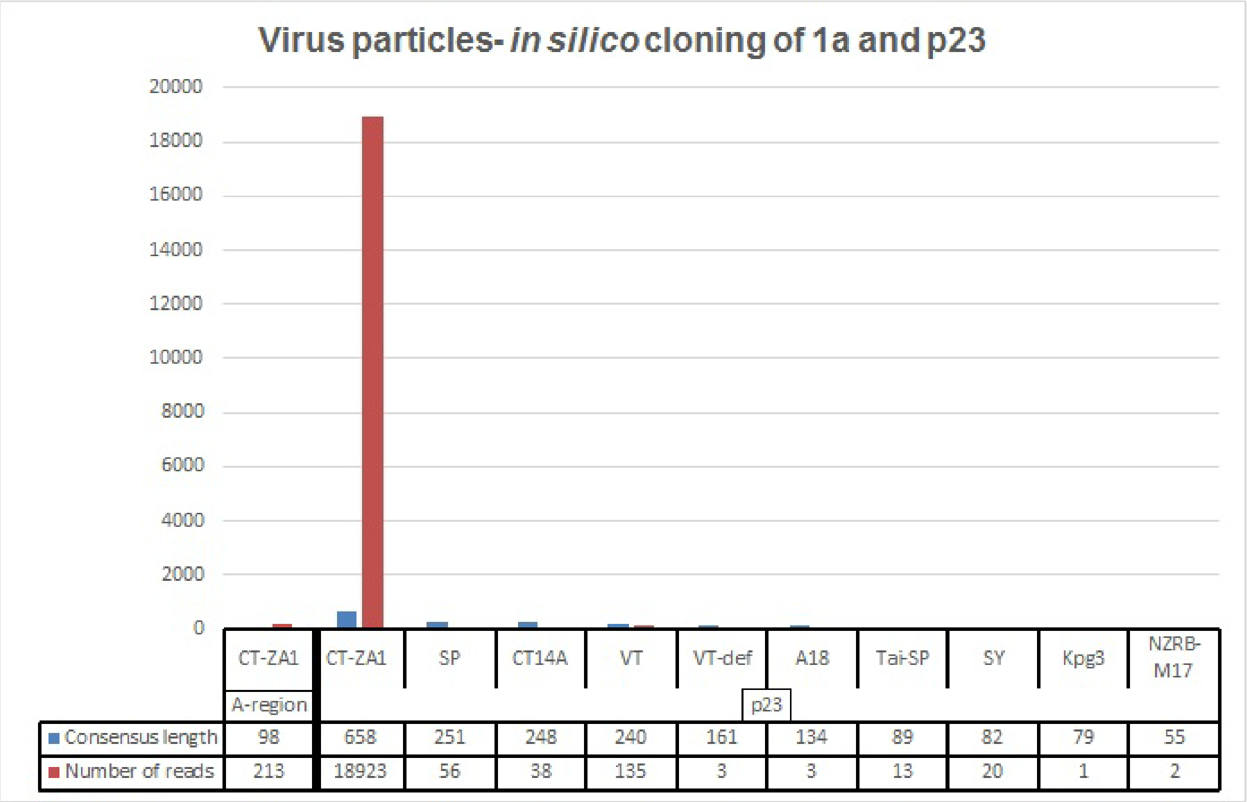


Figure 17: Genotype distributions according to *in silico* cloning for both the A-region and p23 gene for each template dataset. The mapped data was ranked from largest to smallest in terms of consensus length. The vertical axis scale represents the consensus length and read count. Genotypes identities that are missing from these graphs meant that they obtained values of zero during the mapping process.



	CT-ZA1	CT-ZA1	SP	CT14A	VT	VT-def	A18	Tai-SP	SY	Kpg3	NZRB-M17
	A-region	p23									
🟦 Consensus length	98	658	251	248	240	161	134	89	82	79	55
🟥 Number of reads	213	18923	56	38	135	3	3	13	20	1	2

Figure 17: Continued.

For the virus particles template, the 1a region results were unequivocal, since mapping showed CT-ZA1 as the only match (albeit at very low number, n= 213). However, the p23 region contained reads of various genotype identities. CT-ZA1 was the dominant component with 18923 reads, while VT was the second most common match with 135 reads. The remaining genotypes obtained moderate consensus lengths, but were all represented with very low read counts.

Determination of optimal template choice

One of the goals of this chapter was to determine which template would function best for genotyping CTV isolates on a whole genome level. This can be a complex decision to make, as it depends on which values one chooses to look at (Table 5).

As tested previously, mapping metrics (read count, consensus length, and coverage) can drastically fluctuate depending on how stringent read mappings are calibrated to. It is therefore suggested that for the consideration of which template surpasses the others, it should be based on the final parameter set tested (#4), which was used throughout this analysis. In this regard, whether or not an appropriate reference genome was present, the values remained in a near identical ratio. But for the sake of comparison, mapping metrics *before*- and *after*- the inclusion of CT-ZA1 are displayed in the table. In terms of consensus length, dsRNA yielded the lengthiest, followed by total RNA and virus particles. For total read counts specific for any CTV genotype, dsRNA was also best, followed by either total RNA or virus particles, depending if CT-ZA1 was/was not included. For average coverage values, dsRNA achieved the best values, followed by virus particles and total RNA. From this data, dsRNA seemed to be the template of choice. However, due to the ability of reads from the total RNA dataset to assemble into a lengthy *de novo* sequence, this template may also be considered as an alternative to dsRNA or may be used in concert.

Table 5: Read mapping metric comparison between templates in terms of consensus length, total read counts and average coverage. The comparison is further extended by showing these values *before*- and *after*- the addition of CT-ZA1 in the reference set.

Mapping timeline	Template	Average consensus length (min/max)[1]	Total read count (% from dataset)[2]	Average coverage (min/max)[1]
Pre- CTZA1	dsRNA	947.5 (18/7281)	106535 (0.98)	14.6 (0.0/188.4)
	Total RNA	643.2 (12/4057)	15765 (0.13)	2.2 (0.0/21.7)
	Virus particles	543.6 (8/3027)	77343 (0.70)	11.1 (0.0/121.3)
Post-CTZA1	dsRNA	1139.9 (18/13491)	286687 (2.66)	45.5 (0.0/1190.2)
	Total RNA	922.3 (9/17727)	73303 (0.61)	10.97 (0.0/325.8)
	Virus particles	679.7 (0.0/8910)	162386 (1.47)	24.1 (0.0/602.1)

[1]= minimum and maximal value for a given metric (min/max).

2.4 DISCUSSION

In this chapter, a high-throughput sequencing pipeline was developed to characterize a known CTV-infected source previously analyzed by the cloning and sequencing of multiple genes. In this new pipeline, five key components were assessed: 1) the most appropriate template enrichment method, 2) a whole-genome amplification protocol, 3) Illumina paired-end sequencing technology, 4) the analytic capabilities of a single software suite (CLC Genomics) and 5) validation of NGS data by comparison with cloning results.

The overall methodology and bioinformatics pipeline used in this chapter is summarized in Figure 18. The main software suite that was used for handling raw reads, CLC Genomics version 6.0.1, was found to be very efficient and user-friendly. A major advantage of this software lies in its multi-functionality, as it encompasses raw reads processing, trimming (quality- and adapter-based) and assembly (reference-guided or *de novo*). Another useful feature is the software's visualization capabilities, which form a crucial part in the analysis. In terms of output, multiple identical analyses were repeated several times, which were found to be highly reproducible. The main disadvantage of CLC is its high demand in computational resources (e.g. amount of available RAM), specifically for *de novo* assembly involving more than ~10 million reads. Other *de novo* algorithms such as Velvet (Zerbino and Birney, 2008) are more resource friendly, but take longer processing periods.

A crucial first step for the development of the pipeline was the calibration of read mapping parameters to optimal levels. This was very important, as results derived from each parameter set can highly vary and therefore influence greatly an investigator's interpretations of the data. Four parameter combinations were tested, and drastic effects in terms of read mapping behavior were observed. In the least stringent parameter set (#1, with random read mapping allowed, length fraction to 0.5 and similarity to 0.8) , many reference genome obtained a high number of reads, long consensus lengths and high coverage, which made distinction between genotypes nearly impossible. This situation was probably because reads could map identical regions across all CTV reference genomes due to conserved genomic regions. With the most stringent parameter set (#4, removal of read mapping redundancy, length fraction to 0.99 and similarity to 0.99), most reference genotypes yielded counts of zero in all metrics, which indicated that this parameter was best to eliminate “background noise genomes”, therefore providing enhanced resolution.

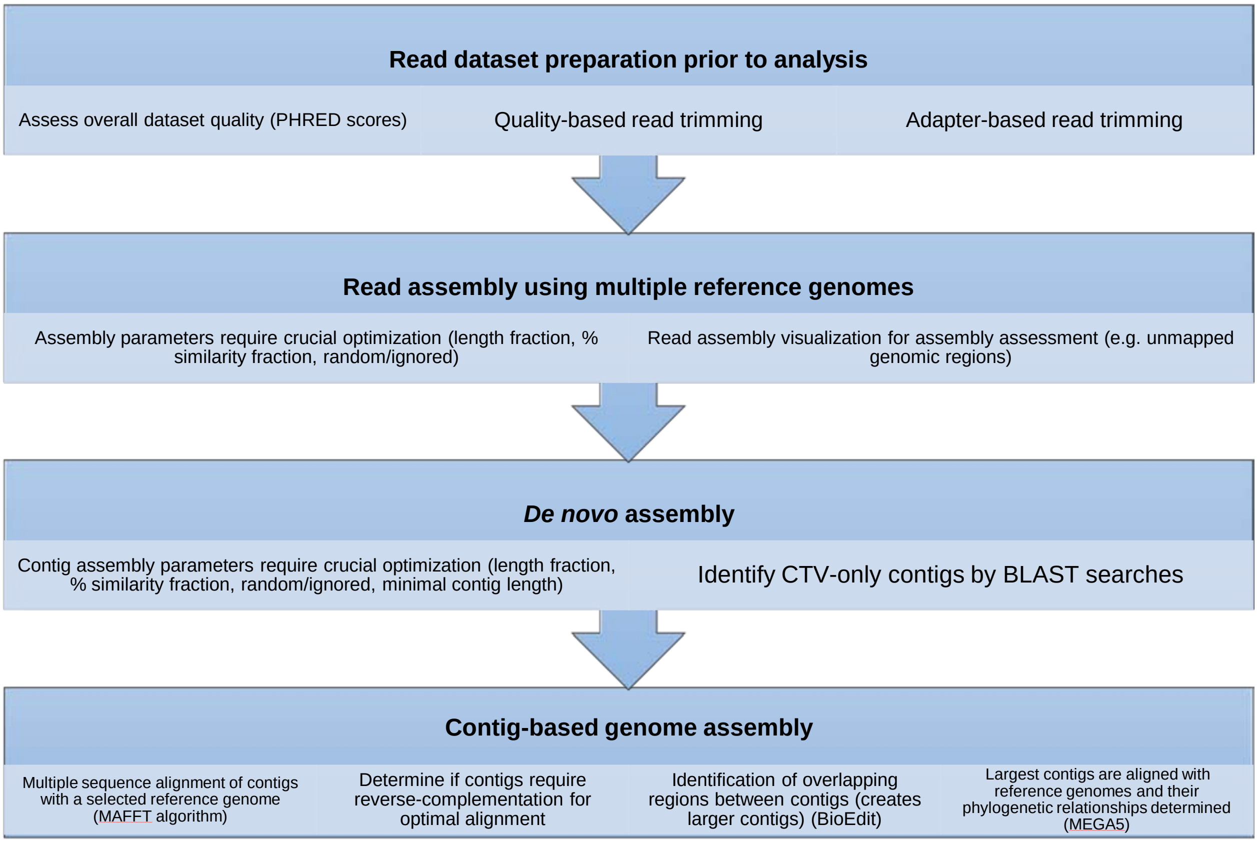

Read dataset preparation prior to analysis
Assess overall dataset quality (PHRED scores)
Quality-based read trimming
Adapter-based read trimming
Read assembly using multiple reference genomes
Assembly parameters require crucial optimization (length fraction, % similarity fraction, random/ignored)
Read assembly visualization for assembly assessment (e.g. unmapped genomic regions)
De novo assembly
Contig assembly parameters require crucial optimization (length fraction, % similarity fraction, random/ignored, minimal contig length)
Identify CTV-only contigs by BLAST searches
Contig-based genome assembly
Multiple sequence alignment of contigs with a selected reference genome (MAFFT algorithm)
Determine if contigs require reverse-complementation for optimal alignment
Identification of overlapping regions between contigs (creates larger contigs) (BioEdit)
Largest contigs are aligned with reference genomes and their phylogenetic relationships determined (MEGA5)

Figure 18: General bioinformatics pipeline used to process and analyze Illumina datasets.

This phenomenon, where several genotypes still obtained decent mapping metric values despite stringent parameters, may be interpreted in various ways. Since reads that were mapped could only be specific to each genotype (due to stringent parameters), and their consensus lengths were low, it could be hypothesized that the actual CTV genotype present in the sample was absent from the reference genome set, and that it possibly constituted perhaps a novel genotype. Therefore the observed mappings may have represented pieces of the putative novel genome, but that contained identical nucleotide sequences to the mapped references. Another possibility may be due to the excessive stringency of the mapping parameters. This prevented any sequence variability (compared to the reference sequence set), and therefore any slight variants of each genotype reference was omitted, which in turn produced short consensus lengths and read counts. Finally, it may simply be that the nucleic acid extraction, its random amplification and the nature of the template itself may have omitted regions of the genome(s) present in the sample.

Following parameter optimization, each template was analyzed with parameter set #4 (read mapping redundancy not permitted, fraction length to 0.99 and similarity to 0.99). For dsRNA, several genotypes obtained moderate mapping metrics, such as Tai-SP (an RB-like genotype), VT, SP and T68-1. A similar hierarchy was recorded for the other two templates (total RNA and virus particles), except that Tai-SP-like reads were absent. VT and SP consistently obtained the highest read count across all templates. Due to the relatively short consensus lengths obtained (Tai-SP was the longest, with 7281 nt), we attempted to *de novo* assemble reads to attempt the reconstruction of a putative distinct genotype, which appeared to be absent from our initial reference set. Reads from the total RNA template proved to be very useful in this regard, as it was the only dataset that yielded a consensus sequence (18801 nucleotides), which approached the full genome length of a typical CTV genome. It is not fully understood why the other templates failed to provide lengthier and more useful consensuses, but it may be that lower titers of the virus in these forms (dsRNA or virus particles) were present in the plant, or a failure from the extraction protocol occurred. The consensus sequence that was derived from the total RNA dataset was named CT-ZA1. Alignment of the novel construct with a set of 29 CTV reference sequences placed CT-ZA1 closest to T68-1, within the HA16-5 clade. Relative to whole genomes of CTV, four short regions were missing, including the 3' UTR, probably due to random priming which failed to amplify them. A recombination analysis of the aligned reference set, which included CT-ZA1 was conducted it was observed that the 5' half from CT-ZA1 was unique, and was not the result of recombination with T68-1, its closest relative. The fact that CT-ZA1 was not the result of a T68-1 recombination was unexpected (due to their sharing of the same node, see Figure 9) and the reasons for this are unclear. However, despite non-recombination events between the two genotypes, they shared a nucleotide similarity of 95%, which could explain the initial high mapping counts to T68-1 before the inclusion of CT-ZA1. Additionally, the fact that T68-1 was almost not represented during *in silico* cloning of the 1a region (11 matched reads in only the dsRNA template) further correlated how unique the CT-ZA1 genomic sequence was despite a high degree of similarity with T68-1. The analysis further demonstrated that the majority of the 3' half of the CT-ZA1 genome was mostly VT and SP in origin. This might explain why a number of reads and consensus lengths mapped to these two genotypes in the initial read mappings. To confirm this assumption, all template datasets were subjected to reference mapping, which this time included CT-ZA1. In all template datasets, results showed an overwhelming majority of reads mapping against CT-ZA1, as well as in consensus lengths and average coverages. The lengthiest consensus was reached using the reads from the total RNA template (which was expected) but reached lower values for dsRNA and virus particles, which suggested that these templates of CT-ZA1 were not well represented.

A comparison with NGS data (whole-genome) and cloning data (two gene regions) was made to assess if the respective results correlate. The A-region failed to provide accurate homology between clone sequences and reference genomes, even if CT-ZA1 clustered closely to the clones. For the p23 gene region however, where clones had previously closest homology with the VT cluster, specifically the SP genotype, CT-ZA1 was shown to be their closest match. Therefore, at least when compared to the results from the p23 gene, NGS data correlated well, whether it was before- or after- the inclusion of CT-ZA1.

An "*in silico* cloning" strategy was also devised to further assess not only if the results correlated, but also to determine if using two genes could actually infer population composition instead of at a whole-genome level. This strategy allowed the simulation of a conventional cloning experiment, but on a much larger scale and in a fraction of the usual time required. Using dsRNA, both gene regions showed CT-ZA1 as the main component of the population. A second main component was found, but whose identity varied according to the gene region. In the 1a fragment, Tai-SP was found whereas in the p23 gene, VT was present. This demonstrated a well-documented problem typically encountered when characterizing CTV isolates using multiple gene regions (Scott *et al.*, 2010). In this study (but not limited to this one alone), different genotype identities were obtained depending on the gene region used. The fact that this phenomenon also occurred in the simulation proved how similar the experiment was from actual gene cloning. For total RNA, CT-ZA1 and Tai-SP were the only genotypes found based on the A-region. This was the first time that Tai-SP appeared in this template (2 reads only), as it was previously only identified in the dsRNA dataset. The reason(s) for this are unclear, but it may be due the bias that gene-specific identifications induce, and since the NGS pipeline included the whole-genome, this bias was not seen in whole-genome read mappings. With the p23 region, a larger diversity of hits were obtained, which still showed CT-ZA1 as the dominant component. However, with the dsRNA template, VT replaced Tai-SP as the secondary component. Using virus particles as template, the A-region sequences suggested that CT-ZA1 is the sole component. Based on p23 region though, a wider genotype diversity was obtained, similar to that of total RNA. CT-ZA1 was still the dominant component, but was accompanied by reads to SP, CT14A, VT, VT-defective, A18, Tai-SP, SY, Kpg3 and NZRB-M17. These genotypes yielded extremely low read counts, ranging from 1 to 135, and it is hypothesized that these additional read mappings also arose by the same potential mechanisms of appearance as proposed for the total RNA dataset mappings. In summary, *in silico* cloning correlated well (but not exactly) in both gene regions and whole-genome comparisons.

The genotype identities determined during the NGS pipeline analysis, combined with those from the cloning experiments, brought up the question: “Does GFMS12 sub-isolate 12-7 contain one or several CTV genotypes?” Based on read counts, consensus lengths and average coverage, it is suggested that CT-ZA1 is the dominant genotype. While its nucleotide sequence was not fully mapped against reads from dsRNA and virus particles templates, it still yielded the best mapping metrics compared to other genotypes. This was also apparent in visual inspections of mapped reads. On a whole-genome level, it may be that the CTV population also contains a secondary component, a close variant of the Tai-SP genotype (RB-like), since its genome was well represented amongst reads, despite not providing a satisfactory consensus length. However, there remains an uncertainty, due to Tai-SP only being observed in the dsRNA template, and not in the other two (although 2 reads matched while *in silico* cloning of total RNA, it is considered not significant enough). If a variant of the Tai-SP genotype was indeed present, it would have showed more significant results in at least a second template, like total RNA. It does not make sense in terms of the virus's life cycle that the replicative form be present but not its single-stranded form. It may therefore be considered not as a secondary component, but a representation of a recombinant, which was constituted of both CT-ZA1 and Tai-SP genomic regions. Alternatively, it may be that the single-stranded form of Tai-SP does exist (and therefore constitutes a secondary dominant component), but was in sub-detectable levels for total RNA and virus particle templates. If there are indeed two dominant components, it would coincide with previously recorded results for this sub-isolate (Scott *et al*., 2010), whereby cloning the A-region of 12-7 a B165/VT-like recombinant and a RB/VT-like recombinant were found. It may be that the B165/VT-like recombinant was actually CT-ZA1 and that the second recombinant was the CT-ZA1/Tai-SP variant previously mentioned, but because CT-ZA1 and Tai-SP reference genomes were not available at the time of study, they were misidentified.

A more elaborate theory might explain the overall results across templates as well as in the cloning simulation: the fact that high read counts and moderate consensus lengths remained in addition to one dominant component (which in this sample appeared to be CT-ZA1). In assuming sub-isolate 12-7 homogeneous for CT-ZA1, there should not have been any remaining reads mapping to other genotypes. However, this was not the case. In all templates VT, SP, T68-1 and CT14A were consistently seen in moderate mapping metrics. To explain these observations, we hypothesize that the cumulative effects, over time, of individual polymorphisms induced by the virus's quasi-species behavior (mainly due to its replicase enzyme) has led to the creation of genomic regions sufficiently different from the original master sequence (in this case, CT-ZA1). This induced evolutionary models and mapping parameters to infer a different genotypic identity. The concerted effect of these mutations could effectively be called "genotype-shifting mutations". As observed in the graphs from Figure 11, reads that have mapped to other "minor components" or genotypes (such as VT and SP) may have been derived from highly polymorphic regions and thus have acquired sufficient nucleotide distance from CT-ZA1 to become sufficiently closer to be mapped against other genotypes (the minor components). Therefore, the additional mappings observed may be a representation of the quasi-species population as well as its extent. Furthermore, it might not be a coincidence that none of these additional mappings were attributed to distant genotypes, such as T36 and the NZRBs. They were from genotypes that were not very far off, nucleotide distance wise, from CT-ZA1 (dendrogram in Figure 9). This meant that a reduced amount of mutations would be required to become closer to genotypes belonging to close clades relative to CT-ZA1 and increase the probability of a genotype shift to occur. It could also be that recombination played a role, due to parts of genomes originating directly from a given genotype (e.g. VT). Most probably, it is a combination of the two mechanisms of novel sequence creation that contributed to these mapping patterns. This theory is visually depicted in Figure 19. This would also explain why very low metrics were achieved before the inclusion of CT-ZA1. As this genotype was the dominant component, only its polymorphic regions were represented in the original mappings that did not include it. In this regard, the use of next-generation sequencing was invaluable into gaining insights into the extent of genetic diversity that the virus is able to generate. The data also permitted to generate *de novo* the dominant component, CT-ZA1 and to show all its “faulty replication- induced” sequences. This was made possible also by the fact that the analysis focused at the whole genome level, which would have been almost impossible to discern if only specific gene regions were used.

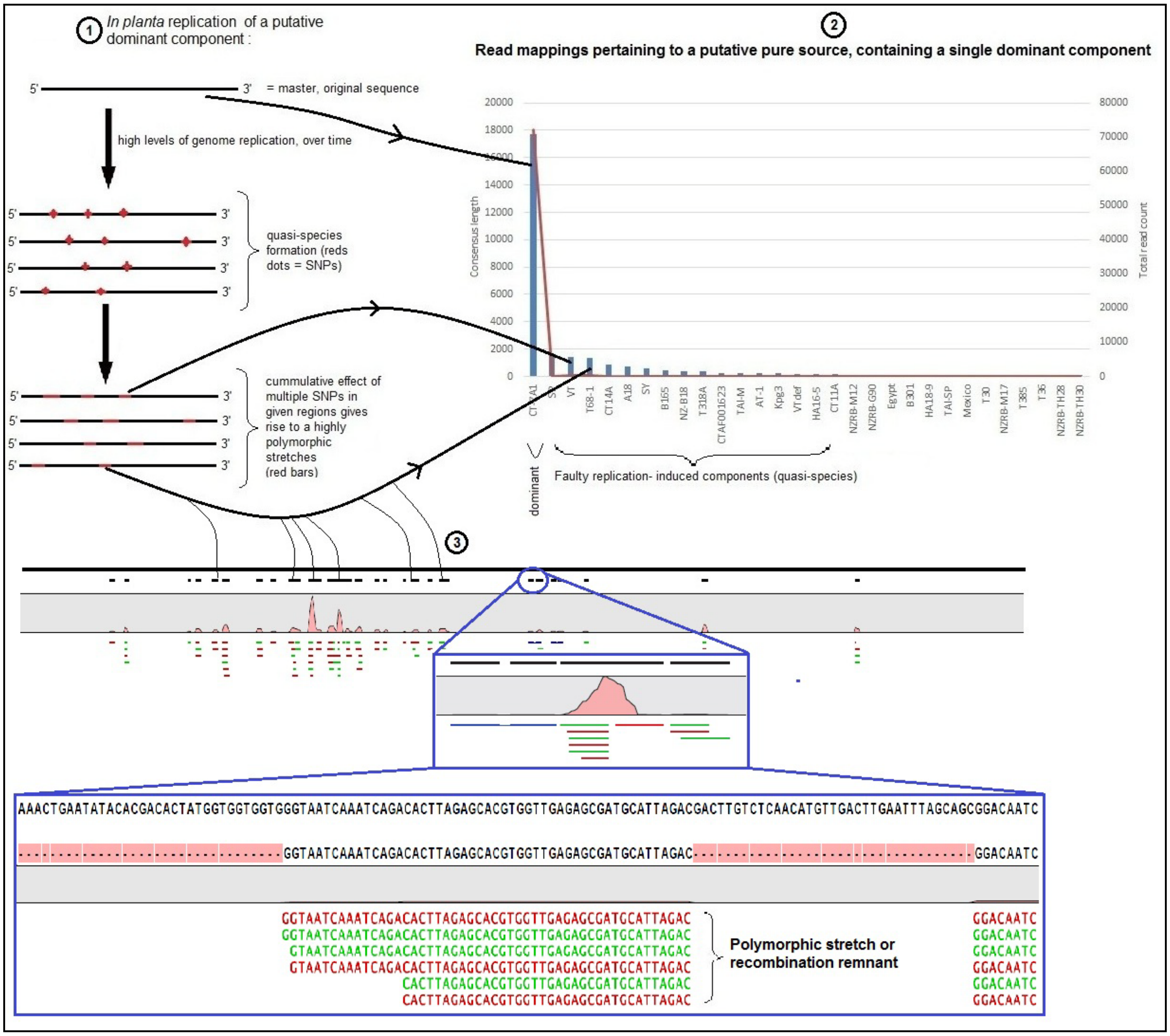

1
In planta replication of a putative dominant component :
5'
3'
= master, original sequence
high levels of genome replication, over time
quasi-species formation (reds dots = SNPs)
cummulative effect of multiple SNPs in given regions gives rise to a highly polymorphic stretches (red bars)
2
Read mappings pertaining to a putative pure source, containing a single dominant component
Consensus length
Total read count
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
80000
70000
60000
50000
40000
30000
20000
10000
0
CT-7A1
SP
VT
T68-1
CT14A
A18
SY
B165
NZ-B18
T318A
CTAF001623
TAI-M
AT-1
Kpg3
VT def
HA16-5
CT11A
NZRB-M12
NZRB-G90
Egypt
B301
HA18-9
TAI-SP
Mexico
T30
NZRB-M17
T385
T36
NZRB-TH28
NZRB-TH30
dominant
Faulty replication- induced components (quasi-species)
3
AAACTGAATATACACGACACTATGGTGGTGGTGGGTAATCAAATCAGACACTTAGAGCACGTGGTTGAGAGCGATGCATTAGACGACTTGTCTCAACATGTTGACTTGAATTTAGCAGCGGACAATC
------------------------------------GGTAATCAAATCAGACACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC--------------------------------------GGACAATC
GGTAATCAAATCAGACACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC
GGTAATCAAATCAGACACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC
GTAATCAAATCAGACACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC
GTAATCAAATCAGACACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC
CACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC
CACTTAGAGCACGTGGTTGAGAGCGATGCATTAGAC
Polymorphic stretch or recombination remnant
GGACAATC
GGACAATC
GGACAATC
GGACAATC
GGACAATC
GGACAATC

Figure 19: The theory of genotype shifting mappings. In the figure, area 1 depicts the normal CTV replication cycle, assuming a single, dominant genotype is present in the host plant. As it makes copies of itself, single nucleotide polymorphisms (SNPs) occur throughout newly synthesized copies. Over time, as SNPs in a given genomic region accumulate, its nucleotide sequence has diverged enough from the original to be more related to another genotype. The other genotypes, besides the dominant component, that have been mapped by reads (graph in area 2), may be the results of these polymorphic stretches. Therefore there exist a population of full-length genomes, each bearing the modified regions (or remnants of past recombination) that mapped to minor components (=quasispecies population). In area 3 of the diagram, the polymorphic stretches are shown in the context of actual read mappings, which includes two levels of zoom to illustrate that although it may look as the polymorphic regions are long spans along the mapped genome, it is actually a result of de-zooming too much.

One of the objectives of this chapter was to determine the optimal template choice for future CTV characterization. Based solely on typical mapping metrics (consensus length, read count and average coverage), dsRNA seemed to be the template which enriched for CTV sequences the most. In addition, the Tai-SP could only be detected in this template (except in *in silico* cloning). The next second best enriching template was total RNA. This was not necessarily in terms of mapping metrics (which were lower than that of virus particles), but rather in its ability to generate an almost complete *de novo* CTV genome sequence. Enrichment of virus particles by immunocapture is definitely not recommended for further use, since it performed poorly in both mapping counts and *de novo* assemblies. However, it may be possible to optimize the method further by, for example, affinity chromatography.

Overall, CTV-specific reads across all templates were low, even when non-stringent mapping parameters were used (a maximum of 6.57% of total reads with dsRNA). This was unexpected, especially for dsRNA and virus particles, where clear enrichment was expected. This may be attributed to low titers of the virus, host components retained after the nucleic acid extraction method or un-optimized protocols. Non-viral contamination was observed during the filtration of CTV contigs from the *de novo* analysis (not shown). Upon MultiBLASTn searches, sequences belonging to host mRNAs, ribosomal RNA (18S and 26S) as well genes from the mitochondrial, nuclear and chloroplast genome of various tree species were present in high numbers. Interestingly, one of the most represented host sequences was the *Poncirus trifoliata* CTV resistance gene locus mRNA. This extensive non-viral amplification also resulted from the random priming technique. This was expected, especially for total RNA, but it was a necessary step to eliminate primer bias, which was demonstrated to be very useful to amplify a whole CTV genome, even in AT-rich regions such as the 5' and 3' UTR.

Conclusions:

- Double-stranded RNA was found to be the most enriching template for CTV.
- Calibration of read mapping parameters is a crucial requirement in order to avoid false data interpretations.

- *De novo* assemblies from the total RNA dataset yielded a novel, partially complete CTV genome sequence, named CT-ZA1. Phylogenetic analysis showed CT-ZA1 being a close relative of the T68-1 genotype.
- Across all templates tested, CT-ZA1 was shown to be the dominant component.
- Additional read mappings despite the homogeneity status of GFMS12 12-7 sub-isolate was attributed to the quasi-species nature of CTV.
- Comparison between multiple gene regions versus whole genome characterization of CTV showed that individual gene regions may yield biased results.

2.5 REFERENCES

Kong, P., Rubio, L., Polek, M., and Falk, B.W., 2000. Population structure and genetic diversity within California Citrus tristeza virus (CTV) isolates. Virus Genes 21(3): 139-145.

Le Provost, G., Iskra-Caruana, M., Acina, I., Teycheney, P., 2006. Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. *Journal of Virological Methods* 137 :7-13.

Morris, T.J., Dodds, J.A., Hillman, B., Jordan, R.L., Lommel, S.A. and Tamaki, S.J., 1983. Viral specific dsRNA: diagnostic value for plant virus disease identification. *Plant Molecular Biology Reporter* 1:27-30.

Nolasco, G., de Blass, C., Torres, V. and Ponz, F. 1993. A method combining immunocapture and PCR amplification in a microtiter plate for the detection of plant viruses and subviral pathogens. *Journal of Virological Methods* 45: 201-218.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G. and Roe, B.A. 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology* 19: 81-88.

Rubio, L., Ayllon, M.A., Kong, P., Fernandez, A., Polek, M., Guerri, J., Moreno, P., and Falk, B.W., 2001. Genetic variation of Citrus Tristeza virus isolates from California and Spain: evidence for mixed infections and recombination.

Saitou N. and Nei M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.

Scott, K.A., Hlela, Q., Zablocki, O., Read, D., van Vuuren, S., and Pietersen, G., 2012. Genotype composition of populations of grapefruit-cross-protecting citrus tristeza virus strain GFMS12 in different host plants and aphid-transmitted sub-isolates. *Archives of Virology* DOI: 10.1007/s00705-012-1450-4.

Studholme, D.J., Glover, R.H., and Boonham, N., 2011. Application of high-throughput DNA sequencing in Phytopathology. *Annual Reviews of Phytopathology* 49:87-105.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* 28: 2731-2739.

Van Vuuren, S.P., van der Vyver, J.B. and Luttig, M. 2000. Diversity amongst sub-isolates of Cross-protecting *Citrus Tristeza virus* isolates in South Africa. In “Proceedings of the 14th Conference of the International Organization of Citrus Virologists.” (J.V. da Graca, R.F. Lee, R.K Yokomi, Eds), pp 103-109, IOCV, Riverside, California.

Zerbino, D. R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5): 821-829.

CHAPTER 3:

CHARACTERIZATION OF FIELD AND GLASSHOUSE CTV ISOLATES MEDIATED BY AN UNBIASED ILLUMINA SEQUENCING WORKFLOW

3.1 INTRODUCTION

The aim of this piece of work was to apply the high-throughput sequencing workflow described in chapter 2 to characterize additional CTV isolates, two of which were kept under glasshouse conditions and the remaining ones from commercial orchards. We observed previously that in terms of CTV read counts, consensus lengths and coverage values, double-stranded RNA was the most useful template for future analyses for CTV. The use of random priming was also demonstrated to be efficient at amplifying the whole CTV genome, as seen in high coverage values for the newly assembled CT-ZA1 genotype. Additionally, one of the main problems associated with random priming, i.e. high contamination, could be controlled by effective *in silico* subtraction. Therefore, the isolates under study in this chapter were treated the same way as previously described, but dsRNA was the sole nucleic acid template used.

A total of five samples were processed, two glasshouse-kept GFMS12 sub-isolates, 12-8 and 12-9, as well as three from a commercial orange orchard in Nelspruit, Mpumalanga. A previous study (*Scott et al.,* 2012) performed on sub-isolates 12-8 and 12-9 in our lab showed these being homogeneous for B165/VT and VT-like genotypes, respectively. This was inferred by sequencing multiple clones of the A-fragment and the *p23* gene. Field isolates were selected based on previous surveying (Zablocki *et al.*, unpublished data) which showed two distinct sequence clusters (11-5007 and 11-5009) not closely related to major CTV groups (data not shown), as well as several isolates dominated by the T30 genotype, one of which was selected for this study (11-5000). As demonstrated by Chapter 2, inferring genotype composition for a given isolate based on limited cloning of specific gene regions may lead to erroneous conclusions. Therefore we applied the NGS approach to further characterize isolates for this chapter.

3.2 MATERIALS & METHODS

3.2.1 Virus isolates

CTV isolates used in this study (11-5000; 11-5007; 11-5009) were originally derived from randomly selected field orange (*Citrus sinensis*) trees derived from the Citrus Improvement Program and pre-immunized with GFMS12 (from Nelspruit, Mpumalanga). Sources were established by bud-grafting on Mexican Lime (*Citrus* *aurantifolia*) (L.) Swingle) seedlings and kept under insect-free glasshouse conditions. All samples were extracted from these grafted trees and used for full-genome sequencing. Additionally, two glasshouse-maintained GFMS12 sub-isolates (12-8; 12-9) derived by single-aphid transfers (*Toxoptera citricida*) (Kirkaldy)) from a GFMS12 mother tree (van Vuuren *et al*., 2000) were used.

3.2.2 Double-stranded RNA purification

Purification of double-stranded RNA was performed according to Morris and Dodds (1983) with minor modifications. 0, 5 grams of leaf petioles were pulverized with liquid nitrogen and ground to a fine powder with a mortar and pestle. The powder was transferred to a 2 ml microcentrifuge tube and 1 vol of 2X STE buffer (0.1 M NaCl, 0.05 Tris, 0.001 M EDTA, pH 7.0, supplemented with 1% SDS and 0.1% 2-mercaptoethanol), 1 vol phenol and 0.5 vol of chloroform: pentanol (25:1) was added. The tube was centrifuged at 8,000 g for 10 minutes. The aqueous phase was decanted and transferred to a new tube and its volume measured. 100% ethanol was added to a concentration of 15 % (v/v). The solution was then added to a tube containing CF-11 cellulose pre-equilibrated with 1X STE-E 15% buffer. The mixture was vortexed 3 times during 15 min and centrifuged for 3 min to pack the cellulose. The cellulose was then washed 3 times by adding 1 ml STE-E 15%, vortexing, centrifuging briefly and decanting the supernatant. After the final wash, 0.5 ml of 1X STE was added. The tube was vortexed and centrifuged, and the supernatant transferred to a 1.5 ml Eppendorf tube (Eppendorf, Germany). The dsRNA was precipitated with 100% ethanol and stored at -20ºC overnight. The tube was centrifuged at 14,000 g for 20 min, the supernatant was discarded and the pellet was left to dry. Finally, the dsRNA was resuspended in 50µl of molecular-grade water.

3.2.3 Random- primed, two-step reverse-transcription PCR

Random-primed reverse transcription PCR was conducted according to Roosinck *et al*., (2010) with minor modifications. 12 µl of dsRNA was mixed with 5µl universal RDOD primers (RDOD: $\text{CCTTCGGATCCTCCN}_{12}$; $\text{RDOD}_{\text{comp}}$: CCTTCGGATCCTCC) (IDT, USA) at a concentration of 10 µM in a 0.2 ml microcentrifuge tube. The tubes were incubated for 10 min at 95ºC followed by 4ºC for 5 min. A 7µl cDNA synthesis reaction mix containing 1X Avian Myeloblastosis virus (AMV) buffer (Roche, Mannheim, Germany), 200 µM of each dNTP (Promega, Madison, WI, USA), 5 units of RNAse inhibitor (Roche, Mannheim, Germany) and 10 units of AMV reverse transcriptase (Roche, Mannheim, Germany) was added, for a total volume of 24 µl. The tubes were incubated for 10 min at 25ºC, followed by 60 min at 50ºC. A 10 µl cDNA aliquot was used as template in PCR (Rubio et al, 2001). The reaction mix consisted of 5 µl 1X PCR buffer, 2.5 µl $MgCl_2$(50mM), 2 µl dNTPs (3.5 mM), 10 µl 2 µM $RDOD_{comp}$ complimentary primer (IDT, USA) , 2.5 units of BioTaq DNA polymerase (Bioline, UK), 5 µl BSA (20 µg/ µl) to a final volume of 50 µl with molecular grade water. The samples were loaded into a Bio-Rad T100 thermal cycler (Bio-Rad, CA, USA) under the following cycling conditions: 94ºC for 2 min, followed by 40 cycles of 94 ºC for 30 sec, 42 ºC for 45 sec,72 ºC for 1 min and a final extension at 72 ºC for 10 min.

3.2.4 PCR clean-up & Illumina sequencing

Randomly amplified cDNAs were visualized on an agarose gel to confirm the reaction had taken place. CTV-specific primers were used in RT-PCR to confirm dsRNA extraction. Amplicons were cleaned prior to sequencing with the MinElute PCR purification kit (Qiagen, USA) according to the manufacturer's instruction. Preparation of Illumina compatible libraries were made with the Nextera V2 DNA sample kit (paired-end) and the resulting sample-specific libraries were sequenced individually on 1/15th of a lane in an Illumina HiScanSQ located at the ARC Biotechnology platform in Pretoria, South Africa.

3.2.5 Illumina data analysis

Illumina paired-end datasets were analyzed using CLC Genomics version 5.1. Raw reads were imported as paired-end (distance of 180-250 nucleotides) and quality scores were assessed with the Illumina pipeline 1.8. Reads were filtered by removal of low quality sequences (limit of 0.05), removal of ambiguous nucleotides (maximum 2 nucleotides allowed) and removal of adapter sequences (Nextera v2 transposase 1 forward/reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ; Nextera v2 transposase 2 forward/reverse: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG)

Quality control for each read dataset was performed using the FastQC function imbedded into the software. A set of 29 CTV full genome sequences (NZRB-G90 (FJ525432), NZRB-M12 (FJ525431), NZRB-TH30 (FJ525434), NZRB-M17 (FJ525435), NZRB-TH28 (FJ525433), Qaha (AY340974), T385 (Y18420), B165 (EU076703), CT11A (JQ911664), CT14A (JQ911663), VT (EU937519), T30 (EU937520), T36 (EU937521), A18 (JQ798289), Mexico (DQ272579), B301 (JF957196), Kpg 3 (HM573451), CTV (AF001623), NUaGa (AB046398), T68-1 (JQ965169), SP (EU857538), NZ-B18 (FJ525436), AT-1 (JQ061137), Taiwan-Pum/SP/T1 (JX266712), Taiwan-Pum/M/T5 (JX266713), T318A (DQ151548), VT defective RNA CTV-L5' complete sequence (AY206452), HA18-9 (GQ454869) and HA16-5 (GQ454870)) obtained from the NCBI database imbedded within CLC was imported and used in reference assembly. In addition, the CT-ZA1 genome sequence (Chapter 2) was also included in the reference dataset. Reference assembly was performed with the following settings: mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 1, similarity of 1, conflict resolution by voting (A, T, C, G) and non-specific matches ignored. *De novo* assemblies were also made with the same long-read settings described for reference mapping, except for the length fraction (0.5) and similarity (0.8), which allowed moderate sequence variability during contig creation. In addition, a minimum contig length of 200 nucleotides along with the creation of “full” contigs option was selected. *De novo*-generated contigs were subjected to a multiBLASTn search against the nucleotide collection hosted by the NCBI portal (http://www.ncbi.nlm.nih.gov/) via the CLC interface. Multiple sequence alignments were made with the online version of MAFFT (version 6.952) (http://mafft.cbrc.jp/alignment/software/), applying the following parameters: non-use of structural alignments, “same as input” for the direction of sequences, “aligned” as output order, “automatic” strategy, amino acid scoring matrix to BLOSUMM62, nucleotide matrix scoring of “1PAM / k=2”, a gap opening penalty of 1.53 and an offset value of 0.0. Alignments were imported into BioEdit v7.1.3.0 (Tom Hall, Isis pharmaceuticals, Inc. 1997-2004). Phylogenetic analysis were performed using MEGA version 4 (http://www.megasoftware.net/). Unrooted dendrograms were inferred using the Neighbor-joining method with a bootstrap test of a thousand pseudo-replicates. Evolutionary distances were determined by applying the Jukes-Cantor base substitution model. Recombination analysis was performed by RDP (version beta 4.16) by selecting the “X-over” function (http://darwin.uvigo.es/rdp/rdp.html).

3.3 RESULTS

A total of 6,930,546 raw 100 bp paired-end reads were obtained for all five samples, which equated to ~750Mb of sequence data. After adaptor and quality trimming, 6,927,702 paired-end reads remained, which reflected the overhaul good quality of the dataset, as confirmed by the FastQC analysis, which indicated an average PHRED quality score of 37.4. A summary of the individual sample metrics is shown in Table 6.

Table 6: Illumina data statistics, encompassing raw data, post-filtering and global CTV assembly values.

Sample	***Average PHRED score***	***Total reads (unfiltered)***	***Total reads (post-trim)***	***% reads trimmed***	***Average length post- trim***	***Total reads assembled to CTV [1] (% of total filtered read obtained)***
12-8	38	1,799,620	1,797,164	99.93	83.3	391,455 (21.78%)
12-9	38	2,909,044	2,904,692	99.93	80.6	689,712 (23.74%)
11-5000	37	842,938	841,758	99.93	82.2	70,544 (8.38%)
11-5007	37	699,498	570,357	99.93	82.8	36,681 (6.41%)
11-5009	37	679,446	678,602	99.94	82.4	40,582 (5.98%)

[1]=total read count obtained when reads where mapped against 30 full-genome CTV sequences simultaneously on highest stringency parameters.

Both reference and *de novo* assembly were performed for each sample. Table 7 summarizes contig statistics for all isolates. For all analyses, three key mapping metrics were recorded: read count (RC), consensus length (CL) and average coverage (AC). In terms of *de novo* assemblies, several long contigs were obtained with the longest one’s being 17274 bp, 12648 bp and 6668 bp. For the sake of clarity, results of each sample will be presented separately.

Table 7: Summary statistics pertaining to *de novo* assemblies.

Contig measurements	***Samples***				
	12-8	***12-9***	***11-5000***	***11-5007***	***11-5009***
Min. length	198	200	201	202	200
Max. length	12,648	17,274	5,386	5,386	6,668
Avg. length	535	493	447	451	478
Total count	404	574	567	372	706
CTV-only contigs	3	4	105	116	71

Sub-isolate 12-8

Read assembly against a set of 30 CTV references (which included CT-ZA1) is shown in Figure 20. CT-ZA1 obtained the single highest read count (296211), an almost fully mapped consensus length (18083 against 18801) and the most significant coverage (1369-fold). The remaining reads mapped with relatively high RCs and CLs but very low coverages to the following: T68-1 (RC=3935; CL= 4176; AC=17), VT (RC=2705; CL= 4221; AC=12.45), SP (RC=1529; CL= 3910; AC=7.35) and CT14A (RC=1424; CL= 2410; AC=6.11). A visual representation of the mapped reads for these genotypes is depicted in Figure 21.

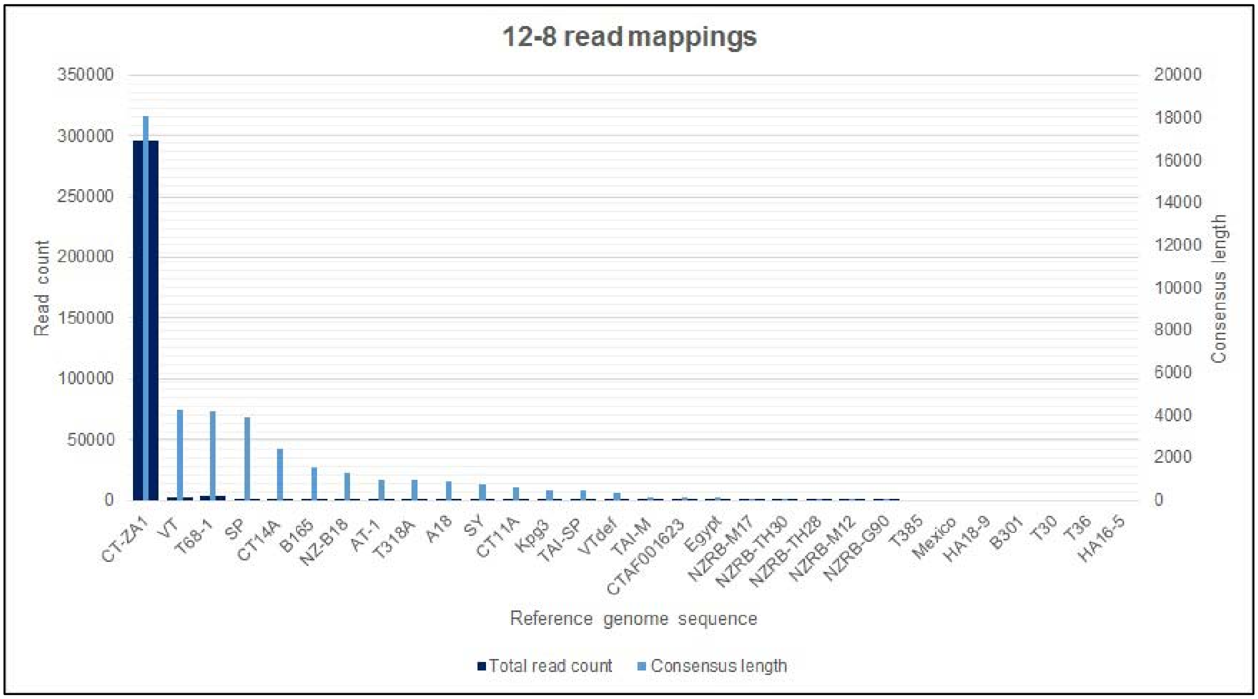


Figure 20: Read mapping results spanning 30 CTV reference genomes for GFMS12 sub-isolate 12-8. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

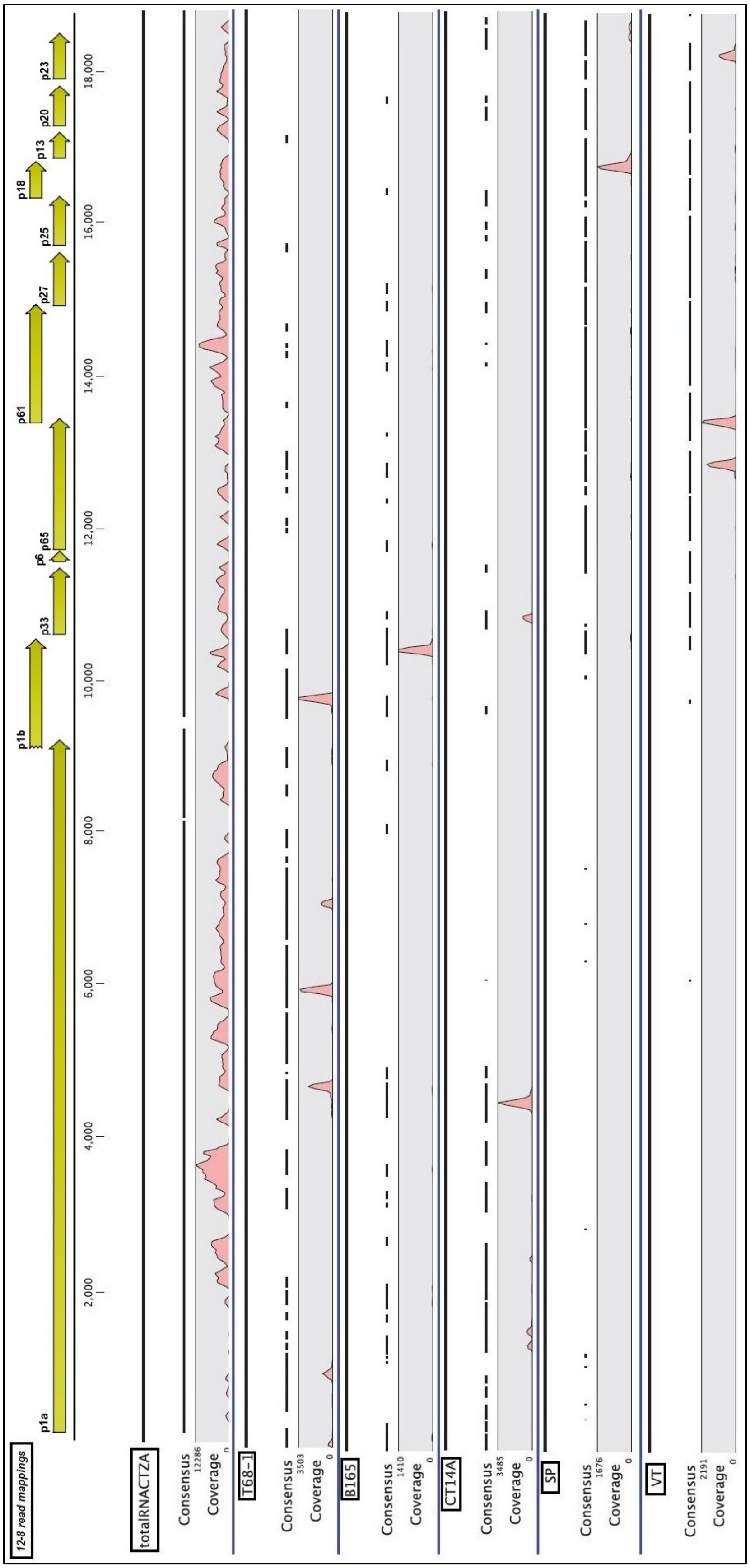


Figure 21: Visual mappings of sub-isolate 12-8 reads against the six mostly represented genotypes (CT-ZA1 or “totalRNACTZA”, VT, T68-1, SP, CT14A and B165). The scale on top of the image represents the full-genome length in nucleotides. The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read counts. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

In figure 21, CT-ZA1 was almost fully mapped by the reads, with some minor gaps present in the 5' half. In addition to CT-ZA1 being mapped, reads matched uniquely (due to stringent parameters, see Chapter 2) to several other genotype reference sequences. For VT and SP genomes, mapping almost exclusively occurred in the 3' half, which contained many gapped regions. In the case of T68-1, CT14A and B165, read mapping was the opposite of VT/SP. Mapping was mostly in the 5' half in a very scattered fashion. Some reads (4 to 5) mapped the 3' half as well.

De novo assembly produced only 3 CTV-specific contigs of 12648 bp, 6643 bp and 108 bp in length. Identification via BLASTn from the NCBI database showed these contigs having closest homology with T68-1. However, due to the absence of CT-ZA1 (Chapter 2) in the GenBank database, it was hypothesized that in actual fact it could have closest homology with CT-ZA1. This would also occur in analysis of the other samples. To confirm this, contigs were re-assembled by means of a multiple sequence alignment to identify missing pieces and close potential gaps. We found that for certain contigs, reverse complementation was needed for the correct alignment to occur. From only two long contigs (12648 nt and 6643 nt), a consensus sequences was created which was enabled by a "linker sequence", which was the result of a single overlapping region between the two contigs (Figure 22). Once this was identified, reference genomes were removed and a consensus sequence between contigs was made, followed by gaps removal. This yielded a putative full genome, 19244 bp in length, which we named CT-ZA3 (GenBank accession number KC 333869). To identify its relationship with other CTV genotypes, a dendrogram was created (Figure 23). CT-ZA3 clustered within the same node as CT-ZA1, and based on branch lengths, were practically identical.

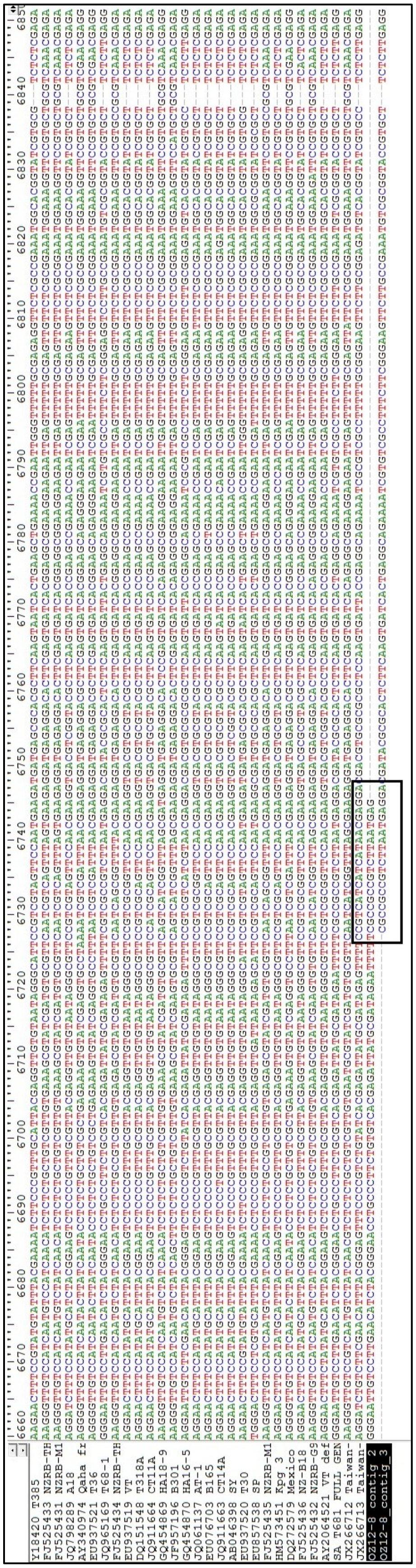


Figure 22: Multiple sequence alignment encompassing all CTV reference genomes and two contigs from 12-8 *de novo* assemblies that allowed the reconstruction of CT-ZA3 from nucleotide 6660 to 6851. The area within the black-edged rectangle highlights the single overlapping region between the two contigs that permitted the final assembly of CT-ZA3.

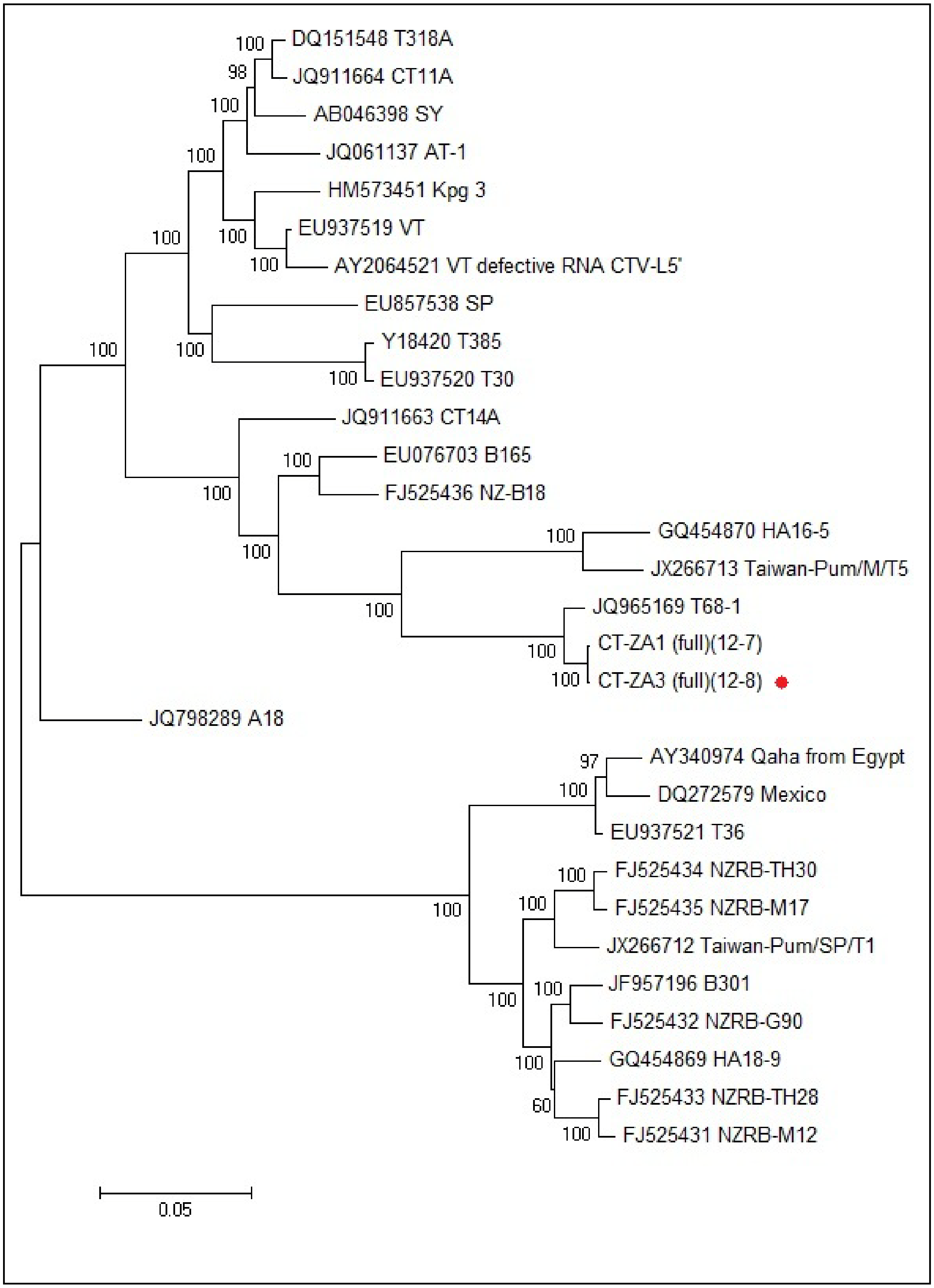


Figure 23: Neighbor-joining dendrogram showing the relationship between CT-ZA3 and the other full CTV genome sequences. CT-ZA3's location is marked by a red dot. Confidence levels are shown as bootstrap values at each node. (Full) means the whole genome was used. (12-7) and (12-8) signifies the sample name from which the sequences were derived.

Another round of reference mapping was conducted, this time including CT-ZA3 in the reference dataset (Figure 24). CT-ZA1 was not included, because it was shown to be very closely related (97.42%) to CT-ZA3, but also because the CT-ZA3 genome was full length, as opposed to CT-ZA1, which would produce enhanced mapping counts due to all genome regions being included. In addition, CT-ZA3 will only be used from this point on for further read mapping in other datasets. An almost identical mapping pattern as previously observed for CT-ZA1 occurred with CT-ZA3. It obtained the highest read count (390525), lengthiest consensus (19046) and coverage (1776.5 –fold). The next most represented genotypes (VT, SP, T68-1 and CT14A) obtained an average read count of 133.75, an average consensus length of 3134.75 and average coverage of 0.579 -fold. A visual representation of these genome wide mappings are shown in Figure 25.

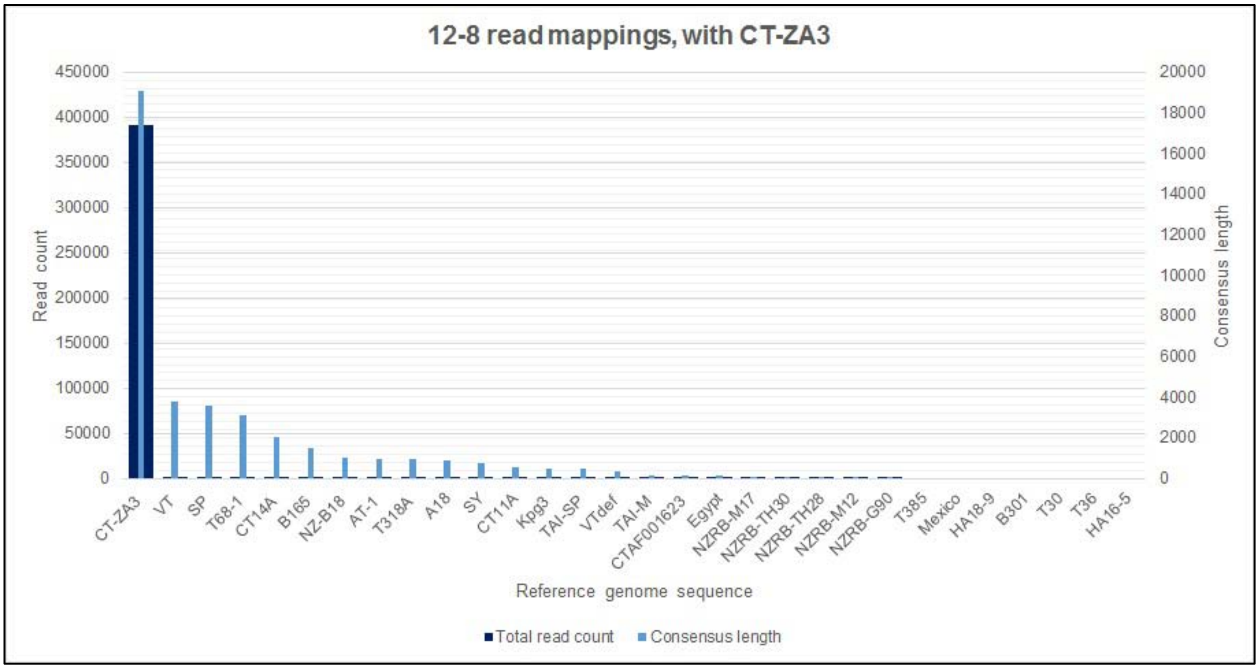


Figure 24: Read mapping results spanning 30 CTV reference genomes (including CT-ZA3) for GFMS12 sub-isolate 12-8. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

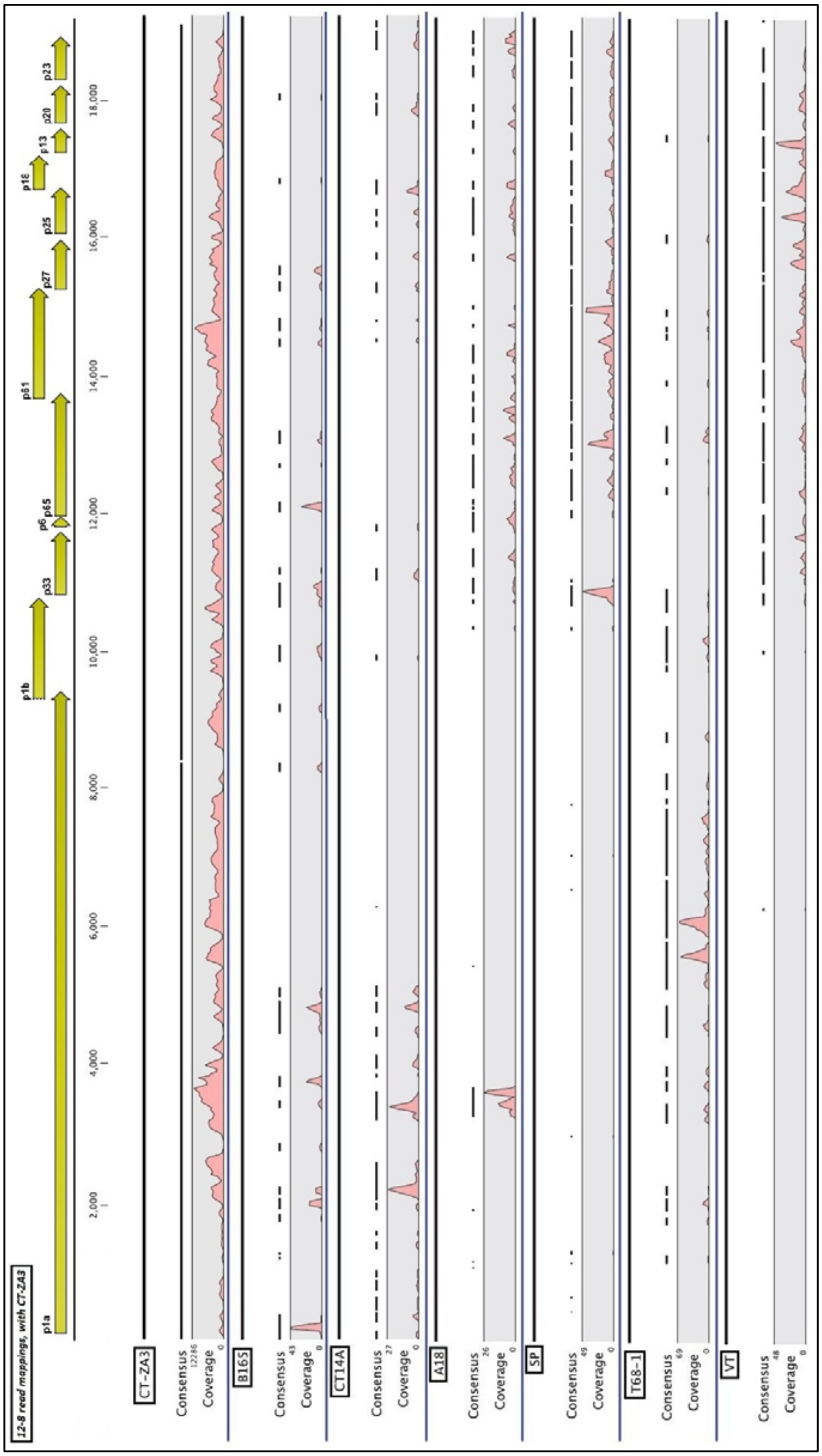


Figure 25: Visual mappings of sub-isolate 12-8 reads against the five mostly represented genotypes (CT-ZA3, VT, SP, T68-1 and CT14A). The scale on top of the image represents the full-genome length in nucleotides. The continuous, solid black lines represent the full length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read counts. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

As in the mappings shown in Figure 21, Figure 25 depicts an almost identical pattern. CT-ZA3 obtained in this case an enhanced read mapping, with only one minor gap remaining. As for the other genomes, VT-SP genotypes were mapped mostly in their 3' half while T68-1/CT-14A were mapped mostly in their 5' half. These reads that mapped to other genotypes in addition to CT-ZA3 were unique to these genotypes.

Sub-isolate 12-9

Results from the read mappings against the CTV reference set is shown in Figure 26. CT-ZA3 achieved similar results as observed for sub-isolate 12-8. It mapped the majority of the reads (415294), obtained had an almost complete consensus length (18763) and a coverage value of 1851-fold. The remaining, most mapped genotypes, included VT (RC=4707; CL= 5340; AC=20.82), SP (RC=9397; CL= 4282; AC=43.8), T68-1 (RC=136; CL= 2742; AC=0.54), CT14A (RC=1859; CL= 2300; AC=7.85), B165 (RC=78; CL= 2038; AC=0.31), and A18 (RC=200; CL= 1559; AC=0.83). A visual representation of the mapped reads for these genotypes is depicted in Figure 27.

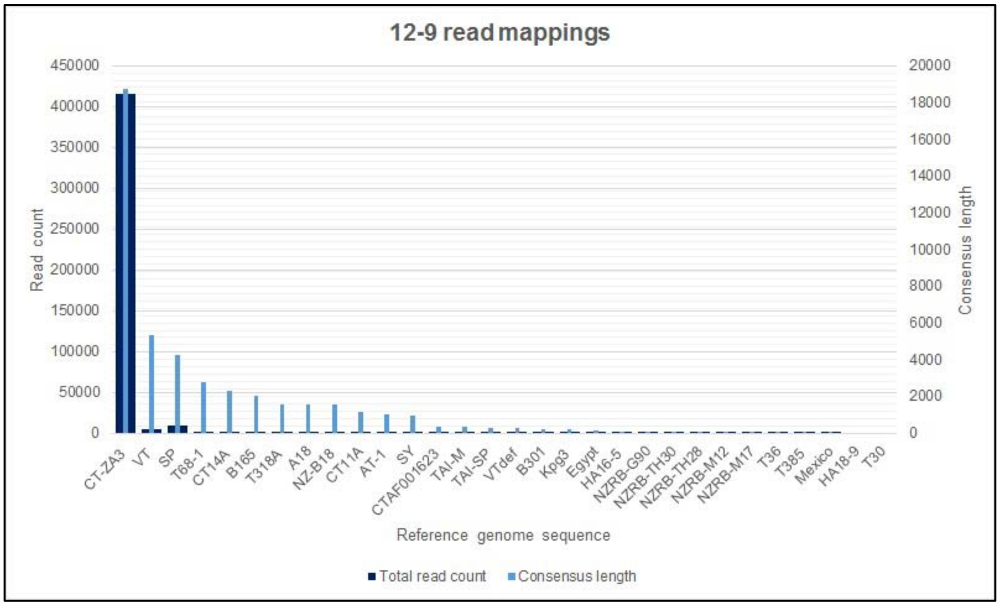


Figure 26: Read mapping results spanning 30 CTV reference genomes (including CT-ZA3) for GFMS12 sub-isolate 12-9. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

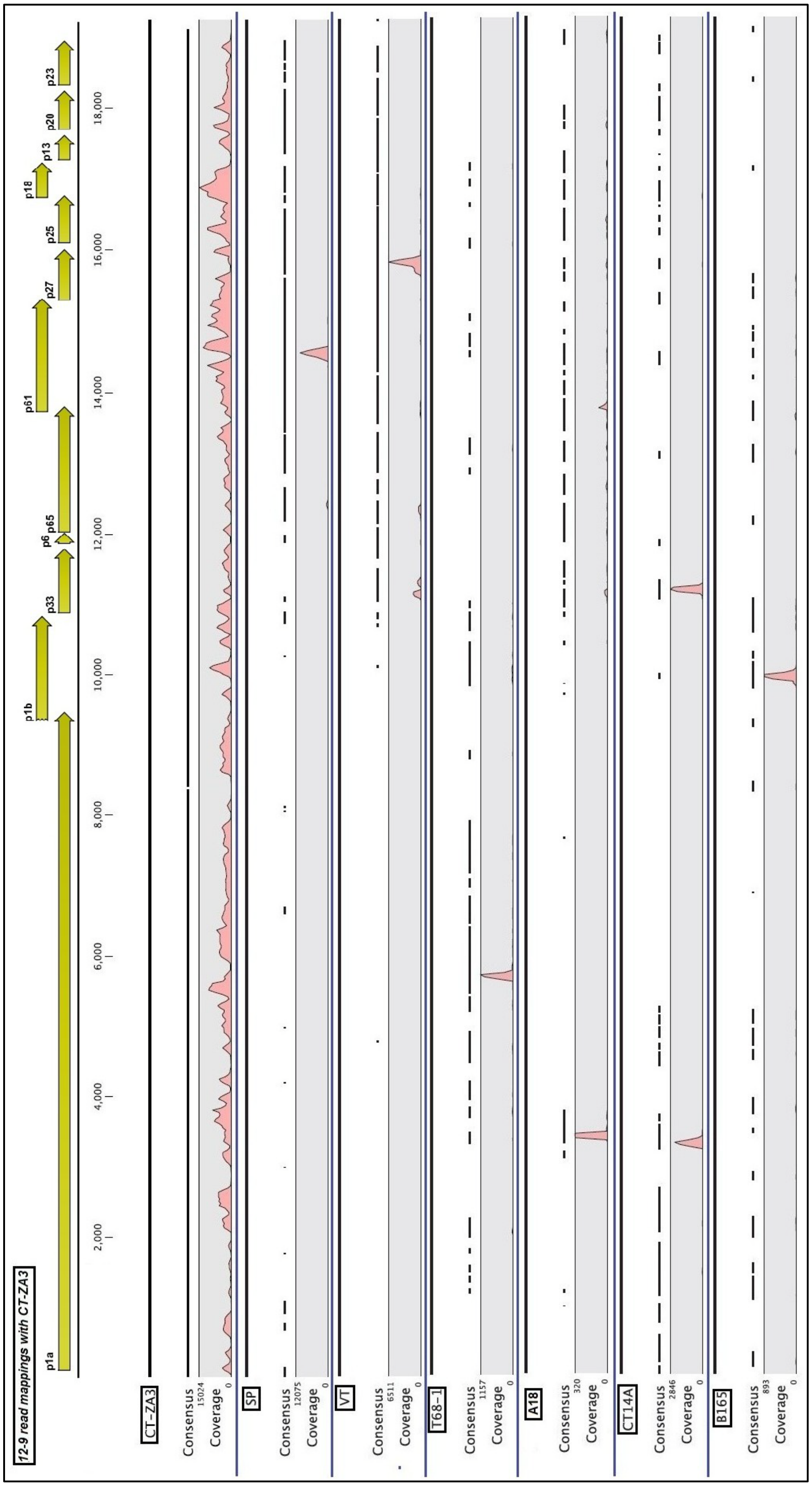


Figure 27: Visual mappings of sub-isolate 12-9 reads against the six mostly represented genotypes (CT-ZA3, VT, SP, T68-1, B165 and T318A). The scale on top of the image represents the full-genome length in nucleotides. The continuous, solid black lines represent the full length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read counts. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

In whole-genome read mappings, CT-ZA3 was almost completely covered, with only 2 minor gaps both the 5' and 3' halves. Genotypes VT/SP/T318A were mapped almost exclusively in their 3' halves while T68-1 was mapped mostly in its 5' half. Lastly, B165 was mapped in a very scattered way across its genome.

Four contigs were shown to belong to CTV following *de novo* assembly that had lengths of 17274 bp, 345 bp, 247 bp and 205 bp. The longest contig was shown to be most closely related to T68-1, and the other three shorter ones were most similar to the Tai-M genotype. To create the lengthiest consensus possible, all contigs, irrespective of their sizes were included in a multiple sequence alignment with a randomly selected CTV reference genome, which acted as guide for contigs to align. As contig #3 (17274bp) was almost a fully recreated genome (89.7% complete compared to VT), we attempted to fill its gaps by aligning it with other CTV-contigs. All the short contigs were able to fill gaps, but a few gaps still remained. Overall, the contigs covered genome regions from nucleotide position 226-17803 (1a, 1b, p33, p6, p65, p61, p27, p25, p18 and p13). Only the last two genes (p20 and p23) and the 5' and 3' UTR were not covered. This resulted in a “partial” genome, 17407 nt long, named CT-ZA2 (GenBank accession number KC 333868. To confirm its phylogenetic relationship, a dendrogram was created using CT-ZA2 and was aligned against the set of reference strains including CT-ZA1 and CT-ZA3 (Figure 28). Despite missing its terminal portions (5' UTR, 3' UTR, p20 and p23), CT-ZA2 clustered within the same node as CT-ZA1/CT-ZA3, which indicated they were essentially identical. Therefore, no further read mapping with CT-ZA2 was necessary.

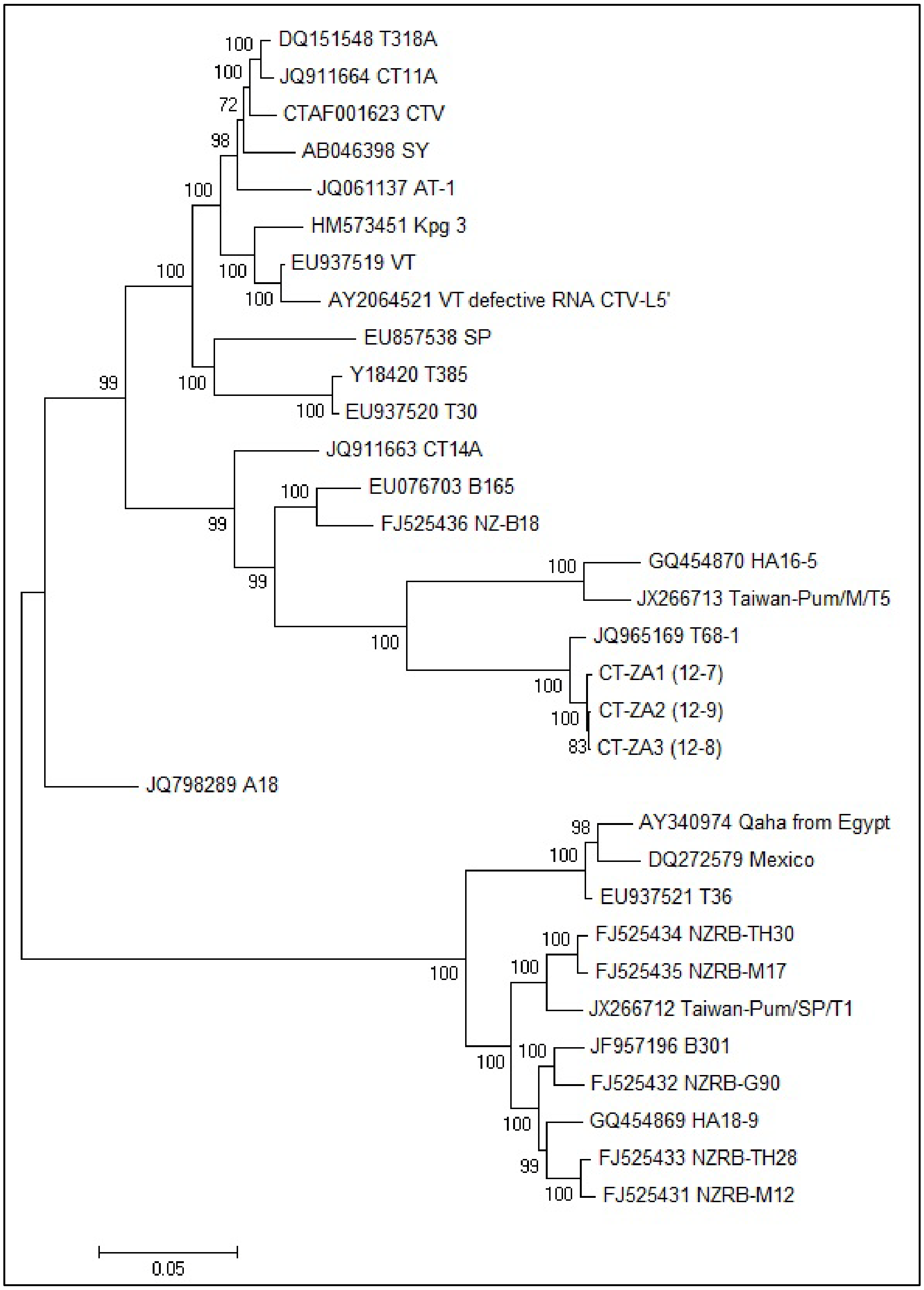


Figure 28: Neighbor-joining dendrogram showing the relationship between CT-ZA2 and the other full CTV genome sequences. CT-ZA2's location is marked by an arrow. The sub-isolate name from which each CT-ZA was derived from are shown in brackets. Confidence levels are shown as bootstrap values at each node.

Field isolate 11-5000

Read mappings against the set of 30 CTV reference genomes for this field isolate are shown in Figure 29. Two distantly related genotypes, Tai-SP (RB-like) and Kpg3 (VT-like) scored the lengthiest CL (18222 nt and 14695 nt, respectively), highest RC (34479 and 20014, respectively) and AC (152.5- and 86.4-fold, respectively). The next most represented genotypes were NZRB-TH28 (RC=2226; CL=9020; AC=10.01), VT (RC=2665; CL=8424; AC=11.99), NZRB-M12 (RC=1777; CL=7799; AC=8.2) and SP (RC=1424; CL=4993; AC=5.51). A visual representation of these mappings are depicted in Figure 30.

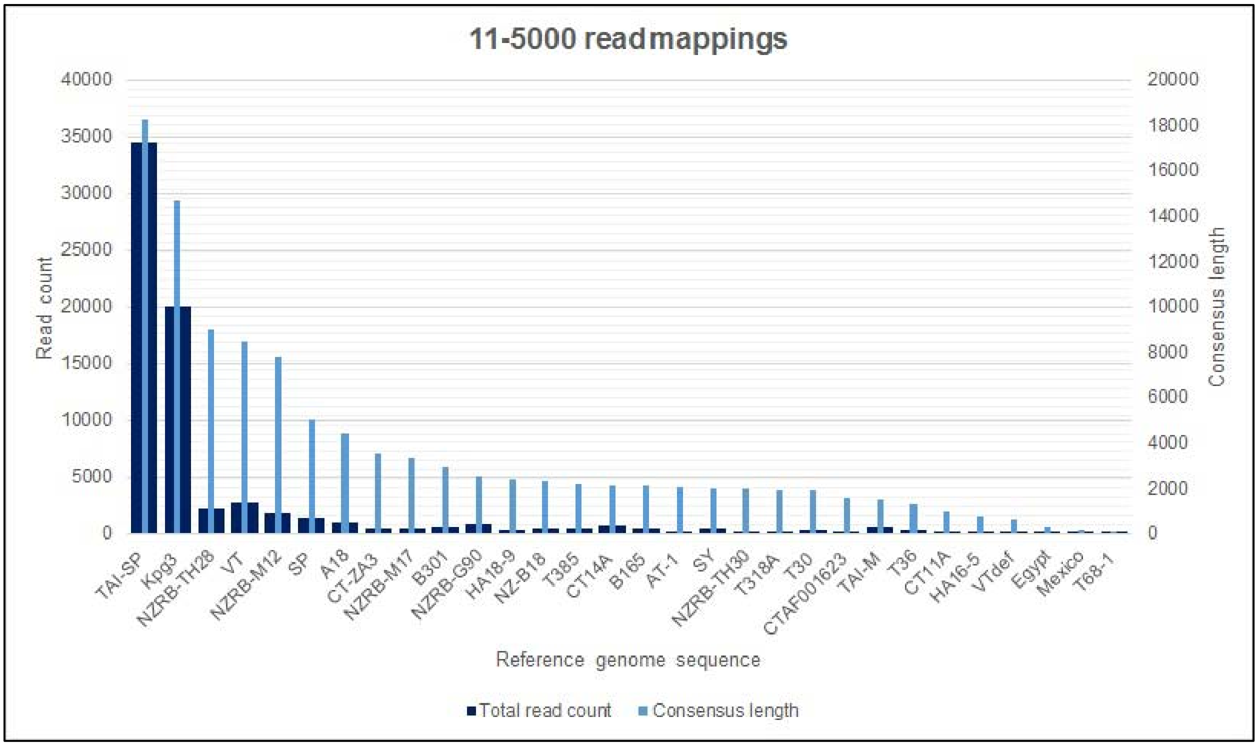


Figure 29: Read mapping results spanning 30 CTV reference genomes for field isolate 11-5000. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

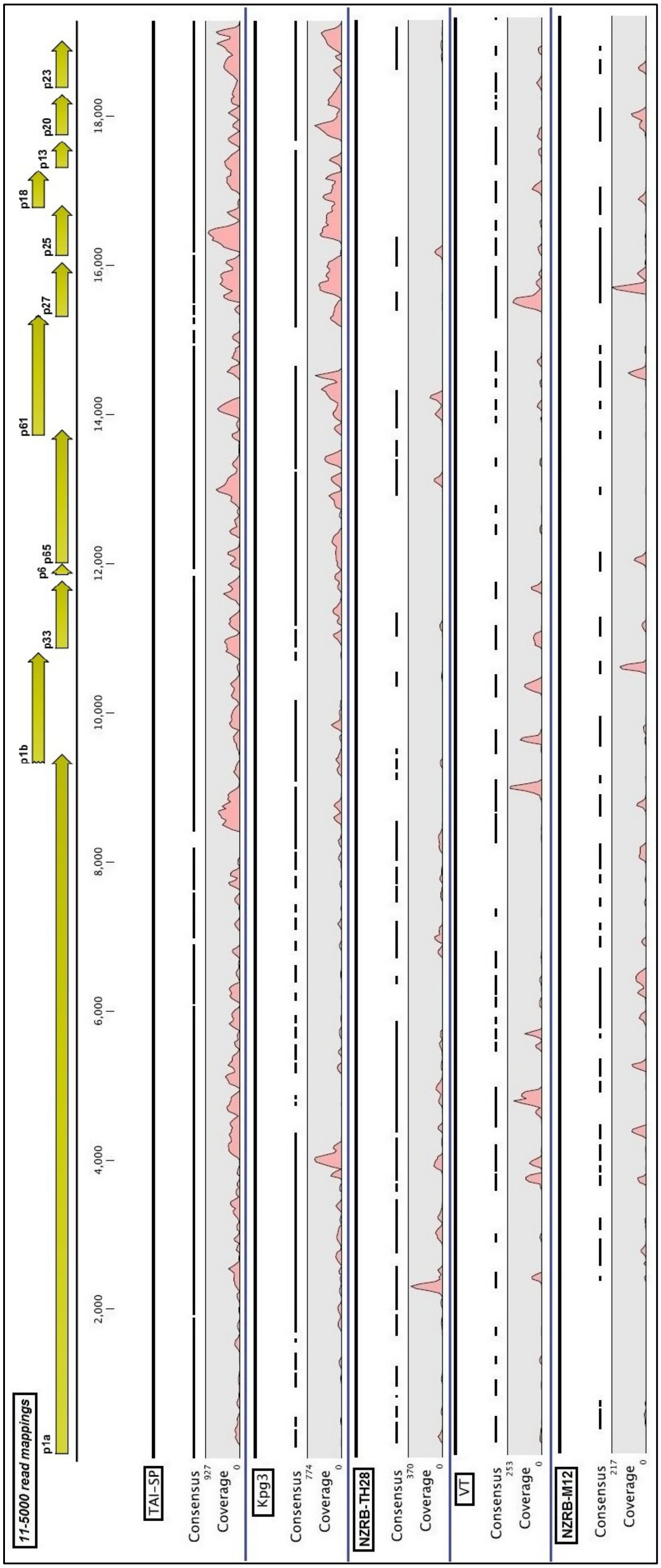
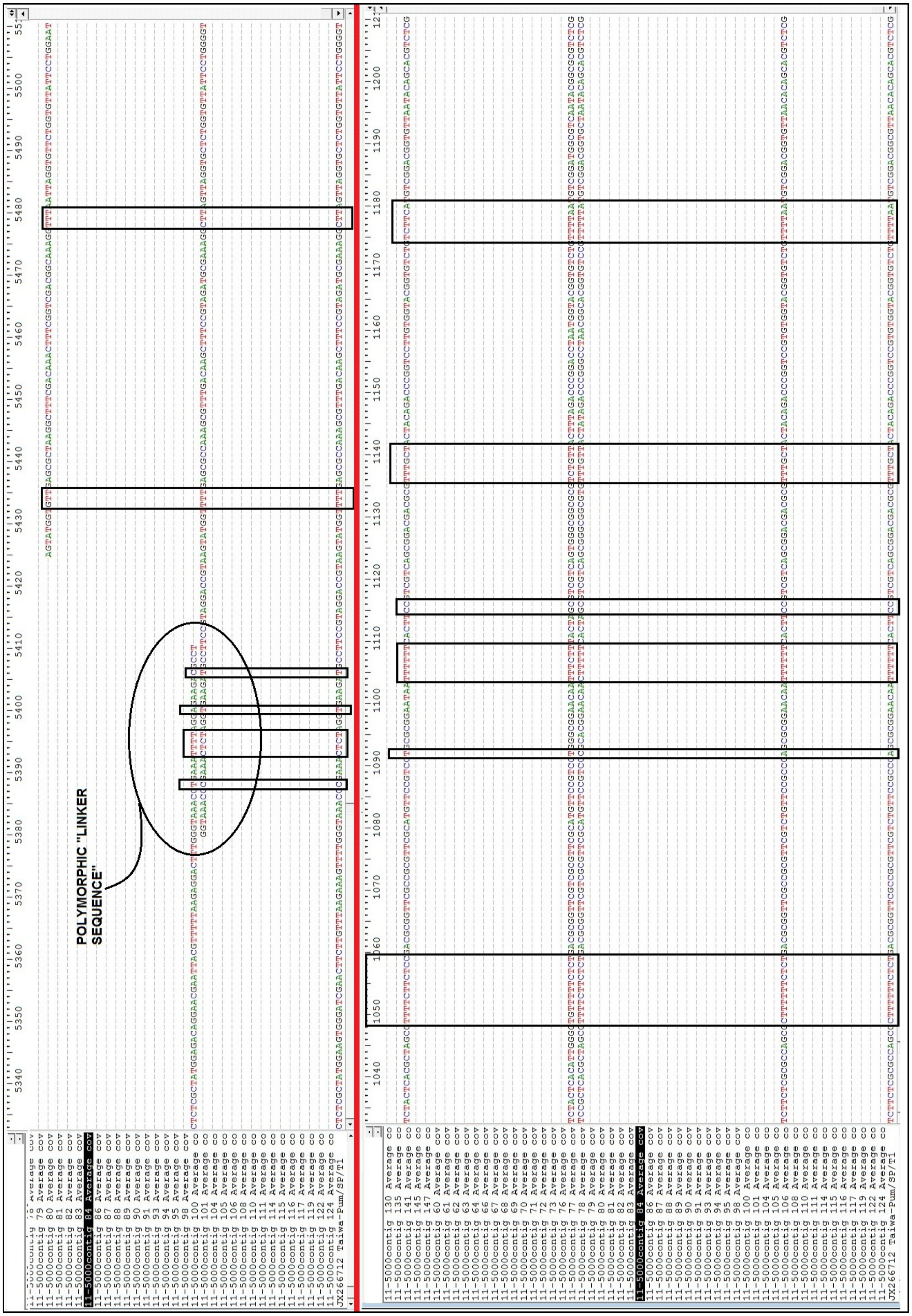


Figure 30: Visual mappings of isolate 11-5000 reads against the five mostly represented genotypes (Tai-SP, Kpg3, NZRB-TH28, VT and NZRB-M12). The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read count. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

The least represented genotypes obtained an average consensus length of 186.75, read count of 331.63 and and average coverage of 1.39. Visual inspection of Figure 30 showed the Tai-SP and Kpg3 genome to be almost completely mapped by the reads, with gaps still remaining mostly in the 1a region of Kpg3. NZRB-TH28 was mapped mostly in the 5' half, with additional minor read mappings in the 3' half. VT and NZRB-M12 achieved almost identical mapping patterns. No specific regions were mapped, but instead was mapped in a much gapped manner, throughout their respective genomes.

De novo assembled contigs were used to generate a full-length genome sequence. When contigs were aligned with a "guide reference genome", there were many overlapping regions ("linker sequences" specifically), which were polymorphic, thus impeding the creation of a full-length consensus. In addition, identical regions of the genome were also covered, but by contigs that differed too much in terms of nucleotide identities as well. Figure 31 depicts these two scenarios in snapshots of a multiple alignment with these contigs. The rectangles highlight single nucleotide polymorphisms across contigs and the guide reference genome. Slight nucleotide difference were acceptable between contigs and genome reference, since the reference is subsequently removed for consensus creation. The major problem lays in high numbers of polymorphisms between contig sequences. Both alignments in Figure 30 only spanned a region 177 nucleotides long, but these results were found throughout the whole genome length (~19200 nt). A full genome could have been assembled, but it would have been composed of many "N" nucleotides and gaps and therefore was not implemented. To the best of our efforts, we tried to counter-act the problem by using individual contigs in reference assembly against a set of 30 CTV reference genomes in efforts to identify towards which genotypes these contigs mapped specifically to (Figure 32). Kpg3 (2 contigs), Tai-SP (2 contigs), NZRB-TH28 (1 contig) and VT (1 contig) were found to be the only genotypes that were mapped by contigs.

Figure 31 (next page): Two separate contig mapping scenarios encountered which prevented the construction of a full-length genome. Both scenarios (alignments) are separated by the red line. In the upper part, a single polymorphic overlapping region (or linker sequence) is showed within the oval. The vertical rectangle show the different nucleotide identites between the contigs and the reference genome. Due these polymorphisms, consensus creation was not possible. In the lower part of the figure, the vertical rectangle shows the nucleotide difference between contigs and reference, except that these overlapping regions are not “linker sequences”, but rather multiple contigs mapping an identical, long region. Many polymorphisms were also observed, which also prevented consensus sequence creation. This snapshot only represents a 177 nucleotide long portion of the genome length, but these observations were found throughout the genome.


POLYMORPHIC "LINKER" SEQUENCE

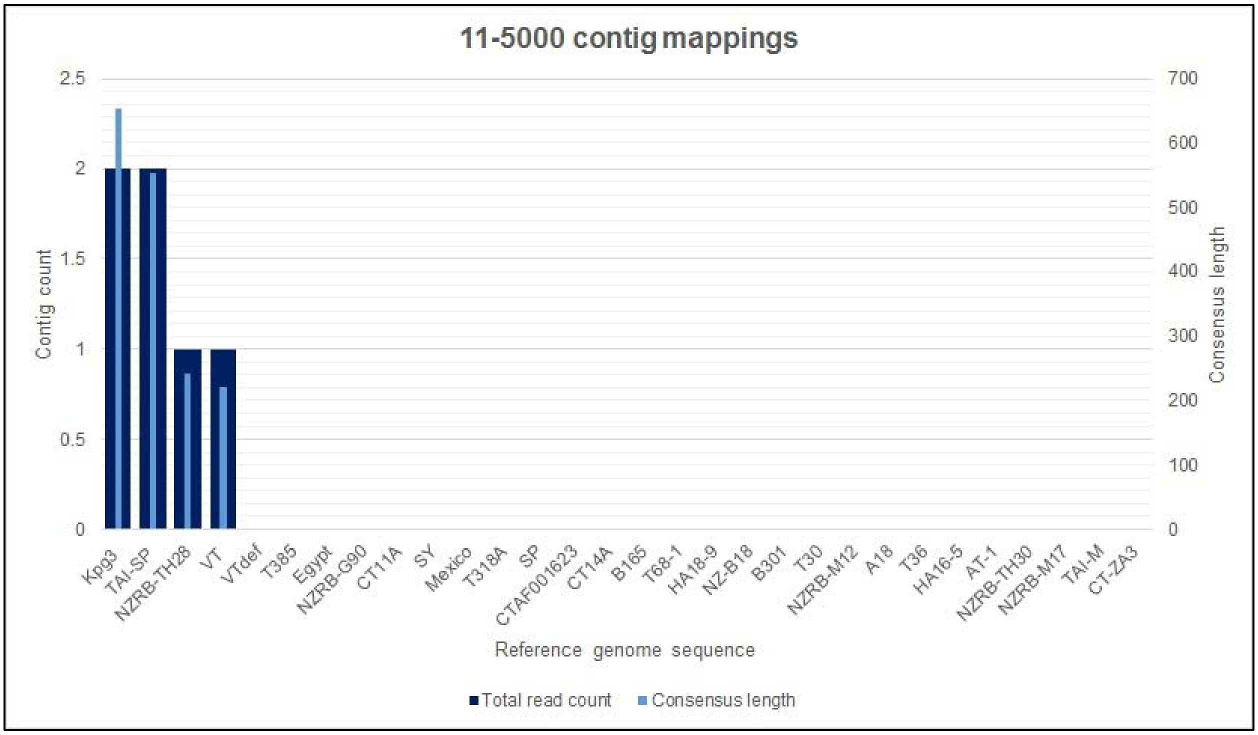


Figure 32: Contig mapping results spanning 30 CTV reference genomes for field isolate 11-5000. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by cumulative contigs. The primary vertical axis (left) scale represents the amount of contigs mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Contig count= thick dark blue bars; Consensus length= thin light blue bars.

Field isolate 11-5007

Read assembly for this isolate is shown in Figure 33. The most represented genotypes mapped were Tai-SP (CL=17452; RC=18054; AC= 80.57), Kpg3 (CL=9464; RC=7930; AC=33.91), SP (CL=9146; RC=3036; AC=12.13) and VT (CL=7698; RC=1494; AC=6.77. The remaining least represented genotypes averaged a consensus length of 1656.75, an average read count of 237.19 and average coverage of 1.007. A visual representation of the read mapping for the most represented genotypes is shown in Figure 34. The Tai-SP genome was almost fully mapped by the reads, with only a few gaps remaining. Kpg3 and CT-ZA3 were mapped mostly in their 3' halves, with some coverage in the 5' half. The opposite was observed for SP, whose 5' half was almost fully mapped, but not as much in the 3'half. VT was mapped in a scattered manner throughout its genome. Pertaining to *de novo* assemblies, the same problem as previously in isolate 11-5000 occurred in which genome-wide consensus generation was not possible due to many unique contig identities. Therefore contig mapping was performed, shown in Figure 35. Three genotypes obtained mapping counts: Tai-SP (2 reads), CTAF001623 (1 read) and Kpg3 (1 read).

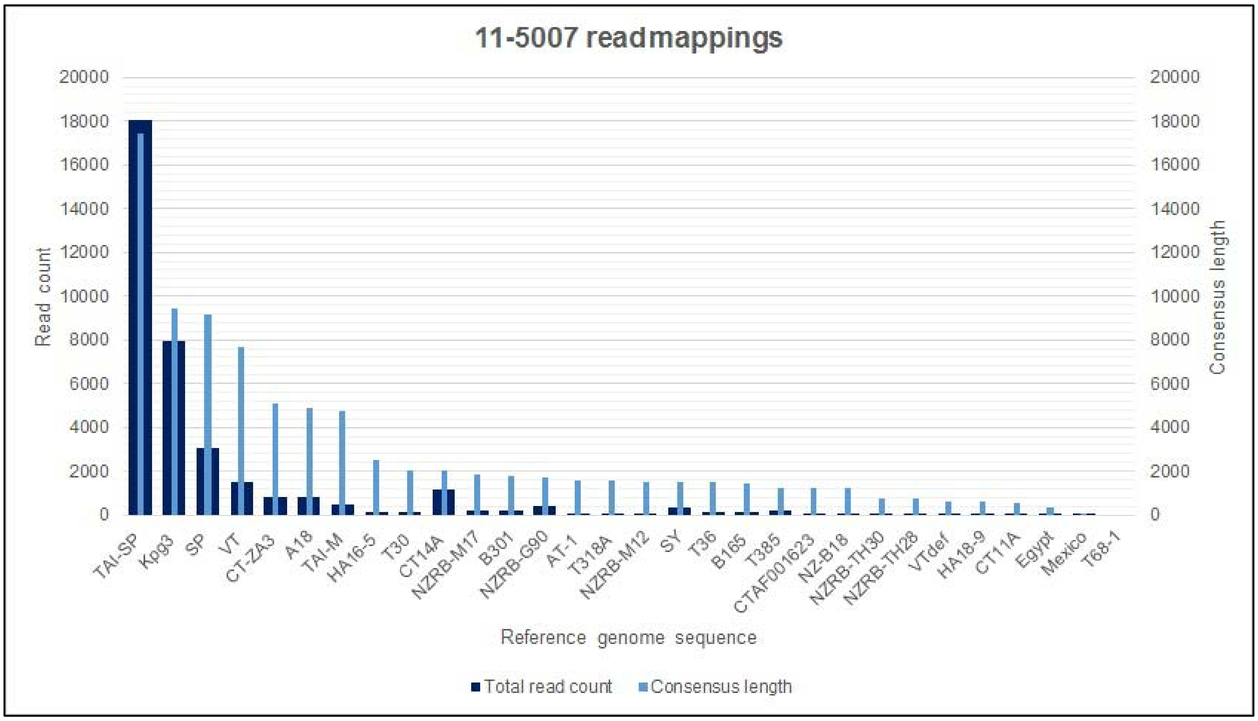


Figure 33: Read mapping results spanning 30 CTV reference genomes for field isolate 11-5007. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

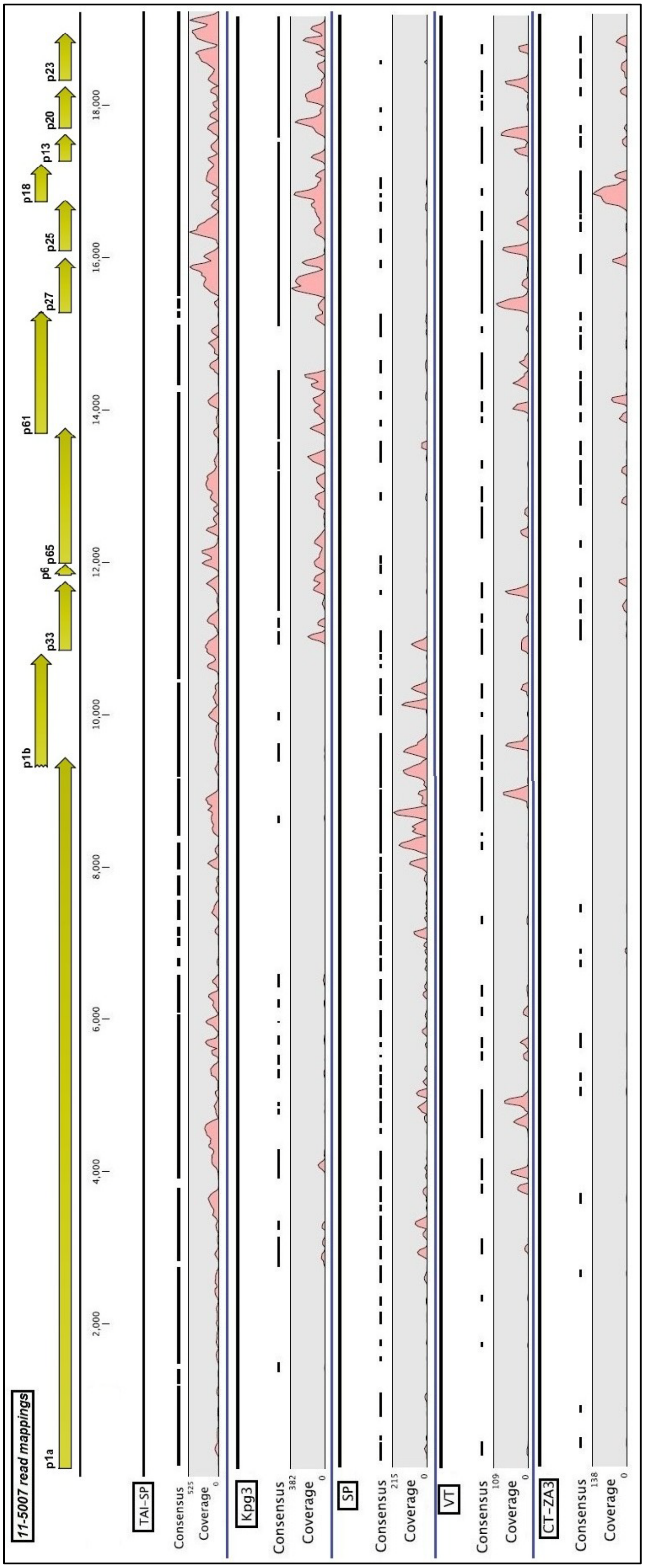


Figure 34: Visual mappings of isolate 11-5007 reads against the five mostly represented genotypes (Tai-SP, Kpg3, SP, VT and CT-ZA3). The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read count. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

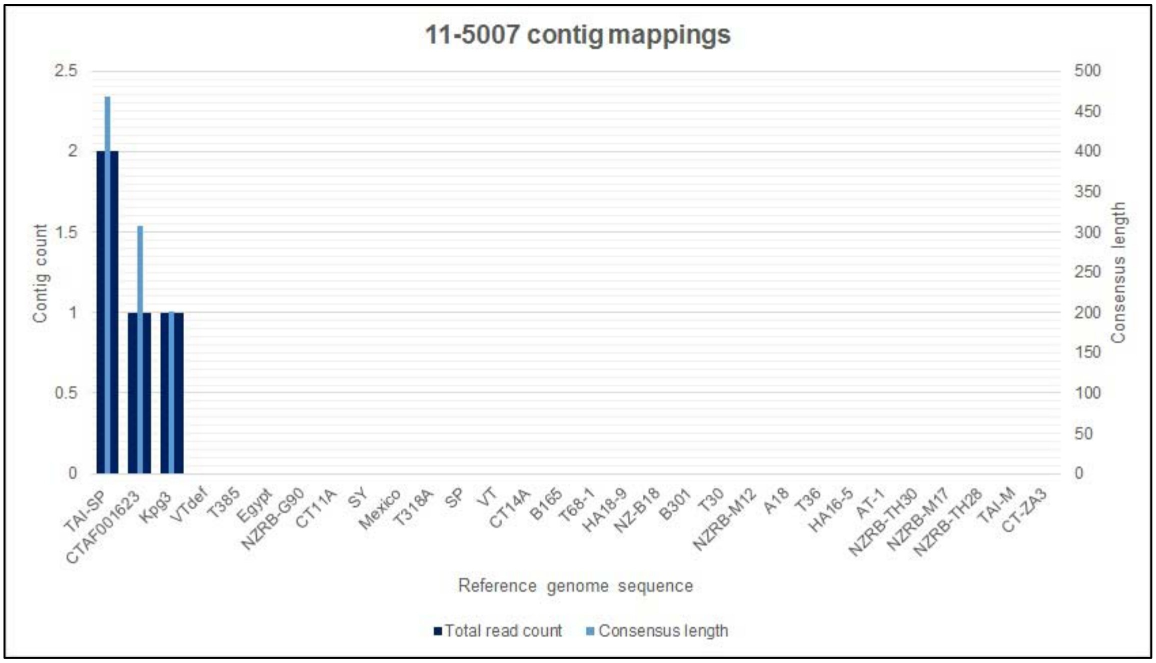


Figure 35: Contig mapping results spanning 30 CTV reference genomes for field isolate 11-5007. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by cumulative contigs. The primary vertical axis (left) scale represents the amount of contigs mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Contig count= thick dark blue bars; Consensus length= thin light blue bars.

Field isolate 11-5009

Reads from this isolate were mapped against a set of CTV reference genomes and is shown in Figure 36. The highest highest metrics were for Tai-SP (CL=18358, RC=17485, AC=78.03) followed by AT-1 (CL=8993, RC=5780, AC=25.6), HA16-5 (CL=8962, RC=2647, AC=11.22), Kpg3 (CL=7863, RC=6006, AC=26.21) and CTAF001623 (CL=4001, RC=2162, AC=9.53). A visual representation of these mappings is depicted in Figure 37. Tai-SP was practically fully mapped. AT-1 and Kpg3 were mapped almost exclusively in their 3’ halves while HA16-5 was almost exclusively mapped in its 5’ half. The CTAF001623 genotype achieved mappings in mostly its 5’ half, in a scattered fashion.

De novo assemblies constructed 71 CTV-specific contigs, which after genome-wide consensus generation, was made not possible by the previously encountered contig diversity described for isolate 11-5000 and 11-5007. As such, the same strategy was applied as before, which matched CTV-only contigs against reference genomes (Figure 38). Tai-SP (6 reads) obtained the highest metrics above all other matched. It was followed by SY (1 read) and AT-1 (3 reads). The remaining minor hits were for genotype CTAF001623 (1 read), T318A (1 read), Kpg3 (1 read), CT14A (1 read), VT (1 read) and HA16-5 (1 read).

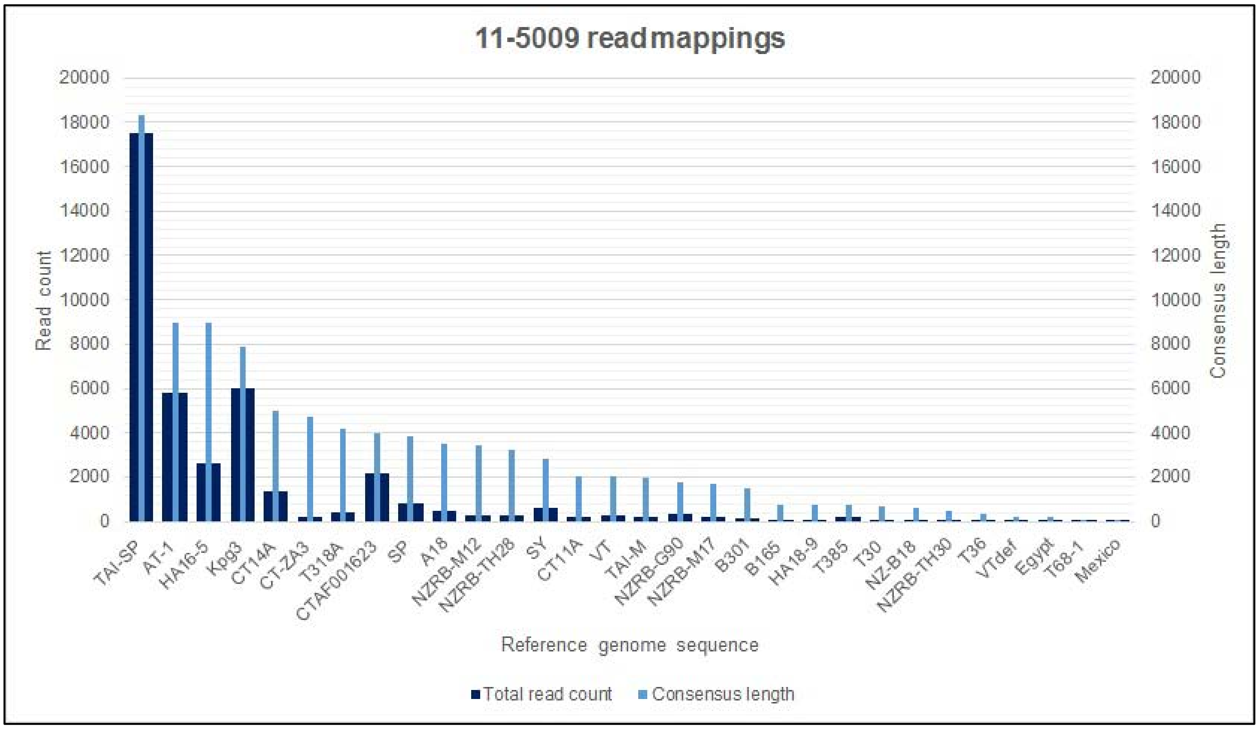


Figure 36: Read mapping results spanning 30 CTV reference genomes for field isolate 11-5009. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

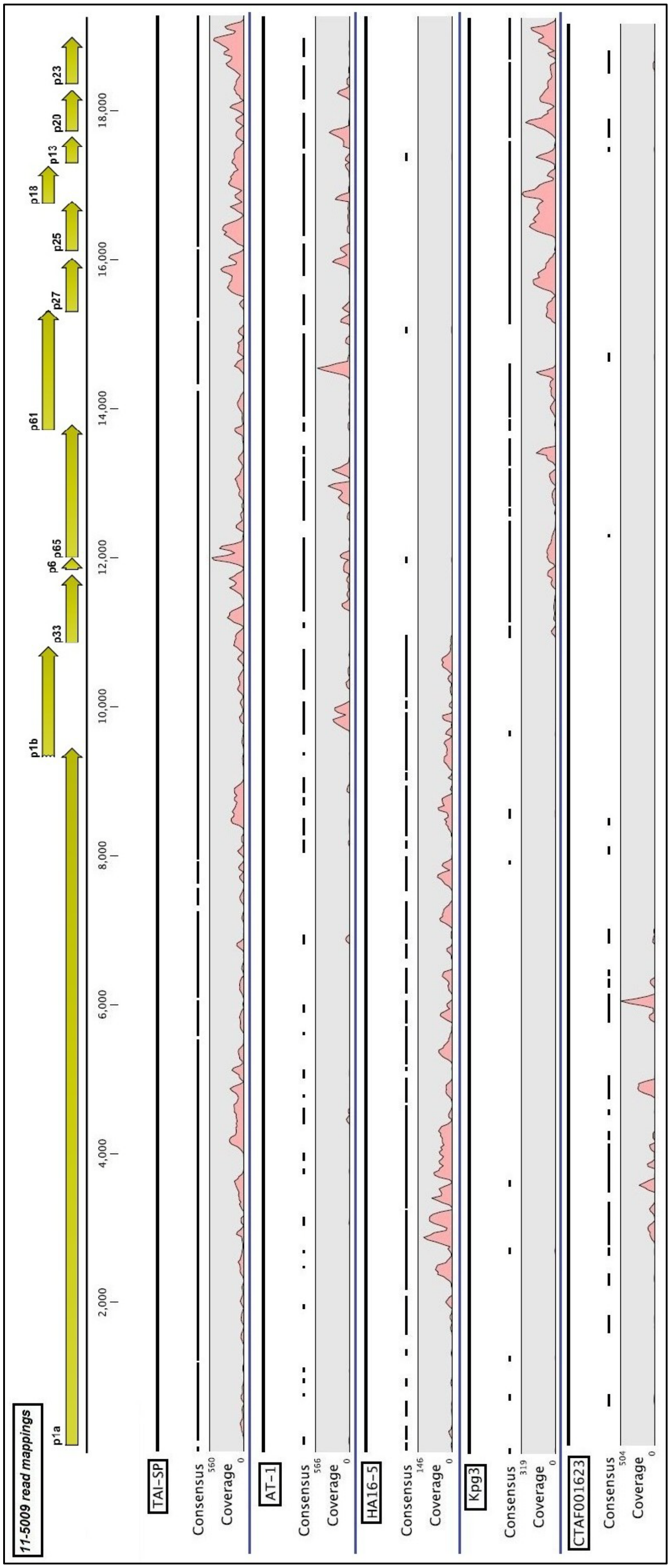


Figure 37: Visual mappings of isolate 11-5009 reads against the five mostly represented genotypes (Tai-SP, AT-1, HA16-5, Kpg3 and CTAF001623). The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the consensus, the coverage scale and graph are represented throughout the mapping of a given reference genome. The curves depicted in the coverage graphs indicate how strong an area of the genome was represented in terms of read count. Individual genome mappings are delineated by blue lines. Note that the scale of coverage values are not constant. Open reading frames are shown on top as yellow arrows with their respective gene products names.

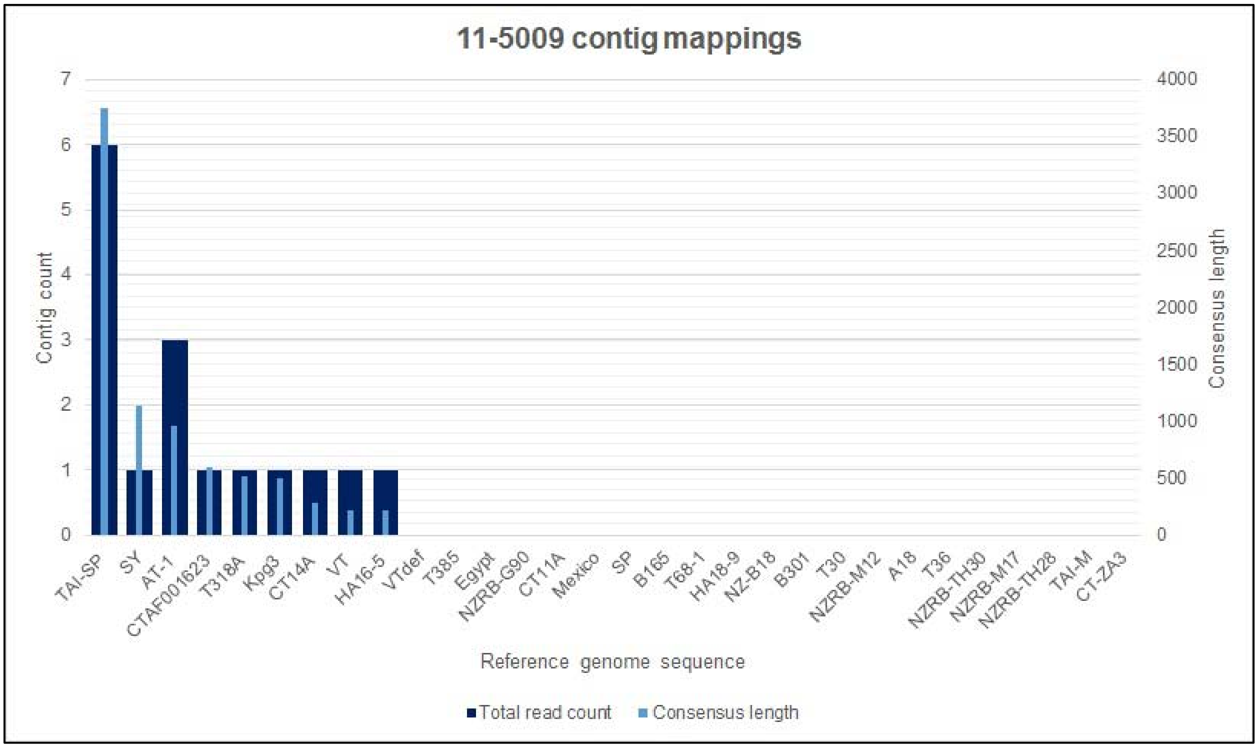


Figure 38: Contig mapping results spanning 30 CTV reference genomes for field isolate 11-5009. The mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by cumulative contigs. The vertical axis scale represents the contig count and consensus length values.

CTV recombination hypotheses

To better understand the sequence diversity and multiple identical mappings observed throughout this study, a recombination analysis was performed using the RDP program. A multiple sequence alignment spanning all CTV genomes as well as CT-ZA3 generated in this study was used as dataset for the analysis. The focus of the analysis aimed to identify recombination patterns between different genotypes. A visual representation showing only the most probable recombination events was generated (Figure 39). From a visual inspection only, several hypotheses regarding CTV recombination behavior could be drawn:

1) *"Like recombines with like".* In most cases, with the only exception of genotype A18, recombination or exchange of genes occurred only between members of

the same clade. For example, the Kpg3 fragment 1b portion (encoding the RdRp) was only shared between members of the VT-like clade only, such as VT, T318A, CT11A, AT-1, SY, and VT-defective. A similar patterns was observed between members of the RB-like clade, where the 5’ NTR was shared between B301, NZ-RB-M12 and NZRB-TH28 only.

2) *Recombination occurs mostly within ORF boundaries*. In other words, recombination is gene-specific or may involve the movement of a whole “gene block”. Moreover, some genes tend to be more exchanged than others. For example, genes such ORF1b, 5’UTR, p65, p27 and p23 were observed to be exchanged extensively between members of the same clade. Three genes, namely p18, p13 and p20, where not exchanged individually, but as a block (=p51 block). However, they did not follow an intra-clade exchange like individual genes. The same p51 block, from Kpg3, was found in HA16-5, HA18-9 and Tai-SP. Genes that were the least shared were p61 and p25, but sometimes (probably due to errors) pieces of adjacent genes (either 5’ or 3’ direction) are recombined along with the original target genes. This was probably the case for p61 and p65.

3) *Gene-specific regions appeared to be hot-spots for recombination.* This was mostly observed for the 5’ half of the p33 gene segment, as well as for portions from the 1a gene.

4) *Genotype A18 was the most diverse recombinant.* More specifically, its 1a portion encompassed portions from genotypes of all clades, including NZ-RB-M17, T68-1, SY, Egypt, and HA18-9. Its 3’ half was unusually conserved, except for its p27 and p23 gene regions.

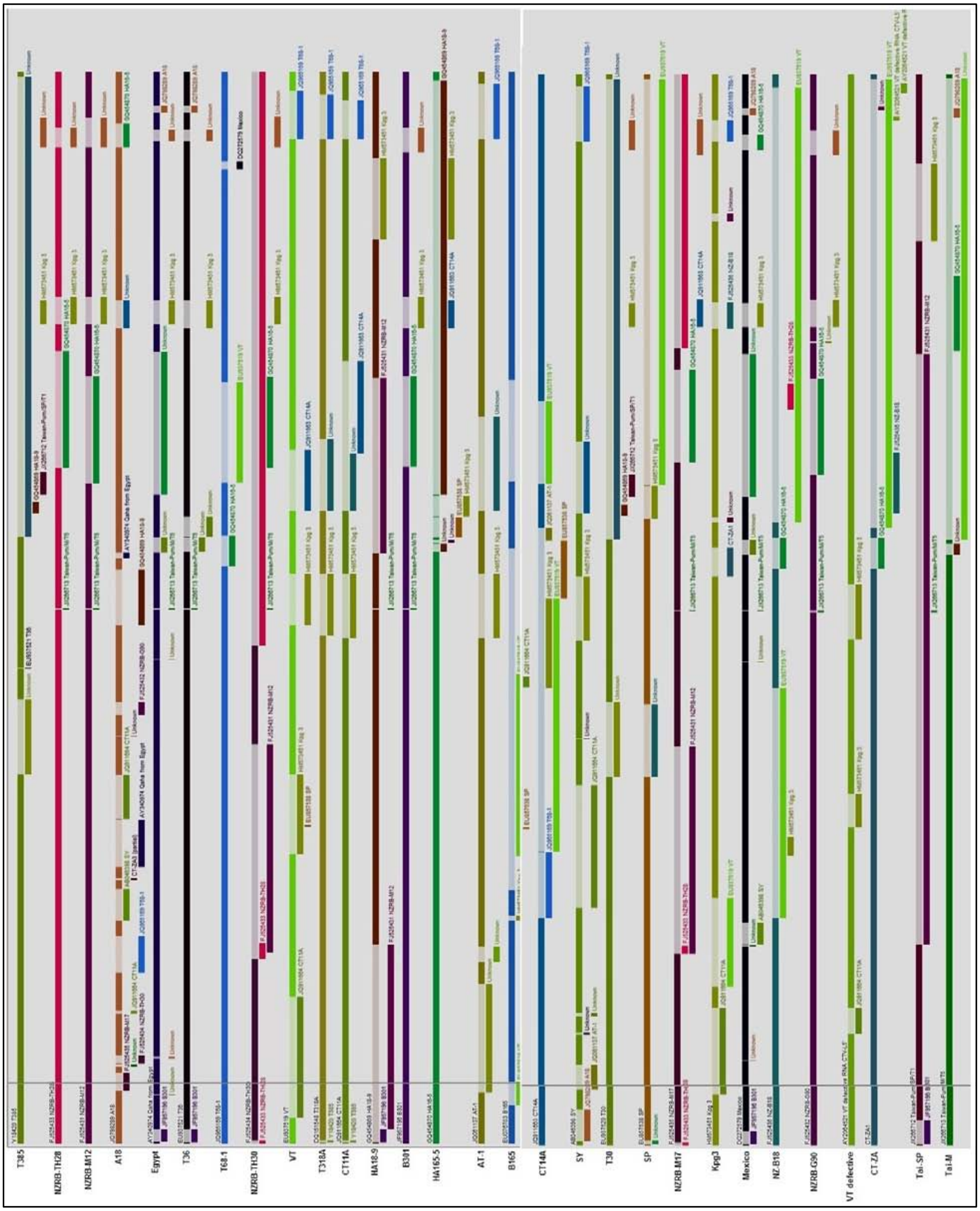


Figure 39: Recombinant analysis across a reference CTV genome set, including CT-ZA3. Each genome is differently color-coded so as to identify each recombinant portion in other other genomes.

4. DISCUSSION

An almost identical analysis pipeline as the one used in the previous chapter was applied on additional GFMS12 sub-isolates, as well as on isolates derived from field conditions. The major difference between this analysis and the previous one resided in dsRNA being the sole template used. Compared to the amount of CTV-specific reads obtained in the previous chapter, the smaller read datasets in this analysis yielded on average 22.76 % enrichment for pure isolates (12-8 and 12-9) and 6.92% for field isolates (11-5000, 11-5007 and 11-5009), which was a good improvement over sub-isolate 12-7 (average of 1.58%).The reason(s) for this increase are not clear, but it may be that the extraction protocol was better performed or/and that virus titers were higher in the samples. These percentages of CTV reads from the total read dataset could have been increased much further if less stringent mapping parameters were used. Despite these percentages, two additional CTV genomes could be assembled *de novo* and were submitted on the GenBank database (accession number KC 333868 and KC 333869). Nonetheless, confirmation of these novel genome sequences should be made with Sanger sequencing via for example, primer walking. As such, dsRNA as a template has overcome its main limitation previously noted, in which it could not yield satisfactory *de novo* assembly results. This may be due the increase in CTV-specific reads obtained and/or higher titers of the virus in these isolates. Regardless of how it performed globally, the nature of dsRNA as a virus-enriching template implied on a biological level a certain bias because it represented only CTV genotypes that were actively replicating ones. Therefore other components of the CTV population might have been missed.

GFMS12 sub-isolates 12-8 and 12-9 were found to be homogeneous for CT-ZA3 or a CT-ZA-like genotype. Although additional genotypes were mapped, it is suspected to be the result of quasi-species reads generated by CT-ZA3 replication. Therefore, it appears that all GFMS12 sub-isolates characterized in this thesis were pure sources of CT-ZA3 variants. This contradicts the results of previous study (Scott *et al.*, 2012). For sub-isolate 12-8, a B165/VT recombinant was suggested to be the dominant component. For 12-9, a mixture of VT-like and B165/VT recombinant was suggested. It is most likely that the B165/VT recombinant was actually CT-ZA3, since it has been shown by our recombination analysis to have most of its 3' half derived from VT. This implies that CT-ZA3 was a main component of the original GFMS12 mother tree population. For all sub-isolates tested, only CT-ZA3 was transferred by aphids, which could imply its dominance within GFMS12 when inoculated in Mexican Lime. Indeed, as far as it known, CT-ZA3 was only detected in Mexican Lime, but not in sweet orange (Zablocki, unpublished data) and grapefruit (Read, unpublished data), as confirmed by dendrograms based on 1a and p23 regions from these Citrus species (data not shown in this study).

Three isolates derived from sweet orange and bud-grafted on Mexican Lime were shown to be a mixture of strains. An RB-like, stem pitting genotype (Tai-SP) was the dominant component of the population across all samples. However, the other co-infecting genotypes could not accurately be assessed due to inconsistencies between reference mapping and *de novo* assemblies. For isolate 11-5000, Kpg3 (VT-like) was the second most represented genotype in terms of read count and consensus length. Visual inspection also showed the Kpg3 genome being almost fully mapped by the reads. However the T30 genotype, originally believed to be present in this isolate based on cloning the A-region, is not believed to be present according to whole-genome characterization. For 11-5007, an additional SP/Kpg3 recombinant is hypothesized to be present, following visual inspection of the read mappings for these two genotypes. For SP, almost all of the 5' half was mapped, and for Kpg3, the same was true for the 3' half. Unfortunately, a full-length recombinant genome could not be generated, probably due to an insufficient amount of reads, high diversity of contig identities and read length being too short. For the last field isolate tested, 11-5009, a putative HA16-5/Kpg3 and/or a HA16-5/AT-1 recombinant may be present, based on visual read mappings. The HA16-5 genotype was almost exclusively mapped in its 5' half (up to the end of the 1b region), while both the AT-1 and Kpg3 genotypes were mostly mapped in their 3' halves (Figure 37). However since Kpg3 mapped more exclusively in its 3' region as compared to AT-1, and that the Kpg3 read mappings started just after the 1b portion of the genome, we are more inclined to believe that the actual recombinant was the HA16-5/Kpg3 combination.

In comparing GFMS12 sub-isolates to field-derived isolates, we could see based on read mapping patterns how clearly a single and mixed infection differed from each other (Figure 40). This in turn confirmed the assumption that GFMS12 sub-isolates 12-7, 12-8 and 12-9 were pure sources, originally obtained through single aphid transmissions. Similarly, this also confirmed the assumption that isolates derived from the field were composed of a mixture of CTV genotypes (isolates 11-5000, 11-5007 and 11-5009).

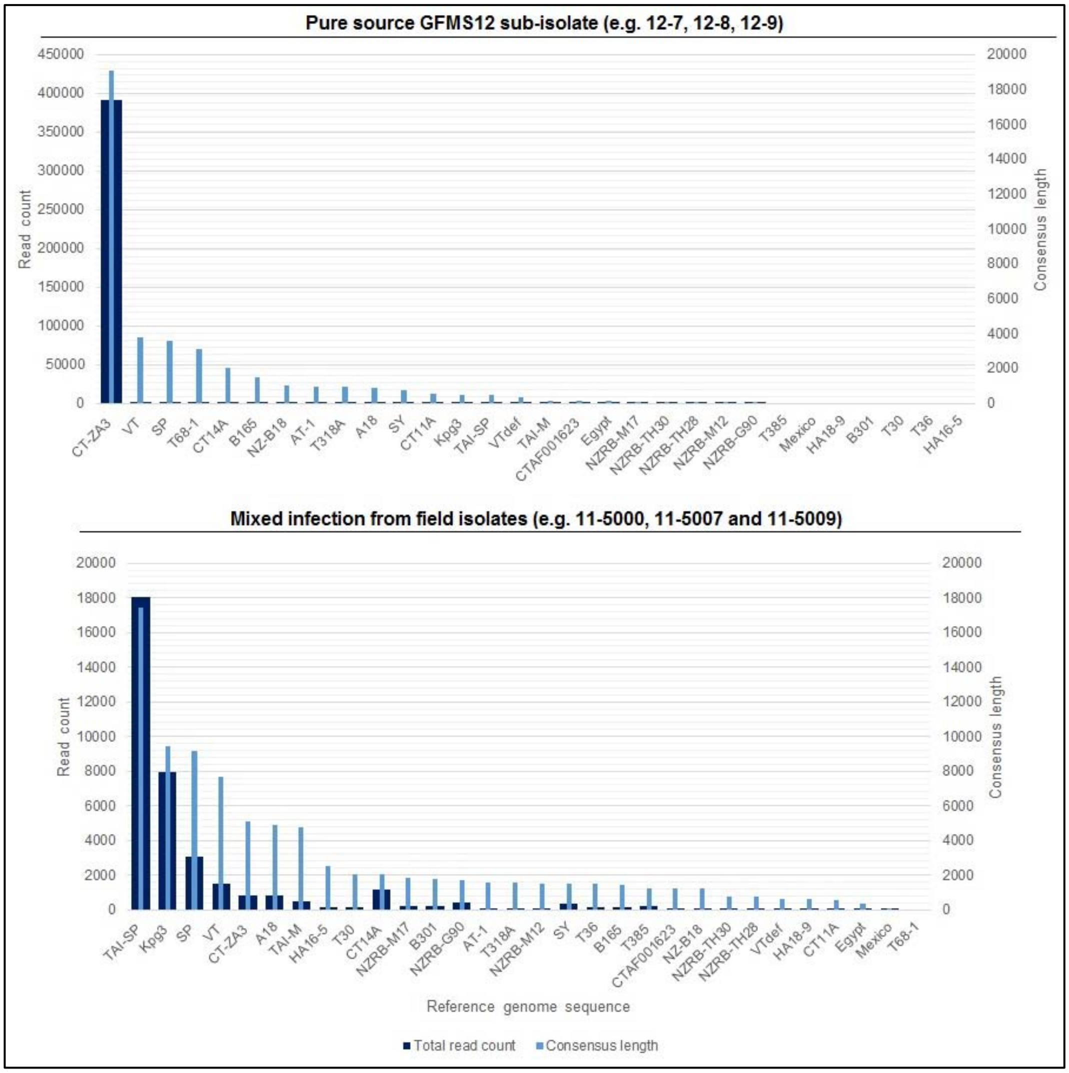


Figure 40: Comparison in terms of read mapping patterns between a single (top) and mixed (bottom) infection sample. Each graph depicts read mapping results spanning 30 CTV reference genomes. Mapped reference genotypes were ranked from largest to smallest in terms of consensus length attained by reads. The primary vertical axis (left) scale represents the amount of reads mapped. The secondary vertical axis (right) scale represents the consensus length in nucleotides. Read count= thick dark blue bars; Consensus length= thin light blue bars.

The genotype-shifting model, used in the previous chapter to explain why reads mapped to additional genotypes despite the source being homogeneous, could also be applied for the mappings patterns observed in this chapter. For pure sources (12-8 and 12-9), very few additional additional reads mapped to other genotypes, which was also the case of sub-isolate 12-7. The effects on read mapping from genotype-shifting mutations and/or past recombinations in a mixture of strains (field isolates) were much more pronounced. This made sense, since many more polymorphic regions would be generated by the combination of two or more genotypes co-infecting its host (this was also reflected in Figure 40, where a greater diversity of genotypes were mapped). This may explain why T30 clones in isolate 11-5000 were originally observed in the A-region. Additionally, it is suggested that for isolate 11-5009, where a putative HA16-5/Kpg3 recombinant is most likely present, we suggest that the other potential recombinant combination (HA16-5/AT-1) may be the result of genotype-shifting mutation and/or recombination originating from the HA16-5/Kpg3 chimeric genome. It may also be that both recombinants exist, but is less unlikely. To test this, it would be necessary to obtain the full-length *de novo* chimeric genome by obtaining longer reads (e.g. 454 pyrosequencing). This genome would then be mapped against the 11-5009 dataset and the resulting mappings should shed light into which genotype “donated” its 3’ half (Kpg3 or AT-1).

Pertaining to *de novo* assemblies in general, full-genome reconstruction could only be achieved for pure isolates but not for field-derived ones due to their high contig diversity (in terms of genotype identity). This diversity, which could be called “the field mapping effect” (Figure 41) could be attributed to not only a population of different genotypes, but also their resulting quasi-species (genotype-shifting theory) due to their cumulated effects of co-replicating together. The resulting contigs, when aligned against a set reference genomes, could not create consensus sequences because no identical overlapping regions were found. This phenomenon can be best visualized in Figure 40. The contig mappings in (a) and (b) show several patterns identified by numbers. For pattern 1, 2 and 4, color-coded contigs mapped to identical regions of the genome, but due to their sequence diversity, could not create an accurate consensus sequence in these regions due to their unique nucleotide sequences, probably because these are genotype specific. Pattern 6, 7 and 8 depict that for some areas of the genome, there was only one contig that spanned that genomic region, and so can be considered a conserved region across genotypes. For some regions, pattern number 3 was observed, were only two to three contigs spanned an identical genome region, and were differing in sequence very mildly. Lastly, pattern 5 and 9 depict the ideal consensus generation scenario were only a small contig overlap occurs, which allows a minimum of ambiguities in consensus generation while spanning a long portion of the genome. This “ideal” scenario was mostly observed in “pure” infection samples, such as in sub-isolate 12-8 where only one genotype was present (b). The opposite of this ideal scenario was mostly found for mixed infections.

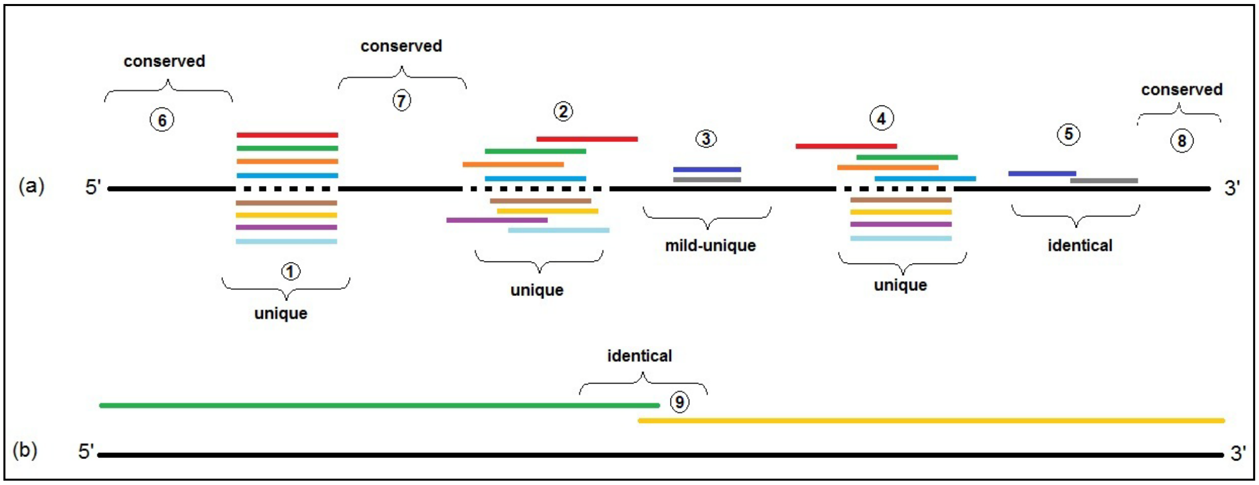


Figure 41: Effects of contig diversity on whole-genome alignments. (a) Contig mapping in a mixture of strains and (b) Contig mapping in a “pure” isolate. Black lines interspersed by dots represent the full-genome length of a typical CTV. The color lines represent pieces of contigs that have mapped to a particular region along the type genome. Numbers 1, 2, 4 indicate regions with high diversity of contigs; numbers 6, 7, 8 represent conserved regions of the genome; numbers 5, 9 indicate identical overlapping contig regions and 3 represent a mildly unique region. The mapped regions depicted in the figure do not reflect the actual mapping portions obtained. A detailed explanation of each numbers indicated is provided in text.

As an alternative to building *de novo* genomes for field isolates, we mapped individual contigs against a set of reference genome to correlate their mapping results with read mapping results. Although it correlated mostly (but not always) with read mapping results, it is not recommended to use this method for future characterization. Contigs were in majority short (between 200 and 4000 nucleotides long), and therefore only spanned small, random regions of the CTV genome. This cannot be used to assume the presence/absence of a given genotype solely based on such a small region.

To elucidate the identities of the observed recombined genome regions, as well as their extent, an RDP analysis was performed, encompassing all known reference genomes as well as the three CT-ZA genomes generated in this study. It was observed that only certain genes (mostly in the 3' half) were recombined, while others were not. In a previous study, Satyanarayana *et al.* (2011) determined that for example, genes involved for replication and assembly tended to be more conserved according to sequence identities, but our RDP analysis suggested that although this is true, it was because the gene (1b) is shared across members of only a certain clade, which have all in origin the same genotype, Kpg3. In this sense, it may be that to increase its survival fitness, the best replication enzyme gene is being shared between closely related genotypes. The extensive range of recombination observed in this study, but also by Weng *et al*. (2007), combined with the quasi-species nature of this virus was the main cause for the observed mapping problems observed with the NGS analysis as well as the incongruences observed in cloning-based characterization studies.

Conclusions:

- A CTV genome was assembled *de novo* in each GFMS12 sub-isolate (12-8 and 12-9). Both genomes were found to be essentially identical to CT-ZA1 (Chapter 1)
- Both GFMS12 sub-isolate 12-8 and 12-9 were found to be homogeneous for a CT-ZA-like genotype
- Field isolates were found to be a mixture of strains:
    - Isolate 11-5000: Tai-SP (RB-like) as a main component and Kpg3 (VT-like) as a minor component. No evidence for the presence of the T30 genotype was found.
    - Isolate 11-5007: Tai-SP (RB-like) as a main component and a putative SP/Kpg3 recombinant as a minor component.

  - Isolate 11-5009: Tai-SP (RB-like) as a main component and a putative HA16-5/Kpg3 or HA16-5/AT-1 recombinant as a minor component.
- In terms of read mapping patterns against multiple reference genomes, a clear distinction can be made between a single and mixed infection.
- The effect of quasi-species on reference-based read assemblies were more pronounced in multiple infections.
- *De novo* assembly of full-length CTV genomes could not be achieved in field isolate datasets

5. REFERENCES

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26, 2462-2463.

Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596-1599.

Scott, K.A., Hlela, Q., Zablocki, O., Read, D., van Vuuren, S., and Pietersen, G., 2012. Genotype composition of populations of grapefruit-cross-protecting citrus tristeza virus strain GFMS12 in different host plants and aphid-transmitted sub-isolates. *Archives of Virology* DOI: 10.1007/s00705-012-1450-4.

Weng, Z., Barthelson, R., Gowda, S., Hilf, M.E., Dawson, W.O., Galbraith, D.W., and Xiong, Z., 2007. Persistent infection and promiscuous recombination of multiple genotypes of an RNA virus within a single host generate extensive diversity. *PLoS One* 2(9): e917.

Chapter 4:

CONCLUDING REMARKS

4.1 INTRODUCTION

Citrus tristeza virus is a complex virus, in terms of its associations with a multitude of *Rutaceae* hosts, symptom expression, pathogenicity determinants but also its high genetic variability. With its error-prone RNA replicase, promiscuous recombination tendencies and quasi-species nature, the virus creates a multitude of closely related genome variants, the result of which proved to be a challenging task to assemble back together in this multi-million piece puzzle. Unbiased, ultra-deep sequencing by means of the Illumina platform for the characterization of CTV isolates required a polyphasic bioinformatics approach, encompassing quality control of the reads, reference assembly, *de novo* assemblies, multiple-sequence alignments, phylogenetic relationship inference and detection of recombination signals. All these techniques were used in concert to validate the results of each other and for pattern detection. This chapter aims to summarize lessons that have been learned throughout this work as well as prescribing potential solutions for future studies involving next-generation characterization of a known virus.

4.2 RECOMMENDATIONS

One of the key features of this work was to remove bias as much as possible, so as to obtain a global view of CTV genetic diversity on a whole-genome level. The use of random priming proved highly efficacious for this task, but came at a certain price. From our data, we saw that up to 23% of the read dataset was virus-derived, where the rest mostly consisted of host-derived components. There are several ways to enrich for the virus besides the choice of template (which in our case was double-stranded RNA). Turner *et al.* (2009) and Althaus *et al.* (2012) have suggested DNase/RNase treatment, filtration, density gradient centrifugation and subtractive hybridization or a combination of several of these as ways to minimize unwanted nucleic acids.

Another major problem was the high degree of single nucleotide variants (SNVs) obtained when we performed reference assembly of the reads, a universal problem in next-generation sequencing projects. Next-generation sequencers require high amounts of starting DNA, which forced a reverse-transcription step to characterize CTV (ssRNA (+) virus). As any amplification step introduce errors (Beerenwinkel *et al.,* 2012), such as reverse-transcription PCR, Illumina library preparation and bridge amplification, it becomes almost impossible to discern between experimental errors from genuine mutations. Many software suites offer quality control and the ability to filter reads based on quality scores and read length (bases towards the end of a read tend to have higher error rates). In our study, we solved the problem by first trimming reads based on quality scores, but mainly by applying the most stringent mapping parameters. This ensured that any mapped read was exactly identical to the reference. This, in effect, was a filtering device which eliminated indels, whether genuine or not. This induced a loss in observable viral diversity, but because the goal of this study was to determine the main genotypes present in the samples, it was acceptable to apply this method. But even with stringent mapping parameters, a substantial amount of reads still mapped to other minor components despite one or two dominant genotypes. To explain this, we suggested the “genotype-shifting theory” (see discussion in Chapter 2).

One of the most powerful by-products from next-generation sequencing is the potential to assemble novel, previously unknown virus genomes by means of *de novo* assembly algorithms. In this study, we generated three CTV genomes (CT-ZA1, CT-ZA2 and CT-ZA3), one from each pure GFMS12 sources. All three genomes were found to be essentially identical as well as being the dominant component in these sources. This is the first report of a fully sequenced Citrus tristeza virus isolate in South Africa. However, the same could not be achieved for field-derived isolates, due to a large contig population diversity which yielded many differing overlapping regions. To solve this problem, we subjected the contigs to reference assembly, but as discussed in Chapter 3, this is not recommended for future analyses due to these contigs only spanning defined regions of the genome, which cannot be used to infer the presence of a genotype. What would be ideal is a quasi-species reconstruction algorithm. This would entail identifying conserved areas of the genome, fixing them into a frame and incorporating every variable regions obtained. Such softwares have recently been released, such as ShoRAH (Zagordi *et al.*, 2011) and QuRe (Prosperi and Salemi, 2012). Since quasispecies reconstruction would take into account variable section of the genome, both softwares include a strong error-correction pipeline, so as to minimize the inclusion of experimentally-induced indels. Once reads have been deemed genuine by the software, it calculates the haplotype frequencies by counting the amount of all variable genomic regions compared to a reference sequence (ShoRAH). QuRe goes one step further, has it is capable of not only calculating haplotype frequencies, but to also include them in the reference genome, thus creating a population of genome variants in all their respective frequencies. Furthermore, previous versions of QuRe were limited to Mac or other Unix-based operating systems only, but has now been implemented into a Java environment. This allowed the program to run under a wider diversity of operating systems, including Windows, mostly used by computer neophytes. This is a good indication that more and more tailor-made softwares for complex problems are beginning to emerge and are starting to be accessible to biologist and not restricted only to bioinformaticists. This trend is expected to grow in the near future.

Another major consideration for the optimization of *de novo* assemblies is read length and amount (Quail *et al.,* 2012). In their review, three third generation sequencers (Ion Torrent, PacBio and Illumina MiSeq) were compared in many aspects, including sequence generation, workflow, genome coverage, error rates and SNP calling. In general, in terms of read length, size matters. More specifically, the longer the read length, the higher the probability to obtain overlapping regions. Therefore, for studies requiring *de novo* reconstruction of virus genomes, a sequencing platform that generates long, low-error rate reads, is preferred. Such sequencers are the 454 GS FLX Titanium XL+ and the Pacific Biosciences (PacBio) RS machines. Both are currently the only platforms capable of generating reads in average lengths of 700 bp and 1500 bp, respectively. However, these two instruments come at a high cost, impeding their wide use. Furthermore, PacBio's instrument, although generating the longest reads, have been shown to generate the most error-prone reads across all platforms (13%), while Illumina have rates below 0.4%. Additionally, *de novo* assembly with very long reads has proven less efficacious than with short ones, due to the current algorithms' design optimized for read lengths of about 100 bp. It is expected that in the near future, these issues will be fixed, as this technology has not had time to mature as much as Illumina or 454 from Roche. In summary, 454 is at present the recommended sequencing platform for the analysis of field isolates in cases where there is a clear lack of an appropriate reference genome, therefore requiring *de novo* analysis.

Lastly, there remains the question of when is it most appropriate to make use of next-generation sequencing technologies. Despite sequencing costs having experienced a major price drop in the last couple of years (a currently ongoing trend), it still remains a major budget not to be taken slightly due to the cost of sequencing kits used for the preparation of sequencing-ready libraries (Rees, J., personal communication). The decision must also be based on the nature of the research question(s). For example, NGS is recommended for in-depth characterization of pure CTV sources only if incongruences arise from an initial screening between clones across multiple regions of the genome. This might be an indication that a putative recombinant is present and therefore require *de novo* assembly to identify it. If both gene regions are in accord, this approach is less useful, unless other aspects are under investigation. Alternatively, NGS could be used not on a whole genome level, but focus only on amplicons derived from specific gene regions. This method is recommended for survey purposes, where hundreds of samples may be involved. Because of the potential high number of samples, too many reads would be required to cover the full genome, even one as small as a virus. The crucial choice would then shift to the selection of the most suitable candidate gene region(s). Moreover, to save time, funds and augment the sequencing throughput, multiple sample should be sequenced in a single run, by means of bar-coding. If this option is chosen, it will entail reducing the amount of reads generated for every sample, thus reinforcing the idea that only multi-gene amplicons should be used as templates so as to achieve decent coverages per sample.

4.3 REFERENCES

Althaus, C.F., Vongrad, V., Niederost, B., Joos, B., Di Giallonardo, F., Rieder, P., Pavlovic, J., Trkola, A., Gunthard, H.F., Metzner, K.J., and Fischer, M., 2012. Tailored enrichment strategy detects low abundant small noncoding RNAs in HIV-1 infected cells. *Retrovirology* 9:27.

Beerenwinkel, N., Gunthard, H.F., Roth, V., and Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology* 3: 329.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341.

Prosperi, M.C.F., and Salemi, M., 2012. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* 28(1): 132-133.

Turner, E.H., Ng, S.B., Nickerson, D.A., and Shendure, J., 2009. Methods for genomic partitioning. *Annual Reviews in Genomics and Human Genetics* 10:263-284.

Zagordi, O., Bhattacharya, A., Eriksson, N., and Beerenwinkel, N., 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12:119.

APPENDIX 1: Agarose gel electrophoresis photographs

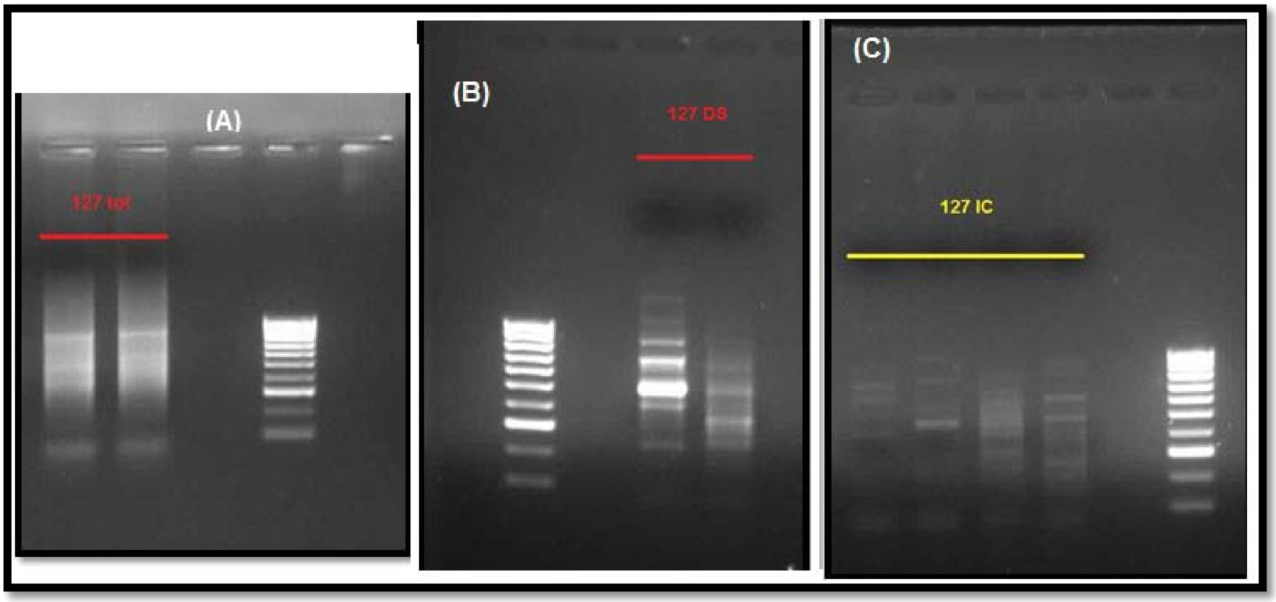


Figure 41: Agarose gel electrophoresis of randomly generated amplicons derived from sub-isolate 12-7. (A) Total RNA; (B) double-stranded RNA; (C) Virus particles through immunocapture. All molecular ladders are 100 bp (1 kb maximum).

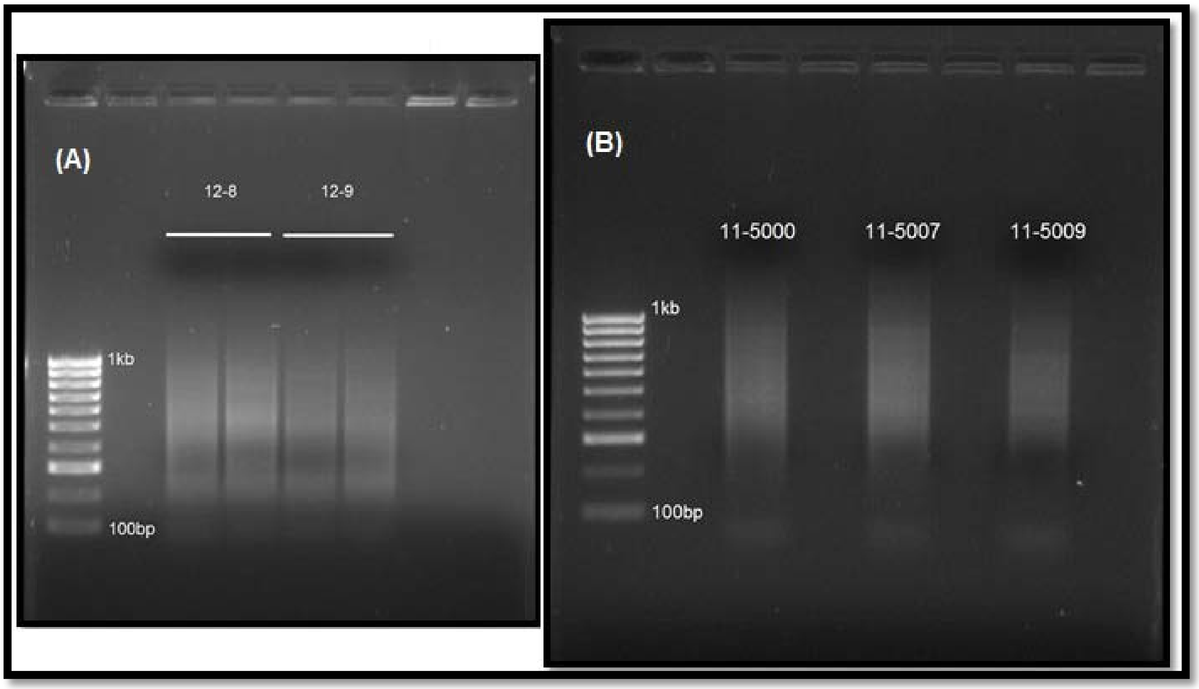


Figure 42: Agarose gel electrophoresis of randomly generated amplicons derived from glasshouse sub-isolates and field isolates. (A) Sub-isolate 12-8 and 12-9; (B) field isolates, from left to right: 11-5000, 11-5007 and 11-5009. All molecular ladders are 100 bp (1 kb maximum).

APPENDIX 2: Clone sequence data

Sub-isolate 12-7, p23 gene:

```
>127.01

ACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAACACTCAGACTCACTGGGTATTCCGACGAA
GTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAACATTGCCTTAGCAGTTATACCGAACTTATA
ATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGCCAATCTCGTCTTCTCCCTTTCAGCGGCCA
AATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGATCGCAAACATCTCTGCGTTAGAAAAAGCT
CTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAACATATGCGCCACCTCGGCCTGAGATTGTGT
ATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTGTTTTCTACCACAATCCACGCACACTCCTA
TTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTATCAAGGCGTCGATAATAAAGGGATTCATT
TTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTGACTTCAGTGCTAGCTGTGTTGCTTTCGTC
AGAAAGGNTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCATCGTAACTCGCAGACTCTTAATAATAAATC
TGTTCGAATTAAAGTTAGTAATCG
```

```
>127.02

AATTCACTAGTGATTACTAACTTTNNTTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGGA
GTTCTCCGTGAA
```

```
>127.04

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCG
```

```
>127.05

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCCCGCGGC

>127.06

AATTCACTAGTGATTACTAACTTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCTAGTGAGTCTGAGTGTTTCAGGCGGA
GTTCTCCGTGAACACCACTTCACCTGATTGAAGTGGACGGA

>127.07

AATTCACTAGTGATTACTAACTTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACGCGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCACAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAAGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGTGATAACGAGGACACTTC

>127.08

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGATGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
```

```
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTT

>127.09

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGACTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATC

>127.10

AATTCACTAGTGATTACTAACTTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGGA
GTTCTCCGTGAACACCACTTCACCTGATTGAAGTGGACGGAATAA

>127.11

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTA

>127.13

AATTCACTAGTGATTACTAACAATAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
```

ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGTGATAACGAGGACACCTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGGA
GTTCTCCGTGAACACCACTTCACCTGATTGAAGTGGACGGAATAAGTTAATC

>127.14

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACCGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAAT

>127.15

AATTCACTAGTGATTACTACCCTTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATA
CTAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTA
AGTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAAC
CAATTATCGGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGG
CATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCAC
GATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGT
TATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGAC
ACCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGAC
ATCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGG
AGTTCTCCGTGAACACCACTTCACCTGATTGAAGTGGACGGAATAAGTTAATCG

>127.16

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCCCG

>127.17

AATTCACTAGTGATTACTAACTTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTACTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA

```
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGGA
GTTCTCCGTGAACACCACTTCACCTGATTGAAGTGGACGGAATAAGTTAATCGAATTCCCGCGGCC

>127.18

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGATGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
GCTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCATC
GTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCCCGCG

>127.19

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGATGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACATTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAATAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAA

>127.20

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGATTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGCTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAA
```

```
>127.21

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCGCGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTA

>127.22

AAAACCGTAAGTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGAT
ACGGAAAACCAATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAAC
ACGATAAGGCATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATG
TTAATGCACGATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGAT
CGATTTGGTTATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGG
CTCGTAGACACCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATG
TTACCGGACATCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGT
TTCAGGCGGAGTTCTCCGTGAACACCACTTCATCTGATTGAAGTGGACGGAATAAGTTAATCGAATTCCCGCGG
CC

>127.23

TTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATACTAGCGGACAAACTTTTATTTC
TGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAAGTTTGGAAGCGGATCGCTTGG
AATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACCAATTATCAGGGTGCTCGCTTT
CGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGCATCGAGGACTGAACGTAAGTG
TAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACGATCCCGTGAAATATTTAAATA
AAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTTATGCATACCAAAGAAAGGCAA
TTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACACCCGATGCGTTCTCCGGAGGA
AACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACATCAACGCTATAGACGTTGGTG
ATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGGAGTTCTCCGTGAACACCACTTC
ACCTGATTGAAGTGGACGGAATAAGTT

>127.24

CCGGTAACATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGT
CTACGAGCCAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAAC
CAAATCGATCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGT
GCATTAACATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCC
TTATCGTGTTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGT
TTTCCGTATCAAGGCGTCGATAGTAAAGGGATTCATTTTCTGTAAAAATTCCAAGCGATCCGCTTCCAAACTTA
CGGTTTTGACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAATAAAAGTTTGTCCGCTAGTA
TCGTCCATCGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCCCGCGGCC
```

```
>127.25

AACCGTAAGTTTGGAGCGGATCGCTTGGAATTTTTACGGGGAAAATGAATCCCTTTATTATCGACGCCTTGATA
CGGAAAACCAATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACA
CGATAAGGCATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGT
TAATGCACGATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATC
GATTTGGTTATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGC
TCGTAGACACCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGT
TACCGGACATCAACGCTATAGACGTTGGTGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTT
TCAGGCGGAGTTCTCCGTGAACACCACTTCACCTGATTGAAGTGGACGGAATAAGTTAATCGAATTCCCGCGGC
C

>127.29

GCCTGAACACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTC
CGGTAACATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTC
TACGAGCCAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACC
AAATCGATCGCAAACATCTCTGCGTTGGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTG
CATTAACATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCT
TATCGTGTTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTT
TTCCGTATCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTAC
GGTTTTGACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAATAAAAGTTTGTCCGCTAGTAT
CGTCCATCGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCCCGCG

>127.30

AATTCACTAGTGATTAACTTATTCCGTCCACTTCAATCAGGTGAAGTGGTGTTCACGGAGAACTCCGCCTGAAA
CACTCAGACTCACTGGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCC

>127.E1

AATTCACTAGTGATTAACTTATTCCGCCCACTTCAATCAGATGAAGTGGTGTTCTCGGAGAACTCCGCCTGAAA
CACTCAGACTCACTAGGTATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAAC
ATTGCCTTAGCAGTTATACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGC
CAATCTCGTCTTCTCCCTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGA
TCGCAAACATCTCTGCGTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAAC
ATATGCGCCACCTCGGCCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTG
TTTTCTACCACAATCCACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTA
TCAAGGCGTCGATAATAAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTG
ACTTCAGTGCTAGCTGTGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCAT
CGTAACTCGCAGACTCTTAATAATAAAATCTGTTCGAATTAAAGTTAGTAATCGAATTCCCGCGG
```

```
>127.E2

AATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAACATTGCCTTAGCAGTTAT
ACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGCCAATCTCGTCTTCTCCC
TTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGATCGCAAACATCTCTGCG
TTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAACATGTGCGCCACCTCGGC
CTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTGTTTTCTACCACAATCCA
CGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTATCAAGGCGTCGATAATA
AAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTGACTTCAGTGCTAGCTGT
GTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCATCGTAACTCGCAGACTCT
TAATAATAAATCTGTTCGAATTAAAGTTAGTAAT

>127.E3

AAATTCCGACGAAGTGTCCTCGTTATCACCAACGTCTATAGCGTTGATGTCCGGTAACATTGCCTTAGCAGTTA
TACCGAACTTATAATGTTCCGGAGTTTCCTCCGGAGAACGCATCGGGTGTCTACGAGCCAATCTCGTCTTCTCC
CTTTCAGCGGCCAAATCAACCGCTAATTGCCTTTCTTTGGTATGCATAACCAAATCGATCGCAAACATCTCTGC
GTTAGAAAAAGCTCTAGCTTTTCTTTTATTTAAATATTTCACGGGATCGTGCATTAACATATGCGCCACCTCGG
CCTGAGATTGTGTATTGTTGACTTTACACTTACGTTCAGTCCTCGATGCCTTATCGTGTTTTCTACCACAATCC
ACGCACACTCCTATTATTCTCGCGCGAAAGCGAGCACCCTGATAATTGGTTTTCCGTATCAAGGCGTCGATAAT
AAAGGGATTCATTTTCCGTAAAAATTCCAAGCGATCCGCTTCCAAACTTACGGTTTTGACTTCAGTGCTAGCTG
TGTTGCTTTCGTCAGAAAGGTTCACAGAAGTAAAAGTTTGTCCGCTAGTATCGTCCATCGTAACTCGCAGACTC
TTAATAATAAATCTGTTCGAATTAAAGTTAGTAATCGAATTC

>127.E4

AATTCACTAGTGATTACTAACTTTAATTCGAACAGATTTATTATTAAGAGTCTGCGAGTTACGATGGACGATAC
TAGCGGACAAACTTTTATTTCTGTGAACCTTTCTGACGAAAGCAACACAGCTAGCACTGAAGTCAAAACCGTAA
GTTTGGAAGCGGATCGCTTGGAATTTTTACGGAAAATGAATCCCTTTATTATCGACGCCTTGATACGGAAAACC
AATTATCAGGGTGCTCGCTTTCGCGCGAGAATAATAGGAGTGTGCGTGGATTGTGGTAGAAAACACGATAAGGC
ATCGAGGACTGAACGTAAGTGTAAAGTCAACAATACACAATCTCAGGCCGAGGTGGCGCATATGTTAATGCACG
ATCCCGTGAAATATTTAAATAAAAGAAAAGCTAGAGCTTTTTCTAACGCAGAGATGTTTGCGATCGATTTGGTT
ATGCATACCAAAGAAAGGCAATTAGCGGTTGATTTGGCCGCTGAAAGGGAGAAGACGAGATTGGCTCGTAGACA
CCCGATGCGTTCTCCGGAGGAAACTCCGGAACATTATAAGTTCGGTATAACTGCTAAGGCAATGTTACCGGACA
TCAACGCTATAGACGTTGGAGATAACGAGGACACTTCGTCGGAATACCCAGTGAGTCTGAGTGTTTCAGGCGGA
GTTCTCCGAGAACACCACTTCATCTGATTGAAGTGGACGGAATAAGTTAATCGAATTCCCGCGGC
```

APPENDIX 3: PHRED scores for Illumina datasets

Sub-isolate 12-7, double-stranded RNA

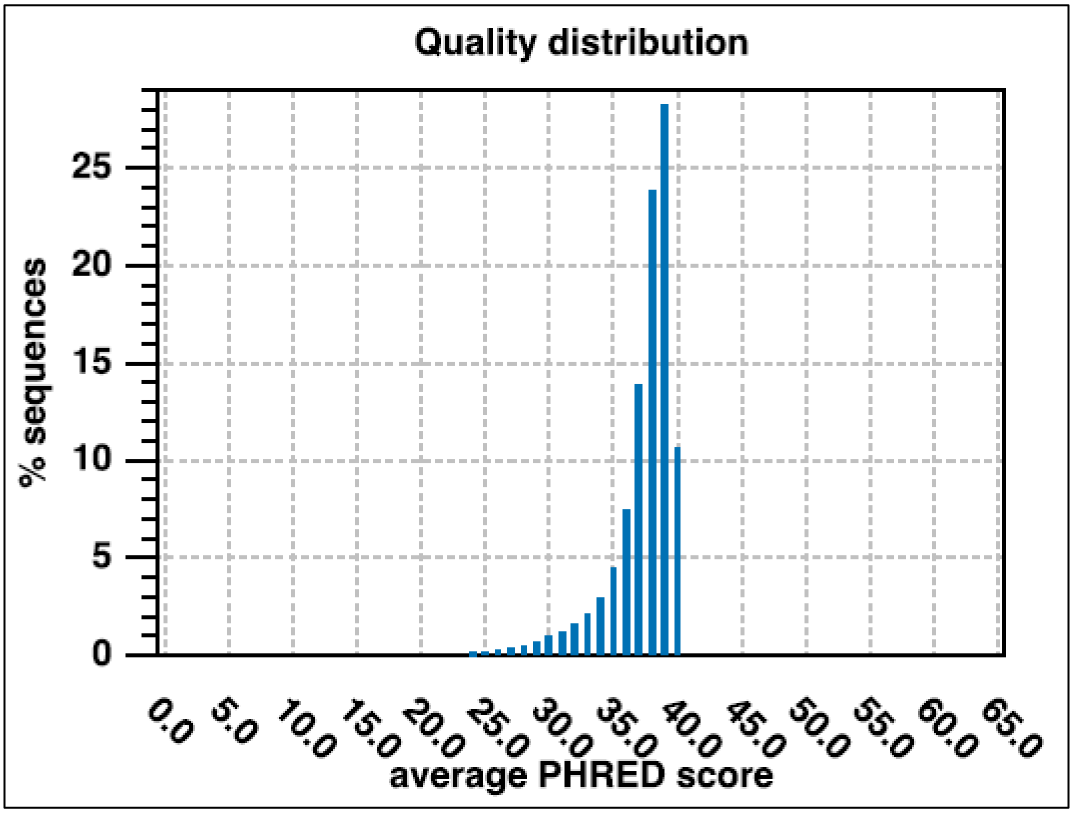


Figure 43: Distribution of average sequence quality scores expressed as PHRED scores for sub-isolate 12-7 (double-stranded RNA). The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Sub-isolate 12-7, total RNA

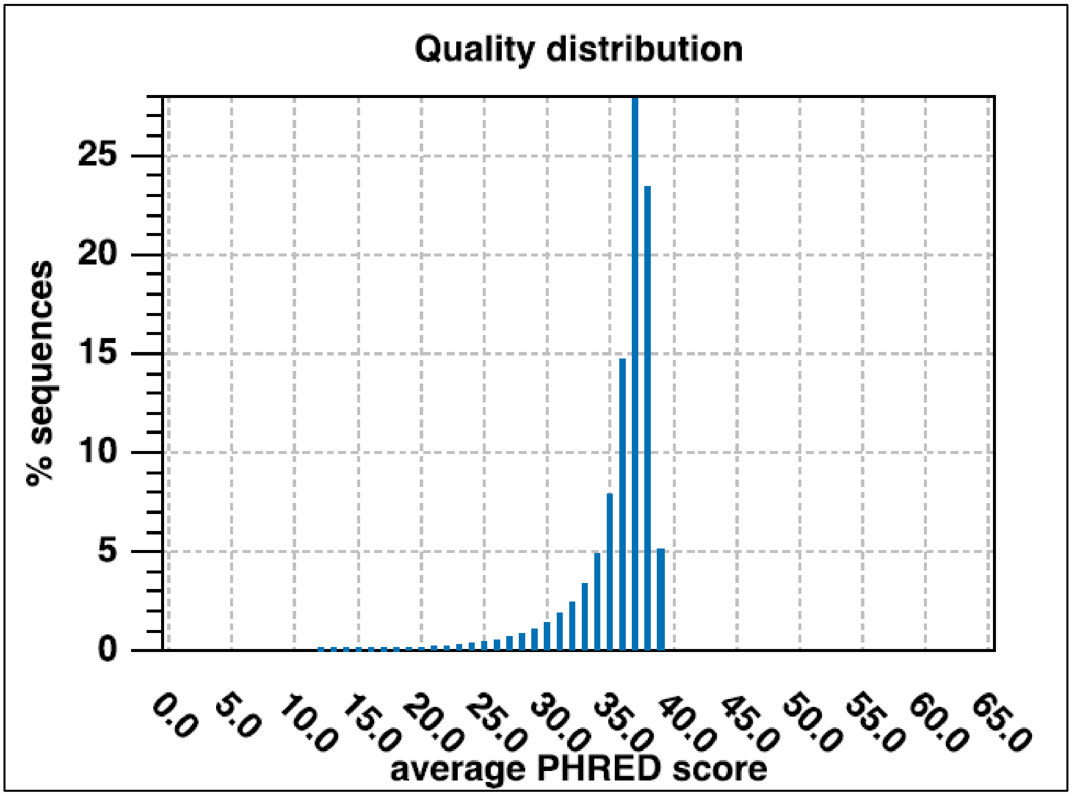


Figure 44: Distribution of average sequence quality scores expressed as PHRED scores for sub-isolate 12-7 (total RNA). The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Sub-isolate 12-7, virus particles by immunocapture

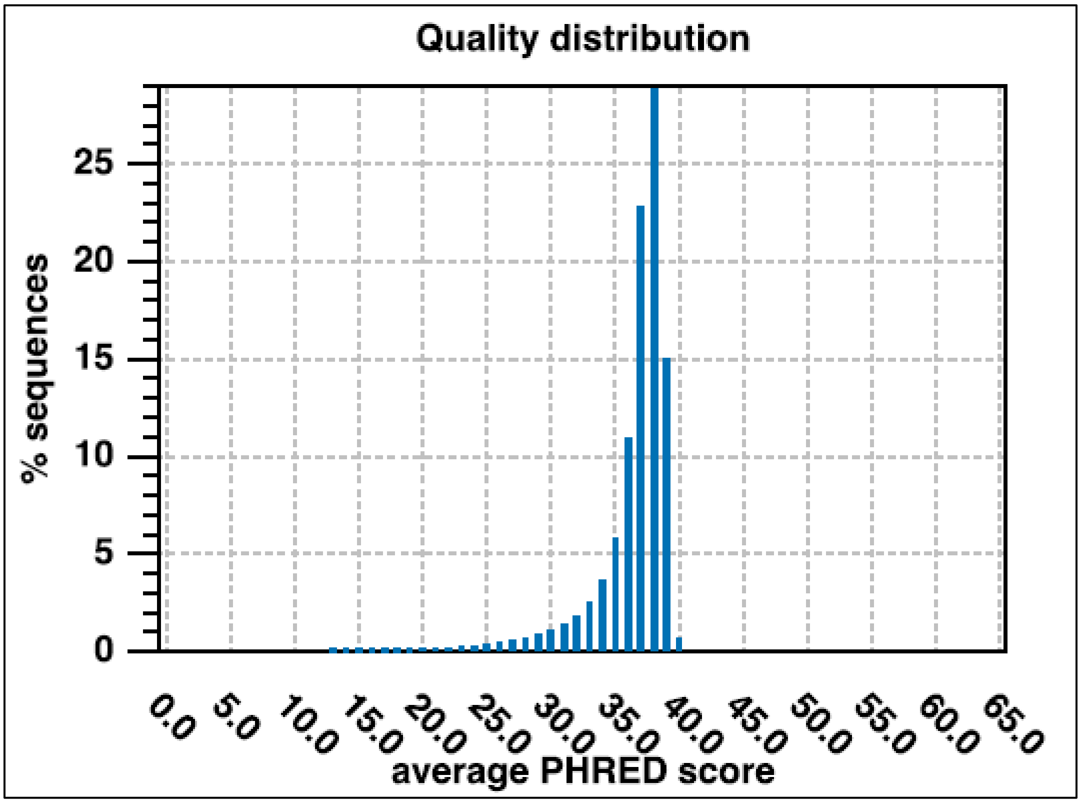


Figure 45: Distribution of average sequence quality scores expressed as PHRED scores for sub-isolate 12-7 (virus particles). The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Sub-isolate 12-8

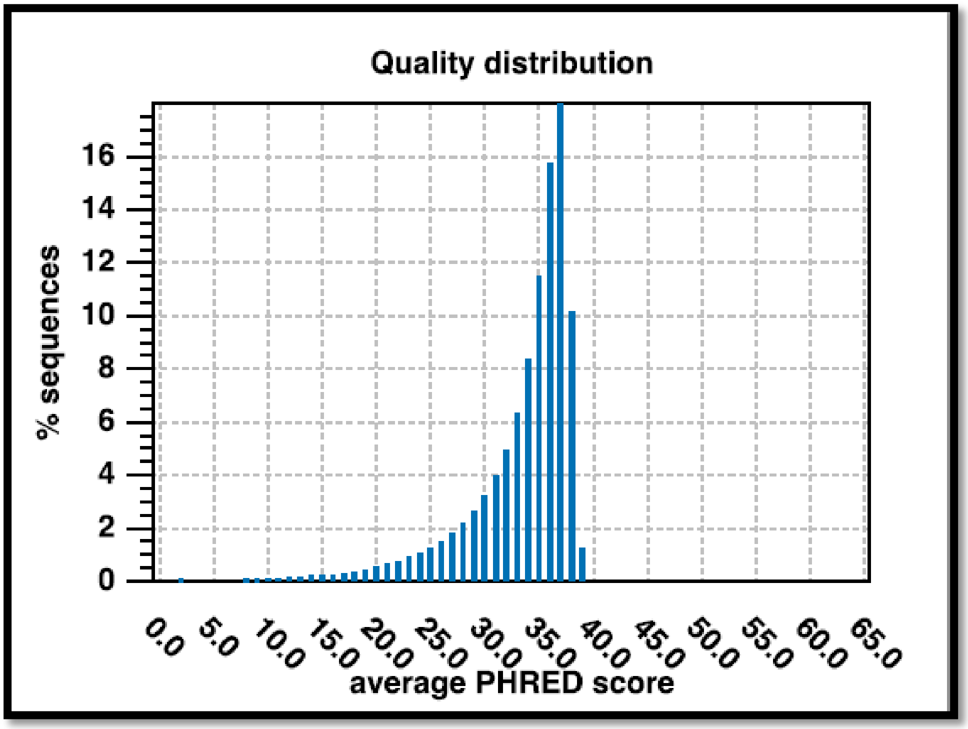


Figure 46: Distribution of average sequence quality scores expressed as PHRED scores for sub-isolate 12-8. The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Sub-isolate 12-9

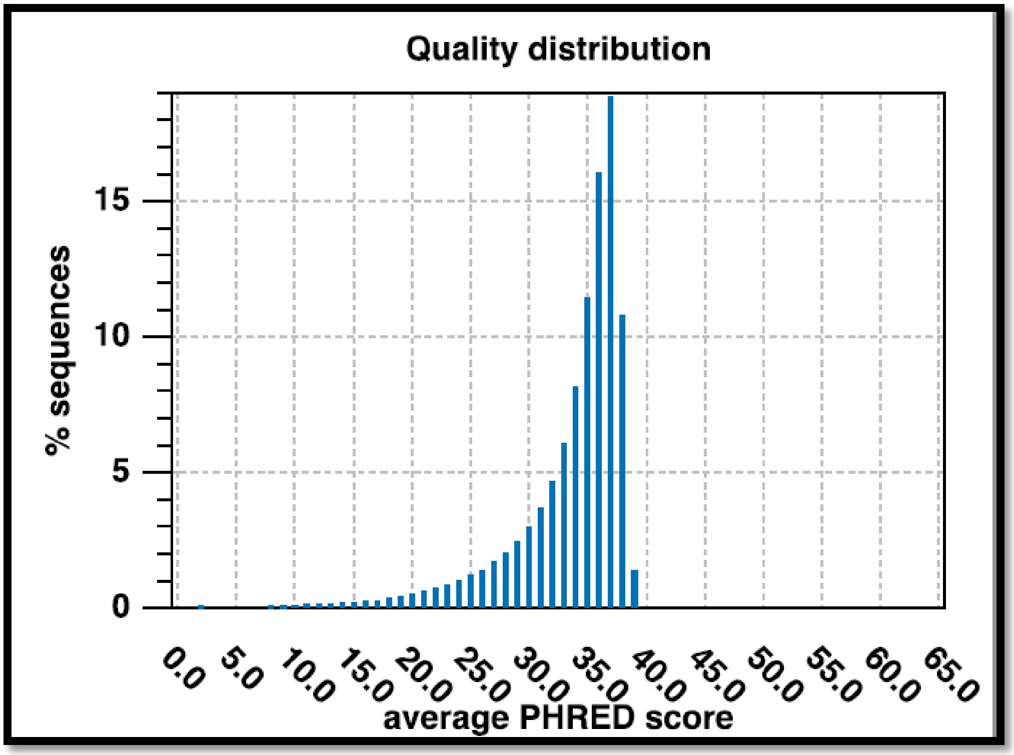


Figure 47: Distribution of average sequence quality scores expressed as PHRED scores for sub-isolate 12-9. The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Field isolate 11-5000

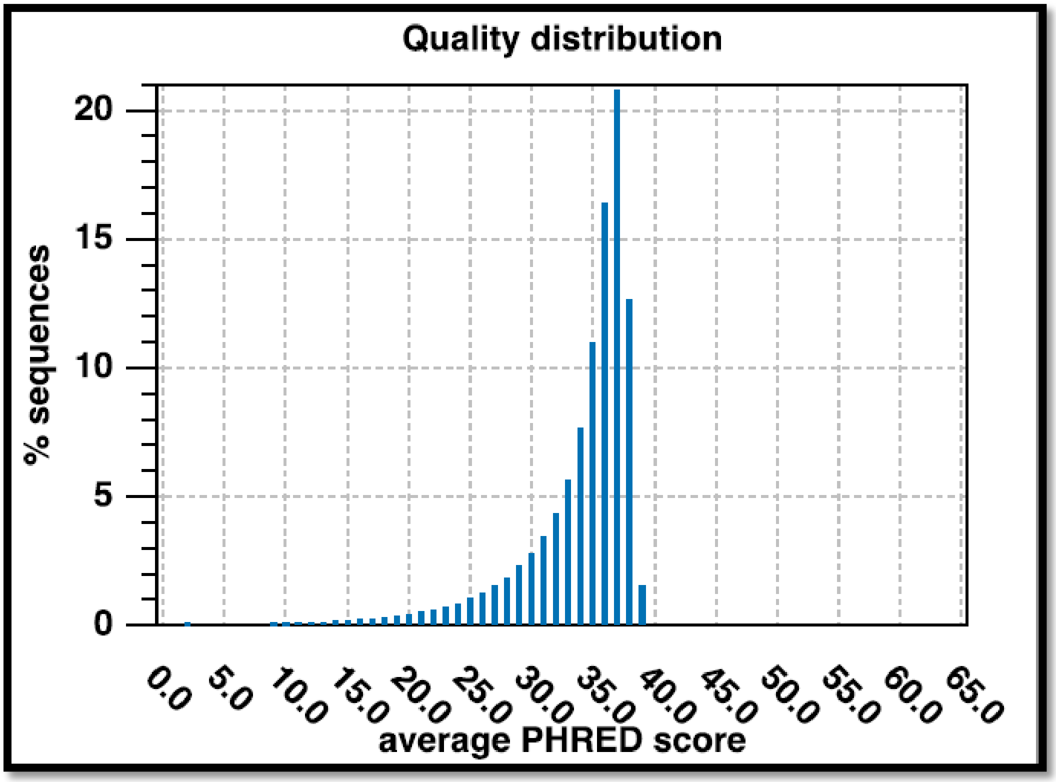


Figure 48: Distribution of average sequence quality scores expressed as PHRED scores for isolate 11-5000. The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Field isolate 11-5007

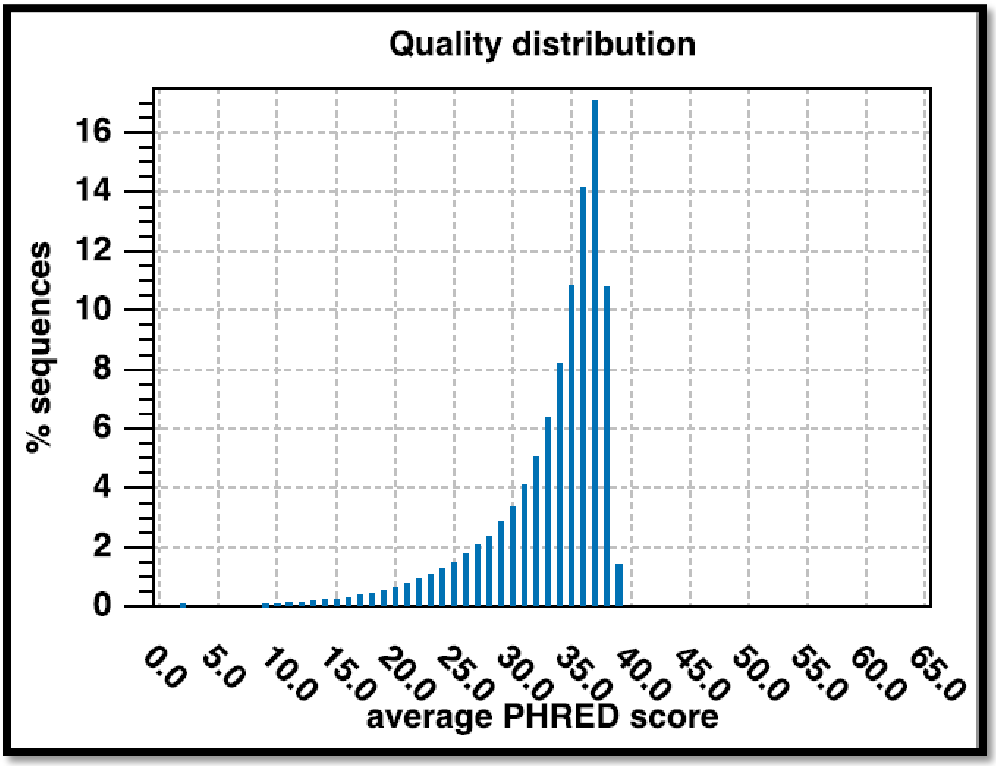


Figure 49: Distribution of average sequence quality scores expressed as PHRED scores for isolate 11-5007. The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Field isolate 11-5009

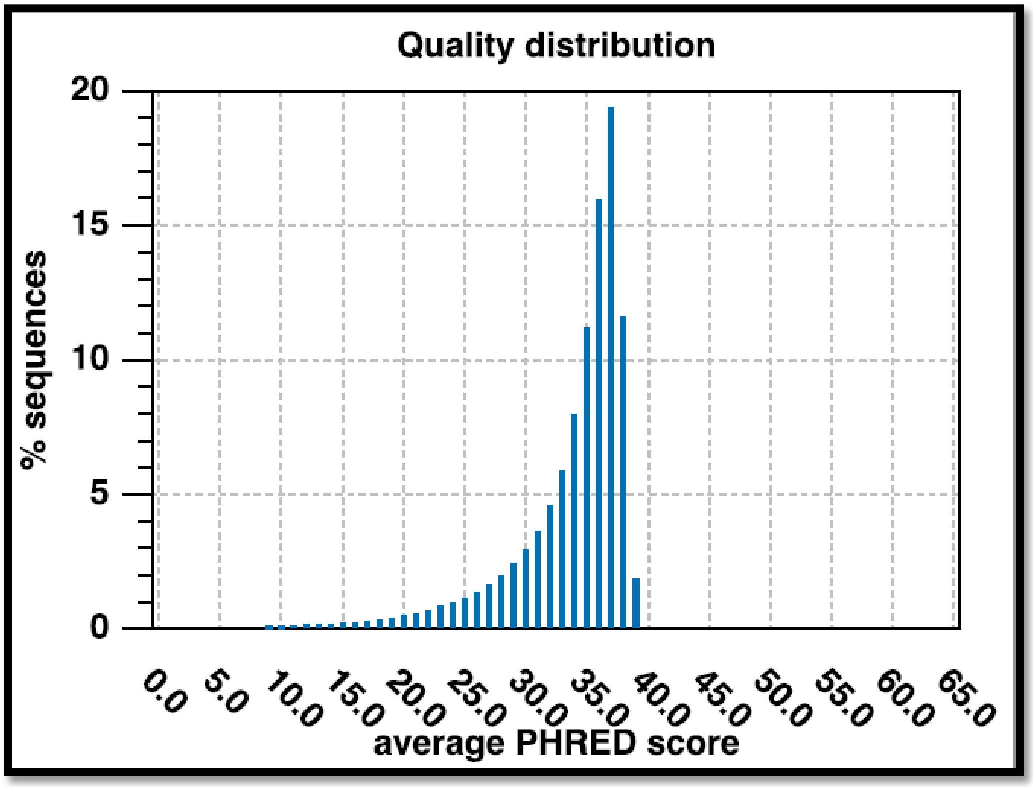


Figure 50: Distribution of average sequence quality scores expressed as PHRED scores for isolate 11-5009. The quality of a sequence is calculated as the arithmetic mean of its base qualities. X-axis: PHRED-score. Y-axis: number of sequences observed at that quality score normalized to the total number of sequences.

Note: **The remaining appendices (4, 5 and 6) can be found in digital format on the complementary DVD attached to the thesis manuscript.**

APPENDIX 4: *de novo* CTV contig sequences

APPENDIX 5: Raw Illumina datasets from all samples

APPENDIX 6: Novel genome sequences