

Overcoming bias in *Citrus tristeza virus* (CTV) genotype detection and a population study of CTV within Southern African Star Ruby grapefruit orchards

David Read

**Submitted in partial fulfilment of the requirements for the degree
Philosophiae Doctor (Microbiology)
In the Faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria
South Africa**

August 2015



DECLARATION

I, David Alan Read, declare that the thesis which I hereby submit for the degree of Philosophiae Doctor at the University of Pretoria, is my own work and has not been previously submitted by me for a degree at this or any other tertiary institution.

Signature

Date

ACKNOWLEDGEMENTS

I would like to thank the following people and organisations for their invaluable contributions made towards the completion of this study.

Citrus Research International (CRI), University of Pretoria (UP), the National Research Foundation (NRF) and the Technology and Human Resources for Industry Program (THRIP) funded the study.

My supervisor, Prof. Gerhard Pietersen, for being an excellent mentor and whose in depth experience and knowledge of plant virology carried this study to completion.

My parents for their endless support throughout my academic career, which has allowed me to reach this important milestone in my academic career.

My fiancé, Cornelia Warschkuhl for her support and understanding.

Kirsti Snyders for her assistance with collecting the necessary material from the field.

Students in the Department of Microbiology and Plant Pathology, particularly Gaby Carstensen, Sarah MacRae, Helen Walsh, Jackie Lubbe, Kirsti Snyders, Jennifer Wayland and Megan Harris for their friendship and support.

CONTENTS

Acknowledgements.....	iii
List of Tables	vi
List of Figures.....	vi
Abstract.....	vii
Chapter 1: Introduction.....	1
Chapter 2: Literature Review.....	7
<i>(2.1) Biology, pathology and epidemiology of Citrus tristeza virus.....</i>	<i>7</i>
<i>(2.2) Control of CTV and the use of cross-protection and in Southern Africa.....</i>	<i>9</i>
<i>(2.3) CTV detection, diagnostics and phylogenetics</i>	<i>11</i>
<i>(2.4) Next Generation Sequencing chemistries and platforms</i>	<i>14</i>
<i>(2.5) Characterising viral populations using NGS technologies.....</i>	<i>18</i>
Chapter 3: PCR bias associated with primers targeting conserved sequences for genotype diversity studies within <i>Citrus tristeza virus</i> populations.....	37
<i>Introduction.....</i>	<i>37</i>
<i>Materials and Methods.....</i>	<i>39</i>
<i>Results.....</i>	<i>43</i>
<i>Discussion.....</i>	<i>47</i>
Chapter 4: Diversity of Citrus tristeza virus populations in commercial Star Ruby orchards in Southern Africa, using Illumina MiSeq technology.....	62
<i>Introduction.....</i>	<i>62</i>
<i>Materials and Methods.....</i>	<i>64</i>
<i>Results.....</i>	<i>71</i>
<i>Discussion.....</i>	<i>82</i>
Chapter 5: Concluding Remarks.....	97

List of appendices

Appendix A.....	104
Appendix B.....	108
Appendix C.....	111
Appendix D.....	118
 Supplementary data (supplied electronically)	
Section 1.....	128
Section 2.....	194

LIST OF TABLES

Chapter 3

Table number	Page
1. Genotype specific primer sequences	40
2. Cloned sequence and Illumina MiSeq data	45

Chapter 4

1. Details of the sites from which samples were collected.....	64
2. The identities of the direct sequencing data.....	69
3. Representation of the Illumina sequencing data of field samples.....	76
4. Illumina MiSeq data of the original GFMS 12 populations.....	78
5. Illumina MiSeq data of the non-preimmunised Star Ruby trees in the Letsitele production area.....	79
6. Comparison between p33 direct sequencing and Illumina datasets.....	80
7. Comparison between cloned sequence data and Illumina datasets.....	81

LIST OF FIGURES

Figure number	Page
---------------	------

Chapter 4

1. CTV diversity based on cloned sequence data.....	72
---	----

Abstract

Citrus tristeza virus (CTV) is present in almost all of the major citrus production areas where it continues to reduce the profitability of citriculture worldwide. Severe stem-pitting strains are endemic to Southern Africa. In addition to making use of decline-tolerant rootstocks, the implementation of a mild-strain cross-protection program has sought to increase the citrus production in the region. Of all the commercially cultivated citrus species, grapefruit cultivars are among the most susceptible to the severe stem-pitting strains and infections often lead to a breakdown of cross-protection. Recent research has shown that CTV cross-protection operates through a strain-specific mechanism, which relies on a virus-specific protein, expressed from the p33 gene. The specificity of this mechanism has highlighted the need for determining the CTV diversity within production areas as well as accurately characterising CTV cross-protection sources that will be capable of preventing secondary inoculations of severe strains in the field. The accurate characterisation of CTV populations, which are usually made up of a number of disparate strains, requires the use of robust PCR protocols. Mismatches between primers and their corresponding binding sites may introduce primer-associated bias during amplification. The primer-associated bias of four sets of CTV specific primers, targeting the A and F regions and the p23 and p33 genes, were evaluated. This was done through the amplification of defined templates followed by their characterisation using the sequencing of multiple clones, as well as Illumina next generation sequencing. High levels of bias were found to be associated with the primer pairs targeting the A and F regions. The p33 gene primers were found to be biased against two genotypes and suggestions for preventing this apparent bias are discussed. The primer pair targeting the conserved p23 gene was found to have very little associated bias. The second major aspect of this study was the large scale survey of CTV diversity of pre-immunised Star Ruby grapefruit orchards in Southern Africa. Samples were collected from eight Star Ruby production sites throughout Southern Africa, namely Hoedspruit 1, Hoedspruit 2, Malelane 1, Malelane 2, Swaziland (Mananga), Northern Cape (Kakamas), Sundays River Valley and Nkwalini Valley, where between 16 and 32 samples per site were collected. The p33 gene was amplified for each of the collected samples and subjected to direct Sanger sequencing. A protocol making use of Illumina MiSeq sequencing was established

and used to sequence 96 samples. A subset of six samples was selected for cloning, which resulted in a total of 218 sequenced clones and compared with that of the Illumina sequencing data. High levels of CTV diversity were observed between orchards, as well as between different trees within the same orchard. Most of the populations were made up of a single dominant group, sometimes with several minor sequence types. Sequence reads corresponding to strains within the Resistance Breaking (RB) genotype were most numerous, especially in the most recently planted orchards and were present within all of the populations analysed. The Kpg3/SP/T3 group appeared to be represented the second most prevalent genotype and seems to become more common as the orchard ages. Genotypes of the HA 16-5, VT, AT-1, T36, Taiwan-Pum/M/T5 and T30 types, were represented sporadically and at variable levels in populations from numerous collection sites. This has been one of the most extensive diversity studies of CTV to date and has provided unprecedented baseline knowledge of CTV diversity in Southern Africa.

Chapter 1

Introduction

Citrus tristeza virus (CTV) has been established in almost all of the commercially important citrus production areas around the world (Niblett *et al*, 2000) and is responsible for the most devastating viral disease of citrus (Bar-Joseph *et al*, 1989). Virus strains are able to elicit symptoms varying from being asymptomatic for mild strains, to decline, stem-pitting, stunting, seedling yellows and quick decline (Bar-Joseph *et al*, 1989) for severe strains. The only economically viable control strategy is to quarantine imported material (Moreno *et al*, 2008) to prevent new strains of the virus being introduced and to employ cross-protection in regions where the virus is endemic (Rubio *et al*, 1996). Cross-protection involves selecting a population of CTV strains, known to elicit only mild symptoms and use it to inoculate CTV-free material (van Vuuren and Manicom, 2005). Cross-protection is described as the partial or complete protection of a plant against a severe strain of a virus subsequent to being inoculated with a mild strain of the same virus (Gal-On and Shibolet, 2005). Grapefruit Mild Strain 12 (GFMS 12) was used as the routine cross-protecting source for *Citrus x paradisi* (grapefruit) cultivars in South Africa, however severe cross-protection breakdown was observed in motherblock trees and in the field (van Vuuren and Manicom, 2005) and for these reasons GFMS 35 was approved as the interim cross-protecting source for all grapefruit cultivars.

It is thought that cross-protection breakdown occurs due to superinfection, which occurs when a mild protecting population fails to protect against the infection of a severe strain of the same virus (Fulton, 1986). The basis of cross-protection in the case of CTV apparently works through the mechanism of superinfection exclusion. In the pioneering work by Folimonova *et al*, (2010), it was shown that superinfection exclusion within CTV populations can only occur between strains of members of the same genotype, namely T36, T30, T3, T68 and VT. This study has revolutionised the way in which cross-protecting isolates need to be evaluated. In many cases, potential cross-protecting sources are evaluated, through laborious trials on indicator hosts, which yield no information about the genetic composition of the viral population. Folimonova *et al*, (2010) have therefore proposed a logical workflow, which can be used to improve the efficacy of cross-protection within a growing area.

Firstly, severe strains present within a particular growing area must be identified and earmarked for control. Secondly, mild strains, corresponding to the prevalent severe genotypes in an area must be obtained either through their isolation from natural populations or by synthesising mild strains through the substitution of mild sequence determinants. Adding to the complexity of this task, the dominant CTV strains within a region can be influenced by different climatic conditions (Broadbent *et al*, 1996) suggesting that the composition of populations from a number of different production areas would need to be determined.

A multitude of molecular tools are available for the determination of the composition of CTV populations. However, a number of these techniques, such as Single Stranded Conformation Polymorphism (SSCP) and Restriction Fragment Length Polymorphism (RFLP), are only able to indicate that variations in strain composition exist between different populations. Nucleotide sequencing of genes and gene regions allows for the identification of particular strains within a population (Read and Pietersen, 2015). Automated Sanger sequencing was the gold standard for over 30 years (van Dijk *et al*, 2014) however, due to the demand for ever increasing throughput and drawbacks such as bacterial cloning, Sanger sequencing has been gradually replaced by so-called Next Generation Sequencing (NGS) technologies. The main objective of this study was to determine the CTV population diversity within the various *Citrus x paradisi* cv. Star Ruby production areas around Southern Africa and to establish a robust protocol for the analysis of these populations, using NGS technologies that could be used as an alternative to existing protocols based on Sanger sequencing. Various template preparations can be sequenced using NGS, such as randomly amplified DNA (Roossinck *et al*, 2010), double-stranded (ds) RNA (Zablocki and Pietersen, 2014) and template amplified using gene-specific primers. Due to the putative association of the p33-gene with genotype specificity in cross-protection in CTV (Folimonova *et al*, 2010), this gene was selected as the phylogenetic marker for this study. A total of 192 leaf samples were collected from the Star Ruby production areas of Hoedspruit, Malelane, Swaziland, Northern Cape, Sundays River Valley and Nkwalini Valley. Three samples harbouring the original GFMS 12 populations were also collected from various greenhouse sources. All 195 amplicons were sequencing using direct Sanger sequencing, a randomly selected subset of 95 samples were sequenced

using Illumina MiSeq technologies and an additional subset of six amplicons, were selected for mass bacterial cloning and Sanger sequencing.

To achieve the above aim we also evaluated the degree of primer-directed bias of genotypes among various existing PCR protocols, including the p33 gene utilised in the survey. The genome sequences of the various CTV strains are highly heterogeneous when compared with one another (Ayllón *et al*, 2001), especially toward the 5' end (Hilf *et al*, 1999). Due to this heterogeneity, universal PCR primers are designed in order to target the maximum number of variants within a population. However, strains with primer binding sites of greater homology to the universal primer sequences may have their target genes preferentially amplified within a mixed template. This results in the skewing of population data (Engelbrektson *et al*, 2010). In this study the primer-associated biases of three previously published primer pairs, targeting the A-region, F-region (Rubio *et al*, 2001) and the p23 gene (Sambade *et al*, 2003) as well as the those of the unpublished p33 gene (this study). This was carried out through designing sets of genotype-specific primer pairs that were identical to the primer binding sites of cloned sequences corresponding to these gene regions. Templates containing equimolar concentrations of each genotype-specific amplicon were produced and then subjected to reamplification, using the original universal primer pairs. These resulting products were then used in both mass cloning experiments and subjected to Illumina MiSeq sequencing. In the following literature review we explore previously generated studies and information, relevant to the objectives of the current study.

REFERENCES

- Ayllón M.A., Lopez C., Navas-Castillo J., Garnsey 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* vol. 146 pp27-40
- Bar-Joseph M., Marcus R. and Lee R.F. (1989). The continuous challenge of Citrus Tristeza Virus Control. *Annual Reviews in Phytopathology* vol. 27 pp291-316
- Broadbent P., Bransky R.H. and Indsto J. (1996). Biological Characterization of Australian Isolates of Citrus Tristeza Virus and Separation of Subisolates by Single Aphid transmissions. *Plant Disease* vol. 80 pp329-333
- Engelbrektson A., Kunin V., Wrighton K.C., Zvenigorodsky N., Chen F., Ochman H. and Hugenholtz P. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal* vol. 4 pp642-647
- Folimonova S.Y., Robertson C.J., Shilts T., Folimonov A.S., Hilf M.E., Garnsey S.M. and Dawson W.O. (2010). Infection with Strains of Citrus Tristeza Virus Does Not Exclude Superinfection by Other Strains of the Virus. *Journal of Virology* vol. 84 pp1314-1325
- Gal-On A. and Shibolet Y.M. (2005). Cross-protection, pp261-288. In G. Loebenstein and J.P. Carr (ed.), *Natural resistance mechanisms of plants to viruses*. Springer, Dordrecht, The Netherlands
- Hilf M.E., Karasev A.V., Albiach-Marti M.R., Dawson W.O. and Garnsey S.M. (1999). Two Paths of Sequence Diversion in the Citrus Tristeza Virus Complex. *Phytopathology* vol. 89 pp336-342
- Moreno P., Ambros S., Albiach-Marti M.R., Guerri J. and Pena L. (2008). '*Citrus tristeza virus*: A pathogen that changed the course of the citrus industry'. *Molecular Plant Pathology* vol. 9 pp251-268

Niblett C.L., Genc H., Cevik B., Halbert S., Brown L., Nolasco G., Bonacalza B., Manjunath K.L., Febres V.J., Pappu H.R. and Lee R.F. (2000). Progress on strain differentiation of *Citrus tristeza virus* and its application to the epidemiology of citrus tristeza disease. *Virus Research* vol. 71 pp97-106

Read D. and Pietersen G. (2015). Genotypic diversity of *Citrus tristeza virus* within red grapefruit, in a field trial site in South Africa. *European Journal of Plant Pathology* vol. 142 pp531-545

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* vol.19 pp81-88

Rubio L., Ayllón M.A., Guerri J., Pappu H., Niblett C. and Moreno P. (1996). Differentiation of citrus tristeza virus (CTV) isolates by single-strand conformation polymorphism analysis of the coat protein gene. *Annals of Applied Biology* vol. 129 pp479-489

Rubio L., Ayllón M.A., Kong P., Fernandez A., Polek M., Guerri J., Moreno P. and Falk B. (2001). Genetic variation of *Citrus tristeza virus* isolates from California and Spain: Evidence for mixed infections and recombination. *Journal of Virology* vol. 75 pp8054-8062

Sambade A., Lopez C., Rubio L., Flores R., Guerri J. and Moreno P. (2003). Polymorphism of a specific region in gene *p23* of *Citrus tristeza virus* allows for the discrimination between mild and severe isolates. *Archives of Virology* vol. 148 pp2325-2340

van Dijk E.L., Auger H., Jaszczyszyn Y. and Thermes C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics* vol. 30 pp418-426

van Vuuren S.P. and Manicom B.Q. (2005). The Response of Star Ruby Grapefruit to Different Citrus tristeza virus isolates. In: *Proc. 16th Conf. IOCV*, 112-116 IOCV, Riverside, California, United States

Zablocki O. and Pietersen G. (2014). Characterization of a novel citrus tristeza virus genotypewithin three cross-protecting source GFMS12 sub-isolates in South Africa by means of Illumina sequencing. *Archives of Virology* vol. 158 pp2133-2139

Chapter 2

Literature Review

(2.1) Biology, pathology and epidemiology of *Citrus tristeza virus*

CTV causes some of the most devastating and economically important virus-associated diseases of citrus and reduces the profitability of citriculture worldwide (Moreno *et al*, 2008). Once established, it is able to spread by a number of aphid vectors (Moreno *et al*, 2008). *Tristeza* is the Spanish word for sadness and is a description of the decline of infected citrus plants when sour orange is used as rootstock (Bertolini *et al*, 2008). CTV elicits three major host disease responses, depending on the virus population composition, the host and the rootstock and scion combinations thereof (Moreno *et al*, 2008). Two of the three symptoms are of concern to commercial citrus production and these are stem-pitting and decline (Harper, 2013). The third symptom that may be expressed is seedling yellows, which is characterised by, mostly transient, stunting and chlorosis of leaves (Albiach-Marti *et al*, 2010). Seedling yellows is for the most part an economically unimportant disease, with occasional incidence under nursery conditions.

CTV has become established in almost all major citrus growing regions (Niblett *et al*, 2000) and in the past century has been responsible for the complete destruction of citrus industries in a number of countries (Albiach-Marti *et al*, 2000). Estimates reveal that CTV has killed approximately 38 million trees in the Americas (Niblett *et al*, 2000), more than 55 million in the Mediterranean countries of which 40 million of these were in the Valencia region of Spain (Cambra *et al*, 2000) and a further 5 million killed elsewhere around the world (Niblett *et al*, 2000). The dissemination of CTV within and between countries and continents has mainly been the result of illegally imported budwood or seedlings that are infected with CTV (Niblett *et al*, 2000, Herron *et al*, 2006), with the subsequent local spread of the virus through the resident aphid populations.

Aphids spread the virus in a semi-persistent manner (Bertolini *et al*, 2008, Ruiz-Ruiz *et al*, 2007) and a number of aphid species, namely *Toxoptera citricida* (brown citrus aphid), *Aphis gossypii*, and *Aphis spiraecola* are the most important vectors, with the brown citrus aphid (BrCA) being the most efficient transmitter of the virus

(Yokomi *et al*, 1994; Černi *et al*, 2008). Despite the major economic impact of this virus-aphid association, the CTV-BrCA relationship and the subsequent spread of the disease is still poorly understood (Herron *et al*, 2006). This paucity of information makes studying CTV epidemiology, as well as controlling the disease, a monumental task.

The best methods of CTV control is through effective quarantine in regions that are free of CTV or complete eradication of infected trees in regions with low incidence (Moreno *et al*, 2008). Eradication is not a feasible option in regions with widespread and high incidence of CTV infections that also have effective vector populations to facilitate its spread. In this case, use of decline-tolerant rootstocks is an option to curb negative economic effects (Moreno *et al*, 2008). Stem-pitting isolates however, are able to elicit symptoms regardless of the rootstock scion combinations used (Moreno *et al*, 2008). Severe stem-pitting isolates are present in the South African CTV population (van Vuuren *et al*, 1993) and the only means of protecting against them is through cross-protection (Moreno *et al*, 2008). The use of resistant plant varieties is more favourable than implementing cross-protection schemes due to the reduced costs and labour required (Fulton, 1986) however the only way of producing resistant citrus varieties is through transgenic modification, which, in the case of citrus is not feasible due to the large number of cultivars and consumer concerns.

Cross-protecting sources are selected from trees that have been exposed to many years in the field but are still healthy looking, producing acceptable quantities of large fruit (van Vuuren and Manicom, 2005) and are surrounded by other trees showing severe symptoms (Fulton, 1986). Potential cross-protecting sources have usually been evaluated through biological trials (Folimonova, 2013), with no knowledge of the actual strain assemblage of the source. This evaluation is time consuming and often takes many years, with most sources failing to protect (Folimonova, 2013) and those that do prove effective are usually suitable for only a few varieties.

CTV is a member of the family Closteroviridae, is one of the largest known plant viruses (Karasev *et al*, 1995) and form filamentous thread-like particles approximately 2000 x 11nm in size (Price, 1966; Bar-Joseph *et al*, 1972). The ~20kb single stranded positive sense RNA genome is one of the largest RNA genomes known discovered to date (Karasev *et al*, 1995) and is organised in 12 open reading frames (ORF's) (Albiach-Marti *et al*, 2000). The capsid is made up of one major and

one minor protein encoded by the p25 and p27 genes respectively (Bar-Joseph *et al*, 1972; Febres *et al*, 1996). ORF 1a and 1b code for replication enzymes (Tatineni *et al*, 2008) and are expressed from the genome directly (Karasev *et al*, 1995). The remaining ten ORF's are expressed via subgenomic RNA's and encode for a hydrophobic membrane protein (p6), an HSP70 homolog (p65), an HSP90 homolog (p61) (Satyanarayana *et al*, 2000), a protein involved in inclusion body formation (p20) (Gowda *et al*, 2000) and an RNA binding protein (p23) (Satyanarayana *et al*, 2002). The protein product of the p33 gene has been shown to play a major role in strain-specific superinfection exclusion (Folimonova, 2012). Collectively, the p33, p18 and p13 genes function to enhance the host range of CTV strains (Tatineni *et al*, 2011) and it was subsequently shown that the deletion of these genes increases the stem-pitting capabilities of CTV (Tatineni and Dawson, 2012). Strains show asymmetric sequence identity with low sequence conservation toward the 5' end, while the conservation of sequence identity increases progressively toward the 3' end (Yang *et al*, 1999).

(2.2) Control of CTV and the use of cross-protection and in Southern Africa

The eradication of CTV, subsequent to its introduction to a production area, is the most preferable form of control. This method is only feasible if the number of trees and the area of the spread are relatively small (Moreno *et al*, 2008). The most efficient CTV vector, *T. citricida*, has become established in most of the major citrus producing countries (Moreno *et al*, 2008) except those of the Mediterranean basin where only an isolated population has been found in Spain and Portugal, far removed from major production areas. In cases where efficient vector populations have become established along with the spread of the virus, the use of CTV decline-tolerant rootstocks becomes more preferable to eradication of infected trees. However, severe stem-pitting isolates of CTV cause disease symptoms on scions regardless of rootstock combination (Ruiz-Ruiz *et al*, 2006). This often limits the range of cultivars that can be grown within production areas that have endemic stem-pitting isolates circulating (Albiach-Marti *et al*, 2010). Once introduced, the spread of stem-pitting isolates becomes exacerbated by the fact that *T. citricida* spreads them preferentially (Brlanksy and Roy, 2011). In order to curb the effects of

severe stem-pitting infections, cross-protection becomes the only feasible option for their control.

Cross-protection is described as the phenomenon where a plant shows complete or partial resistance to a severe strain of a virus, after being infected by a mild or attenuated variant of the same virus (Gal-On and Shibolet, 2005 and Yoon *et al*, 2006). The use of mild-strain cross-protection involves the intentional infection of virus free material with a CTV isolate known to exhibit only mild symptoms (van Vuuren and Manicom, 2005). This approach is applied to a number of different crop or ornamental plant varieties in order to curb the damage caused by any subsequent infection of their associated viruses in the field (Fulton, 1986). A number of terms have been used to describe the technique, such as acquired immunity, prophylactic inoculation and antagonism but the term cross-protection was deemed most accurate in describing the reasons for its use, which is the protection against more severe strains (Fulton, 1986). A number of concerns against the use of cross-protection have arisen over the years. Firstly, the effectiveness of cross-protection to completely protect against certain severe plant virus strains has been called into question. However, Fulton, (1986) argues the concept of “relative rather than complete” protection where a certain level of protection is superior to none at all, especially in cases where there is no other feasible alternative. Fulton, (1986) lists a number of other concerns against cross-protection, among these are that propagating a virus may result in its rapid spread to other sensitive host plants. Secondly, synergism may result between the cross-protecting virus and another unrelated virus, resulting in severe and unforeseen disease symptoms. Potential synergism between CTV and another pathogenic agent may be the cause of a disease called Citrus sudden death (CSD) (Gomes *et al*, 2008) however, further evidence is needed to substantiate this. A third concern adding to the uncertainty is the cost and labour required for the inoculation of an entire crop with a mild-strain virus. This concern is of particular importance to annual crops grown from seed and of least concern to perennial crop plants multiplied through vegetative propagation, such as citrus.

In South Africa, CTV cross-protection in commercial citrus began in 1972 with the establishment of the South African Citrus Improvement Scheme (SACIS). Together with this, shoot-tip grafting (STG) was employed to eliminate graft-transmissible diseases from commercial budwood source material (von Broembsen and Lee

1988). Prior to STG, a number of CTV populations were collected from grapefruit budwood source trees that still generated high yields after 15 years (van Vuuren *et al*, 1993). Subsequent to satisfactory performance in both greenhouse and field trials, three of these populations were considered mild and designated GFMS (Grapefruit Mild Strain) 12, 27 and 35 (van Vuuren and Moll 1987; van Vuuren *et al*, 1993). Ultimately the GFMS 12 population was selected for use within the improvement scheme, however severe-stem pitting symptoms were observed in mother block trees in 1993 (van Vuuren and Manicom, 2005). Subsequent field trials on various grapefruit cultivars confirmed that pre-immunisation with either GFMS 12 or GFMS 35 led to significant losses in yield, although the performance of GFMS 35 was superior to that of GFMS 12 (van Vuuren and van der Vyver 2000). GFMS 35 was selected as the replacement cross-protecting population within the improvement project (Luttig *et al*, 2002) and has been used to pre-immunise all red cultivars within the mother block since 1998 (van Vuuren, personal communication), however a number of commercial nurseries continued to use GFMS 12 for a number of years subsequent to this.

(2.3) CTV detection, diagnostics and phylogenetics

CTV is unique among members of the *Closteroviridae* in that it consists of a multitude of different strains that share a common evolutionary history (Harper, 2013). It is generally accepted that CTV consists of two main branches of sequence divergence, one containing T36-like sequences and the other consisting of VT/T30-like sequences (Hilf *et al*, 1999). Within these two major branches of divergence, Harper, (2013) proposed a standardized system of classification for CTV isolates, based on the analysis of a collection of whole genome sequences. This classification separates CTV isolates into six major strains, namely T36, VT, T3, RB, T68, and T30. Harper (2013) acknowledges the fact that HA16-5 (Melzer *et al*, 2010) could represent a seventh lineage, however cautions against this until additional homologs are discovered in order to ensure that HA16-5 is not merely a product of an isolated recombination event. It could also be assumed that within these strains are swarms of quasispecies (Domingo, 2002) that arise due to the error prone nature of RNA dependant RNA polymerase (Sanz-Ramos *et al*, 2008). It is clear that CTV represents a uniquely complex phylogenetic history among plant viruses and the

detection of individual strains within a plant is complicated by the fact that most populations of CTV are mixtures of multiple strains (Scott *et al*, 2012)

Prior to the advent of molecular techniques, the detection of CTV was done through the graft indexing onto sensitive indicator host cultivars such as Mexican Lime (Moreno *et al*, 2008). Indexing of isolates yields little information regarding the genetic identity of particular strains and due to the fact that many populations are mixtures and the phenotype observed may be the sum of interactions between the components and not of individual strains. The use of single aphid transfers (SATs) is a method for inducing a genetic bottleneck that purifies strains of CTV from a population (Brlansky *et al*, 2003). van Vuuren *et al*, (2000) made use of SATs to evaluate individual components of the GFMS 12 and Lime mild-strain (LMS) 6 pre-immunising populations. When used in isolation, indexing provides little value to the characterisation of CTV populations and the sub-isolates thereof. However, it is still widely used in CTV research (Moreno *et al*, 2008) since this is still the only method for determining the phenotypes of strains. When used in conjunction with various molecular tools, indexing provides a more complete characterisation, where sequences can be linked to phenotypic traits. More recently, indexing makes use of a number of different citrus cultivars to phenotypically characterise a single CTV isolate. Garnsey *et al*, (2005) proposed the use of five different cultivars to determine whether an isolate is able to induce various disease symptoms on these different hosts.

The development of an Enzyme-linked immunosorbent assay (ELISA) (Bar-Joseph *et al*, 1979) was the first step toward molecular characterisation of CTV and provided a much more rapid tool for the detection of CTV. This contributed greatly to more improved studies of CTV epidemiology, eradication and quarantine efforts (Moreno *et al*, 2008). A variation of the ELISA used to detect CTV was developed by Permar *et al*, (1990), known as MCA13, to detect the presence or absence of a specific CTV coat-protein epitope thought to play a role in decline. Nikolaeva and Karasev (1998) developed a serological test for the differentiation of isolates capable of eliciting stem-pitting symptoms.

Cloning fragments of the CTV genome (Rosner *et al*, 1983) allowed for the production of probes (Rosner and Bar-Joseph, 1984) that could be used to characterise CTV isolates at the genetic level for the first time. Sekiya *et al*, (1991) were among the first to sequence part of the CTV genome (coat protein) but it was

not until the first whole CTV genome was sequenced (Karasev *et al*, 1995) that PCR primers could be designed to target multiple gene sequences across the CTV genome. A number of PCR protocols have been developed, targeting a number of gene sequences, namely; the A and F regions, which target within the 5' half of the genome, the coat protein and p20 genes (Rubio *et al*, 2001), p23 gene (Sambade *et al*, 2003), p25, p27 (Iglesias *et al*, 2008), p18 and p13 (Ayllón *et al*, 1999). Martin *et al*, (2009), made use of gene specific primers to sequence and determine the genetic variability of isolates based on seven distinct gene regions, namely p33, p65, p61, p18, p13, p20 and p23. The ability to amplify various CTV genes allowed for the use of various molecular techniques such as SSCP (Rubio *et al*, 1996; Gago-Zachert *et al*, 2000; Luttig *et al*, 2002) and RFLP (Gillings *et al*, 1993; van Vuuren *et al*, 2000). Hilf and Garnsey, (2000) and Hilf *et al*, (2005), developed a set of genotype-specific primer pairs, based on existing sequence data of the 5' end of various strains that could be used to amplify these molecular markers. These PCR products can then be analysed on agarose gels and the presence or absence of particular bands gives an indication of which strains are present within a population. However, these techniques suffer from a number of disadvantages. RFLP and SSCP are only able to resolve differences between populations of CTV and the Multiple Molecular Markers of Hilf *et al*, (2005) are only capable of indicating the presence or absence of strains and not their relative numbers within a population. The use of nucleotide sequencing provides more definitive data for determining the identity of genotypes within a population and when coupled with a mass-cloning strategy, can roughly determine the relative population composition (Iglesias *et al*, 2008; Scott *et al*, 2012; Read and Pietersen, 2015).

The characterisation of CTV strains within populations is complicated by the asymmetry of divergence between the 5' and 3' halves of the genome, meaning that certain strains, that may have distinct phenotypes, can only be differentiated from one another when sequences within the 5' half of the genome are used for genotyping (Harper, 2013). The strain-specific cross-protection mechanism of CTV (Folimonova *et al*, 2010) has shown the importance for the accurate characterisation of CTV populations since "mild" protecting strains will now need to be produced or discovered to protect against cognate severe strains in the field.

Universal primers are often designed to target gene regions in the 5' heterogeneous half of the genome and results in inevitable nucleotide differences

between these universal primers and the target primer binding sites of some strains. When amplifying target gene regions from a mixed population, universal primers have the potential to preferentially amplify targets that show a greater homology to their own sequences. Among the most widely used universal primers are the 16S and 18S ribosomal RNA primers for determining microbial diversity in bacterial and fungal populations respectively. A number of studies have determined that PCR primer-associated bias can play a role in distorting the apparent microbial populations, when universal primer pairs are used to amplify these targets (Bru *et al*, 2008; Hong *et al*, 2009; Engelbrektson *et al*, 2010; Polz and Cavanaugh, 1998; Sipos *et al*, 2007). Little research has been done to determine whether universal primers targeting viral sequences suffer from similar drawbacks; however there are no technical reasons why they should not. Universal primers should therefore be designed with care, with minimal mismatches, especially towards to critical 3' end of the primer (Dieffenbach *et al*, 1993; Read and Pietersen, 2015).

The use of the Sanger method to sequence mass numbers of clones containing CTV sequences has been used effectively to make rough estimates of the CTV strain composition of populations (Iglesias *et al*, 2008; Scott *et al*, 2012; Cevik *et al*, 2013; Read and Pietersen, 2015). However, due to the labour involved in isolating very large numbers of individual clones required, the per base cost of Sanger sequencing and the relatively low level of data output, there has become an essential need for characterising CTV populations with novel protocols that utilise high throughput sequencing technologies (Beerenwinkel and Zagordi, 2011).

(2.4) Next Generation Sequencing chemistries and platforms

Only a decade has passed since the introduction of the first so-called Next-Generation Sequencing (NGS) technologies and in this short amount of time, it has revolutionised the way in which researchers sequence genomes and genes thereof. The rapid adoption of these technologies has been mainly attributed to the following. NGS library preparation does not require bacterial cloning and secondly, millions of sequencing reactions can be carried out in parallel as opposed to hundreds in the case of automated Sanger sequencing (van Dijk *et al*, 2014), which was previously considered to be the gold standard in sequencing (Harismendy *et al*, 2009).

The first NGS technology available on the market in 2005 was the now defunct 454 Life Sciences platform, which makes use of the pyrosequencing or sequencing by synthesis (SBS) technology (Margulies *et al*, 2005). The pyrosequencing reaction, first described by Nyrén *et al*, (1993), makes use of a chemiluminescent enzyme, present in the reaction mix. The four dNTPs are added, one at a time and pyrophosphate is released when a complementary nucleotide is incorporated by a polymerase, leading to a detectable light signal through the catalysis of pyrophosphate by the chemiluminescent enzyme. The library preparation for the 454 platform makes use of an emulsion PCR reaction described by Tawfik and Griffiths, (1998). The emulsion is made up of streptavidin beads, which form the “droplets”. Sheared DNA fragments are bound to individual beads by means of adapters (Morozova and Marra, 2008). Bound fragments are amplified on the beads that can reach up to 10^7 copies of individual templates per bead (Margulies *et al*, 2005). Beads with bound, clonally amplified template are added to a well of a picotiter plate, which contains hundreds of thousands of wells, each of which can be analysed in parallel via a sequencing by synthesis reaction (Morozova and Marra, 2008).

In 2006 the Solexa/Illumina platform was commercialised and also makes use of the pyrosequencing/SBS chemistry (van Dijk *et al*, 2014). Solid phase amplification is used in contrast to the emulsion PCR amplification used by the 454 platform (Metzker, 2010). DNA samples to be sequenced are sheared to an appropriate length and 5' and 3' adapters are ligated to each DNA molecule (Song *et al*, 2010). ssDNA is then bound to a flow cell, using the ligated adapters. Complementary adapters are attached to the flow cell and allow for the amplification of the attached DNA molecules through bridge amplification (Metzker, 2010). Clusters of clonally amplified DNA are bound to the flow cell after the amplification step, with close to a thousand copies per cluster and up to 200 million clusters per flow cell (Metzker, 2010; Morozova and Marra, 2008).

The third major NGS technology was released in 2007 and was known as Sequencing by Oligo Ligation Detection or SOLiD. This platform makes use of different chemistry dubbed Sequencing by Ligation (SBL), of which multiplex polony sequencing (Shendure *et al*, 2005) by ligation reaction forms the basis. The amplification of the target DNA is performed by an emulsion PCR reaction that is similar to the one used by 454 platforms (Morozova and Marra, 2008). Nucleotide sequences are determined by the ligation of what are known as 1,2-probes.

Sequencing is initiated through the annealing of a universal primer. A library of probes, with 16 dinucleotide combinations and each labelled with one of four different fluorescent dyes (Morozova and Marra, 2008), is added to the sequencing mix (Metzker, 2010). The base directly adjacent to the binding site of the universal primer is the first to be decoded and is denoted as x. The probes are then cleaved and an n-1 primer is used, which decodes the next base, one position to the left. Since an “adjacent valid colour change” (Metzker, 2010) is required, each base is essentially decoded twice, which is one of the inherent advantages of using the SOLiD platform because each base is called with high accuracy.

An offshoot of the 454 platform, called the Personal Genome Machine (PGM) by Ion Torrent was released in 2010. This was later acquired by Life Technologies (van Dijk *et al*, 2014). It was developed by the founders of the 454 platform and uses many of the same technologies in terms of chemistry. The main difference between the PGM and the original 454 technology is its use of semiconductors as opposed to the optical detection used by its parent technology. Within this semiconductor a change in the pH is detected when a proton is released through the incorporation of a nucleotide into a DNA chain by a polymerase enzyme (Liu *et al*, 2012). If the chip is flooded with the incorrect nucleotide, no signal is detected. However, repeats of the same nucleotide will be detected by a signal which will be in proportion to the number of repeated nucleotide.

Another chemistry, which is said to be at the interface of the so-called NGS and third-generation technologies, is that of single molecule sequencing (van Dijk *et al*, 2014). With single molecule sequencing, no amplification of the template is required prior to the sequencing reaction. Instead base determination is done through the incorporation of fluorescently labelled nucleotides that are incorporated during DNA synthesis, by a single DNA polymerase enzyme (van Dijk *et al*, 2014). The first platform to make use of this chemistry was that produced by Pacific Biosciences (PacBio).

At the time of its commercialisation in 2005, the 454 sequencing platform, despite its advantages of massively increased throughput, was received with scepticism by the scientific community. Concerns about fidelity, read length and the redundancy of automated Sanger infrastructure (Schuster, 2008) were abounding. However, just 10 years later, it seems that the scientific community has fully embraced NGS technology with over 40 000 peer-reviewed publications, relating to NGS, published

between the years of 2005 and 2014 (ISI Web of Science database, topic search, January 2015). Each of these platforms and their associated chemistries has their own set of advantages and disadvantages depending on the required applications (Metzker, 2010).

The major advantage of using the 454 platform was its long read lengths (~400bp), relative to other NGS platforms, which was very close to the gold standard length of the Sanger platform (~750bp) (Schuster, 2008). After an updated version of the platform was released in 2008, 454 sequencing could obtain read lengths of 700bp, a figure that is still unsurpassed among the mainstream NGS technologies, with the exception of the PacBio platform, which is capable of producing read length in excess of 25kb (Brown *et al*, 2014). Long read length is of particular importance to *de novo* sequence assemblies, such as sequencing new genomes, where no reference data is available for the accurate construction of contigs (Morozova and Marra, 2008). This platform also comes with an array of disadvantages, chief of which is the high cost of reagents (Metzker, 2010). Dames *et al*, (2010) determined that running samples on the 454 GS FLX was almost 4 times as expensive as running the same samples on the Illumina Genome Analyser platform or US\$10/M bases (Liu *et al*, 2012). Furthermore, it has been shown that short read sequence data, such as those produced by the Illumina and SOLiD platforms, can be used for *de novo* sequence assembly (Farrer *et al*, 2009; Hernandez *et al*, 2008; Mizrachi *et al*, 2010). The high per base cost of the 454 platform and the improvement in the read length figures of the Illumina platforms especially, led to the inevitable phasing out of the 454 platform in favour of most cost effective options.

The SOLiD sequencing technology has become an attractive option for NGS-based research as it provides high throughput per run of up to 120Gb (Liu *et al*, 2012). The per base costs are also many times less than when compared to that of 454 at ~US\$0.17/M bases. The SOLiD platforms have undergone multiple updates over a short period of time (Liu *et al*, 2012), with significant improvements in read length, although this figure is still only 75bp (van Dijk *et al*, 2014), making short read length one of SOLiD technologies biggest drawbacks (Liu *et al*, 2012). To add to these disadvantages of this platform, it requires complex physical and computational infrastructure as well as highly trained personnel to operate (Liu *et al*, 2012). The length of time required to complete a single run can vary between 7 and 14 days depending on the chemistry used (Metzker, 2010).

Illumina is currently considered to be the leader in the field of NGS technologies (van Dijk *et al*, 2014). This can be attributed to improvements which have allowed Illumina platforms to provide the greatest throughput per run (600Gb) and the lowest per base cost of ~US\$0.07/M bases (Liu *et al*, 2012). A number of variations of the Illumina platforms and associated library preparation kits are available, which yield a range of sequence throughputs and read lengths (www.illumina.com). An appropriate Illumina platform should be selected according to the needs of individual projects, in terms of the data output and read length required. Read lengths associated with the MiSeq platform in particular are approaching those required for *de novo* sequencing projects. As of the first quarter of 2015 read length for this platform stood at 2 x 300bp paired-end (http://www.illumina.com/systems/miseq/performance_specifications.html). One of the disadvantages, is that platforms such as the HiSeq 2000 have short read length of 50-100bp paired-end (Liu *et al*, 2012) and loading of the sample onto the flow cells requires skilled personnel due to the fact that overloading of template can lead to overlapping clusters and poor sequence quality (van Dijk *et al*, 2014). Despite this, the advantages of these platforms have led them to become the most widely used with the NGS research environment, including plant virology research (Barba *et al*, 2014).

(2.5) Characterising viral populations using Next Generation Sequencing technologies

Of all the disciplines that require sequencing at their core, NGS has had the greatest impact on the field of genomics (Koboldt *et al*, 2013). Genome size and complexity are no longer the barriers that they were before the advent of NGS technologies and genome projects are now within the reach of small laboratories and research groups (van Dijk *et al*, 2014), due to the massively reduced per base costs and the fact that much less prior knowledge of the genome is required. NGS has also been used for other applications in molecular biology, many of which were previously not possible with the technology. Microarrays are being phased out in favour NGS methods, especially in the area of transcriptome analyses (Morozova and Marra, 2008), since throughput is rapid and there is no need for existing knowledge of specific transcriptomes (Wang *et al*, 2009) and involves sequencing

different RNAs such as mRNA, microRNA and rRNA (Wang *et al*, 2009; Barba *et al*, 2014). Others fields where NGS methods are currently being employed are the discovery of small RNAs (sRNA) in plants, where the mechanisms of RNA interference are being elucidated in the cases of viral infection (Pantaleo *et al*, 2010a) as well as endogenous plant gene expression (Pantaleo *et al*, 2010b). The NGS sequencing of PCR amplified products of specific genes or regions of the genome has been termed amplicon sequencing and has previously been used to pinpoint rare mutations in populations of HIV quasiespecies (Barba *et al*, 2014).

Viral diagnostics and discovery typically relies on a set of conventional and modern techniques (Adams *et al*, 2009). Electron microscopy, PCR protocols using degenerate primers and microarrays are used in cases where novel viruses are suspected to be present but in the case of electron microscopy, small round viruses often go undetected and PCR and microarrays, although sensitive, sometimes have a limited detection capacity for detecting novel viruses (Adams *et al*, 2009). Roossinck *et al*, (2010) suggested that metagenomics should be used as a tool in virus diagnostics and discovery. In the past five years, as with all disciplines of microbiology, NGS has had a major impact on the field of plant virology (Barba *et al*, 2014). An increase in throughput has made the sequencing of whole viral genomes simpler, along with studies in viral epidemiology, detection and discovery.

Total RNA extracts can be used as templates (Adams *et al*, 2009) in diversity studies of known plant viruses as well as in virus discovery. However, this approach yields sequence data, not only of RNA and DNA viruses (from the RNA stage) but also viroids, the plant transcriptome, as well as transcriptome data of any bacteria, fungi or phytoplasmas present in the sample. Although this approach is suitable for metagenomic studies, the vast range of heterogeneous templates present in a sample will lead to relatively low overall coverage of specific sequences, which may be too low if a specific plant virus is being studied. Total RNA extracts are an attractive approach since protocols are usually simple and rapid. Coverage of sequences of interest, generated from total RNA extracts can be increased through enrichment methods, such as the subtractive hybridisation method used by Adams *et al*, (2009). This method removes unwanted host related cDNA from the sample. An alternative to enrichment subsequent to nucleic acid extraction was shown by Day *et al*, (2010). In this study, virus and virus-like particles were isolated by ultracentrifugation and subjected to total RNA extraction. The isolation of intact viral particles may be a

source of concentrated CTV template and maybe be useful for increasing the number of sequence reads returned from an Illumina sequencing run. A protocol for CTV viral particle purification has previously been described by Robertson *et al*, (2005). Template isolated from purified virions will be considered for use as a template for Illumina sequencing.

The majority of NGS projects involving plant viruses have relied on using small interfering RNAs (siRNAs), microRNAs and dsRNAs (Barba *et al*, 2014). siRNAs and microRNAs are typically 21 – 25 nt in length and are derived from dsRNA, which is a product of plant virus replication in their host (Zamore, 2004). The deep sequencing of siRNAs has been used for the sequencing of whole viral genomes (Kreuze *et al*, 2009) with the advantage of a significant reduction in the proportion of host-specific sequence reads (Harper, personal communication). In addition to the advantage of a reduction in host-specific sequence reads, the use of siRNAs as template for deep sequencing studies allows for the discovery of novel viruses (Giampetruzzi *et al*, 2012). In cases where dsRNA is used as template, sequence data will be limited to RNA and circular DNA viruses (Kreuze *et al*, 2009). The use of siRNA templates should yield sequence information on all virus families present within a host. While the discovery of novel viruses may not be the objective of many deep sequencing studies, sequence information of these previously unknown viruses may be of interest for further research. The use of small RNAs as templates is advantageous for studies directed at virus discovery in plants, since even viruses present in very low titres should be detected (Barba *et al*, 2014). Viral metagenomes and genome sequences have been characterised through the sequencing of small RNAs from a variety of agriculturally important crops, such as grapevine (Pantaleo *et al*, 2010a; Giampetruzzi *et al*, 2012) and citrus (Ruiz-Ruiz *et al*, 2011; Loconsole *et al*, 2012; Harper, 2013).

Double stranded RNA (dsRNA) extraction is an alternative method to total RNA and siRNA for plant virus studies. dsRNA species are usually unique to viruses (Roossinck *et al*, 2010), which is produced during replicative reproduction. The use of this method should eliminate most of the background plant transcriptome. However, deep sequencing studies on CTV have yielded large proportions of host specific sequence reads when dsRNA was used as template (Harper and Dawson, personal communication, Zablocki and Pietersen *et al*, 2014). Subsequent to dsRNA extraction, cDNA can be produced using random primers (Adams *et al*, 2009,

Roossinck *et al*, 2010) or specific primers for targeting specific viral sequences. The use of specific primers does have the potential to create biases toward some sequence types. In the viral metagenomics study by Roossinck *et al*, (2010), dsRNA extraction was performed and cDNA was synthesised using random dodecamer primers with specific sequences at their 5' ends. The use of random priming allowed for primer annealing at many sites along either of the RNA strands. This approach reduces the bias that is introduced by using sequence specific primers and could further enhance the value of other studies by yielding viral sequences of viruses other than the virus of interest. An alternative to specific PCR amplifications is the use of ultra-long RNA oligonucleotides as bait, to select for sequences of interest (Gnirke *et al*, 2009). These long oligos are much more tolerant of mismatches when compared with primers and the bias toward certain sequence types is reduced. Roossinck *et al*, (2010) found that a cDNA synthesis temperature of 50°C was ideal, as this favoured the single stranded forms of the RNA. Random hexamer priming was also used but it was found that cDNA products were relatively short and hence random dodecamers were chosen as they produced long cDNA products of up to 3kb. Adams *et al*, (2009), have also used random octomers with success. However, recent work on CTV whole genome data generation, using the random priming method of Roossinck *et al*, (2010), returned a low number of CTV sequence reads, relative to the host, as well as a bias toward the amplification and sequencing of discreet sections of the genome (Zablocki, unpublished data). Therefore, an alternative method of amplification would be the use of CTV specific primers to reverse transcribe and amplify CTV genomic template.

NGS platforms are able to produce increasingly large quantities of data at a rate that exceeds Moore's law (Smith *et al*, 2010). While this increase in data output is desirable for metagenomic studies, lower complexity samples will experience oversampling where the increased number of reads results in no increase in the quality of the data (Smith *et al*, 2010). When low numbers of taxa are analysed, multiple samples can be pooled into a single run, which maximises the useful data output through the process known as multiplexing. When pooling samples, individual reads must be related back to specific samples within the pool (Timmermans *et al*, 2010).

Multiplexing can be achieved through various approaches, the first of which is to physically subdivide the sequencing device (Timmermans *et al*, 2010), for example

the flow cell for Illumina sequencing. Other methods involve tagging the sample sequences themselves. Tagging can be achieved by using either indexed primers or indexed adapters (Cronn *et al*, 2008). Tagged initial primers are used when specific homologous gene regions of samples are targeted (Binladen *et al*, 2007). These primers allow software models to recognise specific tags and relate them back to specific samples within a pool. However, the use of initial primers will bias sequencing data toward the gene region being targeted. Therefore tagged specific primer approaches are only amenable to targeted gene analysis, such as variant discovery in gene regions of interest (Harakalova *et al*, 2011). Tagged random primers, such as those used by Roossinck *et al*, (2010), can be used to reduce the bias introduced by specific primers.

Tagging can also be introduced during library preparation where tagged adapters are ligated to DNA fragments. Tagged adapters have the following advantages over tagged primers. Firstly, the use of tagged adapters alleviates the need to constantly synthesise sample specific tagged primers, which can become time-consuming and expensive (Meyer *et al*, 2007). Secondly, the presence of the known adapter sequence adjacent to the tag sequence provides improved confidence when cross-referencing samples from pools. Roossinck *et al*, (2010) showed that up to 96 samples could be pooled, using a set of 4bp primer tags. The upper limit of numbers of samples that can be pooled will be sample specific and will have to be optimised for each experiment, depending on the amount of coverage required.

REFERENCES

- Adams I.P., Glover R.H., Monger W.A., Mumford R., Jackeviciene E., Navalinskiene M., Samuitene M. and Boonham N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology* vol. 10 pp537-545
- Albiach-Marti M.R., Mawassi M., Gowda S., Satyanarayana T., Hilf M.E., Shanker S., Almira E.C., Vives M.C., López C., Guerri J., Flores R., Moreno P., Garnsey S.M. and Dawson W.O. (2000). Sequences of Citrus tristeza virus Separated in Time and Space are Essentially Identical. *Journal of Virology* vol. 74 pp6856-6865
- Albiach-Marti M.R., Robertson C., Gowda S., Tatineni S., Belliure B., Garnsey S.M., Folimonova S.Y., Moreno P. and Dawson W.O. (2010). The pathogenicity determinant of Citrus tristeza virus causing the seedling yellows syndrome maps at the 3'-terminal region of the viral genome. *Molecular Plant Pathology* vol.11 pp55-67
- Ayllón M.A., Rubio L., Moya A., Guerri J. and Moreno P. (1999). The Haplotype Distribution of Two Genes of Citrus Tristeza Virus Is Altered after Host Change or Aphid Transmission. *Virology* vol. 255 pp32-39
- Barba M., Czosnek H. and Hadidi A. (2014). Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* vol.6 pp106-136
- Bar-Joseph M., Loebenstein G. and Cohen J. (1972). Further Purification and Characterization of Threadlike Particles Associated with the Citrus Tristeza Disease. *Virology* vol. 50 pp821-828.
- Bar-Joseph M., Garnsey S.M., Gonsalves D., Moscovitz M., Purcifull D.E., Clark M.F. and Loebenstein G. (1979). Use of Enzyme-linked Immunosorbent Assay for detection of Citrus tristeza virus. *Phytopathology* vol. 69 pp190-194

Beerenwinkel N. and Zagordi O. (2011). Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology* vol. 1, pp413-418

Bertolini E., Moreno A., Capote N., Olmos A., de Luis A., Vidal E., Pérez-Panadés J. and Cambra M. (2008). Quantitative detection of Citrus tristeza virus in plant tissues and single aphids by real-time RT-PCR. *European Journal of Plant Pathology* vol. 120 pp177-188

Binladen J., Gilbert M.T.P., Bollback J.P., Panitz F., Bendixen C., Nielsen R. and Willerslev E. (2007). The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. *PLoS ONE* vol. 2 e197

Brlansky R.H., Damsteegt V.D., Howd D.S. and Roy A. (2003). Molecular analyses of Citrus tristeza virus subisolates separated by aphid transmission. *Plant Disease* vol. 87 pp397-401.

Brlansky R.H. and Roy A. (2011). Stem-Pitting *Citrus tristeza virus* Predominantly Transmitted by the Brown Citrus Aphid from Mixed Infections Containing Non-Stem-Pitting and Stem-Pitting Isolates. *Plant Disease* vol. 95 pp913-920

Brown S.D., Nagaraju S., Utturkar S., De Tissera S., Segovia S., Mitchell W., Land M.L., Dassanayake A. and Köpke M. (2014). Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnology for biofuels* vol. 7 art.40

Bru D, Martin-Laurent F. and Philippot L. (2008). Quantification of the Detrimental Effect of a Single Primer-Template Mismatch by Real-Time PCR Using the 16S rRNA Gene as an Example. *Applied Environmental Microbiology* vol. 74 pp1660-1663

Cambra M., Gorris M.T., Marroquin C., Roman M.P., Olmos A., Martínez M.C., de Mendoza A.H., López A. and Navarro L. (2000). Incidence and epidemiology of Citrus tristeza virus in the Valencian Community of Spain. *Virus Research* vol. 71 pp85-95

Černi S., Ruščić J., Nolasco G., Gatin Ž., Krajačić M. and Škorić D. (2008). Stem pitting and seedling yellows symptoms of Citrus tristeza virus infection may be determined by minor sequence variants. *Virus Genes* vol. 36 pp241-249

Çevik B., Yardimci N. and Korkmaz S. (2013). The First Identified *Citrus tristeza virus* Isolate of Turkey Contains a Mixture of Mild and Severe Strains. *Plant Pathology Journal* vol. 29 pp31-41

Cronn R., Liston A., Parks M., Gernandt D.S., Shen R. and Mockler T. (2008). Multiplex sequencing of plant chloroplasts using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* vol. 36 art.122

Dames S., Durtschi J., Geiersbach K., Stephens J. and Voelkerding K.V. (2010). Comparison of the Illumina Genome Analyser and Roche 454 GS FLX for Resequencing of Hypertrophic Cardiomyopathy-Associated Genes. *Journal of Biomolecular Techniques* vol. 21 pp73-80

Day M.J., Ballard L.L., Duke M.V., Scheffler B.E. and Zsak L. (2010). Metagenomic analysis of the turkey gut RNA virus community. *Virology Journal* vol.7 art.313

Dieffenbach C.W., Lowe T.M. and Dveksler G.S. (1993). General concepts for PCR primer design. *Genome Research* vol. 3 ppS30-37

Domingo E. (2002). Quasispecies; Theory in Virology. *Journal of Virology* vol. 76 pp463-465

Engelbrektson A., Kunin V., Wrighton K.C., Zvenigorodsky N., Chen F., Ochman H. and Hugenholtz P. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal* vol. 4 pp642-647

Farrer R.A., Kemen E., Jones J.D.G. and Studholme D.J. (2009). De novo assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina Solexa short sequence reads. *FEMS Microbiology Letters* vol. 291 pp103-111

Febres V.J., Mawassi M., Frank A., Bar-Joseph M., Manjunath K.L., Lee R.F. and Niblett C.L. (1996). The p27 Protein Is Present at One End of Citrus Tristeza Virus Particles. *Phytopathology* vol. 86 pp1331-1335

Folimonova S.Y., Robertson C.J., Shilts T., Folimonov A.S., Hilf M.E., Garnsey S.M. and Dawson W.O. (2010). Infection with Strains of Citrus Tristeza Virus Does Not Exclude Superinfection by Other Strains of the Virus. *Journal of Virology* vol. 84 pp1314-1325

Folimonova S.Y. (2012). Superinfection exclusion is an active virus-controlled function that requires a specific viral protein. *Journal of Virology* vol. 86 pp5554-5561

Folimonova S.Y. (2013). Developing an understanding of cross- protection by *Citrus tristeza virus*. *Frontiers in Microbiology* vol 4 art.76

Fulton F.W. (1986). Practices and precautions in the use of cross protection for plant virus disease control. *Annual Reviews in Phytopathology* vol 24 pp67-81

Gago-Zachert S.L., Costa S.N. and Grau O. (2000). Variability of p27 gene from several isolates of citrus tristeza virus as analyzed by single-strand conformation polymorphism. In: *Proc. 14th Conf. IOCV*, 338-342. IOCV, Riverside, California, United States

Gal-On A. and Shibolet Y.M. (2005). Cross-protection, pp261-288. In G. Loebenstein and J.P. Carr (ed.), *Natural resistance mechanisms of plants to viruses*. Springer, Dordrecht, The Netherlands

Garnsey S.M., Civerolo E.L., Gumpf D.J., Paul C., Hilf M.E., Lee R.F., Brlansky R.H., Yokomi R.K. and Hartung J.S. (2005). Biological Characterization of an International Collection of Citrus tristeza virus (CTV) Isolates. In: Proc 16th Conf. IOCV, 75-93, IOCV, Riverside, California, United States

Giampetruzzi A., Roumi V., Roberto R., Malossini U., Yoshikawa N., La Notte P., Terlizzi F., Credi R. and Saldarelli P. (2012). A new grapevine virus discovered by deep sequencing of virus and viroid-derived small RNAs in Cv Pinot gris. *Virus Research* vol. 163 pp262-268

Gillings M., Broadbent P., Indsto J. and Lee R. (1993). Characterisation of isolates and strains of citrus tristeza Closterovirus using restriction analysis of the coat protein gene amplified by the polymerase chain reaction. *Journal of Virological Methods* vol.4 pp305-317

Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., Fennell T., Giannoukos G., Fisher F., Russ C., Gabriel S., Jaffe D.B., Lander E.S. and Nusbaum C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* vol. 27 pp182-189

Gomes C.P.C., Nagata T., de Jesus W.C., Nito C.R.B., Pappas G.J. and Martin D.P. (2008). Genetic variation and recombination of RdRp and HSP 70h genes of Citrus tristeza virus isolates from orange trees showing symptoms of citrus sudden death disease. *Virology Journal* vol. 5 pp1-7

Gowda S., Satyanarayana T., Davis C.L., Navas-Castillo J., Albiach-Marti M.R., Mawassi M., Valkov N., Bar-Joseph M., Moreno P. and Dawson W.O. (2000). The p20 Gene Product of *Citrus Tristeza Virus* Accumulates in the Amorphous Inclusion Bodies. *Virology* vol. 274, pp246-254

Harakalova M., Nijman I.J., Medic J., Mokry M., Renkens I., Blankensteijn J.D., Kloosterman W., Baas A.F. and Cuppen E. (2011). Genomic DNA Pooling Strategy for Next-Generation Sequencing-Based Rare Variant Discovery in Abdominal Aortic Aneurysm Regions of Interest – Challenges and Limitations. *Journal of Cardiovascular Translational Research* vol. 4, pp271-280

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

Harper S.J. (2013). Citrus tristeza virus: evolution of complex and varied genotypic groups. *Frontiers in Microbiology* vol. 4 art.93

Hernandez D., Francois P., Farinelli L., Østerås M. and Schrenzel J. (2008). De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research* vol. 18 pp802-809

Herron C.M., Mirkov T.E., da Graca J.V. and Lee R.F. (2006). *Citrus tristeza virus* transmission by the *Toxoptera citricida* vector: In vitro acquisition and transmission and infectivity immunoneutralization experiments. *Journal of Virological Methods* vol. 134 pp205–211

Hilf M.E., Karasev A.V., Alibiach-Marti M.R., Dawson W.O. and Garnsey S.M. (1999). Two Paths of Sequence Diversion in the Citrus Tristeza Virus Complex. *Phytopathology* vol. 89 pp336-342

Hilf M.E. and Garnsey S.M. (2000). Characterization and Classification of Citrus Tristeza Virus Isolates by Amplification of Multiple Molecular Markers. In: *Proc 14th Conf. IOCV*, 18-27, IOCV, Riverside, California, United States

Hilf M.E., Mavrodieva V.A. and Garnsey S.M. (2005). Genetic Marker Analysis of a Global Collection of Isolates of *Citrus tristeza virus*: Characterization and Distribution of CTV Genotypes and Association with Symptoms. *Virology* vol. 95 pp909-917

Hong S., Bunge J., Leslin C., Jeon S. and Epstein S.S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal* vol.3 pp1365-1373

Iglesias N.G., Gago-Zachert S.P., Robledo G., Costa N., Plata M.I., Vera O., Grau O. and Semorile L.C. (2008). Population structure of *Citrus tristeza virus* from field Argentinean isolates. *Virus Genes* vol. 36 pp199-207

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., Garnsey S.M., Lewandowski D.J. and Dawson W.O. (1995). Complete Sequence of the Citrus Tristeza Virus RNA Genome. *Virology* vol. 208 pp511-520

Koboldt D.C., Steinberg K.M., Larson D.E., Wilson R.K., and Mardis E.M. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell* vol. 155 pp27-38

Kreuze J.F., Perez A., Untiveros M., Quispe D., Fuentes S., Barker I. and Simon R. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* vol. 388 pp1-7

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L. and Law M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*. art.251364

Loconsole G., Saldarelli P., Doddapaneni H., Savino V., Martelli G.P. and Saponari M. (2012). Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* vol. 432 pp162–172

Luttig M., van Vuuren S.P. and van der Vyver J.B. (2002). Differentiation of single aphid cultured sub-isolates of two South African *Citrus tristeza virus* isolates from grapefruit by single-strand conformation polymorphism. In: *Proc 15th Conf. IOCV*, 186-196, IOCV, Riverside, California.

Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A, Berka J. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* vol. 437 pp376-380.

Martin S., Sambade A., Rubio L., Vives M.C., Moya P., Guerri J., Elena J. and Moreno P. (2009). Contribution of recombination and selection to molecular evolution of *Citrus tristeza virus*. *Journal of General Virology* vol. 90 pp1527-1538

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* vol. 40 pp111-118

Metzker M.L. (2010). Sequencing technologies – the next generation. *Nature Reviews* vol. 11 pp31-45

Meyer M., Stenzel U., Myles S., Prüfer K. and Hofreiter M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research* vol. 36 e97

Mizrachi E., Hefer C.A., Ranik M., Joubert F. and Myburg A.A. (2010). *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* vol. 11 issue681

Moreno P., Ambros S., Albiach-Marti M.R., Guerri J. and Pena L. (2008). '*Citrus tristeza virus*: A pathogen that changed the course of the citrus industry'. *Molecular Plant Pathology* vol. 9 pp251-268

Morozova O. and Marra M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* vol. 92 pp255-264

Niblett C.L., Genc H., Cevik B., Halbert S., Brown L., Nolasco G., Bonacalza B., Manjunath K.L., Febres V.J., Pappu H.R. and Lee R.F. (2000). Progress on strain differentiation of *Citrus tristeza virus* and its application to the epidemiology of citrus tristeza disease. *Virus Research* vol. 71 pp97-106

Nikolaeva O.V. and Karasev A.V. (1998). Serological Differentiation of the Citrus Tristeza Virus Isolates Causing Stem Pitting in Sweet Orange. *Plant Disease* vol.82 pp1276-1280

Nyrén P., Pettersson B. and Uhlén M. (1993). Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Analytical Biochemistry* vol. 208 pp171-175

Pantaleo V., Saldarelli P., Miozzi L., Giampetruzzi A., Gisel A., Moxon S., Dalmay T., Bisztray G. and Burgyan J. (2010a). Deep sequencing of viral short RNAs from an infected Pinot Noir grapevine. *Virology* vol. 408 pp49-56

Pantaleo V., Szittyá G., Moxon S., Miozzi L., Moulton V., Dalmay T. and Burgyan J. (2010b). Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. *The Plant Journal* vol. 62 pp960-976

Permar T.A., Garnsey S.M., Gumpf D.J. and Lee R.F. (1990). A monoclonal antibody that discriminates strains of citrus tristeza virus. *Phytopathology* vol. 80 pp224-228.

Polz M.F. and Cavanaugh C.M. (1998). Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied Environmental Microbiology* vol.64 pp3724-3730

Price W.C. (1966). Flexuous Rods in Phloem Cells of Lime Plants Infected with Citrus Tristeza Virus. *Virology* vol. 29 pp285-294

Read D. and Pietersen G. (2015). Genotypic diversity of Citrus tristeza virus within red grapefruit, in a field trial site in South Africa. *European Journal of Plant Pathology* vol. 142 pp.531-545

Robertson C.J., Garnsey S.M., Satyanarayana T., Folimonova S. and Dawson W.O. (2005). Efficient Infection of Citrus Plants with Different Cloned Constructs of *Citrus tristeza virus* Amplified in *Nicotiana benthamiana* Protoplasts. In: *Proc 16th Conf. IOCV*, 187-195 IOCV, Riverside, California, United States

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* vol.19 pp81-88

Rosner A., Ginzburg I. and Bar-Joseph M. (1983). Molecular cloning of complementary-DNA sequences of Citrus tristeza virus RNA. *Journal of General Virology* vol.64 pp1757-1763

Rosner A. and Bar-Joseph M. (1984). Diversity of Citrus tristeza virus-strains indicated by hybridisation with cloned cDNA sequences. *Virology* vol.139 pp189-193

Rubio L., Ayllón M.A., Guerri J., Pappu H.R., Niblett C.L. and Moreno P. (1996). Differentiation of citrus tristeza Closterovirus (CTV) isolates by single-strand conformation polymorphism analysis of the coat protein gene. *Annals of Applied Biology* vol.129 pp479-489

Rubio L., Ayllón M.A., Kong P., Fernandez A., Polek M., Guerri J., Moreno P. and Falk B. (2001). Genetic variation of *Citrus tristeza virus* isolates from California and Spain: Evidence for mixed infections and recombination. *Journal of Virology* vol. 75 pp8054-8062

Ruiz-Ruiz S., Moreno P., Guerri J. and Ambrós S. (2006). The complete nucleotide sequence of a severe stem pitting isolate of *Citrus tristeza virus* from Spain: comparison with isolates from different origins. *Archives of Virology* vol. 151 pp387-398

Ruiz-Ruiz S., Moreno P., Guerri J. and Ambros S. (2007). A real-time RT-PCR assay for detection and absolute quantitation of *Citrus tristeza virus* in different plant tissues. *Journal of Virological Methods* vol. 145 pp96-105

Ruiz-Ruiz S., Navarro B., Gisel A., Peña L., Navarro L., Moreno P., Di Serio F. and Flores R. (2011). Citrus tristeza virus infection induces the accumulation of viral small RNAs (21–24-nt) mapping preferentially at the 3'-terminal region of the genomic RNA and affects the host small RNA profile. *Plant Molecular Biology* vol. 75 pp607-619

Sambade A., Lopez C., Rubio L., Flores R., Guerri J. and Moreno P. (2003). Polymorphism of a specific region in gene *p23* of *Citrus tristeza virus* allows for the discrimination between mild and severe isolates. *Archives of Virology* vol. 148 pp2325-2340

Sanz-Ramos M., Segundo F.D., Escarmis C., Domingo E. and Sevilla N. (2008). Hidden Virulence Determinants in a Viral Quasispecies In Vivo. *Journal of Virology* vol. 82 pp10465-10476

Saponari M., Manjunath K. and Yokomi R.K. (2007). 'Quantitative detection of *Citrus tristeza virus* by real-time reverse trascription-PCR (TaqMan®)'. *Journal of Virological Methods* vol. 147 pp43-53

Satyanarayana T., Siddarama G., Mawassi M., Albiach-Marti M.R., Ayllón M.A., Robertson C., Garnsey S.M. and Dawson W.O. (2000). Closterovirus Encoded HSP70 Homolog and p61 in Addition to Both Coat Proteins Function in Efficient Virion Assembly. *Virology* vol. 278 pp253-265

Satyanarayana T., Gowda S., Ayllón M.A., Albiach-Marti M.R., Rabindran S. and Dawson W.O. (2002). The p23 Protein of *Citrus tristeza virus* controls asymmetrical RNA accumulation. *Journal of Virology* vol. 76 pp473-483

Schuster S.C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods* vol. 5 pp16-18

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*, vol. 158, pp27-37

Sekiya M.E., Lawrence S.D., McCaffery M. and Cline K. (1991). Molecular cloning and nucleotide sequencing of the coat protein gene of citrus tristeza virus. *Journal of General Virology* vol. 72 pp1013-1020

Shendure J., Porreca G.J., Reppas N.B., Lin X., McCutcheon J.P., Rosenbaum A.M., Wang M.D., Zhang K., Mitra R.D. and Church G.M. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* vol. 309 pp1728-1732

Sipos R., Székely A.J., Palatinszky M., Révész S., Márialigeti K and Nikolausz M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiology Ecology* vol. 60 pp341-350

Smith A.M., Heisler L.E., St.Onge R.P., Farais-Hesson E., Wallace I.M., Bodeau J., Harris A.N., Perry K.M., Giaever G., Pourmand N. and Nislow C. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Research* vol. 38 e142

Song C., Wang C., Zhang C., Korir N.K., Yu H., Ma Z. and Fang J. (2010). Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (*Citrus trifoliata*). *BMC Genomics* vol. 11 iss.431

Tatineni S., Robertson C.J., Garnsey S.M., Bar-Joseph M., Gowda S. and Dawson W.O. (2008) Three genes of *Citrus tristeza virus* are dispensable for infection and movement throughout some varieties of citrus trees. *Virology* vol. 376 pp297-307

Tatineni S., Robertson C.J., Garnsey S.M. and Dawson W.O. (2011). A plant virus evolved by acquiring multiple nonconserved genes to extend its host range. *Proceedings of the National Academy of Sciences USA* vol. 180 pp17366–17371.

Tatineni S. and Dawson W.O. (2012). Enhancement or Attenuation of Disease by Deletion of Genes from *Citrus Tristeza Virus*. *Journal of Virology* vol. 86 pp7850-7857

Tawfik D.S. and Griffiths A.D. (1998). Man-made cell-like compartments for molecular evolution. *Nature Biotechnology* vol.16 pp652-656

Timmermans M.J.T.N., Dodsworth S., Culverwell C.L., Bocak L., Ahrens D., Littlewood D.T.J., Pons J. and Vogler A.P. (2010). Why Barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research* vol. 38 e197

van Dijk E.L., Auger H., Jaszczyszyn Y. and Thermes C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics* vol. 30 pp418-426

van Vuuren S.P and Moll J.N. (1987). Glasshouse evaluation of citrus tristeza virus isolates. *Phytophylactica* vol. 19, pp219–221

van Vuuren S.P., Collins R.P. and da Graca J.V. (1993). Evaluation of Citrus Tristeza Virus Isolates for Cross Protection of Grapefruit in South Africa. *Plant Disease* vol. 77 pp24-27

van Vuuren S.P. and van der Vyver J.B. (2000). Comparison of South African Pre-Immunizing Citrus Tristeza Virus Isolates with Foreign Isolates in Three Grapefruit Selections. In: *Proc. 14th Conf. IOCV*, 51-56 IOCV, Riverside, California

van Vuuren S.P., van der Vyver J.B. and Luttig M. (2000). Diversity among sub-isolates of Cross-Protecting Citrus Tristeza Isolates in South Africa. In: *Proc. 14th Conf. IOCV*, 103-110 IOCV, Riverside, California, United States

van Vuuren S.P. and Manicom B.Q. (2005). The Response of Star Ruby Grapefruit to Different Citrus tristeza virus isolates. In: *Proc. 16th Conf. IOCV*, 112-116 IOCV, Riverside, California, United States

von Broembsen L. A. and Lee A.T.C. (1988). South Africa's Citrus Improvement Program. In: *Proc. 10th Conf. IOCV*, 407-416 IOCV, Riverside, California

Wang Z., Gerstein M. and Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* vol. 10 pp57-63

Yang Z.N., Mathews D.M., Dodds J.A. and Mirkov T.E. (1999). Molecular Characterization of an Isolate of Citrus Tristeza Virus that Causes Severe Symptoms in Sweet Orange. *Virus Genes* vol. 19 pp131-142

Yokomi R.F., Lastra R., Stoetzel M.B., Damsteegt V.D., Lee R.F., Garnsey S.M., Gottwald T.R., Rocha-Peña M.A. and Niblett C.L. (1994). Establishment of the Brown Citrus Aphid (Homoptera: Aphididae) in Central America and the Caribbean Basin and Transmission of Citrus tristeza virus. *Journal of Economic Entomology* vol. 87 pp1078-1085

Yoon J.Y., Ahn H.I., Kim M., Tsuda S. and Ryu K.I. (2006). '*Pepper mild mottle virus* pathogenicity determinants and cross protection effect of attenuated mutants in pepper'. *Virus Research* vol. 118 pp23-30

Zablocki O. and Pietersen G. (2014). Characterization of a novel citrus tristeza virus genotypewithin three cross-protecting source GFMS12 sub-isolates in South Africa by means of Illumina sequencing. *Archives of Virology* vol. 158 pp2133-2139

Zamore P.D. (2004). Plant RNAi: How a Viral Silencing Suppressor Inactivates siRNA. *Current Biology* vol.14 ppR198-200

Chapter 3

PCR bias associated with primers targeting conserved sequences for genotype diversity studies within *Citrus tristeza virus* populations

INTRODUCTION

Polymerase chain reaction (PCR) based amplification and subsequent sequencing of genes has been a valuable tool for determining the diversity and taxonomic relationships of populations of fungi, bacteria and viruses. Universal primers are usually designed for amplifying multiple templates present in samples, often composed of a variety of organisms (Farris and Olsen, 2007). Anderson *et al*, (2003) state that it is critical to design universal primers that are specific enough to target certain taxonomic groups but at the same time to beware that the specificity does not bias the PCR reaction toward certain sequences within the population.

The potential for bias and the effect that it has on the distortion of microbial population data has been most widely documented on PCRs targeting regions of bacterial 16S rRNA genes (Bru *et al*, 2008; Hong *et al*, 2009; Engelbrektson *et al*, 2010; Polz and Cavanaugh, 1998; Sipos *et al*, 2007) with few examples illustrating similar issues in viral populations. There are no technical reasons why viral population data sets cannot be biased in the same way as fungal and bacterial populations.

Citrus tristeza virus (CTV) is a *Closterovirus*, whose virions are flexuous, thread-like (Karasev *et al*, 1995) and contains a positive sense genome of ~20kb (Bar-Joseph *et al*, 1985), which has 12 open reading frames (Karasev *et al*, 1995). Sequence heterogeneity increases from the 3' to 5' end of the genome. At the 5' end, as little 44% sequence homology may be observed among genotypes, while sequence homology usually remains much greater within the 8kb toward the 3' ends of the genomes (Mawassi *et al*, 1996). CTV is the most severe viral pathogen of citrus (Ghorbel *et al*, 2001) and has become well established within almost all of the major citrus production areas worldwide, along with its aphid vector species (Moreno *et al*, 2008). Infections may remain symptomless or manifest as a variety of symptoms such as stem-pitting, decline and seedling yellows (Moreno *et al*, 2008),

which result in yield losses and shortened productive lifespans of trees. This translates to reduced profitability for producers.

The virus exists as a complex of strains based on six distinct lineages (Harper, 2013) that can exist as mixtures of strains within a single infection (Rubio *et al*, 2001). Different strains of the virus are able to elicit various disease symptoms. Some severe strains of CTV, which are endemic to southern Africa (da Graça *et al*, 1984) elicit stem-pitting. In order to increase the economically productive lifespan of grapefruit trees in the field, budwood trees are inoculated with mild-strain populations of CTV (Lee and Keremane, 2013). This is known as cross-protection and is intended to prevent the secondary infections of potentially severe strain of the virus (Gal-On and Shibolet, 2005) through aphid inoculations. Folimonova *et al*, 2010 showed that CTV cross-protection functions via strain specific mechanism in which the expression of a functional p33 protein is essential (Folimonova, 2012). Previously, cross-protecting sources were characterised on an empirical basis through grafting onto indicator hosts (Moreno *et al*, 2008). Due to the genotype specific mechanism of CTV cross-protection, it is essential to not only characterise candidate cross-protection sources prior to use but to also determine which CTV genotypes are present in the field. This will allow for the improved selection of specific cross-protecting sources that have the capacity to protect against severe genotypes circulating within a production area. A number of gene regions have been used previously, to characterise CTV populations. These include the A-region and the F-region (Rubio *et al*, 2001), p23 gene (Sambade *et al*, 2003), p27 gene (Iglesias *et al*, 2008), p25 gene (Pappu *et al*, 1993), p18 and p13 gene, (Ayllón *et al*, 2006) and the coat protein gene (CP) (Sekiya *et al*, 1991). Due to the genotype specific mechanism of cross-protection, the need for the accurate determination of CTV population compositions has become apparent. Due to the heterogeneity associated with the 5' end, primer targeting regions in this half of the genome are designed to target consensus sequences and may have a number of nucleotide differences when compared with the cognate binding sites of other known CTV genotypes.

In this study, the primer-directed selection bias of primer sets targeting four different genomic regions was evaluated. These primer sets include those targeting the A and F regions in the variable 5' end of the genome (Rubio *et al*, 2001), the p23 gene toward the conserved 3' end of the genome (Sambade *et al*, 2003) and the p33 gene near the mid-point of the genome (this study). A set of genotype-specific

primers were designed for each gene region and then used to amplify plasmid inserts of known identity. Mixtures containing equimolar concentrations of each, now genotype-true template, were produced and then reamplified, using the original consensus primers. The resulting amplicons were then used in both mass cloning experiments and subjected to Illumina MiSeq sequencing. The MiSeq reads and clone sequences were mapped back or aligned with their corresponding reference sequences.

MATERIALS AND METHODS

Gene selection and primer design

Four CTV gene regions were selected for the PCR bias experiment. The A-region and F-region (Rubio *et al*, 2001) are located toward the 5' end of the genome in ORF1a. The p33 gene is located near the midpoint of the genome and the p23 gene is located near the 3' end of the genome (Sambade *et al*, 2003).

The published primer sequences, used to amplify these gene regions, were aligned against reference sequences, using BioEdit Sequence alignment editor 7.1.3 (Hall, 1999). Whole genomes of 46 reference sequences were obtained from GenBank and are as follows: NC_001661 (T36); AY_340974 (Qaha); U16304 (T36); DQ272579 (Mexico); AY170468 (T36); EU937521 (T36); KC517485 (FS674-T36); KC517486 (FS701-T36); KC517487 (FS703-T36); KC517488 (FS577); JX266713 (Taiwan-Pum/M/T5); AF001623 (SY568); AF260651 (T30); Y18420 (T385); KC517489 (FS701-T30); KC517490 (FL278-T30); KC517491 (FS703-T30); JF957196 (B301); FJ525432 (NZRB-G90); GQ454869 (HA 18-9); FJ525435 (NZRB-M17); JX266712 (Taiwan-Pum/SP/T1); FJ525431 (NZRB-M12); FJ525433 (NZRB-TH28); FJ525434 (NZRB-TH30); JQ798289 (A18); KC525952 (T3); HM573451 (Kpg3); EU857538 (SP); GQ454870 (HA 16-5); DQ151548 (T318A); AB0463981 (NUagA); JQ911664 (CT11A); KC517493 (FL202-VT); U56902 (VT); KC517492 (FS703-VT); EU937519 (VT); KC517494 (FS701-VT); KC262793 (L192GR); JQ911663 (CT14A); FJ525436 (NZ-B18); JQ965169 (T68); EU076703 (B165); JQ061137 (AT-1); KC333868 (12-8); KC333869 (12-9).

Table 1: Genotype specific primer sequences used to amplify the insert of each specific plasmid to generate genotype true amplicons. Bases that differ from the published generic CTV primers are shown in bold text, with the last column indicating the total number of differences between genotype specific primers and the generic primers

Gene/region	Primer orientation	Primer name	Strain represented	Primer sequence (5' to 3')	Number of differences
A-region	Forward	A-F standard (Rubio et al, 2001)	Consensus /VT	AGCTGTTTCGTGAAACGCGG	-
		A-F T3	T3	ACGTGTT T GTGAAACGCGG	1
		A-F T30	T30	ACGTGTTTCGTGAAACG T GG	1
		A-R HA16-5	HA16-5	ACGTGTT TA TGAAACGCGG	2
		A-F RB	RB1 and RB2	ACGTGTT T GTGAA G CG T GG	3
		A-F B165	B165	ACGTGTT TA C GAAACG T GG	4
		A-F T36	T36	ACGTGTT T GT A AA G CG T GG	4
	Reverse	A-R standard (Rubio et al, 2001)	Consensus/T3 85	GTCGATAACTCGACAAACGAGC	-
		A-R VT	VT	G C CGATAACTCGACAAACGAGC	1
		A-R T30	T30	GTCGATAACTCGACA G ACGAGC	1
		A-R RB2	RB-TH30	GTC TG TAACTCGACAAACGAGC	2
		A-R T3	T3	G C CGATAACTCGA T AAACGAGC	2
		A-R RB1	RB-TH28	G C C TG TAACTCGACAAACGAGC	3
		A-R T36	T36	GTC TG TAAC C CGACAAACGAGC	3
		A-R B165	B165	AG C TG TAACTCGACAGACGAGC	4
		A-R HA16-5	HA16-5	CC C TG TA G CTCGACA G ACG G GC	7
F-region	Forward	F-F standard (Rubio et al, 2001)	Consensus	GTGTTATCATGCGTCTGAAGCG	-
		F-F T30	T30	GTGTT T TCATGCGTCTGAAGCG	1
		F-F VT	VT	GTGTT T TCATGCGTCTGA T GCG	2
		F-F B165	B165	A TGTT T TCATGCGTCTGA T GCG	3
		F-F HA16-5	HA16-5	GTGTT T TCATGCGTC C GA T GCG	3
		F-F RB-TH28	RB-TH28	GTGTTATCATG A GTC G GA G GCG	3
		F-F T36	T36	GTGTTATCATGC A TC G GA G GCG	3
	Reverse	F-R standard (Rubio et al, 2001)	Consensus	GGAATCTTAATCCTAATCAAG	-
		F-R T30	T30	GGAATCTTAATCCTAATCAAG	0
		F-R RB-TH28	RB-TH28	GGAATCTT G AT T CTAATCAAG	2
		F-R B165	B165	GGAATCTT G ATCCT G ATCAAG	2
		F-R VT	VT	GGAATCTT G ATCCT G ATCAAG	2
		F-R T36	T36	GGAAT T TT G AT T CTAATCAAG	3
		F-R HA16-5	HA16-5	GG G ATCTTAAT T CTAAT T AA A	4
p33 gene	Forward	Univ p33 F	Kpg3 HA 16-5 T36 CT-3719 AT-1	GATGTTTGCTTCGCGAGC	-
		p33-F-RB	Taiwan-Pum/SP/T1 RB	GATGTTTGC T TTCGCGAGC	1
		P33-F-CTZA	CTZA	GATGTTTGCTTCGCGAG T	1
	Reverse	Univ p33 R	Kpg3 HA 16-5 T36 CT-3719 AT-1 Taiwan-Pum/SP/T1 RB CTZA	CCCGTTTAAACAGAGTCAAACGG	-
p23 gene	Forward/ Reverse	PM50/PM51 (Sambade et al, 2003)	T36 T30 SP Kpg3 RB1	ACTAACTTTAATTCGAACA / AACTTATCCGTCCTCACTTC	-

A total of 7 genotype-specific forward primers and 9 reverse primers targeting the same region as the genotype non-specific primers (A-regions) were used to incorporate genotype-true primer binding sites into amplicons that were generated

through the amplification of plasmid clones of the genotypes VT, T3, T30, HA 16-5, RB1, RB2, B165 and T36. Six forward and six reverse F-region genotype- specific primers were used to target the inserts of plasmid clones representing the T30, VT, B165, HA 16-5, RB1 and T36 genotypes. The p33 gene genotype specific forward primer binding sites of the RB genotype was incorporated into the amplicons representing Taiwan-Pum/SP/T1 and RB, while the CTZA specific forward primer binding site was incorporated into an amplicons representing CTZA. The original reverse primer was used in all amplification steps. The sequences of the primers are listed in *Table 1*. In all instances, these amplifications created amplicons with genotype-true sequences at the genotype non-specific primer binding sites.

Plasmid insert amplification

The inserts of each of these plasmids was amplified with genotype specific primers (*Table 2*), by adding 1µl of each recombinant plasmid to a reaction mix containing: 5µl of 1x PCR reaction buffer, 2.5mM MgCl₂, 2.5 U DNA polymerase (Bioline, London, United Kingdom), 0.14mM dNTP mix, 2µM forward primer, 2µM reverse primer and PCR grade water to a final reaction volume of 50µl, using the following PCR conditions: 1 cycle of 92°C for 2min. 35 cycles of 92°C for 30s, 55°C for 45s and 72°C for 1min. 1 cycle of 72°C for 10min. The concentration of each of these PCR products was determined by using the Quant-IT™ dsDNA BR Assay Kit in conjunction with the Qubit fluorometer (Life Technologies, Grand Island, NY, United States). Equimolar concentrations of each genotype specific template were added together to produce a single defined artificial template for each gene region. These templates were then amplified with the standard A, F and p23 gene primers, by adding 5µl of template to each reaction mix and using the same PCR reaction conditions that were used for the genotype specific primer amplifications. The products were then purified using the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI, United States) and quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, United States).

Cloning of amplified products

The quantified amplicons were then subjected to T/A cloning reactions, using the pGEM-T Easy vector (Promega, Madison, WI, United States), according to the

manufacturer's specifications and used to transform competent *E. coli* JM109 cells. Putative recombinants were selected using blue/white selection and plasmid extractions were performed using alkaline lysis.

Each plasmid insert was amplified using the T7 (5'- TAA TAC GAC TCA CTA TAG GG-3') and SP6 (5'- ATT TAG GTG ACA CTA TAG AA -3') vector specific primers and the same PCR conditions as those used to amplify the original plasmid inserts. Amplicons displaying the correct amplicon size were purified by adding 2µl of FastAP and 0.5µl ExoI enzymes (Thermo Scientific, Vilnius, Lithuania) to 19µl of amplified PCR product.

Sanger sequencing reactions were carried on each purified product, in a single direction, using the vector specific primer T7. Sequencing reactions contained: 1µl BigDye® Terminator mix v3.1 (Applied Biosystems, Foster City, CA, United States), 2.25µl 5x BigDye® v3.1 sequencing buffer (Applied Biosystems, Foster City, CA, United States), 0.75µl T7 primer (2µM), and molecular grade water to a final reaction volume of 10µl. Sequencing reaction conditions were carried out as follows: 94°C for 1 minute, 25 cycles of 94°C for 10 seconds, 50°C for 5 seconds, 60°C for 4 minutes.

Illumina MiSeq sequencing of amplified products

Next generation sequencing (NGS), using the Illumina MiSeq platform (Illumina, San Diego, CA, United States), was carried out on each of the purified PCR amplicons that were used for the cloning experiment. The A-region amplicon sample was sequenced twice in two separate library preparations and sequencing runs. Paired-end DNA libraries were prepared using the Nextera V2 sample kit (Epicentre, Madison, WI, United States). Each of these libraries was run on 1/8th of a lane on an Illumina MiSeq flow cell. The samples were sequenced at the Agricultural Research Council (ARC), Biotechnology Platform, Pretoria, South Africa.

Data analysis

Sanger sequencing chromatograms were viewed and amended using Chromas Lite 2.1 (Technelysium, Brisbane, Australia). Sequences found to be in the reverse orientation were reverse complemented using CLC Main Workbench 6 (CLC bio, Aarhus, Denmark). Alignments of sequences were carried using the CLUSTAL W alignment software (EBI, Cambridgeshire, England) within the BioEdit Sequence

alignment editor 7.1.3 (Hall, 1999). Phylogenetic dendrograms were produced using the MEGA 4.1 software package (Tumara *et al*, 2007), as neighbour-joining trees with a bootstrap value of 1000. To determine whether the numbers of clones recovered for each genotype suggested over or under-representation, significance of the deviation from the expected equivalence of numbers was determined using the Chi-squared test.

The NGS data was analyzed using the CLC Genomics Workbench 5.5.1 (CLC bio, Aarhus, Denmark). Data was imported as paired-end reads with a distance range of 180-250. Adapter and quality trimming was done using the default program settings with Nextera V2 transposase adapter sequences (Transposase 1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG; Transposase 2: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG). Reference assemblies were performed on each of the trimmed datasets, using the sequences of their respective clones, listed in *Table 3* as references. The reference assemblies were carried out using the program defaults, except for the length and similarity fractions, which were both set to 0.9. The program default settings were set as follows: “Ambiguous trim” yes, “ambiguous limit” 2, “quality trim” yes, “quality limit” 0.05, “use colorspace” no, “create report” yes, “also search on reversed sequences” no, “save discarded sequences” no, “remove 5' terminal nucleotides” no, “discard short reads” no, “remove 3' terminal nucleotides” no, “discard long reads” no, “save broken pairs” no.

RESULTS

Multiple clone sequence data

The sequence data of multiple clones from each amplification is listed in *Table 2*. *Table 1* shows the genotype-specific primer sequences for all four of the gene regions that were analysed. A total of 6 genotype-specific forward and reverse primers were used for the F-region primer bias experiment, in which a total of 156 clone sequences were recovered. The clones representing T30 were significantly ($P=0.05$) over-represented with 88% of the clones corresponding to this genotype. The remainder of the genotypes were at roughly the same levels and were significantly underrepresented with VT, HA 16-5, B165 and RB-TH28 represented at levels of 8%, 5%, 3% and 0.5% respectively. No plasmids representing T36 were recovered.

A total of three genotype-specific p33 gene forward primers were used to amplify the cloned inserts of plasmids representing the genotypes listed in *Table 1* and *2*. The same reverse p33 gene primer was used in combination with the genotype specific forward primer. The cloned sequence, putatively termed CT-3719, was obtained from a sample from of South Africa Star Ruby trees (unpublished data) and forms a unique phylogenetic branch (supported by bootstrapping) that contains none of the currently available reference sequences. The Univ-p33-F forward primer was used to amplify the cloned inserts of Kpg3, HA 16-5, T36, CT-3719 and AT-1 (*Figure 1*). The p33-F-RB primer was used to amplify the cloned inserts of Taiwan-Pum/SP/T1 and RB. The cloned insert of CTZA was amplified using the p33-F-CTZA forward primer. The cloning of the p33 gene indicated that a significant primer-directed bias existed for the p33 primer pair. This bias was directed against the Taiwan-Pum/SP/T1, RB and CTZA sequence types. This coincides with the single nucleotide difference near the mid-point of the forward primer binding site of the Taiwan-Pum/SP/T1 and RB and the single nucleotide difference at the 3' position of the CTZA primer. The remainder of the p33 sequence types contained forward primer binding sites that were identical to those of the universal p33 forward primer. The T36 sequence type was also significantly under-represented with only 6 clones (9% of the total) recovered. The remainder of the sequence types were present at roughly similar levels, being over-represented at 27%, 21%, 21% and 18% respectively for Kpg3, HA16-5, AT-1 and CT-3719. *Figures 1-4* under Appendix B illustrate the phylogenetic dendrograms that were used to determine the identity of cloned sequences.

Table 2: Cloned sequence and Illumina MiSeq data showing the primer-directed bias, associated with the A and F region, p33 and p23 gene PCR amplification protocols. The Illumina MiSeq data for the A-region amplicon is an amalgamation of two separate sequencing runs. The first column shows the particular primer pairs that were analysed. Column 3 and 5 show the number of sequenced clone inserts and Illumina MiSeq sequencing reads respectively that correspond back to their particular sequence type. Data is ranked by diminishing number of either cloned sequences that clustered with their particular reference sequence or the number of Illumina MiSeq reads that mapped back to particular references (column 4).

Region/gene	Number of differences within F and R primers	Strain	Nr. of seqs grouping with ref clone (% of total number of recombinants)	Strain	Nr. Of reads mapping to reference (% of total reads)
A-region	3	T3	44 (33)	T3	41043 (32)
	1	VT	44 (33)	VT	31369 (24)
	8	B165	19 (14)	B165	22919 (18)
	2	T30	9 (7)	T30	18831 (15)
	6	RB-TH28	6 (5)	RB-TH28	6800 (5)
	5	HA 16-5	6 (5)	RB-TH30	3967 (3)
	7	RB-TH30	4 (3)	T36	2656 (2)
	9	T36	0 (0)	HA16-5	604 (0.5)
F-region	1	T30	134 (88)	T30	402861 (54)
	4	VT	11 (8)	VT	239504 (32)
	7	HA16-5	6 (5)	HA16-5	49839 (7)
	5	B165	4 (3)	B165	30305 (4)
	6	RB-TH28	1 (0.5)	T36	13218 (2)
	5	T36	0 (0)	RB-TH28	9360 (1)
p33 gene	-	Kpg3	18 (27)	AT-1	195802 (24)
	-	HA 16-5	14 (21)	CT-3719	172031 (21)
	-	AT-1	14 (21)	T36	166143 (20)
	-	CT-3719	12 (18)	Kpg3	113290 (14)
	-	T36	6 (9)	HA 16-5	100392 (12)
	1	Taiwan-Pum/SP/T1	2 (3)	CTZA	40167 (5)
	1	RB	1 (2)	RB	14129 (2)
	1	CTZA	0	Taiwan-Pum/SP/T1	12140 (2)
p23 gene	-	T36	41 (35)	T30	32210 (23)
	-	SP	21 (18)	T36	30100 (22)
	-	Kpg3	19 (16)	RB-TH28	26854 (19)
	-	RB-TH28	19 (16)	SP	25486 (18)
	-	T30	18 (15)	Kpg3	25138 (18)

Illumina MiSeq data

A total of 128 459 reads mapped back to the A-region reference sequences. The number of reads mapping back to each specific reference sequence can be seen in *Table 2*. The rank order of dominance among genotypes by MiSeq read numbers was the same as that of the multiple clone sequence data for the first 5 genotypes. *Tables 1-4* within Appendix A contain the output data from the CLC Genomics Workbench 5.5.1 analyses of each dataset. 32% of all the mapped reads mapped back to the T3 reference, which correlates to the 33% of total clones mapping back to T3. Differences were observed between the MiSeq data and cloned sequences for VT, B165 and T30, where 24%, 18% and 15% of the MiSeq reads and 33%, 14% and 7% of the cloned sequences mapped back to their respective reference sequences. Differences in the order of dominance were observed for the last three genotypes of the A-region, where RB TH-30 is ranked as the sixth most numerous sequence-type with MiSeq, while the cloned sequence data suggests that RB TH-30 is the seventh most numerous sequence type. T36 is the seventh most numerous A-region genotype, representing 2% of the mapped reads, according to the MiSeq data, while no clones were recovered (Read and Pietersen, 2015). Lastly, while 6 HA 16-5 clones (ranked sixth) were recovered from the cloning experiment, this genotype ranked lowest in terms of the number of reads mapping back to this reference (0.5% of mapped reads).

A total of 754 723 reads mapped back to the F-region reference sequences. The ranking in terms of the numbers of reads mapping back to their respective reference sequences was the same, except for some very minor differences, as the ranking of sequence types suggested by the F-region cloning experiment. The minor differences between the two datasets was for RB TH-28 and T36, which were the fifth and sixth most numerous sequence types according to the clones recovered, while this ranking is reversed when considering the MiSeq dataset. In the MiSeq dataset the T30 sequence type was significantly ($P=0.05$) over represented with 54% of the mapped reads mapping back to the T30 reference sequence. This is far greater than the expected 16.67%, if no primer directed bias existed for the F-region primers. The multiple clone sequence data suggests an even greater bias toward T30, with 88% of the recovered cloned sequences, grouping with T30. The VT sequence type was over-represented at 33% of mapped reads, while the multiple clone sequence data suggests that VT was under-represented, with only 8% of

cloned sequences grouping with the cognate reference sequence. The remainder of the genotypes, namely HA 16-5, B165, T36 and RB TH-28 were represented by 7%, 4%, 2% and 1% of the mapped reads respectively, while HA 16-5, B165, RB TH-28 and T36 were represented by 5%, 3%, 0.5% and 0% of the sequences recovered from the cloning experiment.

The primer binding sites for the five genotype sequences used in the p23 gene analysis are all identical in both the forward and the reverse sites. The MiSeq data for the p23 gene amplicon sequencing suggests that this region has reduced inherent primer-directed PCR bias. Assuming no inherent primer-directed bias associated with the conserved p23 gene primers, each sequence type should be represented by 20% of the mapped reads. T30, T36, RB TH-28, SP and Kpg3 were represented at similar levels of 23%, 22%, 19%, 18% and 18% of the mapped reads respectively. Data relating to the mapping of reads for each Illumina MiSeq dataset is present in *Tables 1-4* under Appendix A.

DISCUSSION

This study set out to evaluate the ability of several sets of PCR primer pairs for their ability to amplify target sequences from an artificially assembled equimolar heterogeneous template, in a nonbiased manner. These primer pairs targeted the A, F (Rubio *et al*, 2001), p23 gene (Sambade *et al*, 2003) and p33 gene (this study). Two different technologies were used to analyse the apparent population compositions of the resulting PCR products, namely the Sanger sequencing of multiple clones and Illumina MiSeq sequencing of the PCR products amplified from equimolar templates containing templates with genotype-true primer sites.

Sets of genotype-true primer pairs were designed, conspecific to the published consensus primer binding sites for each gene region, which were then used to incorporate genotype-true primer binding sites into unique PCR amplicons. Artificial equimolar templates were then assembled for each gene region through the addition of precise amounts of amplicons containing these genotype-true primer binding sites. These templates were then re-amplified with the consensus primer pairs and the “population genotype composition” of the resulting PCR products analysed, in order to determine which sequences were preferentially represented.

Between 67 and 156 cloned sequences were analysed for each of the primer pairs, which is considerably more than that usually used in viral quasispecies analysis (Mullan *et al*, 2004). Using Chi-squared analyses, all of the primer sets showed significant biases to multiple sequences, when conducting Sanger sequencing of multiple clones. The analyses of the clone sequences of the A-region and the p23 gene were previously published by Read and Pietersen, (2015). The A-region consensus primers provided an over-representation of sequences containing T3 and VT true primer target sites, accounting for 66% of sequences, while those containing the remaining six genotype-true primer target sites only made up only 34 % of the total. The p23 analysis showed that only the sequence containing the T36-true primer sites was over-represented while the remaining four sequences were represented at roughly similar levels. The F-region consensus primers resulted in the greatest bias of any one sequence within the whole study with T30 being represented at 88% of the total and the remaining five categories represented by only 12% of the sequenced clones. Six out of the eight categories for the p33 gene analysis should have been represented at equal levels, since their primer target sites contained no differences when compared with the consensus primers. This was reflected to a certain degree with Kpg3, HA 16-5, AT-1 and CT-3719 being represented at roughly equal levels, however T36 was significantly ($P=0.05$) underrepresented at 6% of the total. The Taiwan-Pum/SP/T1 and RB sequence types, which both contained single differences in their forward primer binding sites, represented 5% of the total cloned sequences recovered.

Generally speaking, the results of this study support the idea that an increasing number of nucleotide differences between a primer binding site and its conspecific consensus will lead to a decrease in the primers ability to efficiently amplify the template, especially within a multi-template PCR reaction (Ibarbalz *et al*, 2014). A number of exceptions to this trend were observed, such as the in the A-region analysis, where the B165 sequence were represented at 19% of the total clone sequences despite having eight combined differences within the forward and reverse primers. The possible reasons for this could be due to the fact that an insufficient number of clones were selected to accurately quantify the sequences within the A-region PCR amplicon (Read and Pietersen, 2015), although data generated from the Illumina MiSeq run corroborates the observation of B165 being represented at an unexpectedly high level. Therefore the more likely reason is due to the fact that only

one of the eight differences occurred within the critical 5 bp of the 3' end of these primers, the others which are thought to have only moderate effects on amplification efficiencies (Lefever *et al*, 2013; Waud *et al*, 2014). This was further supported by the F-region data, where all of the forward primers, except that specific to T30, contained a nucleotide that differed from the consensus, 4 bp from the 3' end of each primer. This was the probable cause of the severe over-representation of T30 (88% of sequences cloned) and the under-representation of the remainder of the sequences (8% or less).

Four of the eight sequences of the p33 gene analysis, were represented at similar frequencies, between 18% and 27%, of clone inserts sequenced. The under-representation of T36 (6%) cannot be attributed to primer-associated amplification bias, since the T36 primer target sites matched the sequences of the consensus primers exactly. If the Illumina MiSeq data was not available, this under-representation might have been ascribed to experimental errors, such as pipetting or quantification errors during the assembly of the artificial template. However, the Illumina sequence data suggests that T36 was amplified to much higher levels (20%). The under-representation of T36 in the clone sequence dataset must have been due to the under sampling of clones. Due to the much greater data output of the Illumina MiSeq datasets, it is assumed that they provide a much more accurate quantification of the sequences within a PCR amplification product. Given this assumption, the large disparity (12-24%) between the sequences that contained primer target sites totally complementary to the p33 gene consensus primers, can probably be attributed to errors during the assembly of the artificial template, as these sequences should have been detected at the same frequencies. Ihrmark *et al*, (2012) determined the bias associated with various primers targeting fungal ITS regions. In this study, artificial templates were amplified in 20 technical replicates, while our study only made use of a single amplification of the artificial templates. While technical replicates should allow for the normalisation of errors introduced during the amplification of the artificial template, as well as random PCR distortion (Ihrmark *et al*, 2012), any errors introduced during the production of the artificial template itself will only be amplified in all of the replicates. Therefore, in primer validation studies such as these, the practice of performing technical replicates should be taken a step back to the point of producing the artificial templates, where the components of the artificial template can be quantified and then assembled multiple times and the

resulting replicates pooled in order to normalise the effect that individual errors may have during the production of an artificial heterogeneous template.

The apparent under-representation of the CTZA sequence by the p33 gene primer pair can be attributed to the single nucleotide difference within the CTZA-specific forward primer binding site, specifically at the terminal 3' binding position of the primer. The forward primer binding site of Taiwan-Pum/SP/T1 and RB both contained a single nucleotide difference, 10 bp from the 3' terminus of the binding site. As mentioned previously, differences located outside of the crucial 5 bp of the 3' terminal binding region of a primer are considered to have a minor impact on PCR amplification efficiency (Lefever *et al*, 2013; Waud *et al*, 2014). However, Ihrmark *et al*, (2012) showed that a mismatch up to 14 bp away from the 3' terminus of a binding primer can reduce amplicon abundance by up to two orders of magnitude when amplifying from a heterogeneous template. Therefore, the single mismatch at the midpoint of the Taiwan-Pum/SP/T1 and RB forward primer binding sites could have been responsible for the reduction in amplification efficiencies of these templates. The data obtained for the clone sequencing data was confirmed by the Illumina MiSeq data for the p33 gene, where CTZA, RB and Taiwan-Pum/SP/T1 had the fewest numbers of reads mapping back to their corresponding reference sequences. This has shown that the generally accepted tolerances for primer mismatches during amplification (Lefever *et al*, 2013) should not be considered a constant. In most population diversity studies, primers designed in previous studies are used and sometimes combined arbitrarily (Waud *et al*, 2014), without any considerations of the effects that potential bias may have in skewing the population data. Therefore, in order to reduce primer-associated bias, primer sets should be subjected to studies such as this one and assumptions about potential bias should not be made by merely estimating the potential effects of primer-template mismatches as it has become apparent that the effects of mismatches, even when distant from the 3' end of the primer, can be unpredictable.

The obvious effects that even single nucleotide mismatches can have on PCR efficiencies when amplifying from a heterogeneous template suggests that the goal of obtaining completely bias free amplification protocols is unrealistic. Any population study that requires preliminary amplification of a target gene/region can, at best, be considered semi-quantitative (Ihrmark *et al*, 2012). A major move forward to reducing bias in population diversity studies would be to remove the need for amplification

prior to sequencing (Urich *et al*, 2008; Ihrmark *et al*, 2012). This approach has already been used to generate consensus level whole genome sequences of RNA viruses (Logan *et al*, 2014). In their study, Logan *et al*, (2014) produced cDNA using gene specific primers as well as random hexamers. Double stranded cDNA was synthesised using a second strand synthesis kit. This was then used directly as material in an Illumina MiSeq library preparation. However, bypassing of amplification prior to library preparation does not necessarily translate into a totally bias free system. Library preparations also require large quantities (~1µg) of input DNA for a successful sequencing run (van Dijk *et al*, 2014a) and producing this amount of double stranded cDNA may not be possible for all RNA viruses and reverse transcription protocols. When dealing with RNA viruses, the reverse transcription step itself is not immune to the effects of primer bias, although levels are orders of magnitude less than those associated with PCR (Mullan *et al*, 2004). The PCR amplification step during NGS library preparation can also introduce biases into sequencing datasets (van Dijk *et al*, 2014a). This is however not related to primer mismatches, since adapter sequences are identical, but rather to the variable PCR efficiencies during the amplification of AT or GC rich sequences (van Dijk *et al*, 2014a), which can even be lost entirely during library preparation. The potential for bias at this point in data generation should only be considered when sequences might not be close to being GC neutral. With the rapid evolution of technologies for analysing microbial population compositions, sequencing RNA or DNA without the need for any amplification may soon be possible. However, the current focus should be the production of reliable PCR primers able to amplify genetic markers with minimal bias (Ihrmark *et al*, 2012). While most authors acknowledge that the goal of attaining a totally bias free protocol for community characterisation might not be possible, Ibarbalz *et al*, (2014) validates the use of amplicon sequencing by suggesting that while significant biases may exist during amplification, confidence can be placed in the relative abundances of members from samples that have been subject to identical biases during their analyses.

The introduction of Sanger sequencing allowed for an unprecedented transformation in the biological sciences (van Dijk *et al*, 2014b), since it provided researchers with the tools to analyse entire genes and genomes. The technological advancements within the field of next-generation sequencing technologies have yielded an equally, if not more significant transformation in biology due to the sheer

increase in throughput and reduction in per base costs. In this study, the primer directed bias of four CTV-specific primer sets was determined through mass cloning and Sanger sequencing of recovered clones as well as Illumina MiSeq technology. This also allowed for an opportunity to compare the relative performance of NGS technology, against the established method of sequencing clones. In most cases, agreement existed between the ranking of clone sequences and that of the MiSeq datasets, however substantial differences between the percentages of total number clone sequences or MiSeq reads for each sequence type, was observed. This suggests that the practice of sequencing clones is still capable of providing reliable semi-quantitative data relating to sequences present within a PCR amplicon. However, a sufficient number of clones should be sequenced, since sequence types present at very low levels are likely to be overlooked when using this method. An example of significant discrepancy between clone sequence data and Illumina MiSeq datasets from this study was the apparent over-representation of T36 suggested by the clone sequences, for the p23 gene analysis. The Illumina sequence data suggested that all of the sequence types for the p23 gene analysis were roughly equally represented. Due to the significance of the orders of magnitude higher data throughput for the Illumina MiSeq datasets, it can be assumed that they provide a much more accurate quantification of amplicons than that of the sequencing of clones. NGS should therefore be considered as the superior alternative to the isolation and sequencing of clone inserts. Waud *et al*, (2014) suggest that entire communities should be assessed rather than determining the presence or absence of specific members and that the sequencing of a limited number of clone inserts is no longer sufficient for resolution for characterising communities. In the case of CTV, this advocates the use of NGS as a replacement for previously used technologies, which mostly relied upon cloning or the use of multiple molecular markers (MMM) (Hilf and Garnsey, 2000), which is only capable of detecting the presence or absence of a limited number of genotypes. In addition to massively increased throughput, NGS has the advantages of greatly reduced per sample labour input, since there is no need for bacterial cloning and isolation of hundreds of plasmids (van Dijk *et al*, 2014b).

The genotype-specific nature of CTV cross-protection (Folimonova *et al*, 2010) has shown that sustained protection against severe CTV strains in the field can only be achieved through the use of mild-protecting sources of the same genotype

(Folimonova, 2013). This has highlighted a previously unprecedented need for the reliable detection of CTV strains with citrus production areas, where the identification of severe-strains is essential as part of an effective “recipe” for their control through cross-protection (Folimonova, 2013). Primers used for the routine amplification and subsequent estimation of population compositions should be under constant revision and must be compared against available genomes and gene sequences to ensure that primer-directed selection does not cause a significant bias during the PCR amplification of sequences. An example of this is the p33 gene primers from this study, which were designed before the release of the reference sequence CTZA. The p33 forward primer should therefore be revised to exclude the nucleotide mismatch at its 3' terminus and the CTZA primer binding site. This should result in significantly reduced biases against this sequence type. Revisions should be considered as newly released reference sequences become available.

Since its inception in 1983, the PCR reaction has changed the way in which molecular biology and the biological sciences as a whole are conducted. Many variations of the basic PCR reaction have manifested themselves since its inception; however all of these rely on a pair of priming oligonucleotides that target specific sequences of a DNA template. The recent technological advances within the realm of amplicon sequencing has allowed for a massively increased output of data, providing researchers to analyse microbial communities at an unprecedented depth (Waud *et al*, 2014). This in turn has increased the demand on primers to be able to amplify molecular markers in an unbiased fashion, in order to accurately represent the microbial population diversity (Logan *et al*, 2014). Despite this, the majority of primers generally used for microbial community analyses have not been evaluated in terms of their specificity ability to amplify target sequences (Waud *et al*, 2014).

While primer-associated bias during the initial amplification of target sequences, is one of the major contributors to skewed population datasets, a number of other factors related to the PCR reaction can also contribute to bias. While these were not considered in this study, many of these can be simply rectified. Mullan *et al*, (2001) showed that the choice of DNA polymerase used to generate amplicons for community analyses can cause resulting data sets to be biased and unrepresentative. In their study Mullan *et al*, (2001) showed that the use of *Taq* DNA polymerase was probably responsible for increasing the apparent diversity of Hepatitis Virus C quasispecies, through the introduction of spurious nucleotide

incorporations. Furthermore, it was shown that a proofreading enzyme such as *Pwo* polymerase reduced the spurious incorporations by up to three times. Mullan *et al*, (2001) also noted that the editing of primer sites provides another major advantage to using a proofreading DNA polymerase when amplifying targets with heterogeneous primer binding sites. Therefore, after an initial round of amplification, all primer binding should be homogeneous and no longer subject to the biased amplification when using *Taq* based amplification. As mentioned previously, bias can also be introduced during the template amplification of an NGS library preparation, where GC neutral templates will be preferentially amplified. There are no technical reasons why the PCR reactions for the initial production of the amplicons should not suffer from the same problems. Quail *et al*, (2012) compared various proofreading enzymes for their ability to amplify template with variable GC contents and found that the Kapa Hifi enzyme displayed reduced bias compared to the previous gold standard, Phusion. The potential for bias to be introduced in this way should be considered when producing community data sets that are heterogeneous and expected to have variable GC content. However, proofreading enzymes cost roughly five times more than *Taq* DNA polymerase, making their use in large scale population surveys of microbes financially unviable for many laboratories. PCR conditions can be manipulated to improve the fidelity of *Taq* polymerase (Mullan *et al*, 2001).

While the literature cites multiple potential sources of bias, this study has addressed what is probably the most significant of these, which is primer-associated PCR bias. Efforts have been made to assess the bias associated with widely used universal primers for the characterisation of bacterial and fungal communities respectively. Very little data exists with regard to the bias associated with primers targeting viral sequences. This is probably because no truly universal primers exist for viral population characterisation due to the lack of common molecular markers among viruses. Therefore, primers specific to a particular virus are usually used by a limited number of research groups which have not given consideration to the potential influence primer-associated bias can impose in population datasets. This is particularly apparent in the case of plant virology, where factors contributing to artificial biases are rarely considered. This study represents a significant step forward for the improvement of the quantification of components within CTV populations. While the primer pair targeting the p23 gene appears to have less

associated bias than any of the others, the p23 genes location within the 3' half of the genome means that it has inherently less capacity for resolving genotypes when compared with regions located within the variable 5' half of the genome. The primer pair targeting the p33 gene, proposed for routine use in this study, has significantly reduced associated bias, when compared with primers targeting regions closer to the 5' end and the ability to resolve the major CTV genotypes at a phylogenetic level, for use in the reliable assessment of CTV population. This study should also bring the potential effects that unverified primer pairs can have on population datasets, to the attention of plant virologists and to the field of virology as a whole.

REFERENCES

- Anderson I.C., Campbell C.D. and Prosser J.I. (2003). Potential bias of fungal 18S rDNA and internal transcribed spacer polymerase chain reaction primers for estimating fungal biodiversity in soil. *Environmental Microbiology* vol. 5 pp36-47
- Ayllón M.A., Rubio L., Sentandreu V., Moya P., Guerri J. and Moreno P. (2006). Variation in the population of two genes of *Citrus tristeza virus* after aphid transmission or passage to a new host. *Virus Genes* vol.32 pp119-128
- Bar-Joseph M., Gumpf D.J., Dodds J.A., Rosner A. and Ginzberg I. (1985). A Simple Purification Method for Citrus Tristeza Virus and Estimation of its Genome Size. *Phytopathology* vol. 75 pp195-198
- Bru D., Martin-Laurent F. and Philippot L. (2008). Quantification of the Detrimental Effect of a Single Primer-Template Mismatch by Real-Time PCR Using the 16S rRNA Gene as an Example. *Applied Environmental Microbiology* vol. 74 pp1660-1663
- da Graça J.V., Marias L.J. and von Broembsen L.A. (1984). Severe Tristeza Stem Pitting Decline of Young Grapefruit in South Africa. In: *Proc 9th Conf. IOCV*, 62-65 IOCV, Riverside, California
- Engelbrektson A., Kunin V., Wrighton K.C., Zvenigorodsky N., Chen F., Ochman H. and Hugenholtz P. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal* vol. 4 pp642-647
- Farris M.H. and Olsen J.B. (2007). Detection of Actinobacteria cultivated from environmental samples reveals bias in universal primers. *Letters in Applied Microbiology* vol. 45 pp376-381

Folimonova S.Y., Robertson C.J., Shilts T., Folimonov A.S., Hilf M.E., Garnsey S.M. and Dawson W.O. (2010). Infection with Strains of Citrus Tristeza Virus Does Not Exclude Superinfection by Other Strains of the Virus. *Journal of Virology* vol. 84 pp1314–1325

Folimonova S.Y. (2012). Superinfection Exclusion Is an Active Virus-Controlled Function That Requires a Specific Viral Protein. *Journal of Virology* vol. 86 pp5554-5561

Folimonova S.Y. (2013). Developing an understanding of cross-protection by *Citrus tristeza virus*. *Frontiers in Microbiology* vol. 4 art.76

Gal-On A. and Shibolet Y.M. (2005). Cross-protection, pp261-288. In G. Loebenstein and J.P. Carr (ed.), *Natural resistance mechanisms of plants to viruses*. Springer, Dordrecht, The Netherlands

Ghorbel R., López C., Fagoaga C., Moreno P., Navarro L., Flores R. and Peña L. (2001). Transgenic citrus plants expressing the citrus tristeza virus p23 protein exhibit viral-like symptoms. *Molecular Plant Pathology* vol. 2 pp27-36

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* vol. 41 pp95-98

Harper S.J. (2013). Citrus tristeza virus: evolution of complex and varied genotypic groups. *Frontiers in Microbiology* vol. 4 art.93

Hilf M.E. and Garnsey S.M. (2000). Characterization and Classification of Citrus Tristeza Virus Isolates by Amplification of Multiple Molecular Markers. In: *Proc. 14th Conf. IOCV*, 18-27 IOCV, Riverside, California

Hong S., Bunge J., Leslin C., Jeon S. and Epstein S.S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal* vol.3 pp1365-1373

Ibarbalz F.M., Pérez M.V., Figuerola E.L.M. and Erijman L. (2014). The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of Bacterial Community Dynamics. PLOS One vol. 9 e99722

Iglesias N.G., Gago-Zachert S.P., Robledo G., Costa N., Plata M.I., Vera O., Grau, O. and Semorile L.C. (2008). Population structure of *Citrus tristeza virus* from field Argentinean isolates. Virus Genes vol. 36 pp199-207

Ihrmark K., Bödeker I.T.M., Cruz-Martinez K., Friberg H., Kubartova A., Schenck J., Strid Y., Stenlid J., Brandström-Durling M., Clemmensen K.E. and Lindahl B.D. (2012). New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. FEMS Microbiology Ecology vol. 82 pp666-667

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., Garnsey S.M., Lewandowski D.J and Dawson W.O. (1995). 'Complete Sequence of the Citrus Tristeza Virus RNA Genome'. Virology vol. 208 pp511-520

Lee R.F and Keremane M.L. (2013). Mild-strain cross protection of tristeza: a review of research to protect against decline on sour orange in Florida. Frontiers in Microbiology vol. 4 art.259

Lefever S., Pattyn F., Hellermans J. and Vandesompele J. (2013). Single-Nucleotide Polymorphisms and Other Mismatches Reduce Performance of Quantitative PCR Assays. Clinical Chemistry vol. 59 pp1470-1480

Logan G., Freimans G.L., King D.J., Valdazo-González B., Bachanek-Bankowska K., Sanderson N.D., Knowles N.J., King P.J. and Cottam E.M. (2014). A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. BMC Genomics vol. 15 art.828

Mawassi M., Mietkiewska E., Gofman R., Yang G. and Bar-Joseph. (1996). Unusual sequence relationships between two isolates of citrus tristeza virus. *Journal of General Virology* vol. 77 pp2359-2364

Moreno P., Ambros S., Albiach-Marti M.R., Guerri J. and Pena L. (2008). '*Citrus tristeza virus*: A pathogen that changed the course of the citrus industry'. *Molecular Plant Pathology* vol. 9 pp251-268

Mullan B., Kenny-Walsh E., Collins J.K., Shanahan F. and Fanning L.J. (2001). Inferred Hepatitis C Virus Quasispecies Diversity Is Influenced by Choice of DNA Polymerase in Reverse Transcriptase-Polymerase Chain Reactions. *Analytical Biochemistry* vol. 289 pp137-146

Mullan B., Sheehy P., Shanahan F. and Fanning L. (2004). Do *Taq*-generated RT-PCR products from RNA viruses accurately reflect viral genetic heterogeneity?. *Journal of Viral Hepatitis* vol. 11 pp108-114

Quail MA., Otto T.D., Gu Y., Harris S.R., Skelly T.F., McQuillan J.A., Swerdlow H.P. and Oyola S.O. (2012). Optimal enzymes for optimizing sequencing libraries. *Nature Methods* vol 9 pp10-11

Pappu H.R., Pappu S., Niblett C.L., Lee R.F. And Civerolo E.L. (1993). Comparative Sequence Analysis of the Coat Protein of Biologically Distinct Citrus Tristeza Closterovirus Isolates. *Virus Genes* vol. 7 pp255-264

Polz M.F. and Cavanaugh C.M. (1998). Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied Environmental Microbiology* vol.64 pp3724-3730

Read D. and Pietersen G. (2015). Genotypic diversity of Citrus tristeza virus within red grapefruit, in a field trial site in South Africa. *European Journal of Plant Pathology* DOI 10.1007/s10658-015-0631-x

Rubio L., Ayllón 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. *Journal of Virology* vol. 75 pp8054-8062

Sambade A., Lopez C., Rubio L., Flores R., Guerri J. and Moreno P. (2003). Polymorphism of a specific region in gene *p23* of *Citrus tristeza virus* allows for the discrimination between mild and severe isolates. *Archives of Virology* vol. 148 pp2325-2340

Sekiya M.E., Lawrence S.D., McCaffery M. and Cline K. (1991). Molecular cloning and nucleotide sequencing of the coat protein gene of citrus tristeza virus. *Journal of General Virology* vol. 72 pp1013-1020

Sipos R., Székely A.J., Palatinszky M., Révész S., Márialigeti K. and Nikolausz M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiology Ecology* vol. 60 pp341-350

Tamura K., Dudley J., Nei M. & Kumar S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* vol. 24 pp1596-1599.

Urich T., Lanzén A., Qi J., Huson D.H., Schleper C. and Schuster S.C. (2008). Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. *PLOS One* vol. 3 e2527

van Dijk E.L., Jaszczyszyn Y. and Thermes C. (2014a). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* vol. 322 pp12-20

van Dijk E.L., Auger H., Jaszczyszyn Y. and Thermes C. (2014b). Ten years of next-generation sequencing technology. *Trends in Genetics* vol. 30 pp418-426

Waud M., Busschaert P., Ruyters S., Jacquemyn H. and Lievens B. (2014). Impact of primer choice on characterization of orchid mycorrhizal communities using 454 pyrosequencing. *Molecular Ecology Resources* vol. 14 pp679-699

Chapter 4

Diversity of *Citrus tristeza virus* populations in commercial Star Ruby orchards in Southern Africa, using Illumina MiSeq technology.

INTRODUCTION

Citrus tristeza virus (CTV) produces large filamentous flexuous virions with two coat proteins (CP and CPm) (Febres *et al*, 1996) enclosing a positive sense, non-segmented, single stranded RNA genome of ~20kb in size. CTV is the largest known plant virus (Karasev *et al*, 1995), causes one of the most economically devastating disease of citrus (Bar-Joseph *et al*, 1989) and has become endemic to almost all of the major citrus production areas worldwide (Moreno *et al*, 2008). The severity of the disease varies from being asymptomatic to severe stem-pitting, reduced fruit size and decline (Niblett *et al*, 2000). Grapefruit cultivars are among the most sensitive to CTV infections (van Vuuren and Manicom, 2005) with severe stem-pitting being one of the most important symptoms. Strains known to cause severe stem-pitting are endemic to Southern Africa (van Vuuren *et al*, 2000).

CTV strains are introduced into an area through infected budwood and subsequent spread through aphid vector populations. *Toxoptera citricida* is the most efficient vector (Moreno *et al*, 2008) and is also the most important vector species in Southern Africa. It has been shown that areas with endemic populations of *T. citricida* can experience rapid spread of the virus with an up to 95% rate of infection over 2-4 years (Gottwald *et al*, 1996).

In 1972 the South African Citrus Improvement Program (SACIP) was established and one of its functions was to implement mild-strain cross-protection for South African citrus (von Broembsen and Lee, 1988). Fundamentally, cross-protection in plant viruses is the partial or complete resistance of a plant to infection of a severe strain of a virus after the intentional inoculation of a mild strain of the same virus (Gal-On and Shibolet, 2005; Yoon *et al*, 2006). In South Africa candidate mild-strain sources were obtained from trees that were still producing large quantities of high quality fruit, after a minimum of 15 years (van Vuuren *et al*, 1993). Three sources, named GFMS 12, GFMS 27 and GFMS 35 were selected based on their performance in greenhouse and field trials (van Vuuren and Moll 1987; van Vuuren

et al, 1993). The GFMS 12 source was used to pre-immunise all grapefruit cultivars, however in 1993, severe stem-pitting symptoms were observed on a number of mother block budwood trees (van Vuuren and Manicom, 2005). This prompted the approval of GFMS 35 as the interim cross-protection source in 1998 (Luttig *et al*, 2002).

Cross-protection is used in a number of countries, especially where severe stem-pitting strains are endemic (Souza *et al*, 2002). Despite this, the molecular mechanism of cross-protection has remained relatively obscure until recently. A number of mechanisms have been proposed, however the most convincing evidence has been that of the superinfection exclusion principle suggested by Folimonova *et al*, (2010). Superinfection exclusion of CTV has been shown to be strain specific. This has important implications for commercial cross-protection schemes, since pre-immunising populations may only be able to protect against the same strains that they are themselves composed of. It was shown that a CTV specific gene, p33, and its expressed functional protein are fundamental for the functioning of strain-specific exclusion (Folimonova *et al*, 2012; Folimonova, 2013).

A number of studies have been done to characterise the pre-immunising populations of GFMS 12 and 35, as well as a number of their sub-isolates (van Vuuren *et al*, 2000; Luttig *et al*, 2002; Scott *et al*, 2012; Zablocki and Pietersen, 2014; Read and Pietersen, 2015). Relatively little data is available regarding the CTV population compositions of pre-immunised trees in the field. In this study a total of 195 sample collections were performed in the major Star Ruby production areas of Hoedspruit, Malelane, Swaziland, Northern Cape, Sundays River Valley and Nkwalini Valley. Based on the time of planting, all of these trees should have been pre-immunised with the GFMS 12 source. In order to compare populations that were exposed to field conditions with the original pre-immunising sources, three GFMS 12 sources were collected from greenhouses of the Agricultural Research Council – Institute for Tropical and Subtropical Crops (ARC-ITSC) and Citrus Research International (CRI), where they had been maintained under insect-free conditions.

Due to the association of the p33 gene with the mechanism of cross-protection, this gene region was chosen for the analysis of the population composition of the collected samples. PCR amplicons of this gene were subjected direct Sanger sequencing for all samples that were collected. A subset of 95 of these was randomly selected for Illumina MiSeq sequencing and a further subset was chosen

for mass cloning and sequencing of the cloned inserts to compare the results obtained with two techniques. This data indicated that large diversity of CTV populations exists both within and between the various production areas. This data will assist with the targeted selection of mild-strain sources that are able to protect against strains that are circulating in the field.

MATERIALS AND METHODS

Collection of material

Leaf material was collected, usually from one side of each tree, along the rows of the orchards. *Table 1* indicates the name of each collection site, the GPS coordinates, the number of samples collected and the date (month and year) of collection. The exact position of each sampled tree was mapped, using a Trimble Geo (Trimble Navigation, Sunnyvale, CA, United States).

Table 1: Details of the sites from which samples were collected for characterisation of the CTV populations. Each site was named according to its geographical location. The GPS coordinates are the points close to the centre of the orchard that was sampled.

Collection site	GPS coordinates	Number of samples collected	Year that orchard was planted	Date of collection
Hoedspruit 1	24°24'20.15"S 30°49'47.5"E	16	1996	Feb 2013
Hoedspruit 2	24°24'20.15"S 30°49'47.5"E	16	2007	Feb 2013
Malelane 1	25°32'18.84"S 31°26'26.92"E	16	1990	Feb 2013
Malelane 2	25°25'26.69"S 31°40'32.71"E	16	1999	Feb 2013
Swaziland (Mananga)	25°56'28.02"S 31°42'58.24"E	32	1995	Feb 2013
Northern Cape (Kakamas)	28°47'45.18"S 20°40'06.24"E	32	2002	Mar 2013
Sundays River Valley	33°28'29.55"S 25°35'21.28"E	32	1990/2004	Jun 2013
Nkwalini Valley	28°42'17.75"S 31°40'26.19"E	32	1991	Jul 2013

Greenhouse maintained pre-immunising sources were collected from the Agricultural Research Council, Institute for Tropical and Subtropical Crops (ARC-ITSC). The first from a plant harbouring the original Nartia A (GFMS 12) CTV source, and the second, a population of CTV that was collected from the original Nartia A tree, 10 years after the original collected. Both of these were still maintained on the original Nartia A cultivar. A third sample was collected from the Citrus Research International (CRI) facility, which contained the original GFMS 12 pre-immunising source and was maintained on Duncan grapefruit.

Three samples were collected, toward the end of 2014, from biological replicates of Star Ruby trees that had been planted CTV-free (not pre-immunised), in the production area of Kakamas and Letsitele that were planted in 2004 and 2007 respectively.

RNA isolation, reverse transcription and PCR amplification

Each sample was macerated in liquid nitrogen using a mortar and pestle. 100mg of macerated tissue was used per RNA extraction. These extractions were carried out using a GeneJET plant RNA isolation kit (Thermo Scientific, Vilnius, Lithuania). The amplification of the p33 gene of each sample was carried out using a two-step RT-PCR, with the following protocol. Reverse transcription of the p33 gene was primed using 50pmol of the Univ-p33-R reverse primer (5' CCCGTTTAAACAGAGTCAAACGG 3') in conjunction with the following reagents and reaction conditions: 12µl of total RNA extracts, 50pmol of reverse primer, 10U Avianmyeloblastosis virus (AMV) reverse transcriptase (Roche, Mannheim, Germany), 5U RNase inhibitor (Roche, Mannheim, Germany), 1x RT buffer (Roche, Mannheim, Germany) and 0.1 mM dNTP mix (Promega, Madison, WI, United States). Reactions were held at 42°C for 1 hour.

The PCR reactions for the amplification of the p33 gene of each sample were setup as follows: 10µl Promega GoTaq Hot Start Green Master Mix (Promega, Madison, WI, United States), 0.375µl Univ-p33-F forward primer (10µM) (5' GATGTTTGCCTTCGCGAGC 3') (Read and Pietersen, 2015), 0.375µl Univ-p33-R reverse primer (10µM), 2µl of cDNA product, molecular grade water up to a total volume of 20µl. The PCR reactions conditions were as follows: 1 cycle of 92°C for 2min, 35 cycles of 92°C for 30s, 65°C for 45s, 72°C for 1min and 1 cycle of 72°C for 10min. The final PCR products were visualised on an agarose gel.

Direct sequencing of p33 gene amplicons

The amplicons were purified for sequencing by adding 2µl FastAP and 0.5µl ExoI enzymes (Thermo Scientific, Vilnius, Lithuania) to 19µl of the PCR product. These reaction were incubated at 37°C for 15 min followed by a denaturation step of 85°C for 15 min. Sanger sequencing reactions were setup by adding 2µl of the purified PCR product to the following reagents: 1µl BigDye[®] Terminator mix v3.1 (Applied Biosystems, Foster City, CA, United States), 2.25µl 5x BigDye[®] v3.1 sequencing buffer, 0.75µl Univ-p33-F primer (2µM) and molecular grade water to a total volume of 10µl. The sequencing reactions were performed using the following conditions: 1 cycle of 94°C for 1 min., 30 cycles of 94°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes. Sequencing products were purified using ethanol precipitation, according to Sambrook, (2001). The purified sequencing products were submitted to the African Centre for Gene Technologies (ACGT), Automated Sequencing Facility, Department of Genetics, University of Pretoria, South Africa and sequenced using an ABI Prism[®] 3500xl Genetic Analyser (Applied Biosystems, Foster City, CA, United States).

Cloning and sequencing of p33 gene amplicons

A subset of six p33 gene amplicons, were sourced from the set of amplicons used for generating direct sequencing data. These samples had accession numbers 13-3062, 13-3309, 13-3410, 13-3534, 13-3642 and 13-3719, representing a sample each from, Hoedspruit, Malelane, Swaziland, Northern Cape, Sundays River Valley and Nkwalini valley respectively. The purified products produced previously were cloned into the p-GEM T Easy[®] vector (Promega, Madison, WI, United States), followed by the transformation of competent E. coli JM109. Potential recombinant clones were selected using blue/white selection, followed by an alkaline lysis plasmid extraction protocol (Sambrook, 2001). A total of 48 plasmids were isolated for each of the cloning experiments. Sequencing of the cloned inserts was carried out as described previously, with use of the T7 vector specific primer (5' TAATACGACTCACTATAGGG 3'). A total of 36 plasmid inserts were sequenced for 13-3062 and 43, 34, 22, 37 and 21 for 13-3309, 13-3410, 13-3534, 13-3642 and 13-3719 respectively.

Phylogenetic analysis of the direct and clone sequence data

The direct sequencing data was subjected to an initial BLAST (www.ncbi.nlm.nih.gov/blast) analysis (Altschul *et al*, 1990). Only sequences with a minimum PHRED score of 30 and an identity percentage of at least 95 were considered for further phylogenetic analysis. Chromas Lite 2.1 (Technelysium, Brisbane, Australia) was used to edit and correct errors in chromatograms. CLC Main Workbench 5 (CLC Bio, Aarhus, Denmark) was used to produce reverse complements of clone sequences that were in the incorrect orientation. Alignments of sequences were carried using the CLUSTAL W alignment software (EBI, Cambridgeshire, England) within the BioEdit Sequence alignment editor 7.1.3 (Hall, 1999). A total of 44 full-genome reference sequences were accessed from GenBank (www.ncbi.nlm.nih.gov/genbank/) and have the following accession numbers (strain names are in brackets): NC_001661 (T36); AY 340974 (Qaha); U16304 (T36); DQ272579 (Mexico); AY170468 (T36); EU937521 (T36); KC517485 (FS674-T36); KC517486 (FS701-T36); KC517487 (FS703-T36); KC517488 (FS577); JX266713 (Taiwan-Pum/M/T5); AF001623 (SY568); AF260651 (T30); Y18420 (T385); KC517489 (FS701-T30); KC517490 (FL278-T30); KC517491 (FS703-T30); JF957196 (B301); FJ525432 (NZRB-G90); GQ454869 (HA 18-9); FJ525435 (NZRB-M17); JX266712 (Taiwan-Pum/SP/T1); FJ525431 (NZRB-M12); FJ525433 (NZRB-TH28); FJ525434 (NZRB-TH30); JQ798289 (A18); KC525952 (T3); HM573451 (Kpg3); EU857538 (SP); GQ454870 (HA 16-5); DQ151548 (T318A); AB0463981 (NUagA); JQ911664 (CT11A); KC517493 (FL202-VT); U56902 (VT); KC517492 (FS703-VT); EU937519 (VT); KC517494 (FS701-VT); KC262793 (L192GR); JQ911663 (CT14A); FJ525436 (NZ-B18); JQ965169 (T68); EU076703 (B165); JQ061137 (AT-1). An additional three reference sequences from South Africa were used in the alignments, namely CTZA1, CTZA2 and CTZA3 (Zablocki and Pietersen 2014). The cognate region to the p33 gene PCR of the above references were excised and used in further analysis.

Neighbour-joining phylogenetic trees were constructed for each alignment, using MEGA 6 (Tamura *et al*, 2013) and the Maximum Composite Likelihood substitution model with a 1000 bootstrap replicates.

Illumina MiSeq sequencing

Next generation sequencing (NGS), using the Illumina MiSeq platform (Illumina, San Diego, CA, United States), was carried out on each of the purified PCR amplicons that were used for the cloning experiment. Paired-end DNA libraries were prepared using the Nextera V2 sample kit (Epicentre, Madison, WI, United States). The samples were sequenced at the Agricultural Research Council (ARC), Biotechnology Platform, Pretoria, South Africa.

MiSeq data analysis

All trimming and analyses of the Illumina MiSeq datasets was carried out using CLC Genomics workbench 5.5.1 (Aarhus, Denmark). Data was imported as paired-end reads with a distance range of 180-300. Adapter and quality trimming was performed using the default program settings with Nextera V2 transposase adapter sequences (Transposase1:GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG; Transposase2: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG). Quality control was carried out using the Fast QC function. Each samples dataset was then mapped to the p33 PCR cognate region of the same set of reference sequences listed above. Nine groups called “sequence types” were defined and each of these consisted of the following reference sequences. **RB**: F957196 (B301); FJ525432 (NZRB-G90); GQ454869 (HA 18-9); FJ525435 (NZRB-M17); JX266712 (Taiwan-Pum/SP/T1); FJ525431 (NZRB-M12); FJ525433 (NZRB-TH28); FJ525434 (NZRB-TH30). **Kpg3/SP/T3**: HM573451 (Kpg3); EU857538 (SP); KC525952 (T3). **HA 16-5**: GQ454870 (HA 16-5). **VT**: JQ911664 (CT11A); KC517493 (FL202-VT); U56902 (VT); KC517492 (FS703-VT); EU937519 (VT); KC517494 (FS701-VT); KC262793 (L192GR); DQ151548 (T318A); AB0463981 (NUagA); CTZA1, CTZA2 and CTZA3. **AT-1**: JQ061137 (AT-1). **T36**: NC_001661 (T36); AY 340974 (Qaha); U16304 (T36); DQ272579 (Mexico); AY170468 (T36); EU937521 (T36); KC517485 (FS674-T36); KC517486 (FS701-T36); KC517487 (FS703-T36); KC517488 (FS577). **Taiwan-Pum/M/T5**: JX266713 (Taiwan-Pum/M/T5). **T30**: AF001623 (SY568); AF260651 (T30); Y18420 (T385); KC517489 (FS701-T30); KC517490 (FL278-T30); KC517491 (FS703-T30). **B165**: JQ911663 (CT14A); JQ965169 (T68); EU076703 (B165); FJ525436 (NZ-B18).

Table 2: The identities of the direct sequencing data. Each sequence was subjected to a BLAST analysis and only those sequences having a PHRED score of above 30 and an identity percentage of greater than 95% were considered. The sequences are ordered according to their collection sites. The identity given is the reference sequence to which each of the sample sequences showed the highest homology on the GenBank database.

Collection site	Sample	PHRED	Identity (Acc)	Ident (%) / Query length
Hoedspruit 1	13-3001	34	NZRB-TH30 (FJ525434)	98 (801)
	13-3006	37	NZRB-TH30 (FJ525434)	97 (820)
	13-3009	42	NZRB-TH30 (FJ525434)	98 (816)
	13-3011	38	NZRB-TH30 (FJ525434)	98 (834)
	13-3015	33	NZRB-TH30 (FJ525434)	96 (821)
	13-3016	38	AT-1 (JQ061137)	98 (811)
	13-3023	35	NZRB-TH30 (FJ525434)	98 (860)
	13-3040	41	NZRB-TH30 (FJ525434)	99 (840)
	13-3049	36	NZRB-TH30 (FJ525434)	98 (763)
	13-3059	33	NZRB-TH30 (FJ525434)	97 (863)
Hoedspruit 2	13-3062	40	AT-1 (JQ061137)	98 (840)
	13-3067	39	NZRB-TH30 (FJ525434)	95 (833)
	13-3101	40	NZRB-TH30 (FJ525434)	99 (819)
	13-3105	40	Taiwan-Pum/SP/T1 (JX266712.1)	98 (896)
	13-3107	38	NZRB-TH30 (FJ525434)	98 (821)
	13-3113	36	Taiwan-Pum/SP/T1 (JX266712.1)	96 (900)
	13-3114	31	NZRB-TH30 (FJ525434)	96 (677)
	13-3123	42	NZRB-TH30 (FJ525434)	99 (836)
	13-3125	34	Taiwan-Pum/SP/T1 (JX266712.1)	99 (748)
	13-3129	39	NZRB-TH30 (FJ525434)	95 (878)
Malelane 1	13-3152	37	Taiwan-Pum/SP/T1 (JX266712.1)	96 (875)
	13-3155	36	Taiwan-Pum/SP/T1 (JX266712.1)	97 (818)
	13-3160	39	NZRB M12 (FJ525431)	96 (838)
	13-3162	39	NZRB-TH30 (FJ525434)	95 (880)
	13-3166	38	Taiwan-Pum/SP/T1 (JX266712.1)	95 (881)
	13-3171	39	Taiwan-Pum/SP/T1 (JX266712.1)	96 (841)
	13-3205	36	Kpg 3 (HM573451)	97 (870)
	13-3222	39	SP (EU857538)	97 (874)
	13-3230	37	Kpg 3 (HM573451)	97 (839)
	13-3238	38	T3 KC525952	96 (841)
Malelane 2	13-3242	38	Taiwan-Pum/SP/T1 (JX266712.1)	97 (844)
	13-3251	37	NZRB M12 (FJ525431)	97 (840)
	13-3255	41	T3 (KC525952)	98 (841)
	13-3256	37	Kpg 3 (HM573451)	97 (841)
	13-3261	36	Kpg 3 (HM573451)	98 (839)
	13-3267	37	Kpg 3 (HM573451)	95 (907)
	13-3271	34	T3 (KC525952)	97 (838)
	13-3312	36	NZRB-TH30 (FJ525434)	95 (833)
	13-3320	36	AT-1 (JQ061137)	96 (801)
	13-3321	37	NZRB-TH30 (FJ525434)	97 (822)
Swaziland (Mananga)	13-3322	44	NZRB-TH30 (FJ525434)	99 (839)
	13-3324	40	Kpg 3 (HM573451)	97 (841)
	13-3340	40	T3 KC525952	97 (842)
	13-3343	38	Kpg 3 (HM573451)	97 (821)
	13-3345	33	NZRB-TH30 (FJ525434)	96 (864)
	13-3347	34	SP	96 (812)
	13-3357	43	AT-1 (JQ061137)	98 (905)
	13-3358	39	Kpg 3 (HM573451)	96 (833)
	13-3360	40	Kpg 3 (HM573451)	97 (842)
	13-3362	33	AT-1 (JQ061137)	96 (797)
Swaziland (Mananga)	13-3406	38	T3 KC525952	96 (837)
	13-3408	39	T3 KC525952	96 (841)
	13-3413	38	Kpg 3 (HM573451)	97 (836)
	13-3414	40	Kpg 3 (HM573451)	98 (842)
	13-3417	36	SP	97 (822)
	13-3418	37	Kpg 3 (HM573451)	96 (804)
	13-3425	38	Kpg 3 (HM573451)	97 (833)
	13-3427	35	FS-701 T36 (KC517486)	96 (882)
	13-3429	41	Kpg 3 (HM573451)	97 (905)
	13-3431	37	T3 KC525952	95 (840)
	13-3434	40	Kpg 3 (HM573451)	96 (907)
	13-3436	34	Taiwan-Pum/SP/T1 (JX266712.1)	96 (843)
	13-3439	39	NZRB-TH30 (FJ525434)	96 (825)
	13-3445	36	Taiwan-Pum/SP/T1 (JX266712.1)	95 (849)
	13-3447	38	T3 (KC525952)	96 (838)
	13-3449	36	Kpg 3 (HM573451)	96 (839)
	13-3450	40	Kpg 3 (HM573451)	97 (838)
	13-3451	41	T3 KC525952	96 (909)
	13-3453	36	Taiwan-Pum/SP/T1 (JX266712.1)	95 (821)
	13-3454	39	SP	97 (838)

Northern Cape (Kakamas)	13-3501	38	NZRB-TH30 (FJ525434)	97 (840)
	13-3504	37	NZRB-TH30 (FJ525434)	98 (836)
	13-3508	36	NZRB-TH30 (FJ525434)	98 (833)
	13-3512	34	NZRB-TH30 (FJ525434)	95 (820)
	13-3513	50	NZRB-TH30 (FJ525434)	99 (843)
	13-3516	39	NZRB-TH30 (FJ525434)	99 (868)
	13-3520	34	NZRB-TH30 (FJ525434)	95 (822)
	13-3521	37	NZRB-TH30 (FJ525434)	98 (851)
	13-3522	36	NZRB-TH30 (FJ525434)	98 (849)
	13-3523	38	NZRB-TH30 (FJ525434)	99 (844)
	13-3525	40	NZRB-TH30 (FJ525434)	99 (855)
	13-3526	41	NZRB-TH30 (FJ525434)	99 (843)
	13-3539	39	NZRB-TH30 (FJ525434)	99 (844)
	13-3545	36	NZRB-TH30 (FJ525434)	98 (837)
	13-3549	43	NZRB-TH30 (FJ525434)	99 (845)
	13-3553	34	NZRB-TH30 (FJ525434)	97 (825)
	13-3556	41	Taiwan-Pum/SP/T1 (JX266712.1)	98 (839)
	13-3558	38	NZRB-TH30 (FJ525434)	95 (817)
	13-3567	38	NZRB-TH30 (FJ525434)	98 (842)
	13-3568	33	Taiwan-Pum/SP/T1 (JX266712.1)	95 (870)
Sundays River Valley	13-3570	39	NZRB-TH30 (FJ525434)	95 (839)
	13-3604	37	Kpg 3 (HM573451)	96 (837)
	13-3611	33	NZRB-TH30 (FJ525434)	96 (748)
	13-3613	38	Taiwan-Pum/SP/T1 (JX266712.1)	98 (837)
	13-3618	35	NZRB-TH30 (FJ525434)	97 (842)
	13-3619	41	Kpg 3 (HM573451)	97 (905)
	13-3620	36	NZRB-TH30 (FJ525434)	98 (844)
	13-3621	38	NZRB-TH30 (FJ525434)	98 (836)
	13-3622	42	T3 (KC525952)	97 (843)
	13-3624	42	NZRB-TH30 (FJ525434)	99 (844)
	13-3626	36	Kpg 3 (HM573451)	96 (812)
	13-3636	32	NZRB-TH30 (FJ525434)	95 (815)
	13-3637	32	Taiwan-Pum/SP/T1 (JX266712.1)	94 (779)
	13-3639	38	NZRB-TH30 (FJ525434)	99 (824)
	13-3643	38	NZRB-TH30	97 (834)
	13-3647	36	Taiwan-Pum/SP/T1 (JX266712.1)	99 (815)
	13-3648	40	NZRB-TH30 (FJ525434)	98 (842)
	13-3651	37	NZRB-TH30 (FJ525434)	99 (809)
	13-3667	54	NZRB-TH30 (FJ525434)	99 (844)
	13-3668	44	NZRB-TH30 (FJ525434)	99 (845)
Nkwadini Valley	13-3672	41	NZRB-TH30 (FJ525434)	99 (839)
	13-3700	36	Kpg 3 (HM573451)	97 (868)
	13-3703	30	SP (EU857538)	96 (638)
	13-3705	37	T3 (KC525952)	97 (839)
	13-3706	38	T3 KC525952	97 (840)
	13-3714	38	Kpg 3 (HM573451)	97 (838)
	13-3715	35	Kpg 3 (HM573451)	96 (802)
	13-3733	36	Kpg 3 (HM573451)	97 (833)
	13-3734	33	T3 (KC525952)	95 (864)
	13-3743	37	T3 (KC525952)	96 (839)
	13-3745	40	Kpg 3 (HM573451)	97 (836)
	13-3751	37	Kpg 3 (HM573451)	97 (834)
	13-3757	36	Kpg 3 (HM573451)	95 (838)
	13-3758	38	Kpg 3 (HM573451)	98 (839)
	13-3761	35	Kpg 3 (HM573451)	96 (834)
	13-3768	34	Taiwan-Pum/SP/T1 (JX266712.1)	95 (842)
	13-3762	33	Kpg 3 (HM573451)	96 (836)
	13-3770	37	Taiwan-Pum/SP/T1 (JX266712.1)	99 (811)
	13-3775	33	Taiwan-Pum/SP/T1 (JX266712.1)	95 (786)

Threshold to account for index leaching

The 16S RNA sequence of “Candidatus” *Liberibacter africanus*, corresponding to the GenBank sequence EU921620, was amplified using the OA1 forward (5’ GCGCGTATTTTATACGAGCGGCA 3’) and the OI2c reverse primer (5’ ATGGGTTGCGAAGTCGCGAGGC 3’) (Teixeira *et al*, 2005), with the following PCR

reaction conditions 1 cycle of 92°C for 2min, 40 cycles of 92°C for 30s, 65°C for 45s, 72°C for 1.5min and 1 cycle of 72°C for 10min. The amplicon was gel purified, quantified and prepared for Illumina sequencing, as described previously and sequenced on the same flow cell in parallel with the CTV field survey samples.

RESULTS

Direct sequencing of p33 gene amplicons

The identities of the direct sequences, according to an initial BLAST analysis, are listed in *Table 2*. A number of discrepancies exist between the data from the BLAST results and the neighbour-joining phylogenetic trees. While a number of direct sequences conformed to the minimum quality parameters, they were truncated and could therefore not be included in further phylogenetic analyses. The identities of the direct sequences discussed below were derived from the neighbour-joining phylogenetic trees (Appendix E). The maximum likelihood phylogenetic tree of Hoedspruit 1 samples showed that 2 samples grouped with AT-1 and 10 samples with NZRB-TH30. The Hoedspruit 2 site yielded 10 sequences grouping with the NZRB group and 3 sequences grouping with the Taiwan-Pum/SP/T1 group. The Malelane 1 data showed 1 sample grouping with NZRB-TH30 and Taiwan-Pum/SP/T1 respectively and 10 samples grouping with the SP/Kpg3/T3 branch. The Kpg3/SP/T3 strains cannot be resolved when performing phylogenetic analyses on the p33 gene and any direct/cloning sequencing data falling within this branch will be referred to as SP/Kpg3/T3-like. The Malelane 2 site had 6 samples grouping within the SP/Kpg3/T3 branch, 5 with NZRB-TH30 and 3 with AT-1. The phylogenetic tree of the Swaziland samples showed one sequence grouping away from any of the references (13-3427), with the closest branch being T36. Two of the samples grouped closely to NZRB-TH30, 3 with Taiwan-Pum/SP/T1 and 15 within the SP/Kpg3/T3 branch. All 21 of the Northern Cape (Kakamas) samples grouped with NZRB-TH30. The Sundays River Valley samples showed one sample forming a unique branch in the phylogenetic tree, with high bootstrap support. The closest reference branch was the NZRB group. Within the NZRB group 2 of the sequences grouped with Taiwan-Pum/SP/T1 and 13 with NZRB-TH30. The remaining four sequences grouped within the SP/Kpg3/T3 branch. The Nkwadini Valley sequences showed one of the samples (13-3747) grouping away from any references. This

unique branch represents that of 13-3427, suggesting that these samples may represent a previously undescribed strain/genotype. Of the remainder of the Nkwalini Valley samples, 3 grouped with Taiwan-Pum/SP/T1 and 13 clustered within the SP/Kpg3/T3 branch. The remainder of the direct sequences were discarded due to the fact that they did not conform to the minimum PHRED score of 30 and a sequence identity of 95%.

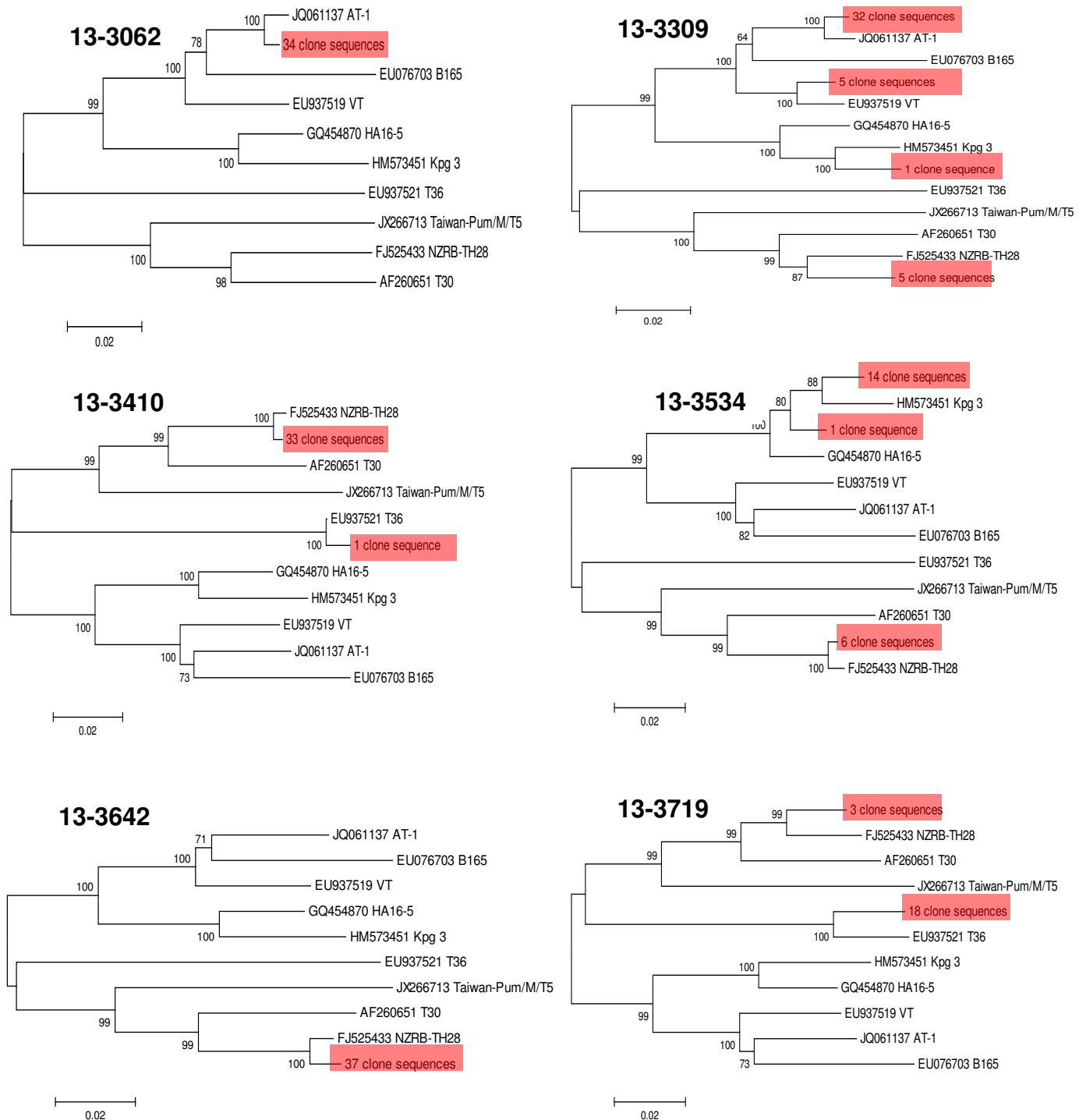


Figure 1: CTV diversity, based on cloned sequence data for samples 13-3062 (Hoedspruit), 13-3309 (Malelane), 13-3410 (Swaziland), 13-3534 (Northern Cape), 13-3642 (Sundays River Valley) and 13-3719 (Nkwalini Valley). Each dendrogram was collated by using a single representative reference sequence from each clade and collapsing the branches representing the clone sequences down to a single branch. The number of clones that were associated with each particular branch is presented. Figures 1-6 in Appendix C illustrate the non-collapsed phylogenetic trees for each of these samples.

Clone sequence data

The CTV diversity, according to clone sequence data, is represented with the six dendrograms in *Figure 1*. The non-collapsed phylogenetic trees for the cloned sequences for each of these samples are illustrated by *Figures 1-6* in Appendix C. Sample 13-3062 from Hoedspruit appeared homogenous for an AT-1-like strain, with the 34 clones grouping most closely to this reference.

Sample 13-3309 from Malelane showed apparently high levels of diversity, with 5 clones falling into a unique branch close to the NZRB references, 1 clone within the Kpg3/SP/T3-like group, while a further 5 clones were found to group in a unique branch close to the VT group. Thirty clones grouped within the AT-1 reference branch with another 2 clones falling within a unique branch close to AT-1. The majority of clones (31) from the 13-3410 sample from Swaziland grouped within the NZRB-TH30 branch, 3 grouped with Taiwan-Pum/SP/T1 (forms part of the RB TH-28 group in the collapsed tree) and 1 clone fell within the T36 branch. The sample from the Northern Cape (13-3534) yielded 14 clones within the Kpg3/SP/T3 group, 1 within the HA 16-5 group and 1 clones, forming a unique group between Taiwan-Pum/M/T5 and T30 and finally 5 clones grouping with the RB group. Sample 13-3642, from Sundays River Valley indicated a mixed infection with 7 clones grouping with Taiwan-Pum/SP/T1 and 30 clones grouping within the NZRB group. Sample 13-3719 also indicated a mixed infection with Taiwan-Pum/SP/T1 and the remainder (18) of the samples fell within a unique group closest to the T36 branch. These clones could represent a previously undescribed strain or genotype.

Illumina MiSeq sequencing data

The Illumina MiSeq data mapped to reference sequence is represented in *Table 3*. The RB sequence type was the most highly represented and was present in the dataset of every sample that was analysed at an incidence of between 1.6 and 100% of mapped reads. Reads of the Kpg3/SP/T3 sequence type was the second most numerous and was found to be present in all datasets except for 8 samples from each of Hoedspruit 2 and Northern Cape and 5 samples from Sundays River Valley. Kpg3/SP/T3 reads were represented at between 0.1% and 93% of CTV mapped reads in samples, and never occurred on its own. HA 16-5 was detected at between 0.1 - 45.7% of total CTV mapped reads and was present in all collection sites except Hoedspruit 2. VT was detected in samples from all collection sites except Malelane 1

and represented low percentages of the total mapped reads with the exception of one of the Swaziland samples, where VT was represented by 47.7% of mapped reads. The remainder of the sequence types appeared occasionally within a few samples, throughout many of the collection sites. AT-1, T36, Taiwan-Pum/M/T5 and T30 were present at incidences of 0.1 – 81%, 0.2 – 45.6%, 0.1 – 2.3% and 0.1 – 0.4% of total mapped reads per sample, respectively. Illustrative analyses of these populations are shown in *Figures 1-8* in Appendix D, while the data relating to the mapping of reads for each sample is shown in *Tables 1.1-8.15* in “Supplementary data” Section A (supplied electronically).

Table 6 compares and serves to determine the agreement that exists between the determination of the dominant sequence type in population that were sequenced directly with Sanger sequencing, as well as Illumina MiSeq sequencing. According to this comparison, 25 out of 52 samples showed agreement between the two sets of data, when the dominant sequence type was represented at levels greater than 90% of the population, according to the Illumina MiSeq data. A total of 21 (40%) populations showed agreement between the direct and Illumina sequencing, when the dominant sequence type, according to the Illumina dataset, was present at less than 90% of the population. A lack of agreement between the two sets of sequence data existed for six of the populations.

The sequencing data of the greenhouse maintained original GFMS 12 sources (*Table 4*) showed that the levels of various CTV components varied significantly between these three populations. The original Nartia A (GFMS 12) source (14-6000) maintained at the ARC-ITSC was dominant for VT-like reads, at 81%. AT-1 reads were presented at a level of 12.3%, while the remainder of the components, namely RB, A18 and CT14A were represented at levels below 10%. The population that was collected 10 years after the GFMS 12 population was collected, but from the same source (14-6001), also maintained at ARC-ITSC, was dominant for RB-like components at 97.2% of mapped reads. The remainder of the components, namely Kpg3/SP/T3, VT, AT-1 and CT14A were minor components and represented at levels below 2% of mapped reads. The original GFMS 12 source maintained at the CRI facility showed an almost equal representation of VT and AT-1-like components at 45.8% and 48.7% of mapped reads respectively. CT14A was represented as a minor component at ~5% of total mapped reads.

The Illumina MiSeq datasets that were generated for the non-pre-immunised Star Ruby trees planted at a trial site in Letsitele are represented in *Table 5*. A total of seven different sequence types were represented across all three samples. 14-6003 was dominant for VT-like components at 81% of total mapped reads. AT-1 was represented at 16.3% and CT14A, RB and A18 were minor components and were represented at levels below 2%. 14-6004 was dominant for RB at 99.6% of mapped reads with AT-1 being a minor component at 0.4%. All seven components were represented in 14-6005. RB-like components were also dominant in this sample at 81.5%. AT-1 was represented at 11.1% of total mapped reads. The remainder were minor components. A comparison between the numbers of reads mapping to a reference sequence and the consensus length of assembly was done for each Illumina MiSeq dataset. These comparisons are illustrated by *Figures 1.1-9.3* in “Supplementary data” Section B (supplied electronically).

Threshold to account for index leaching

The total number of reads within the datasets of each sample mapping back to the “*Candidatus Liberibacter africanus*” 16S RNA reference sequence (EU921620) varied between 0% and 0.1%. The maximum levels of representation, 0.1% of the total mapped reads, was taken as a conservative threshold value whereby all reads map at, and below levels of 0.1% were considered to have potentially originated from index leaching and were eliminated from the dataset.

Table 3: Representation of Illumina MiSeq data produced for the field collected samples. Each set of samples is grouped according to the site at which they were collected. Boxes in the sequence type column containing a † symbol denotes the sequence type associated with the direct sequencing results. A “c” following on from a sequence name shows that sample was subjected to a cloning experiment with the numbers of clones grouping with particular sequence types denoted in brackets. Sample names marked with an ^ symbol were associated with severe, possibly CTV, symptoms in the field. The sequence types represent groups of reference sequences, to which the reads mapped. The logic of grouping the reference sequences in this way, are described in the Materials and Methods section.

Collection site	Age of orchard at time of collection (years)	Sample number	Total number of reads mapping to refs	Percentage of total reads mapping to sequence types							
				RB	Kpg 3/ SP/ T3	HA 16-5	VT	AT-1	T36	Taiwan-Pum/M/T5	T30
Hoedspruit 1	18	13-3023	92768	99.6†	0.15	-	-	-	-	-	-
		13-3067^	174009	98†	0.1	-	0.57	0.2	-	0.86	-
		13-3040	12421	95.5†	0.1	-	2.9	1.1	-	-	-
		13-3042^	73495	92.4	0.4	-	3.9	0.4	-	2.3	-
		13-3015	181749	91.2†	7.5	0.4	-	-	-	0.1	-
		13-3059	133533	73.3†	0.1	-	25.7	-	-	0.2	-
		13-3013	57627	10.9	78.8	6.5	0.2		3.1	0.3	
		13-3062^c	327014	2.7	-	-	14.5	82.2† (34)	-	-	-
Hoedspruit 2	8	13-3113	144570	100†	-	-	-	-	-	-	-
		13-3160	57211	99.9†	-	-	-	-	-	-	-
		13-3152	109041	99.9†	-	-	-	-	-	-	-
		13-3161	58759	99.8†	-	-	-	-	-	-	-
		13-3105	190405	99.7†	-	-	-	-	-	-	-
		13-3155	53385	99.6†	-	-	-	-	-	-	-
		13-3125	25583	97.9†	-	-	2	-	-	-	-
		13-3130	143871	34	62.1†	-	3.8	-	-	-	-
Malelane 1	24	13-3251	69030	90.5†	9	0.4	-	-	-	-	-
		13-3242^	104646	90.2†	8.5	1.2	-	-	-	-	-
		13-3271	74353	48.3	23.6†	27.8	-	-	-	-	-
		13-3257^	19511	35.8	58.7	5.3	-	-	-	-	-
		13-3260^	152311	11.4	85.1	3.4	-	-	-	-	-
		13-3267^	149687	9.4	86.2†	4.3	-	-	-	-	-
		13-3205^	80987	8.9	87.2†	3.8	-	-	-	-	-
		13-3255^	318511	6.4	76.4†	17	-	-	-	-	-
Malelane 2	14	13-3321	129010	99.7†	0.1	-	-	-	-	-	-
		13-3312	158354	99.1†	0.1	-	-	-	-	-	-
		13-3305	39568	86.2†	12	0.7	0.3	0.6	-	-	-
		13-3362^	78835	22.5	4.8	0.7	23.2	48.4†	-	-	0.1
		13-3324	87079	22.8	23.8†	1.9	20.8	30.5	-	-	-
		13-3309^c	70784	12.6 (5)	71.3 (1)	4.1	4.9 (5)	7.1 (32)	-	-	-
		13-3343	125140	2.7	92.6†	4.7	-	-	-	-	-
		13-3347^	284891	1.6	92.7	5.4	-	-	-	-	-

Swaziland (Mananga)	18	13-3410	51624	95.1 (34)	† 2.2	0.2	-	-	2.3 (1)	-	-
		13-3445	62214	88.7†	8.6	0.3	2.2	-	-	-	-
		13-3436^	210778	81.2†	16.9	1	0.5	0.1	-	-	-
		13-3416	105078	80.8	12.6	1.4	4.9	-	-	-	-
		13-3421	50038	73.2	15.7	0.8	0.5	-	9.5	-	-
		13-3427	57164	66.9	2.3	-	-	-	30.1 †	0.1	-
		13-3439^	258754	61.5†	35.4	2.8	-	-	-	-	-
		13-3447^	95957	32.1	29.8 †	37. 9	-	-	-	-	-
		13-3422^	258480	24.2	62.6	12. 8	-	-	-	0.1	-
		13-3406^	103216	10.7	74.7 †	9.9	0.2	-	4.2	-	-
		13-3418^	541918	8.3	89.6 †	2	-	-	-	-	-
		13-3425	86853	8	81.2 †	10. 1	-	-	-	-	-
		13-3454	35808	3.5	71.8 †	24. 7	-	-	-	-	-
		13-3417^	231254	2.2	87.7 †	9.9	-	-	-	-	-
		13-3449^	92851	1.8	93.1 †	5	-	-	-	-	-
Northern Cape (Kakamas)	12	13-3405^	107866	1.5	44.8	5.7	47.7	-	-	-	-
		13-3553	28218	99.9†	-	-	-	-	-	-	-
		13-3545	352081	99.9	-	-	-	-	-	-	-
		13-3512	212758	99.7	-	-	-	-	-	-	-
Collection site		Sample number	Total number of reads mapping to refs	RB	Kpg 3/ SP/ T3	HA 16- 5	VT	AT-1	T36	Taiw an- Pum/ M/T5	T30
Northern Cape (Kakamas) (cond.)		13-3520	105466	99.6†	-	-	-	-	-	0.2	-
		13-3570	81865	98.7†	0.4	0.2	0.2	-	0.2	-	-
		13-3536	103234	97.9	-	-	1.6	-	-	0.2	-
		13-3552	48994	97.8	-	-	-	-	-	-	-
		13-3508	109759	94.5†	-	-	4.6	-	-	0.5	-
		13-3526^	45601	88†	8.8	1.6	1.4	-	-	-	-
		13-3539	55325	79.9†	18.6	0.9	0.4	-	-	-	-
		13-3523	61645	75.3†	20.4	3.5	0.7	-	-	-	-
Sundays River Valley	9	13-3534c	32842	29.2 (6)	66.3 (14)	3.7 (1)	0.7	-	-	-	-
		13-3651	120438	99.5†	0.1	-	-	-	-	0.2	-
		13-3674	117329	98.5	-	-	-	-	-	1.3	-
		13-3655	106127	94.7	-	-	0.9	3.9	-	0.2	-
		13-3658	149419	57.9	35.3	6.4	-	-	-	0.2	-
	23	13-3661	329326	2.3	-	-	19.1	78.3	-	-	-
		13-3621	127293	99.5†	-	-	-	-	-	0.2	-
		13-3637	196489	98.9†	-	-	0.1	-	-	0.3	0.4
		13-3643	26209	97.8†	1.3	0.5	0.3	-	-	-	-
		13-3642c^	34903	97.4 (37)	-	-	1.8	-	-	-	-
		13-3608	67631	97.1	0.3	-	2	-	-	-	-
		13-3612	52122	89.1	5.6	3.2	1.2	0.8	-	-	-

		13-3647^	57568	78.4†	12.2	8.1	1.1	-	-	-	-
		13-3627	180682	63.3†	28.5	5.9	1.9	-	-	-	-
		13-3619	110849	27.8	68.8	3.1	-	-	-	-	-
		13-3626	87694	8.3	91.1	0.4	-	-	-	-	-
		13-3604	110345	4.1	79.4	16.5	-	-	-	-	-
Nkwalini Valley	23	13-3768	66898	93.3†	6.1	0.5	-	-	-	-	-
		13-3776^	56981	86.6	8.8	4.2	-	-	-	-	-
		13-3770^	98996	83.2†	13.4	1.4	1.6	-	-	-	-
		13-3747	87318	76.8	3.2	0.1	-	-	19.3	-	-
		13-3719c^	37751	51.4 (30)	2.7	0.2	-	-	45.6 (18)	-	-
		13-3743	80819	26.8	32.3	40.7	-	-	-	-	-
		13-3705^	109482	24.5	29.4	45.7	0.3	-	-	-	-
		13-3714^	136211	20.7	79.2	-	-	-	-	-	-
		13-3745	93095	20	74.5	5.5	-	-	-	-	-
		13-3732	348076	19.2	77.3	3.3	-	-	-	-	-
		13-3713	271667	18	50.7	31.1	-	-	-	-	-
		13-3748^	290491	16.9	79.1	3.9	-	-	-	-	-
		13-3729	345812	9.2	65.7	24.9	-	-	-	-	-
		13-3708	213313	7.9	87.4	4.2	-	-	0.4	-	-
		13-3727	292114	4.3	93	2.5	-	-	-	-	-

Table 4: Representation of Illumina MiSeq data produced for the greenhouse maintained original GFMS 12 population samples. The sequence types represent groups of reference sequences, to which the reads mapped. The logic of grouping the reference sequences in this way, are described in the Materials and Methods section.

Sample number	Cultivar	Total number of reads mapping to refs	RB	Kpg3/SP/ T3	VT	AT-1	CT14A
14-6000	Nartia	18690	0.8	0.1	80	12.3	6.4
14-6001	Nartia	116779	97.2	0.1	0.7	1.5	0.1
14-6002	Duncan	143372	-	-	45.8	48.7	5.11

Table 5: Representation of Illumina MiSeq data produced for the samples that were collected from non-preimmunised Star Ruby trees in the Letsitele production area. The sequence types represent groups of reference sequences, to which the reads mapped. The logic of grouping the reference sequences in this way, are described in the Materials and Methods section.

Collection site	Age of orchard at time of collection (years)	Sample number	Total number of reads mapping to refs	RB	Kpg3/SP/ T3	HA 16-5	VT	AT-1	A18	CT14A
Letsitele	10	14-6003	278055	0.3	-	-	81	16.3	0.2	1.7
		14-6004	378037	99.6	-	-	-	0.4	-	-
		14-6005	451572	81.5	3	0.1	0.4	11.1	0.4	0.9

Table 6: Comparison between population data from samples for which the p33 gene amplicon was directly sequenced, using Sanger sequencing, as well as by Illumina MiSeq sequencing. The plain text represents cases where agreement between the direct sequencing and Illumina sequencing data exists. Text in bold represents cases where a direct sequencing and Illumina data corresponded to a dominant sequence type that was present at less than 90% of the apparent population. The text in italics represents disagreement between the direct and Illumina sequence data, in terms of the dominant sequence type present within the population.

Sample	PHRED	Identity (Acc)	Dominant genotype according to Illumina MiSeq data
13-3015	33	NZRB-TH30 (FJ525434)	91.2% RB
13-3023	35	NZRB-TH30 (FJ525434)	99.6% RB
13-3040	41	NZRB-TH30 (FJ525434)	95.5% RB
13-3059	33	NZRB-TH30 (FJ525434)	73.3% RB
13-3062	40	AT-1 (JQ061137)	82.2% AT-1
13-3067	39	NZRB-TH30 (FJ525434)	98% RB
13-3105	40	Taiwan-Pum/SP/T1 (JX266712.1)	99.7% RB
13-3113	36	Taiwan-Pum/SP/T1 (JX266712.1)	100% RB
13-3125	34	Taiwan-Pum/SP/T1 (JX266712.1)	97.9% RB
13-3152	37	Taiwan-Pum/SP/T1 (JX266712.1)	99.9% RB
13-3155	36	Taiwan-Pum/SP/T1 (JX266712.1)	99.6% RB
13-3160	39	NZRB M12 (FJ525431)	99.9% RB
13-3205	36	Kpg 3 (HM573451)	87.2% Kpg3/SP/T3
13-3242	38	Taiwan-Pum/SP/T1 (JX266712.1)	90.2% RB
13-3251	37	NZRB M12 (FJ525431)	90.5% RB
13-3255	41	T3 (KC525952)	76.4 % Kpg3/SP/T3
13-3267	37	Kpg 3 (HM573451)	86.2% Kpg3/SP/T3
13-3271	34	T3 (KC525952)	48.3% RB
13-3312	36	NZRB-TH30 (FJ525434)	99.1% RB
13-3321	37	NZRB-TH30 (FJ525434)	99.7% RB
13-3324	40	Kpg 3 (HM573451)	30.5% AT-1
13-3343	38	Kpg 3 (HM573451)	92.6% Kpg3/SP/T3
13-3347	34	SP	92.7% Kpg3/SP/T3
13-3362	33	AT-1 (JQ061137)	48.4% AT-1
13-3406	38	T3 KC525952	74.7% Kpg3/SP/T3
13-3417	36	SP	87.7% Kpg3/SP/T3
13-3418	37	Kpg 3 (HM573451)	89.6% Kpg3/SP/T3
13-3425	38	Kpg 3 (HM573451)	81.2% Kpg3/SP/T3
13-3427	35	FS-701 T36 (KC517486)	66.9% RB
13-3436	34	Taiwan-Pum/SP/T1 (JX266712.1)	81.2% RB
13-3439	39	NZRB-TH30 (FJ525434)	61.5% RB
13-3445	36	Taiwan-Pum/SP/T1 (JX266712.1)	88.7% RB
13-3447	38	T3 (KC525952)	32.1% RB
13-3449	36	Kpg 3 (HM573451)	93.1% Kpg3/SP/T3
13-3454	39	SP	71.8% Kpg3/SP/T3
13-3508	36	NZRB-TH30 (FJ525434)	94.5% RB
13-3520	34	NZRB-TH30 (FJ525434)	99.6% RB
13-3523	38	NZRB-TH30 (FJ525434)	75.3% RB
13-3526	41	NZRB-TH30 (FJ525434)	88% RB
13-3539	39	NZRB-TH30 (FJ525434)	79.9% RB
13-3570	39	NZRB-TH30 (FJ525434)	98.7% RB
13-3604	37	Kpg 3 (HM573451)	79.4% Kpg3/SP/T3
13-3619	41	Kpg 3 (HM573451)	68.8% Kpg3/SP/T3
13-3621	38	NZRB-TH30 (FJ525434)	99.5% RB
13-3626	36	Kpg 3 (HM573451)	91.1% RB
13-3643	38	NZRB-TH30	97.8% RB
13-3651	37	NZRB-TH30 (FJ525434)	99.5% RB
13-3705	37	T3 (KC525952)	45.7% HA16-5
13-3714	38	Kpg 3 (HM573451)	79.2% Kpg3/SP/T3
13-3743	37	T3 (KC525952)	40.7% HA 16-5
13-3768	34	Taiwan-Pum/SP/T1 (JX266712.1)	93.3% RB
13-3770	37	Taiwan-Pum/SP/T1 (JX266712.1)	83.2% RB

Table 7: Comparison between population data from samples for which the p33 gene of six selected amplicons were cloned, as well as by Illumina MiSeq sequencing. The values that are not enclosed within brackets denote the number of clone sequences corresponding to their respective genotypes for each cloned population. The values within the round brackets express the numbers of clones in terms of percentages. The values within the square brackets represent the percentages of Illumina MiSeq reads (Table 3), mapping to each of the genotypes of their corresponding populations.

Collection site	Sample number	Total number of clones isolated	RB	Kpg 3/ SP/ T3 (%)	HA 16-5 (%)	VT (%)	AT-1 (%)	T36 (%)
Hoedspruit 1	13-3062	34	[3]			[15]	34 (100) [82]	
Malelane 2	13-3309	43	5 (12) [13]	1 (2) [71]	[4]	5 (12) [5]	32 (74) [7]	
Swaziland (Mananga)	13-3410	35	34 (97) [95]	[2]			1 (3)	[2]
Northern Cape (Kakamas)	13-3534	21	6 (29) [29]	14 (67) [66]	1 (5) [4]	[1]		
Sundays River Valley	13-3642	37	37 (100) [98]			[2]		
Nkwalini Valley	13-3719	21	3 (14) [51]	[3]				18 (86) [46]

DISCUSSION

The identities of 192 CTV populations, collected from the major Star Ruby grapefruit production areas throughout Southern Africa, were characterised using the established techniques of direct sequencing and the sequencing of multiple clones, and the more recently developed NGS technology. This study has produced one of the largest sets of data pertaining to the diversity of CTV strains in Southern Africa, which should prove very useful in the improvement of CTV cross-protection, which should begin to start targeting specific strains, as more knowledge of diversity becomes available. The main technical advantage provided by this study was the establishment a protocol for characterising CTV population based on NGS technologies, which is a direct replacement of the established direct Sanger sequencing and the sequencing of multiple clones.

Amplicons were generated using the primer pair targeting the p33 gene. All 192 amplicons were characterised using direct Sanger sequencing, while a subset of six of these, each representing a particular production area, was characterised by the sequencing of multiple clone inserts. Direct Sanger sequencing of PCR amplicons generated from an RNA virus quasispecies, such as CTV, is expected to provide the identity of the dominant component of the population (Pawlotsky *et al*, 1998). CTV population are often composed of various disparate strains and direct sequencing data containing consistently ambiguous base calls is usually evidence that a population is made up of variable CTV components (Fontana *et al*, 2014). In this study, direct sequences not adhering to stringent criteria such as PHRED scores were assumed to be a result of populations consisting of more than one detectible sequence type and only the identities of those appearing to contain a seemingly single, clearly dominant sequence are reported. A total of 35 of the direct sequences out of the total of 192 did not conform with the quality parameters, which translates to ~18%.

As shown in the comparison between the two datasets, a total of 52 samples had direct sequence data that conformed to the quality thresholds as well as corresponding Illumina MiSeq datasets (*Table 6*). Agreement between the datasets, in terms of the dominant sequence type, was observed for 46 (88%) of the populations however, in 27 of the populations, the dominant sequence type, made up less than 90% of the population, according to the MiSeq data. The two datasets

showed a lack of agreement for the remaining six populations in terms of the dominant sequence type. In the most extreme case, direct Sanger sequencing detected a sequence type that was represented by only 23.6% of the MiSeq reads and was not the dominant sequence type in the population. Significant components of the population can be overlooked when using direct Sanger sequencing alone. In this study, 40% of direct sequences were of a high quality, which is suggestive of a population that comprises of a clearly dominant single genotype. In addition to sometimes missing critical components of populations, direct sequencing may also misidentify the dominant sequence type in populations that are heterogeneous. Direct sequencing of CTV amplicons should therefore be reserved for the pre-screening of samples prior to more in depth analyses. Due to the unreliability associated with direct sequencing, Mullan *et al*, (2001) advises against using direct sequencing for the characterisation of viral populations.

Prior to the onset of next generation sequencing techniques, the use of cloning and Sanger sequencing of cloned inserts was one of the best methods of determining the composition of heterogeneous viral populations. Due to the labour intensiveness of this approach, a relatively small number of clones were generally used for determining the population composition of plant virus populations (Kim *et al*, 2005; Ge *et al*, 2007; Acosta-Leal *et al*, 2008; Iglesias *et al*, 2008). When compared with the Illumina sequencing data (*Table 7*), the clone sequence data for three out of the six populations (13-3410, 13-3534 and 13-3642) appeared to accurately represent the true composition of the amplified products with as few as 35, 21 and 37 clones respectively. The clone sequence data for 13-3062 suggests a homogenous population of AT-1, however Illumina MiSeq data confirms that, while the population is dominant for AT-1, the clone sequence data did not detect the major component of VT and minor component of RB. The clone sequence data of 13-3719 was capable of detecting both major components, however the levels of these components were significantly different when compared with the Illumina MiSeq datasets. One of the populations, namely 13-3309, showed drastically differing clone and Illumina MiSeq data. While predominant sequence type (Kpg3/SP/T3) was detected thorough cloning (2%) the MiSeq dataset suggested that this sequence type was present at levels above 70%. Read and Pietersen, (2015) showed that selecting and sequencing about 100 clones with as little as 5 variants in a population, was incapable of accurately representing the true levels of each variant

within the population. For this reason, we strongly advocate the use of next generation sequencing technologies for the accurate representation of viral populations.

The majority of studies performed on plant viruses, using NGS, have relied upon a metagenomics approach where small RNAs or dsRNAs are used as templates (Kreuze *et al*, 2009; Wu *et al*, 2010; Giampetruzzi *et al*, 2012; Loconsole *et al*, 2012; Barba *et al*, 2014). Ruiz-Ruiz *et al*, (2011) used Illumina sequence data to profile CTV-derived small RNAs in a number of different citrus hosts. The results of the study by Ruiz-Ruiz *et al*, (2011) shows that the use of small RNAs as a template would probably not be suitable for large scale surveys of CTV populations, since the relative abundance of CTV-derived small RNAs varies between 3.5% and 50% depending on the levels of viral titre among different citrus hosts. Furthermore, it was shown that while whole CTV genome can be reconstructed from small RNAs, mapping of their sequences show an asymmetric distribution focused on the 2500nt at the 3' end of the genome. Therefore the use of small RNAs in the characterisation of CTV populations could result in a loss of coverage to other more abundant small RNAs, if viral titre is low and secondly the asymmetrical accumulation of small RNAs to the 3' end of the genome could result in low coverage of more phylogenetically informative regions. The use of randomly primed dsRNAs, while useful for metagenomic analyses of plant virus populations, also suffers from variable abundances of reads that map to CTV specific references. Zablocki and Pietersen, (2014), showed that between 2.65% and 22.76% of total Illumina reads, derived from different populations from the same cultivar, mapped to CTV-specific references. In a study on another tentative member of the *Closteroviridae* family, *Little cherry virus 1*, by Candresse *et al*, (2013) only 34% of 454 reads, derived from a dsRNA template, mapped back to the viral references with none of the remaining reads appearing to be of viral origin.

The approach of the current study was to implement a NGS protocol that could be used to replace the laborious clone based methods for characterising CTV populations. The concept of using NGS as direct replacement for cloning and Sanger sequencing has already been put forward by Diekstra *et al*, (2015). Metagenomics approaches are widely used for the characterisation of virus populations, since viruses lack universal conserved sequences that can be used for characterisation, such as in other microbes (Roossinck, 2015). When targeting a

specific virus however, the use of amplicon sequencing, comparable to the approaches used for bacterial diversity analyses (Petrosino *et al*, 2009), becomes favourable. The replacement of the Sanger sequencing of multiple clones with NGS-based amplicon sequencing does not negate the need for gene-specific PCR amplification and this practice comes with the risk of introducing primer-associated PCR bias, discussed in *Chapter 3*. However, recommendations for the design of CTV-specific primers, put forward in *Chapter 3*, should reduce the bias associated with the production of amplicons. Protocols used for the large scale survey of populations from multiple samples should yield reproducible datasets that are comparable with one another. The number of reads and coverage of specific viral sequences may be reduced when using small RNAs or dsRNA as template, due to the potential presence of other infecting or persistent viruses (Roossinck, 2015).

An array of techniques has been used to prepare samples containing RNA viruses for Illumina sequencing. These include, the isolation and sequencing of small RNAs (Giampetruzzi *et al*, 2012), or the isolation of total RNA (Adams *et al*, 2009) or double stranded RNA (Coetzee *et al*, 2010; Zablocki and Pietersen, 2014), followed by random, oligo dT or gene specific amplification. While sequencing the whole-genomes of CTV populations would provide the greatest amount of phylogenetic information, the reverse transcription and amplification would require specialised reagents, long amplification times and greater per sample costs for the generation of sequencing data. Whole-genome amplification of CTV using a single primer pair directed at the 5' and 3' terminal nucleotides was attempted and was not achieved (data not shown). For these reasons, the amplification and sequencing of a specific gene was selected, specifically the p33 gene, which has been shown to be essential in the maintenance of strain-specific cross-protection (Folimonova *et al*, 2010).

High levels of CTV diversity were observed between orchards, as well as between different trees within the same orchard. As little as three of the groups were represented at Hoedspruit 2 and Malelane 1, which overall strain diversity increased to a maximum of 8 of the groups being represented at Sundays River Valley. The remainder of the collection sites showed overall strain diversity at levels between these extremes, with 7 groups in Hoedspruit 1 and Swaziland, 6 in Malelane 1 and Northern Cape and 5 in Nkwalini Valley. No definite correlation can be made between the age of the orchards and their overall strain diversity, since both the youngest, Hoedspruit 2 at 7 years and the oldest, Malelane 1 at 23 years, showed

the lowest overall strain diversity. The majority of populations were shown to comprise of a single dominant group, sometimes with several minor sequence types. 76 out of the total 92 samples all had components that comprised 70% or greater of the total population. The more heterogeneous populations were also associated with older trees.

The remainder of the groups were represented sporadically and at variable levels, in populations from numerous collection sites. When present in a population, levels of VT was for the most part negligible, except for five populations from 4 different regions, which had levels of VT greater than 10%. AT-1 was detected in 12 samples from Hoedspruit 1, Malelane 2 and Sundays River Valley. Most of these populations comprised of negligible percentages of AT-1, however 4 populations showed levels of AT-1 of between 30.5 and 82.2%. The incidence of T36 was low with only three populations exceeding levels of 10%, with one origination from Swaziland (30.1%) and two from Nkwalini Valley (19.3% and 45.6%).

The level of potential index leaching was determined by running a single sample, corresponding to the 16S RNA sequence of “Candidatus” *Liberibacter africanus*, together with the other samples from the CTV survey, on the same flow cell. Index leaching results in the false assignment of a sequencing read from one sample, to the dataset of another sample (Kircher *et al*, 2012). The cause of index leaching could be due to sequencing errors incorporated into the index sequence or through index contamination during manufacturing or downstream handling of the indexes themselves. In this experiment, the maximum percentage 16S RNA reads present in any of the CTV survey datasets was 0.1% of the total number of reads. This was therefore a conservative threshold and any reads present at levels of 0.1% or lower may have leached from a different sample and were disregarded within the dataset.

A number of metrics are provided for the output data from the analyses of NGS sequencing reads by CLC Genomics Workbench. The metric known as the “consensus length” provides information on the percentage of the reference sequence to which specific sequencing reads mapped. This metric therefore provides valuable information regarding the confidence that can be put into calling the presence of particular strains/sequence types that have low numbers of mapped reads and truncated consensus lengths. Comparisons between the numbers of Illumina MiSeq read and the consensus length was done for each Illumina MiSeq dataset. While these comparisons might be useful for determining the general

reliability of mapped NGS sequencing reads, variability between the datasets made deriving trends exceedingly difficult. For example, RB TH-28 appears to be one of the most numerous strains within these populations, according to the reference mappings. However, the consensus lengths for this particular reference varied greatly and might be truncated in cases where large numbers of reads have mapped to the references (> 35 000 reads) and full-lengths in cases where small numbers of reads have mapped (<1 000 reads). Pirovano *et al*, (2015) shows that while reference sequence-guided assembly is the most efficient method for analysing virus population data generated by NGS, reads will fail to assemble correctly without a comprehensive reference sequence set that covers all of the possible viral sequences present in a population. References to which large numbers of reads mapped but have a small fraction of their length covered may indicate the presence of novel sequences, not present among the references (Pirovano *et al*, 2015). Consensus sequences produced during the *de novo* assembly of sequence reads could be used as reference sequences to assemble potential novel sequence reads during reference guided assemblies. Trends that could be inferred for these comparisons were the consistently truncated consensus lengths (350-550bp) for the reference sequence Taiwan-Pum/M/T5, as well as a number of VT-like references (450-700bp) and CT14A (400-500bp). This indicates the possible presence of recombinants or novel sequences, which were unable to map to the available reference set, however this is also purely speculative. In the case of CTV, whole genome sequencing data can then be produced through resequencing of randomly amplified products from total RNA or dsRNA extracts (Zablocki and Pietersen, 2014) from populations suspected of containing novel viral variants.

A total of six Star Ruby samples were collected from a trial site in the Kakamas production area, planted in 2004 that bordered the commercial orchard that was sampled in the Northern Cape. None of these samples tested positive for the presence of CTV, with conventional PCR amplification of the p33 gene (data not shown). Due to the rapid spread of the virus in the presence of vectors (Gottwald *et al*, 1996), the presence of uninfected trees that have been exposed to field conditions for 11 years, suggests that there is little influence from aphid inoculations and that the populations of the pre-immunised plants is probably still very similar to the original pre-immunising source. General vector control and the extreme climate of this production area could play a role in the apparent low levels of CTV spread.

Since the majority of the populations from the Northern Cape location were found to contain mostly RB-like sequences, this provides further evidence to suggest that the original pre-immunising source was comprised mainly of similar components. The Illumina MiSeq data of the original GFMS 12 sources, on Nartia and Duncan grapefruit plants, suggests that this pre-immunising population dominant for VT and AT-1 with RB present at exceedingly low levels in Nartia only. The detection of high levels of Kpg3/SP/T3-like components in a small number of populations, collected from the trees within close proximity to the virus free trees in the Northern Cape were probably the result of secondary inoculations by aphid populations, that might have been present in very low numbers, during the first few years that the block was planted.

The p33 gene amplification of a further six samples from non-pre-immunised field trial plants; planted in 2007, in the Letsitele production area suggested that three out of the six samples were negative for the presence of CTV. The amplicons of the three samples that did test positive for the presence of CTV were submitted for Illumina MiSeq sequencing and are considered to be examples of populations that have resulted due to the inoculation from aphid vectors. These populations vary significantly from one another, however; are similar to the populations from pre-immunised plants, in that they all have a dominant component represented at over 80% of the total reads. Two of the populations were dominant for RB-like components, which, is expected since these components are probably present at high levels among pre-immunised populations.

This study has confirmed the complexity of CTV populations, in terms of overall strain diversity between production sites as well as the variability between populations within the same orchard. In the case of CTV, very little data is available, which links genotypes to the phenotypic symptoms that they are potentially capable of inducing. This has been further complicated by the fact that populations of CTV are usually made up of mixed populations of disparate strains, within a single host plant (Harper, 2013). The existence of these disparate strains within single populations also results in the generation of multiple recombinants. While the use of amplicon sequencing of discrete portions of the CTV genome has been advocated in this study, care should be taken when drawing conclusions from this data and symptom expression for example. Whole genome data of individual components of

CTV populations should also provide more concise evidence of recombination events.

The vast majority of molecular characterisations of CTV have been done on populations of the virus, as opposed to the individual components making up these populations. While the study of CTV in the form of populations is essential for determining their role in disease development when strains are mixed, researchers of CTV should take a step back and begin to characterise the individual components of endemic populations, so that better estimations can be made regarding their potential for disease severity. The paucity of data linking genotype and phenotype can probably be attributed to the difficulty associated with isolating individual strains from complex populations. For CTV this is most commonly achieved through single aphid transmissions, which are laborious and time consuming, with transmission rates in the range of 0-55% (Broadbent *et al*, 1996) although are generally toward the lower end of this range (van Vuuren *et al*, 2000).

While the general health and symptoms of each tree that was sampled for this study was recorded (data not shown); limited conclusions can be made regarding the relationship between these symptoms and the components of the populations. Additional disease and physiological factors can contribute to the apparent symptoms in the field. Further work should include the determination of the reaction of a selection of indicator hosts (Garnsey *et al*, 2005) to populations that are collected from the field. Following this, individual components of these populations should be isolated, biologically characterised according to the phenotypes elicited and sequenced at the whole genome level. The pathogenic capacity of sequence types can also be determined through the production of infectious clones. The infectious clone based on T36 (Satyanarayana *et al*, 1999) has also proven to be invaluable as a reverse genetics system for determining the function and pathogenic capacity of individual genes (Folimonova *et al*, 2010; Dawson *et al*, 2013). It has become essential to consider the strain-specific mechanism of cross-protection, suggested by Folimonova *et al*, (2012), which will assist with determining which of the components are responsible for the elicitation of disease symptoms and will allow for the targeting of these components within cross-protection schemes.

REFERENCES

- Acosta-Leal R., Fawley M.W. and Rush C.M. (2008). Changes in the intra-isolate genetic structure of Beet necrotic yellow vein virus populations associated with plant resistance breakdown. *Virology* vol. 376 pp60-68
- Adams I.P., Glover R.H., Monger W.A., Mumford R., Jackeviciene E., Navalinskiene M., Samuitene M. and Boonham N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology* vol. 10 pp537-545
- Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990). "Basic local alignment search tool." *Journal of Molecular Biology* vol. 215 pp403-410
- Barba M., Czosnek H. and Hadidi A. (2014). Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* vol.6 pp106-136
- Bar-Joseph M., Marcus R. and Lee R.F. (1989). The continuous challenge of Citrus Tristeza Virus Control. *Annual Review of Phytopathology* vol. 27 pp291-316
- Broadbent P., Bransky R.H. and Indsto J. (1996). Biological Characterization of Australian Isolates of Citrus Tristeza Virus and Separation of Subisolates by Single Aphid transmissions. *Plant Disease* vol. 80 pp329-333
- Candresse T., Marais A., Fauere C. and Gentit P. (2013). Association of *Little cherry virus 1* (LChV1) with the Shirofugen Stunt Disease and Characterization of the Genome of a Divergent LChV1 Isolate. *Phytopathology* vol. 103 pp293-298
- Coetzee B., Freeborough M.J., Hans J., Maree H.J., Celton J.M., Rees D.J.G. and Burger J.T. (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology* vol. 400 pp157-163

Dawson W.O., Garnsey S.M., Tatineni S., Folimonova S.Y., Harper S.J. and Gowda S. (2013). Citrus tristeza virus-host interactions. *Frontiers in Microbiology* vol. 4 art.88

Diekstra A., Bosgoed E., Rikken A., van Lier B., Kamsteeg E.J., Tychon M., Derks R.C., van Soest R.A., Mensenkamp A.R., Scheffer H., Neveling K. and Nelen M.R. (2015). Translating Sanger-Based Routine DNA Diagnostics into Generic Massive Parallel Ion Semiconductor Sequencing. *Clinical Chemistry* vol. 61 pp154-162

Febres V.J., Mawassi M., Frank A., Bar-Joseph M., Manjunath K.L., Lee R.F. and Niblett C.L. (1996). The p27 Protein Is Present at One End of Citrus Tristeza Virus Particles. *Phytopathology* vol. 86 pp1331-1335

Folimonova S.Y., Robertson C.J., Shilts T., Folimonov A.S., Hilf M.E., Garnsey S.M. and Dawson W.O. (2010). Infection with Strains of *Citrus Tristeza Virus* Does Not Exclude Superinfection by Other Strains of the Virus. *Journal of Virology* vol. 84 pp1314–1325

Folimonova S.Y. (2012). Superinfection Exclusion Is an Active Virus-Controlled Function That Requires a Specific Viral Protein. *Journal of Virology* vol. 86 pp5554-5561

Folimonova S.Y. (2013). Developing an understanding of cross-protection by *Citrus tristeza virus*. *Frontiers in Microbiology* vol. 4 art.76

Fontana A., Debreczeni D.E., Albanese G., Davino S., Flores R. and Rubio L. (2014). Evolutionary analysis of *Citrus tristeza virus* outbreaks in Calabria, Italy: two rapidly spreading and independent introductions of mild and severe isolates. *European Journal of Plant Pathology* vol.140 pp607-613

Gal-On A. and Shibolet Y.M. (2005). Cross-protection, pp261-288. In G. Loebenstein and J.P. Carr (ed.), *Natural resistance mechanisms of plants to viruses*. Springer, Dordrecht, The Netherlands

Garnsey S.M., Civerolo E.L., Gumpf D.J., Paul C., Hilf M.E., Lee R.F., Brlansky R.H., Yokomi R.K. and Hartung J.S. (2005). Biological Characterization of an International Collection of Citrus tristeza virus (CTV) Isolates. In: Proc 16th Conf. IOCV, 75-93, IOCV, Riverside, California, United States

Ge L., Zhang J., Zhou X. and Li H. (2007). Genetic Structure and Population Variability of Tomato Yellow Leaf Curl China Virus. Journal of Virology vol. 81 pp5902-5907.

Giampetruzzi A., Roumi V., Roberto R., Malossini U., Yoshikawa N., La Notte P., Terlizzi F., Credi R. and Saldarelli P. (2012). A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. Virus Research vol. 163 pp262-268.

Gottwald T.R., Garnsey S.M., Cambra M., Moreno P., Irey M. and Borbón J. (1996). Differential Effects of *Toxoptera citricida* vs. *Aphis gossypii* on Temporal Increase and Spatial Patterns of Spread of Citrus Tristeza. . In: Proc. 13th Conf. IOCV, 120-129 IOCV, Riverside, California

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series vol.41 pp95-98

Harper S.J. (2013). Citrus tristeza virus: evolution of complex and varied genotypic groups. Frontiers in Microbiology vol. 4 art.93

Iglesias N.G., Gago-Zachert S.P., Robledo G., Costa N., Plata M.I., Vera O., Grau O. and Semorile L.C. (2008). Population structure of *Citrus tristeza virus* from field Argentinean isolates. Virus Genes vol. 36 pp199-207

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., Garnsey S.M., Lewandowski D.J. and Dawson W.O. (1995). 'Complete Sequence of the Citrus Tristeza Virus RNA Genome'. Virology vol. 208 pp511-520

Kim T., Youn M. Y., Min B. E., Choi S. H., Kim M. and Ryu K. H. (2005). Molecular analysis of quasispecies of Kyuri green mottle mosaic virus. *Virus Research* vol. 110 pp161-167

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

Kreuze J.F., Perez A., Untiveros M., Quispe D., Fuentes S., Barker I. and Simon R. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* vol.388 pp1-7

Loconsole G., Saldarelli P., Doddapaneni H., Savino V., Martelli G.P. and Saponari M. (2012). Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* vol. 432 pp162–172

Luttig, M., van Vuuren, S.P., and van der Vyver, J.B. (2002). Differentiation of single aphid cultured sub-isolates of two South African *Citrus tristeza virus* isolates from grapefruit by single-strand conformation polymorphism. In: *Proc 15th Conf. IOCV*, 186-196, IOCV, Riverside, California

Moreno P., Ambrós S., Albiach-Marti M.R., Guerri J. and Peña L. (2008). ‘*Citrus tristeza virus*: A pathogen that changed the course of the citrus industry’. *Molecular Plant Pathology* vol. 9 pp251-268

Mullan B., Kenny-Walsh E., Collins J.K., Shanahan F. and Fanning L.J. (2001). Inferred Hepatitis C Virus Quasispecies Diversity Is Influenced by Choice of DNA Polymerase in Reverse Transcriptase-Polymerase Chain Reactions. *Analytical Biochemistry* vol. 289 pp137-146

Niblett C.L., Genc H., Cevik B., Halbert S., Brown L., Nolasco G., Bonacalza B., Manjunath K.L., Febres V.J., Pappu H.R. and Lee R.F. (2000). Progress on strain differentiation of *Citrus tristeza virus* and its application to the epidemiology of citrus tristeza disease. *Virus Research* vol. 71 pp97-106

Pawlotsky J.M., Germanidis G., Neumann A.U., Pellerin M., Frainais P.O. and Dhumeaux D. (1998). Interferon Resistance of Hepatitis C Virus Genotype 1b: Relationship to Nonstructural 5A Gene Quasispecies Mutations. *Journal of Virology* vol. 72 pp2795-2805

Petrosino J.F., Highlander S., Luna R.A., Gibbs R.A. and Versalovic J. (2009). Metagenomic Pyrosequencing and Microbial Identification. *Clinical Chemistry* vol. 55 pp.856-866

Pirovano W., Miozzo L., Boetzer M. and Pantaleo V. (2015). Bioinformatics approaches for viral metagenomics in plants using short RNAs: model case of study and application to *Cicer arietinum* population. *Frontiers in Microbiology* vol. 5 article.790

Read D. and Pietersen G. (2015). Genotypic diversity of Citrus tristeza virus within red grapefruit, in a field trial site in South Africa. *European Journal of Plant Pathology* DOI 10.1007/s10658-015-0631-x

Roossinck M.J. (2015). Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. *Frontiers in Microbiology* vol 5 article.767

Ruiz-Ruiz S., Navarro B., Gisel A., Peña L., Navarro L., Moreno P., Di Serio F. and Flores R. (2011). Citrus tristeza virus infection induces the accumulation of viral small RNAs (21–24-nt) mapping preferentially at the 3'-terminal region of the genomic RNA and affects the host small RNA profile. *Plant Molecular Biology* vol. 75 pp607-619

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

Satyanarayana T., Gowda S., Boyko V.P., Albiach-Marti M.R., Mawassi M., Navas-Castillo J., Karasev A.V., Dolja V., Hilf M.E., Lewandowski D.J., Moreno P., Bar-Joseph M., Garnsey S.M. and Dawson W.O. (1999). An engineered closterovirus RNA replicon and analysis of heterologous terminal sequences for replication. *Proceedings of the National Academy of Sciences USA* vol. 96 pp7433-7438

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* vol. 158 pp27-37

Souza A.A., Müller G.W., Targon M.L.P.N., Takita M.A. and Machado M.A. (2002). Stability of the Mild Protective 'PIAC' Isolate of *Citrus tristeza virus*. In: *Proc. 15th Conf. IOCV*, 131-135. IOCV, Riverside, CA

Tamura K., Stecher G., Peterson D., Filipski A. and Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* vol. 30 pp2725-2729.

Teixeira D.C., Lopes S.A., Yamamoto P.T., Eveillard S., Martins E.C., de Jesus Junior W.C., Bassanezi R.B., Ayres A.J., Danet J.L., Saillard C. and Bové J.M. (2005). PCR Detection of the Two Liberibacter Species Associated with Citrus Huanglongbing in São Paulo State, Brazil. In: *Proc. 16th Conf. IOCV*, 432-438. IOCV, Riverside, CA

van Vuuren S.P and Moll J.N. (1987) Glasshouse evaluation of citrus tristeza virus isolates. *Phytophylactica* vol.19 pp219–221

van Vuuren S.P. and Collins R.P. (1993). Evaluation of Citrus Tristeza Virus Isolates for Cross Protection of Grapefruit in South Africa. *Plant Disease* vol. 77 pp24-28

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. In: *Proc. 14th Conf. IOCV*, 103-110 IOCV, Riverside, California

van Vuuren S.P. and Manicom B.Q. (2005). The Response of Star Ruby Grapefruit to Different *Citrus tristeza virus* isolates. In: *Proc 16th Conf. IOCV*, 112-116 IOCV, Riverside, California

von Broembsen L. A, and Lee A.T.C. (1988). South Africa's Citrus Improvement Program. In: *Proc. 10th Conf. IOCV*, 407-416 IOCV, Riverside, California

Wu Q., Luo Y., Lu R., Lau N., Lai E.C., Li W.X. and Ding S.W. (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proceedings of the National Academy of Sciences* vol. 107 pp.1606-1611

Yoon J.Y., Ahn H.I., Kim M., Tsuda S. and Ryu K.I. (2006). '*Pepper mild mottle virus* pathogenicity determinants and cross protection effect of attenuated mutants in pepper'. *Virus Research* vol. 118 pp23-30

Zablocki O. and Pietersen G. (2014). Characterization of a novel citrus tristeza virus genotypewithin three cross-protecting source GFMS12 sub-isolatesin South Africa by means of Illumina sequencing. *Archives of Virology* vol. 158 pp2133-2139

Chapter 5

Concluding Remarks

Citrus tristeza virus (CTV) is considered to be the most severe viral pathogen affecting species of *Citrus* (Moreno *et al*, 2008) and is a continual hindrance to commercial production (Folimonova, 2013). An epidemic of *Phytophthora* led to the exclusive use of sour orange as a rootstock in the Mediterranean and Americas. However, this practice provided the ideal host for decline inducing strains of CTV and an entirely different epidemic took place, with the virus killing millions of trees of a period of 50 years (Moreno *et al*, 2008). While the virus still significantly contributes to reducing production in almost all citrus producing countries around the world, epidemics of tree deaths are prevented through the use of resistant rootstocks as well as rigorous cross-protection programs in countries where severe stem-pitting strains are endemic (Souza *et al*, 2002).

In plant virology, cross-protection is described as the partial or complete resistance inferred to a plant, by a mild-strain of a virus, against a severe strain of the same virus (Gal-On and Shibolet 2005; Yoon *et al*, 2006). While many hypotheses have been proposed for the mechanism of CTV cross-protection, the most substantial research was made by Folimonova, (2013), which showed that CTV cross-protection relies on a CTV specific protein that is expressed from the p33 gene. Furthermore, this mechanism functions in a strain-dependant manner, meaning that strains of the virus are only capable of preventing secondary inoculations of the same strain. This has highlighted the need for determining the CTV diversity within production areas as well as accurately characterising CTV cross-protection sources that will be capable of preventing secondary inoculations of severe strains in the field.

The characterisation of CTV populations on sub-isolates has been done on multiple gene regions such as the high diverse 5' half (Rubio *et al*, 2001) and the more conserved 3' half of the genome (Sambade *et al*, 2003; Iglesias *et al*, 2008). This coupled with a lack of agreement as to what defines a strain of CTV, often led to ambiguous population data that was not comparable. Harper, (2013) provided the most pragmatic approach for defining CTV strains, which should be used in conjunction with any population study. The p33 gene primers, proposed by Read and

Pietersen, (2015) and used in the survey of Star Ruby CTV diversity in this study, target the region apparently responsible for cross-protection. CTV usually exists as a number of disparate strains that often replicate concurrently within a single population (Harper, 2013), which gives rise to numerous, stable recombinants. These recombinants can only be detected through the generation of whole genome sequences or through the sequencing of multiple gene regions. However, this production of these data is time consuming and costly. Due to the importance of the p33 gene in cross-protection, it should be used in large scale surveys of CTV populations, where the sequencing of only a single gene region is feasible.

The characterisation of populations of microorganisms usually requires the amplification of targets that contain varying degrees of variability, which often incidentally exists within the target primer binding sites (Farris and Olsen, 2007). Mismatches between primers and their binding sites have the capacity to introduce primer-associated bias, which results in the preferential amplification of templates whose primer binding sites show the greatest complementarity to the primers. Mismatches that are located toward the critical 3' end of the primers are considered to have the greatest impact on the introduction of primer-associated bias (Dieffenbach *et al*, 1993). Many of the primers targeting CTV specific gene regions were designed when few reference sequences were available. Part of this study involved determining the degree of bias associated with four sets of CTV specific primers. Significant degrees of bias were found to be associated with the primer pairs targeting the A and F regions, toward the 5' end of the genome. The primer pair targeting the conserved p23 gene was found to have very little associated bias, while the primers targeting the p33 gene (used for the survey of CTV diversity in this study), located near to the midpoint of the genome, were found to be significantly biased toward only two genotypes. The removal of the 3' terminal nucleotide from the p33 gene forward primer should greatly reduce the apparent bias toward the CTZA sequence type in future studies. While reduced bias is observed with primers targeting regions in the 3' half of the genome, these regions have limited capabilities for resolving the identities of CTV genotypes, due to sequence conservation within this half of the genome. Therefore it is necessary to characterise strains and populations according to regions in the 5' half of the genome, which are more phylogenetically informative.

The mainstream availability of NGS sequencing platforms has transformed molecular biology (van Dijk *et al*, 2014) and has significantly influenced the way in which plant viruses and their populations are characterised (Barba *et al*, 2014). A number of templates can be isolated and used to characterise plant RNA viruses, which include total RNA (Monger *et al*, 2010), double-stranded RNA (dsRNA) (Candresse *et al*, 2013) and small RNAs (Ruiz-Ruiz *et al*, 2011). However, the use of total RNA and dsRNA can lead to variable and unpredictable abundances of reads that map back to the virus sequences of interest (Candresse *et al*, 2013; Zablocki and Pietersen, 2014), while Ruiz-Ruiz *et al*, (2011) showed that small RNA sequence data of CTV shows an asymmetrical assembly of reads toward the 3' end of the genome. While the use of these templates are useful for virus metagenomics, the reproducibility of existing cloning of amplicons and sequencing inserts is favourable for the characterisation of populations of CTV. The massively increased throughput afforded by NGS technologies has made it an attractive replacement for cloning based characterisations, a concept already put forward by Diekstra *et al*, (2015).

The major aim of the second part of this study was to establish a replacement for the Sanger sequencing of clones, based on Illumina MiSeq technology. Samples were collected from 8 sites in all of the major Star Ruby grapefruit production areas in South Africa, as well as a site in Swaziland. The sites were named Hoedspruit 1, Hoedspruit 2, Malelane 1, Malelane 2, Swaziland (Mananga), Northern Cape (Kakamas), Sundays River Valley and Nkwalini Valley. Between 16 and 32 samples were collected from each site. The primers designed for the amplification of the p33 gene were used to produce amplicons of this gene for each of the collected samples. A variety of techniques were used to characterise the CTV populations of each sample. A total of 192 samples were sequenced directly. 6 samples were selected for cloning, which resulted in a total of 218 sequenced clones. A total of 96 samples were subjected to Illumina MiSeq sequencing. This study has yielded one of the largest single surveys of CTV to date and probably one of the most in depth diversity studies on a plant virus, making use of a Next Generation Sequencing platform and has provided unprecedented baseline knowledge of CTV diversity in Southern Africa. However, few conclusions regarding the pathogenicity of Southern African CTV strains, based on this data alone. Future studies would greatly benefit from the isolation of individual strains from populations and characterising them on multiple

indicator hosts (Garnsey *et al*, 2005), as well as their whole genome characterisation.

REFERENCES

- Barba M., Czosnek H. and Hadidi A. (2014). Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* vol.6 pp106-136
- Candresse T., Marais A., Fauere C. and Gentit P. (2013). Association of *Little cherry virus 1* (LChV1) with the Shirofugen Stunt Disease and Characterization of the Genome of a Divergent LChV1 Isolate. *Phytopathology* vol. 103 pp293-298
- Dieffenbach C.W., Lowe T.M. and Dveksler G.S. (1993). General concepts for PCR primer design. *Genome Research* vol. 3 ppS30-37
- Diekstra A., Bosgoed E., Rikken A., van Lier B., Kamsteeg E.J., Tychon M., Derks R.C., van Soest R.A., Mensenkamp A.R., Scheffer H., Neveling K. and Nelen M.R. (2015). Translating Sanger-Based Routine DNA Diagnostics into Generic Massive Parallel Ion Semiconductor Sequencing. *Clinical Chemistry* vol. 61 pp154-162
- Farris M.H. and Olsen J.B. (2007). Detection of Actinobacteria cultivated from environmental samples reveals bias in universal primers. *Letters in Applied Microbiology* vol. 45 pp376-381
- Febres V.J., Lee R.F., and Moore G.A. (2008). Transgenic resistance to Citrus tristeza virus in grapefruit. *Plant Cell Reports* vol. 27 pp93-104
- Folimonova S.Y., Robertson C.J., Shilts T., Folimonov A.S., Hilf M.E., Garnsey S.M. and Dawson W.O. (2010). Infection with Strains of *Citrus Tristeza Virus* Does Not Exclude Superinfection by Other Strains of the Virus. *Journal of Virology* vol. 84 pp. 1314–1325
- Folimonova S.Y. (2013). Developing an understanding of cross-protection by Citrus tristeza virus. *Frontiers in Microbiology* vol. 4 pp1-9

Gal-On A. and Shibolet Y.M. (2005). Cross-protection, pp261-288. In G. Loebenstein and J.P. Carr (ed.), Natural resistance mechanisms of plants to viruses. Springer, Dordrecht, The Netherlands

Garnsey S.M., Civerolo E.L., Gumpf D.J., Paul C., Hilf M.E., Lee R.F., Brlansky R.H., Yokomi R.K. and Hartung J.S. (2005). Biological Characterization of an International Collection of Citrus tristeza virus (CTV) Isolates. In: Proc 16th Conf. IOCV, 75-93, IOCV, Riverside, California, United States

Gmitter F.G., Xiao S.Y., Huang S., Hu X.L., Garnsey S.M., and Deng Z. (1996). A Localized linkage map of the citrus tristeza virus resistance gene region. Theoretical Applied Genetics vol.92 pp688-695

Harper S.J. (2013). Citrus tristeza virus: evolution of complex and varied genotypic groups. Frontiers in Microbiology vol. 4 art. 93

Iglesias N.G., Gago-Zachert S.P., Robledo G., Costa N., Plata M.I., Vera O., Grau O. and Semorile L.C. (2008). Population structure of *Citrus tristeza virus* from field Argentinean isolates. Virus Genes vol. 36 pp199-207

Inghelbrecht L., Dessein J., and Van Huylenbroeck G. (2015). Explaining The present GM business strategy on the EU food market: The gatekeepers' perspective. New Biotechnology vol.32 pp65-78

Monger W.A., Alicai T., Ndunguru J., Kinyua Z.M., Potts M., Reeder R.H., Miano D.W., Adams I.P., Boonham N., Glover R.H. and Smith J. (2010). The complete genome sequence of the Tanzanian strain of Cassava brown streak virus and comparison with the Ugandan strain sequence. Archives of Virology vol. 155 pp429-433

Moreno P., Ambrós S., Albiach-Martí M.R., Guerri J. and Peña L. (2008). '*Citrus tristeza virus*: A pathogen that changed the course of the citrus industry'. Molecular Plant Pathology vol. 9 pp251-268

Read D. and Pietersen G. (2015). Genotypic diversity of Citrus tristeza virus within red grapefruit, in a field trial site in South Africa. *European Journal of Plant Pathology* DOI 10.1007/s10658-015-0631-x

Ruiz-Ruiz S., Navarro B., Gisel A., Peña L., Navarro L., Moreno P., Di Serio F. and Flores R. (2011). Citrus tristeza virus infection induces the accumulation of viral small RNAs (21–24-nt) mapping preferentially at the 3'-terminal region of the genomic RNA and affects the host small RNA profile. *Plant Molecular Biology* vol. 75 pp607-619

Sambade A., Lopez C., Rubio L., Flores R., Guerri J. and Moreno P. (2003). Polymorphism of a specific region in gene *p23* of *Citrus tristeza virus* allows for the discrimination between mild and severe isolates. *Archives of Virology* vol. 148 pp2325-2340

Souza A.A., Müller G.W., Targon M.L.P.N., Takita M.A. and Machado M.A. (2002). Stability of the Mild Protective 'PIAC' Isolate of *Citrus tristeza virus*. In: *Proc. 15th Conf. IOCV*, 131-135. IOCV, Riverside, CA

van Dijk E.L., Auger H., Jaszczyszyn Y. and Thermes C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics* vol. 30 pp418-426

Yoon J.Y., Ahn H.I., Kim M., Tsuda S. and Ryu K.I. (2006). 'Pepper mild mottle virus' pathogenicity determinants and cross protection effect of attenuated mutants in pepper'. *Virus Research* vol. 118 pp23-30

APPENDIX A

Table 1: Illumina sequence data for the A-region primer associated bias experiment. The two tables represent separate Illumina Miseq runs using the A-region of a single sample as template. The data in Table 2 of Chapter 2 represents an amalgamation of the data in these two tables.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
T3	526	14585	27.86	6913	7672	7912	0	14355	904	8582	7206	53.23	526	1
B165	520	12810	24.47	6806	6004	10276	0	12668	1309	9288	6500	51.73	520	1
VT	526	12499	23.88	6020	6479	6342	0	12300	691	6658	6114	51.52	526	1
T30	526	6934	13.25	4406	2528	4916	0	6803	431	4466	3408	53.99	526	1
RB-TH28	523	2707	5.17	715	1992	1170	0	2531	217	2121	1416	49.90	523	1
RB-TH30	526	1482	2.83	521	961	344	0	1471	95	957	738	51.71	526	1
T36	541	1085	2.07	188	897	282	0	1080	21	906	553	49.35	541	1
HA 16-5	526	244	0.47	164	80	152	0	242	13	168	115	54.56	526	1

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
T3	526	25170	35.01	12170	13000	17452	0	20725	242	12856	9565	53.23	526	1
VT	526	17943	24.96	7929	10014	11004	0	14502	145	9954	6454	51.52	526	1
T30	526	10602	14.75	6437	4165	8192	0	9066	140	6415	3828	53.99	526	1
B165	520	9957	13.85	5020	4937	9328	0	9270	80	5009	4090	51.73	520	1
RB-TH28	523	4040	5.62	1412	2628	2632	0	2986	72	2597	1644	49.90	523	1
RB-TH30	526	2348	3.27	864	1484	894	0	1945	33	1485	921	51.71	526	1
T36	541	1504	2.09	338	1166	606	0	1465	62	1126	589	49.35	541	1
HA 16-5	526	335	0.47	187	148	294	0	294	1	187	116	54.56	526	1

Table 2: Illumina sequence data for the F-region primer associated bias experiment.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
T30	438	402861	54.07	203139	199722	338724	0	389719	4111	203010	138637.69	45.89	438	1
VT	438	239504	32.14	117054	122450	167654	0	96943	2379	121556	82404.16	44.29	438	1
HA16-5	437	49839	6.69	13380	36459	24618	0	27627	914	36026	17183.27	44.62	437	1
B165	438	30305	4.07	17094	13211	7076	0	21497	306	17032	10412.38	45.21	438	1
T36	438	13218	1.77	6233	6985	4958	0	10030	350	6861	4537.13	43.61	438	1
RB-TH28	438	9360	1.26	3381	5979	5386	0	7028	268	5784	3217.94	44.06	438	1

Table 3: Illumina sequence data for the p33 gene primer associated bias experiment.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
AT-1	981	195802	24.05	97880	97922	165776	0	142388	0	42947	27518	43.63	969	0.99
CT-3719	986	172031	21.13	85740	86291	142076	0	113611	0	41215	24141	42.70	974	0.99
T36	988	166143	20.41	82999	83144	138114	0	107191	0	38769	23140	43.62	979	0.99
Kpg3_SP_T3	982	113290	13.92	56478	56812	90096	0	81585	0	37991	17530	44.20	973	0.99
HA16-5	982	100392	12.33	51157	49235	76268	0	70987	0	43693	16737	43.89	966	0.98
CTZA	983	40167	4.93	20006	20161	35300	0	34432	0	11703	5134	43.44	970	0.99
RB	982	14129	1.74	7040	7089	12346	0	11773	0	3401	1863	43.38	968	0.99
Taiwan-Pum/SP/T1	982	12140	1.49	6029	6111	10234	0	10527	0	3303	1780	43.28	968	0.99

Table 4: Illumina sequence data for the p23 gene primer associated bias experiment.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
T30	706	31807	23.12	17077	14730	16590	0	20935	150	18388	9960.439	43.63	708	1
T36	706	30305	22.03	15208	15097	16052	0	20266	0	15784	9465.895	43.91	703	1
RB-TH28	707	25993	18.90	11757	14236	12478	0	16864	174	12949	8129.68	43.99	709	1
SP	707	24884	18.09	12923	11961	11290	0	16757	131	13506	7776.24	44.13	709	1
Kpg3	707	24572	17.86	11794	12778	11444	0	16749	161	12385	7688.406	43.00	709	1

APPENDIX B

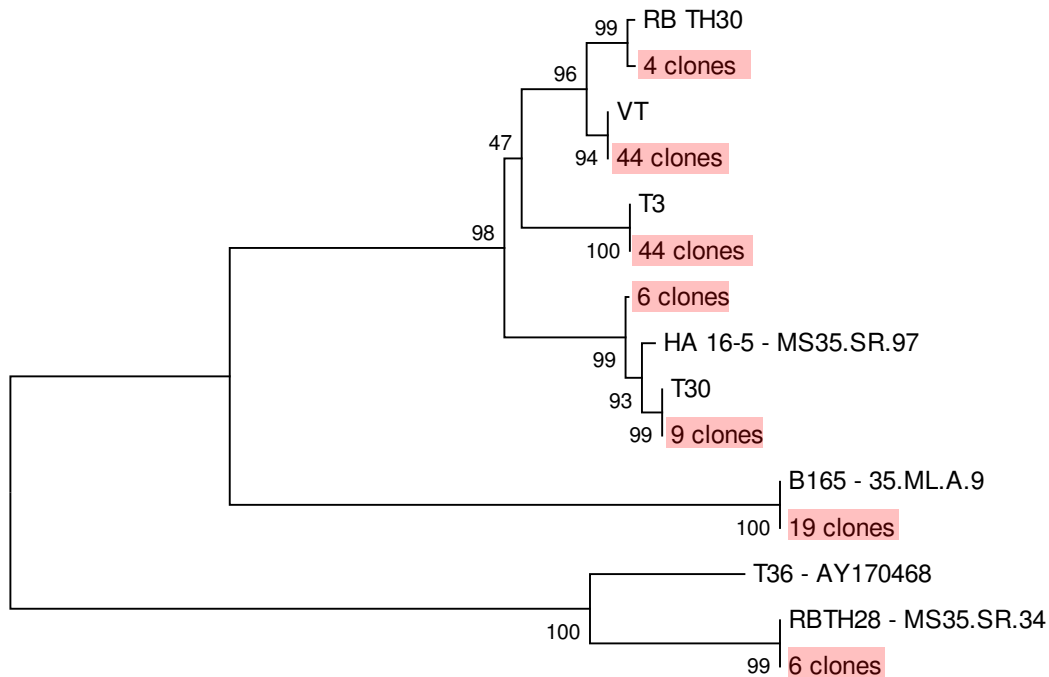


Figure 1: Overview neighbour-joining phylogenetic tree, produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the A-region clones isolated from the bias experiment and their reference sequences. Each dendrogram was collated by using a single representative reference sequence from each group and collapsing the branches representing the clone sequences down to a single branch. Collated branches are highlighted in red. The number of clones that were associated with each particular branch is presented.

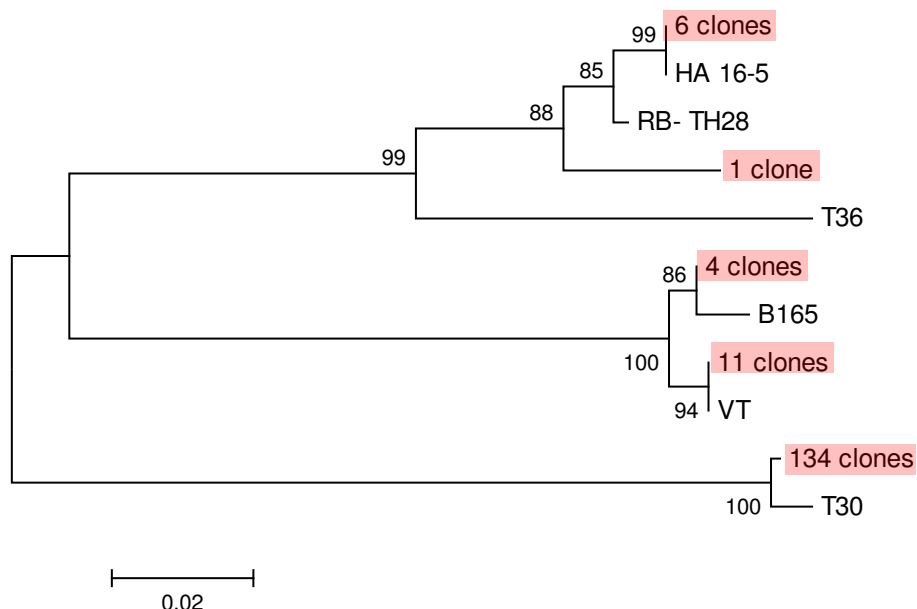


Figure 2: Overview neighbour-joining phylogenetic tree, produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the F-region clones isolated from the bias experiment and their reference sequences. Each dendrogram was collated by using a single representative reference sequence from each group and collapsing the branches representing the clone sequences down to a single branch. Collated branches are highlighted in red. The number of clones that were associated with each particular branch is presented.

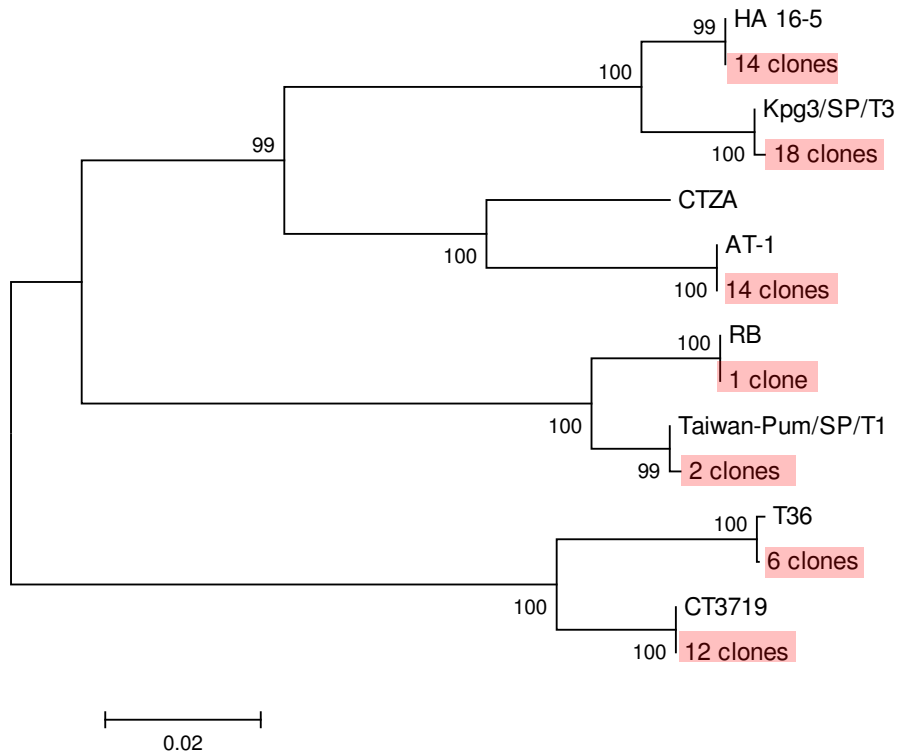


Figure 3: Overview neighbour-joining phylogenetic tree, produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated from the bias experiment and their reference sequences. Each dendrogram was collated by using a single representative reference sequence from each group and collapsing the branches representing the clone sequences down to a single branch. Collated branches are highlighted in red. The number of clones that were associated with each particular branch is presented.

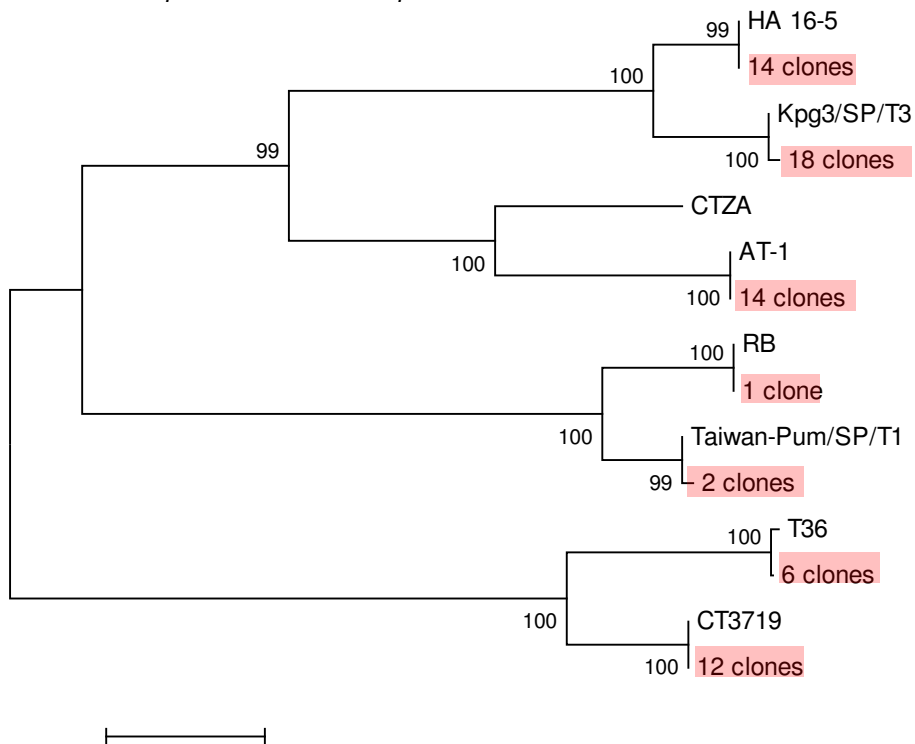


Figure 4: Overview neighbour-joining phylogenetic tree, produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p23 gene clones isolated from the bias experiment and their reference sequences. Each dendrogram was collated by using a single representative reference sequence from each group and collapsing the branches representing the clone sequences down to a single branch. Collated branches are highlighted in red. The number of clones that were associated with each particular branch is presented.

APPENDIX C

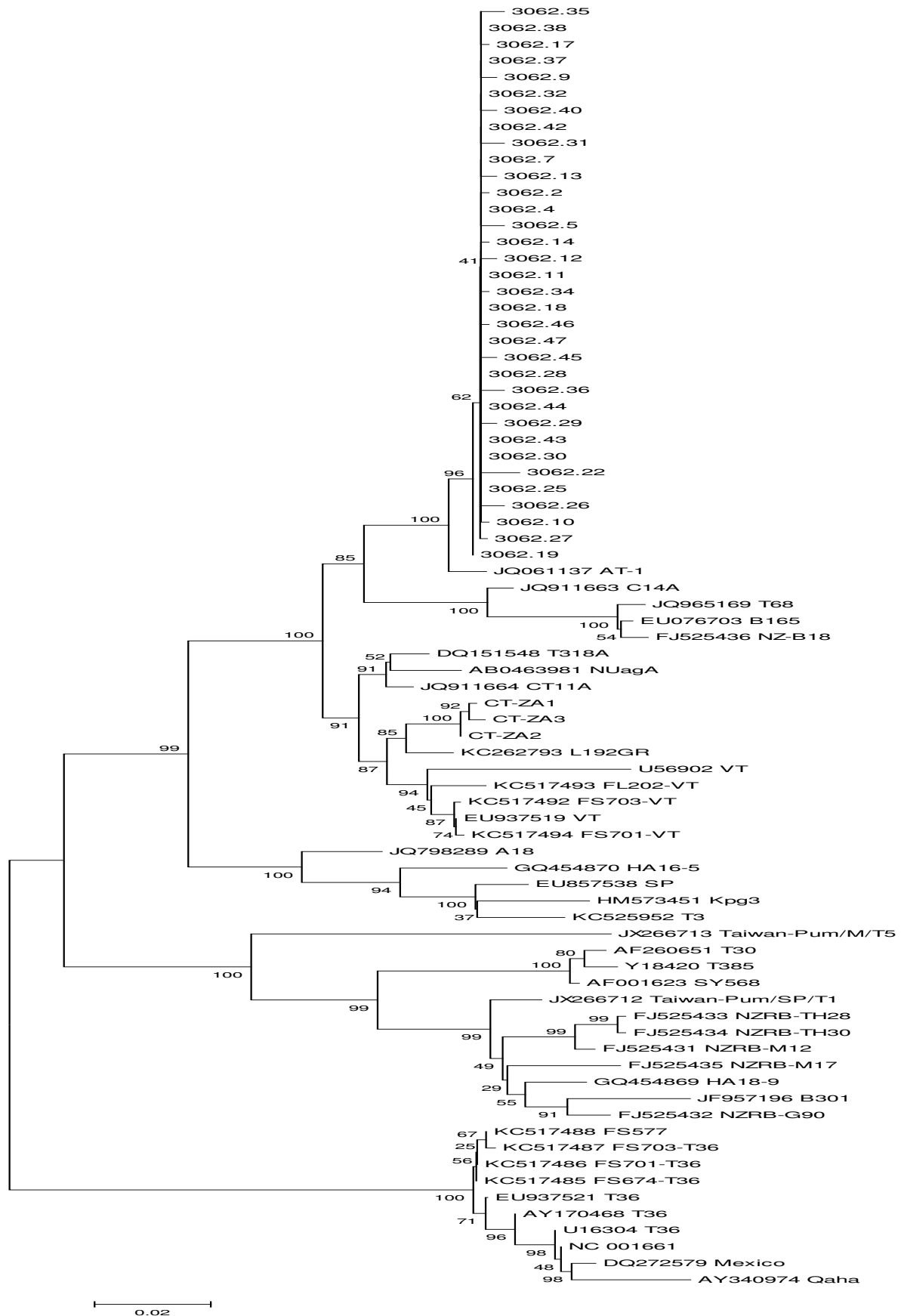


Figure 1: Neighbour-joining phylogenetic tree, produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated for sample 13-3062, representing the Hoedspruit production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

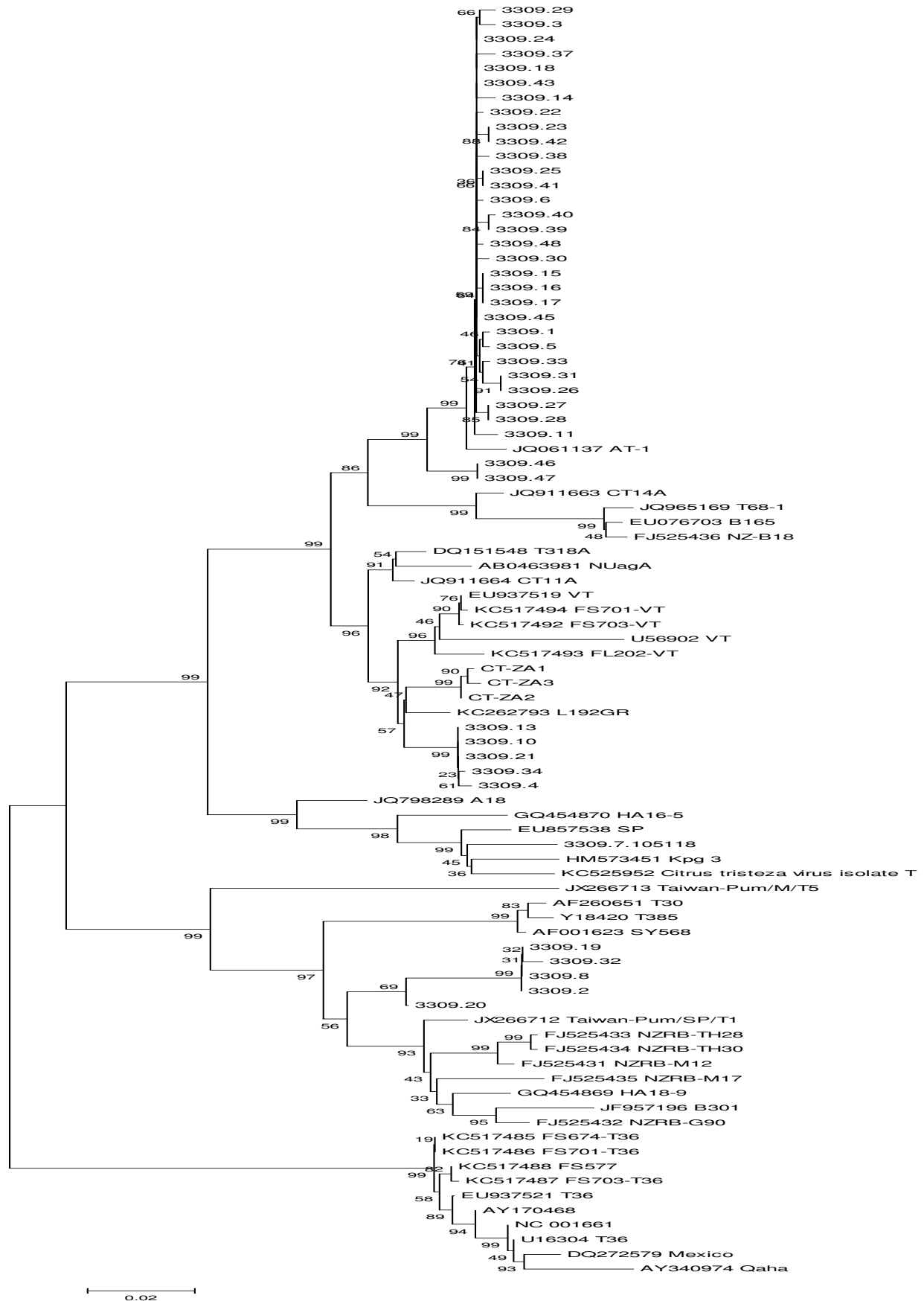


Figure 2: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated for sample 13-3309, representing the Malelane production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

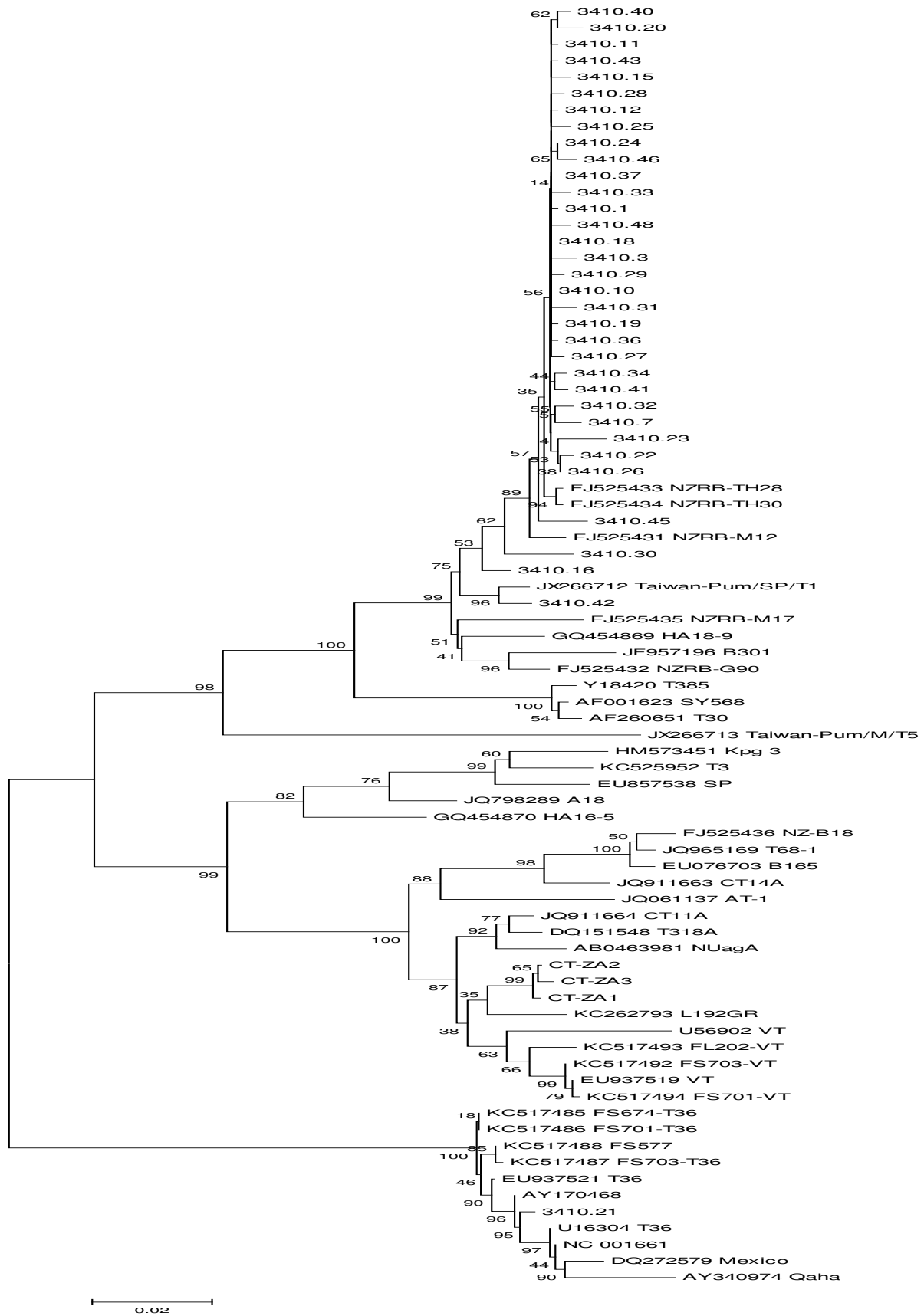


Figure 3: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated for sample 13-3410, representing the Swaziland production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

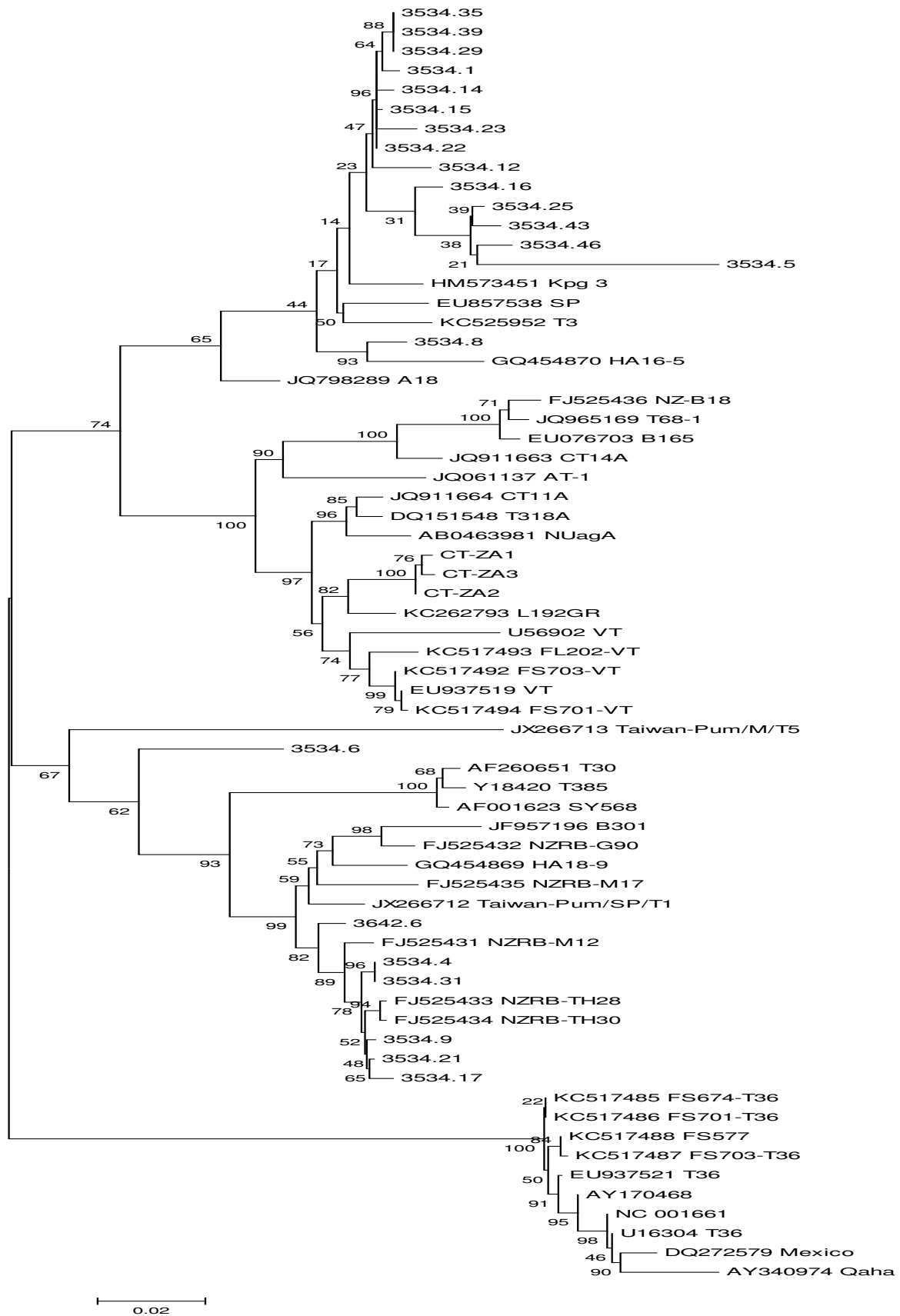


Figure 4: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated for sample 13-3534, representing the Northern Cape (Kakamas) production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

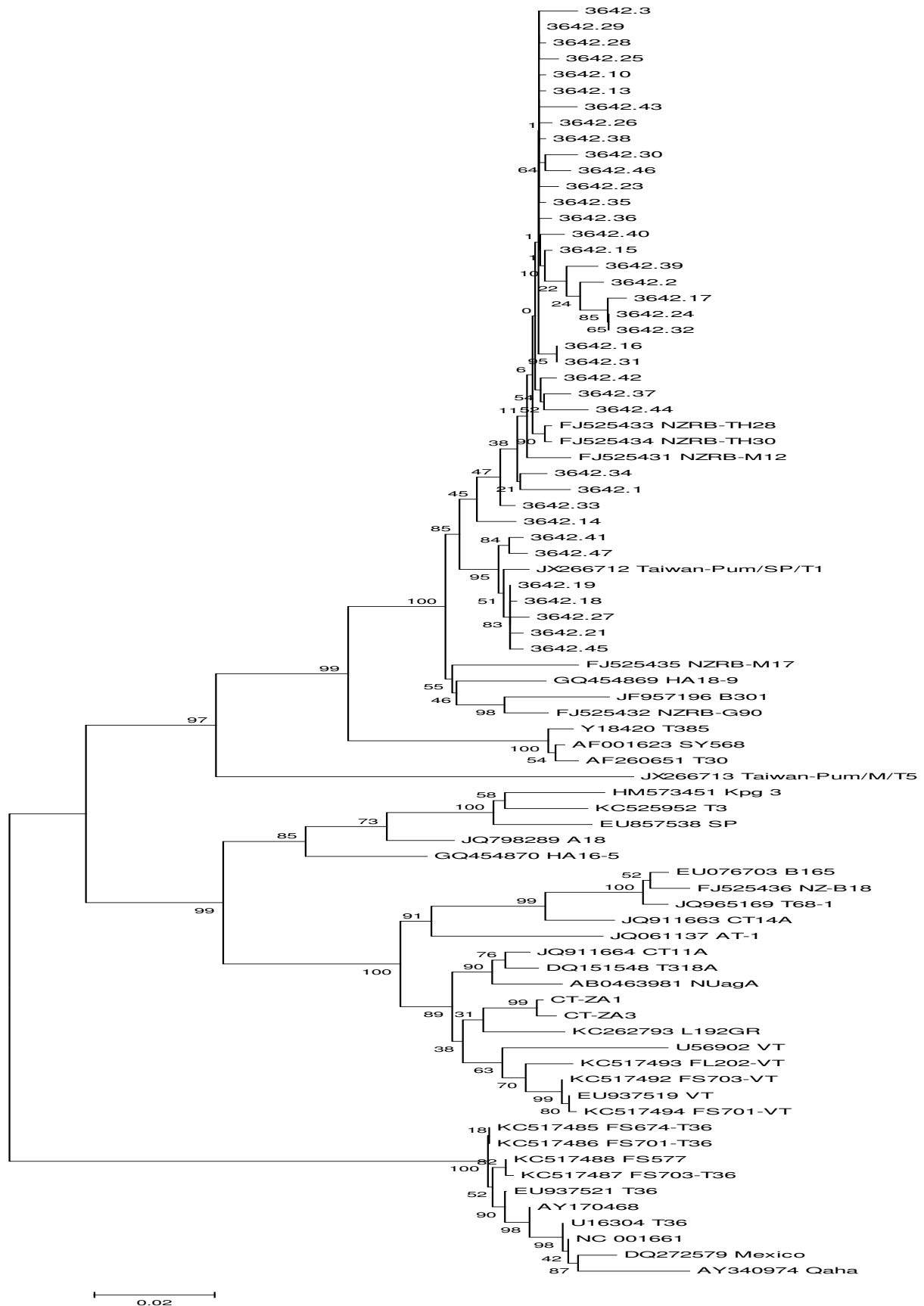


Figure 5: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated for sample 13-3642, representing the Sundays River Valley production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

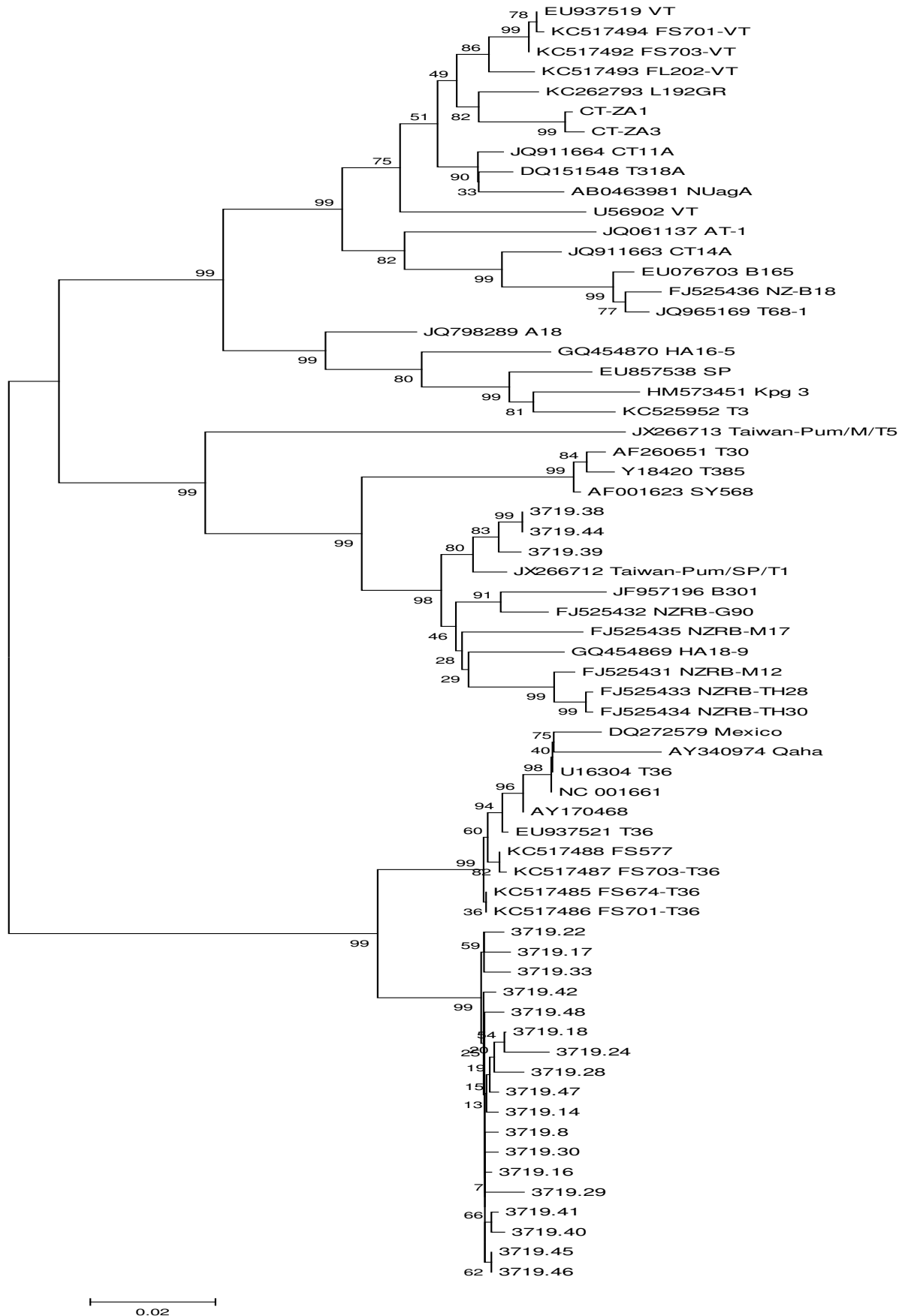


Figure 6: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the p33 gene clones isolated for sample 13-3719, representing the Nkwadini Valley production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

APPENDIX D

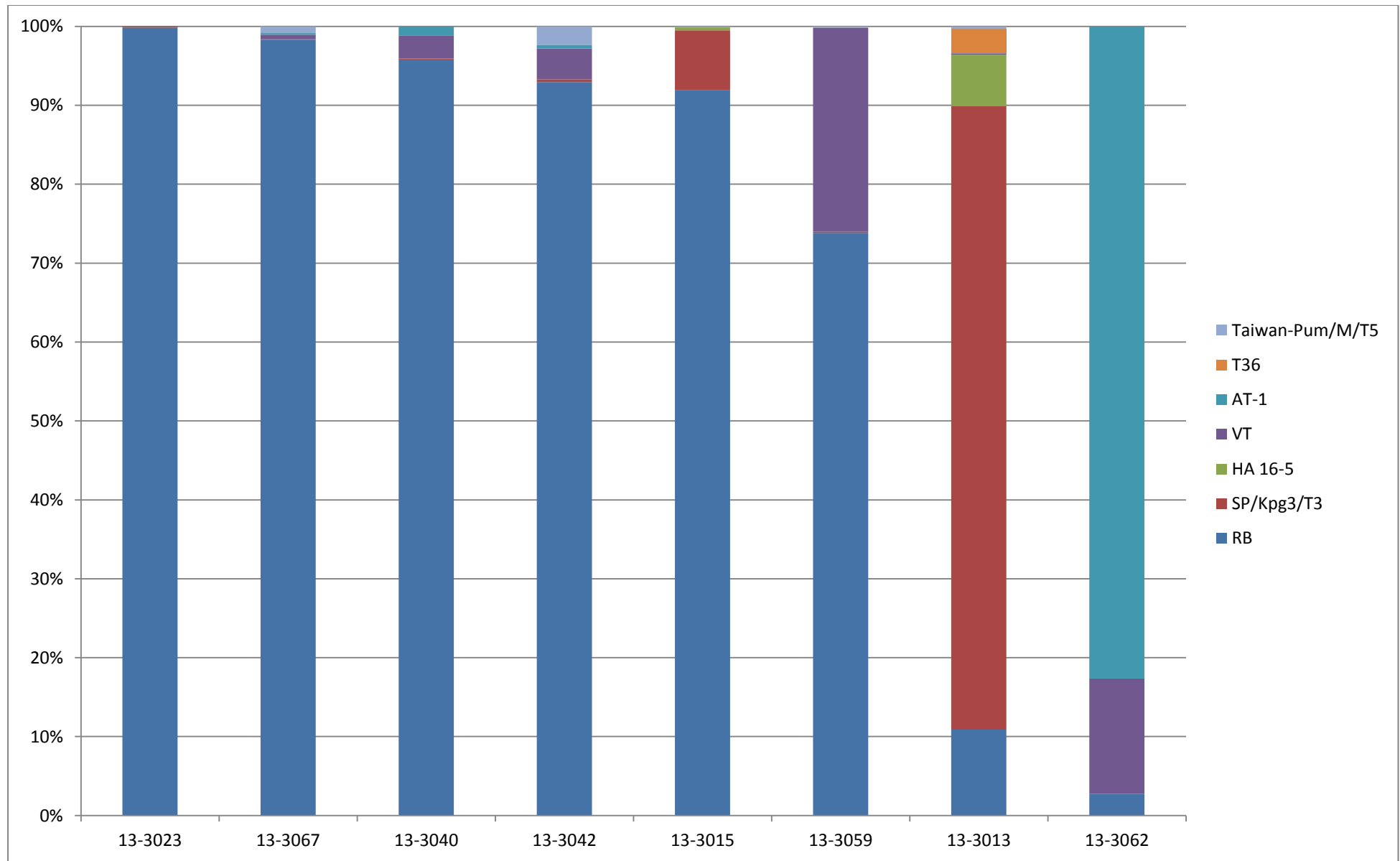


Figure 1: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Hoedspruit 1 orchard, which was 18 years old at the time of collection. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 1 052 616 reads mapped to the reference set.

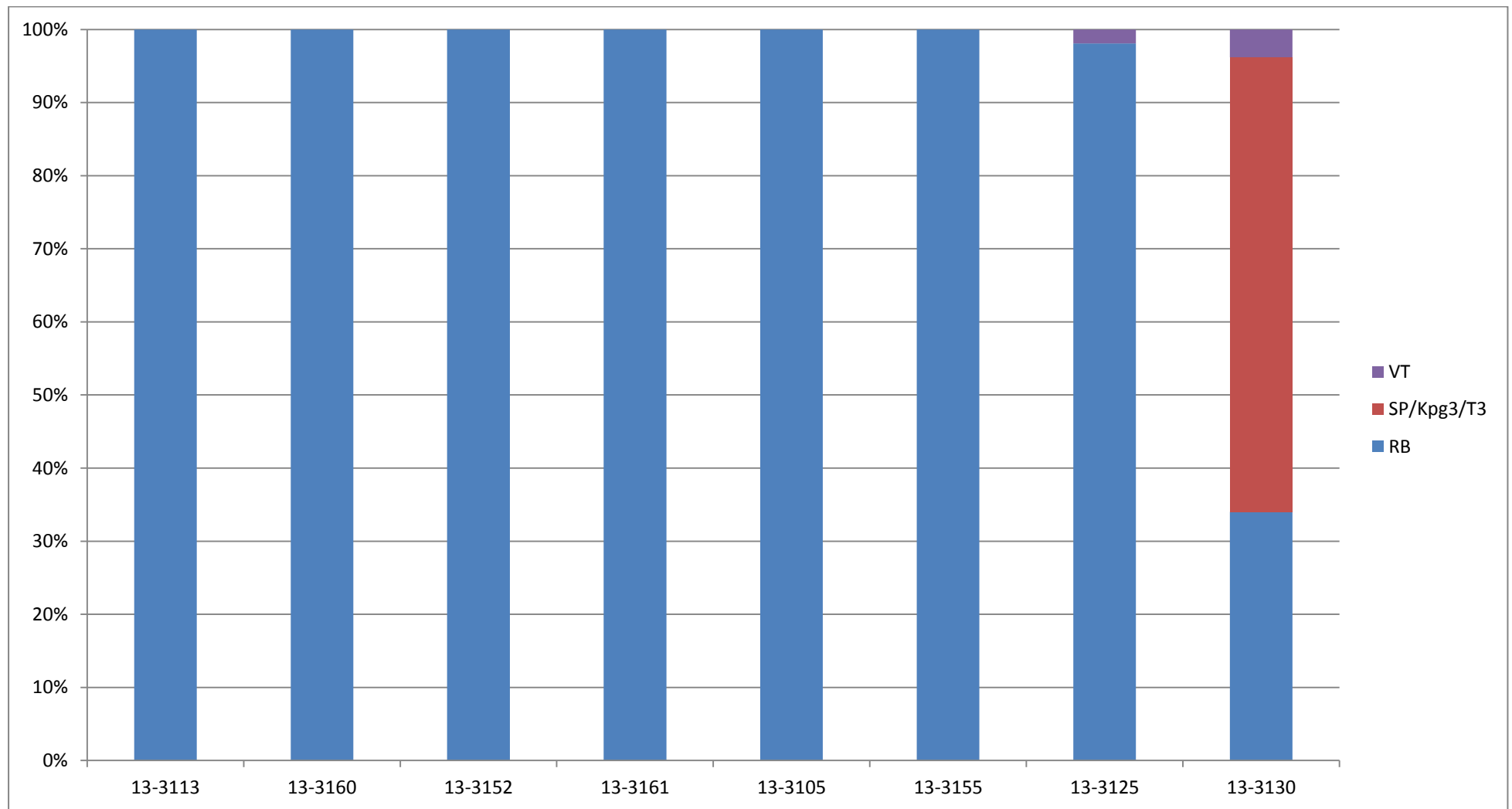


Figure 2: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Hoedspruit 2 orchard, which was eight years old at the time of collection. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 782 825 reads mapped to the reference set.

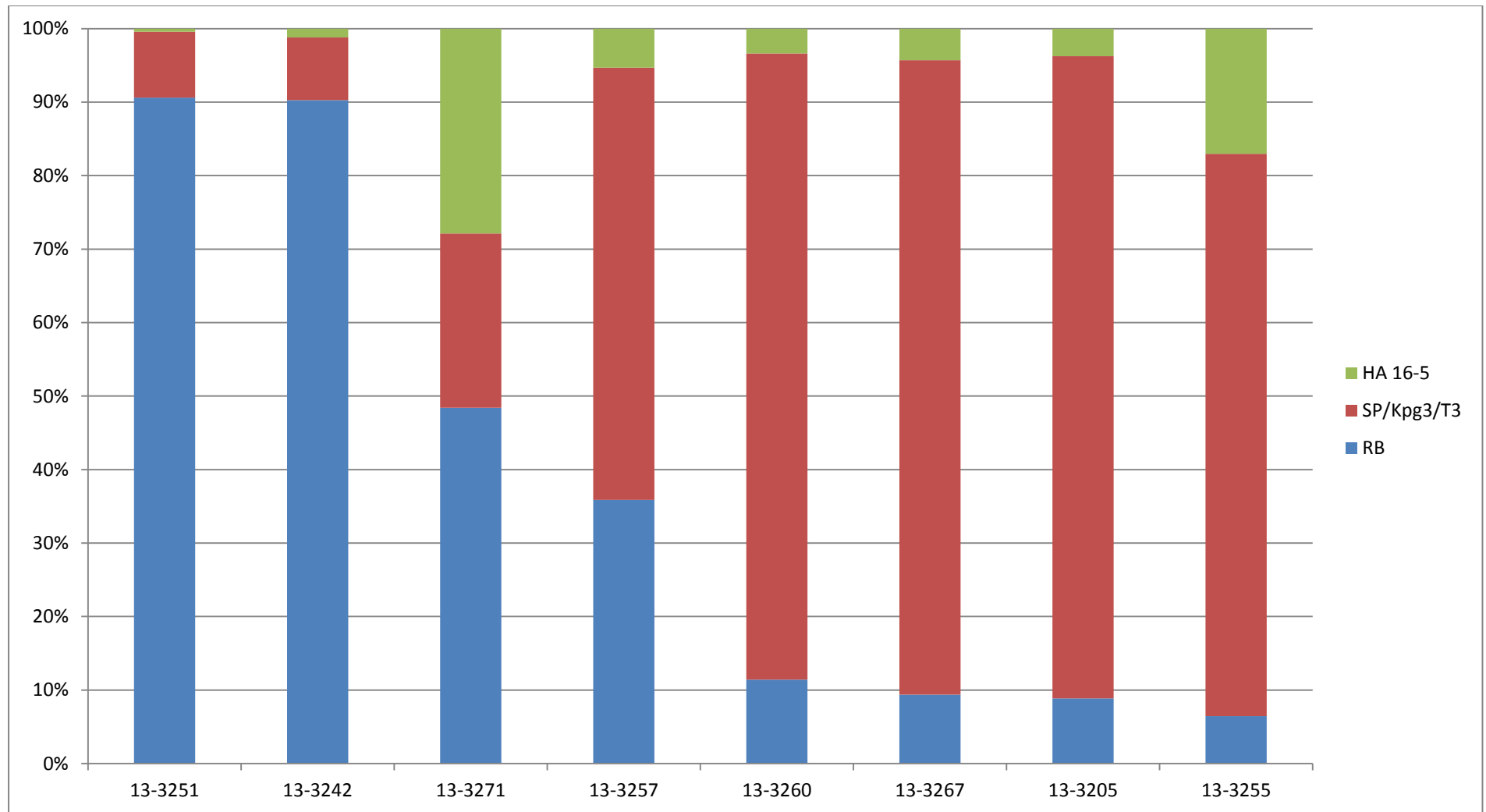


Figure 3: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Malelane 1 orchard, which was 24 years old at the time collection. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 941 736 reads mapped to the reference set.

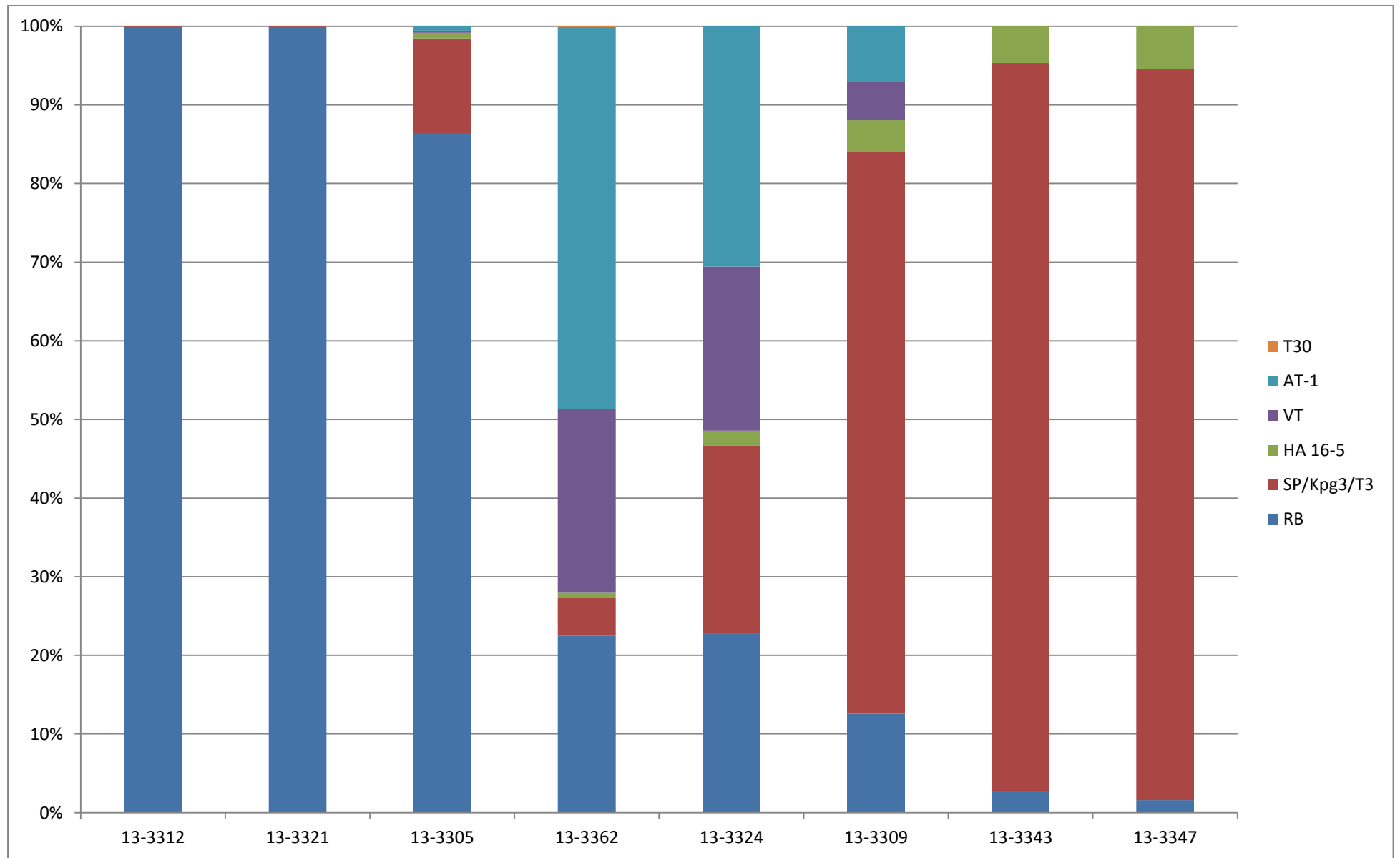


Figure 4: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Malelane 2 orchard, which was 14 years old at the time of collection. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 973 661 reads mapped to the reference set. 122

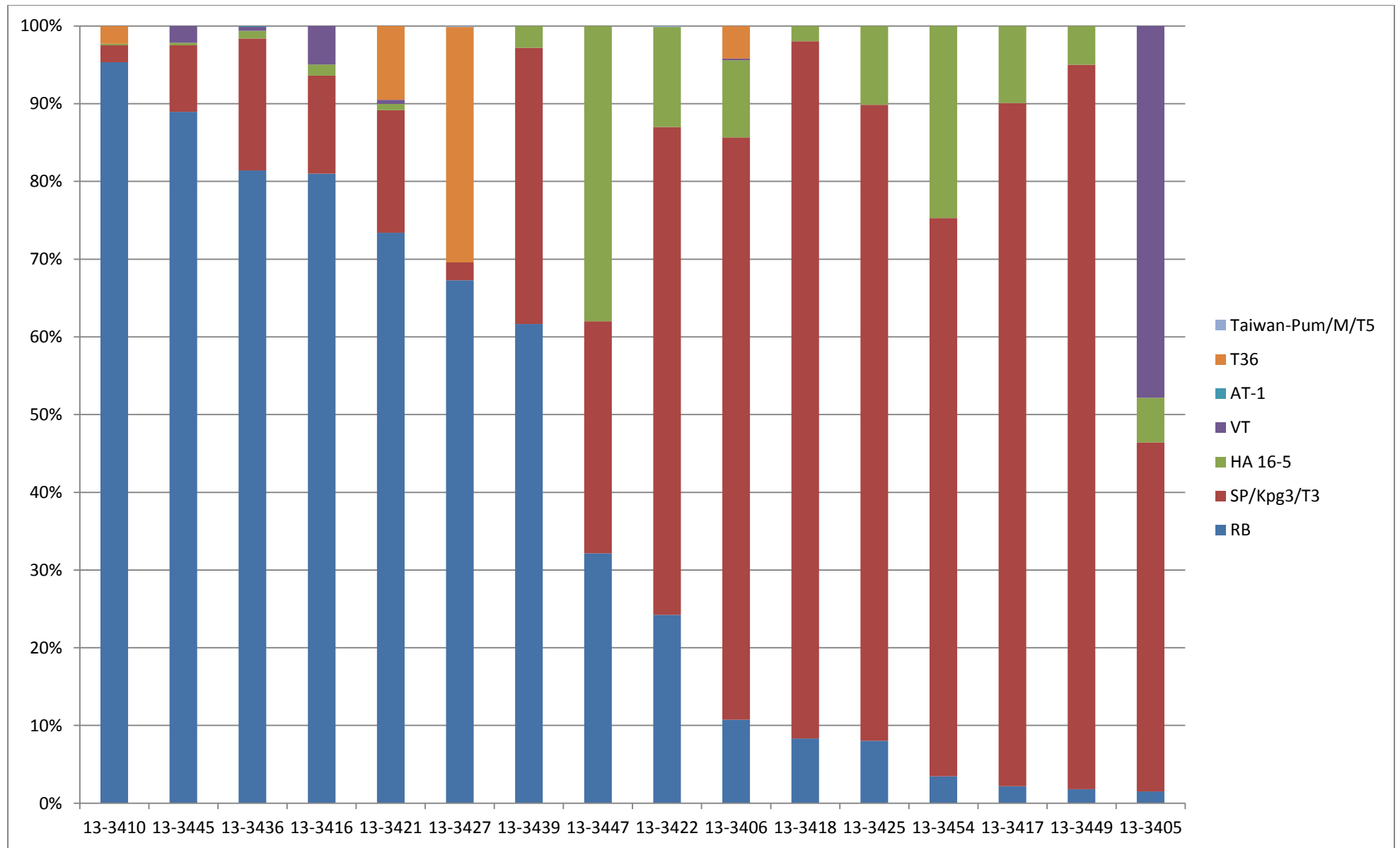


Figure 5: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Mananga, Swaziland orchard, which was 18 years old at the time of collection. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 2 349 853 reads mapped to the reference set. 123

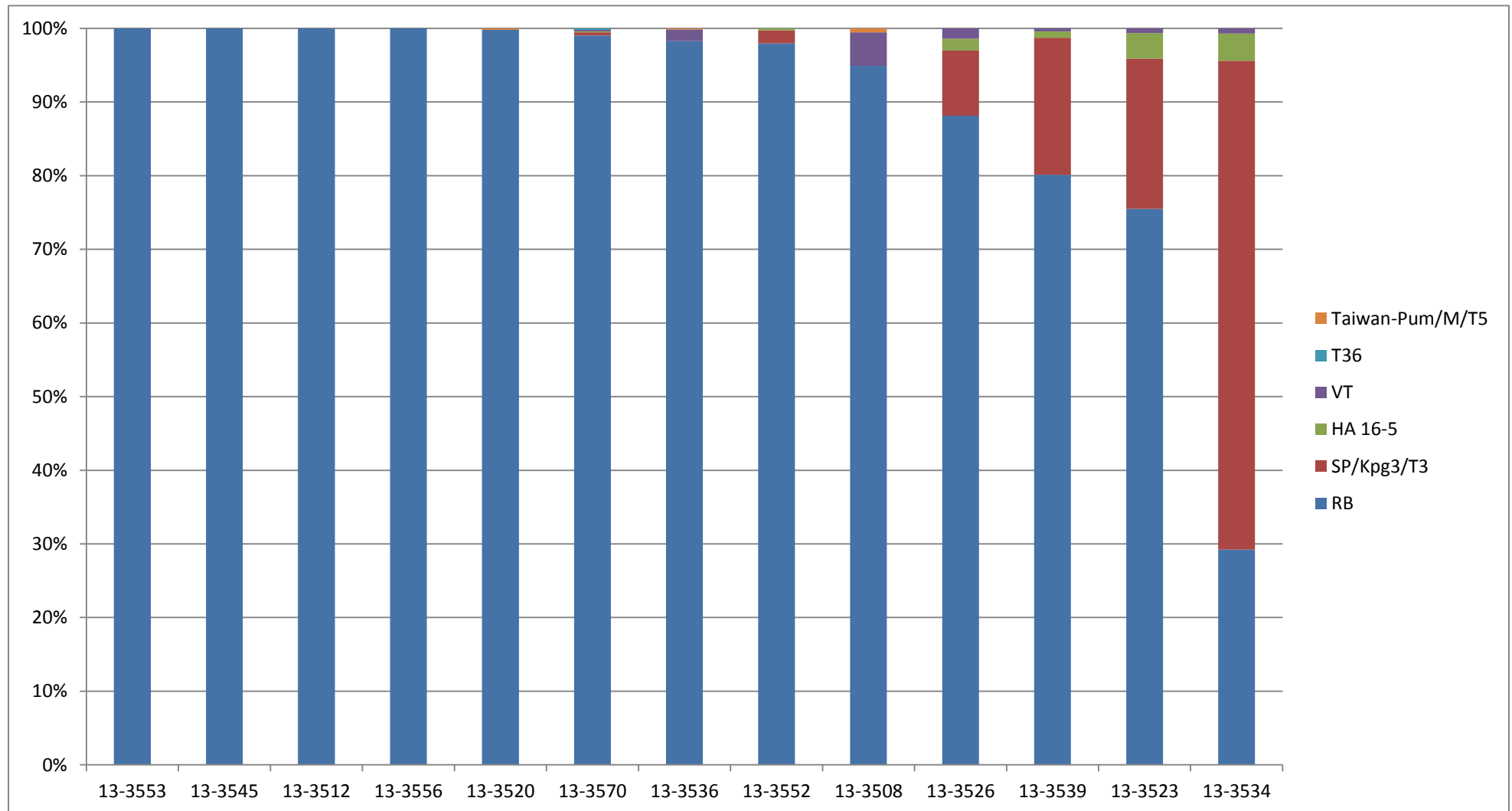


Figure 6: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Kakamas, Northern Cape orchard, which was 12 years old at the time of collection. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 1 361 349 reads mapped to the reference set.

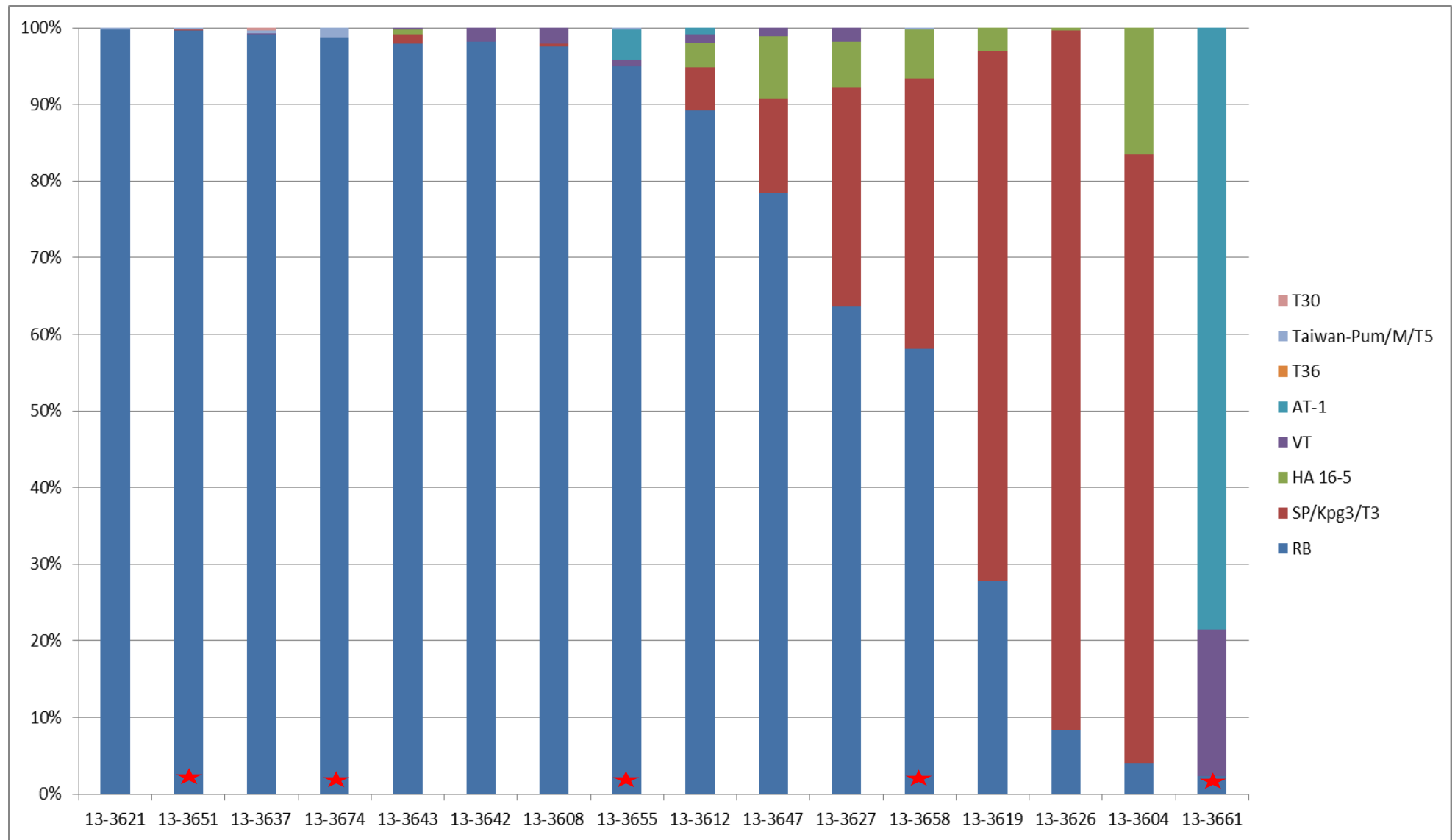


Figure 7: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Sundays River Valley orchards, where trees were aged 9 years (marked with a red star on their corresponding bars) and 23 years. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 1 874 424 reads mapped to the reference set.

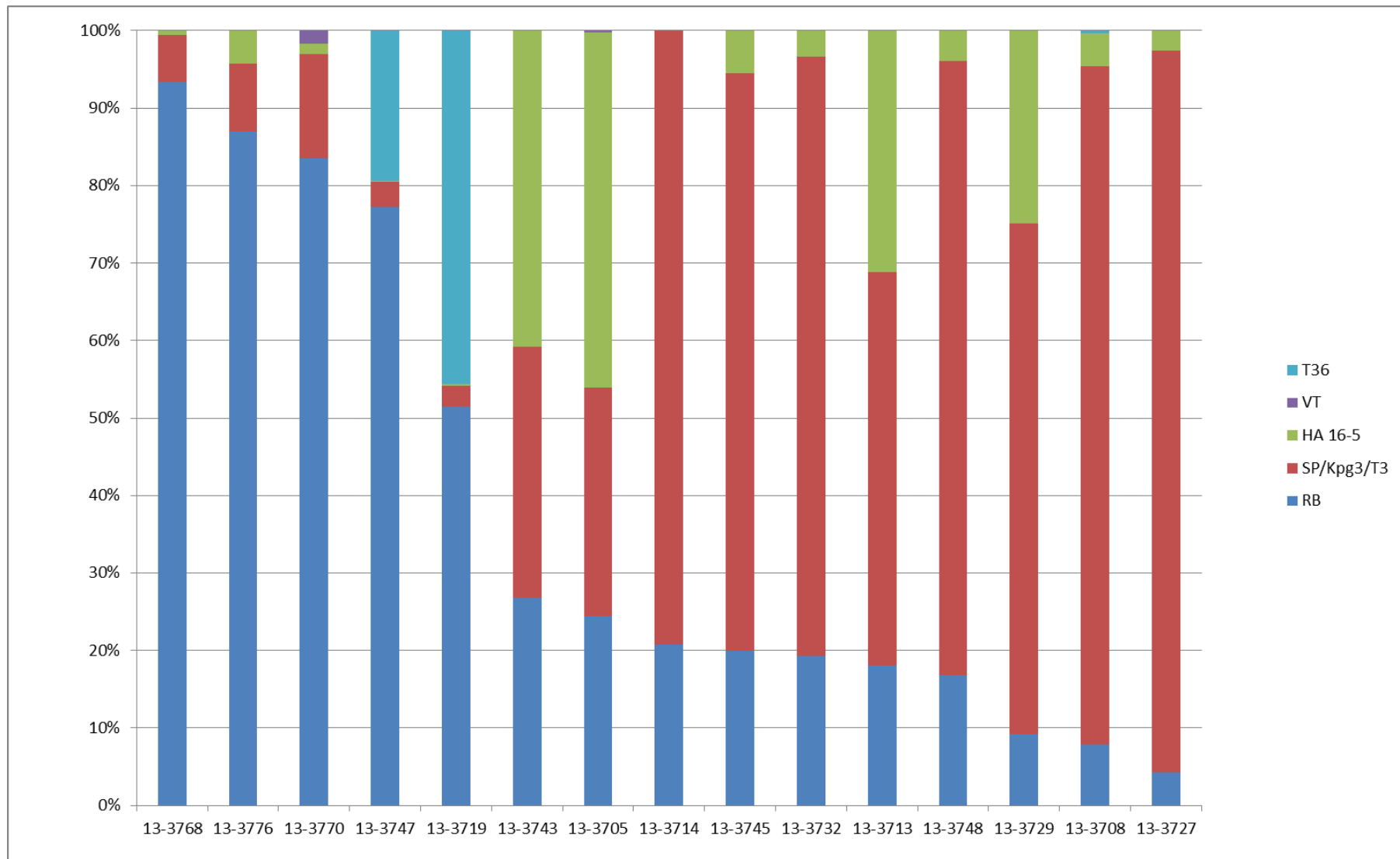


Figure 8: Graph representing the percentage of Illumina sequencing reads that mapped back to each reference group for the samples from the Nkwalini Valley orchard, which was 23 years old at the time of collection. KwaZulu-Natal. Each bar represents the CTV population from a sample collected from a tree in the field. The reference groups are indicated in the figure legend. A total of 2 529 024 reads mapped to the reference set.

APPENDIX E

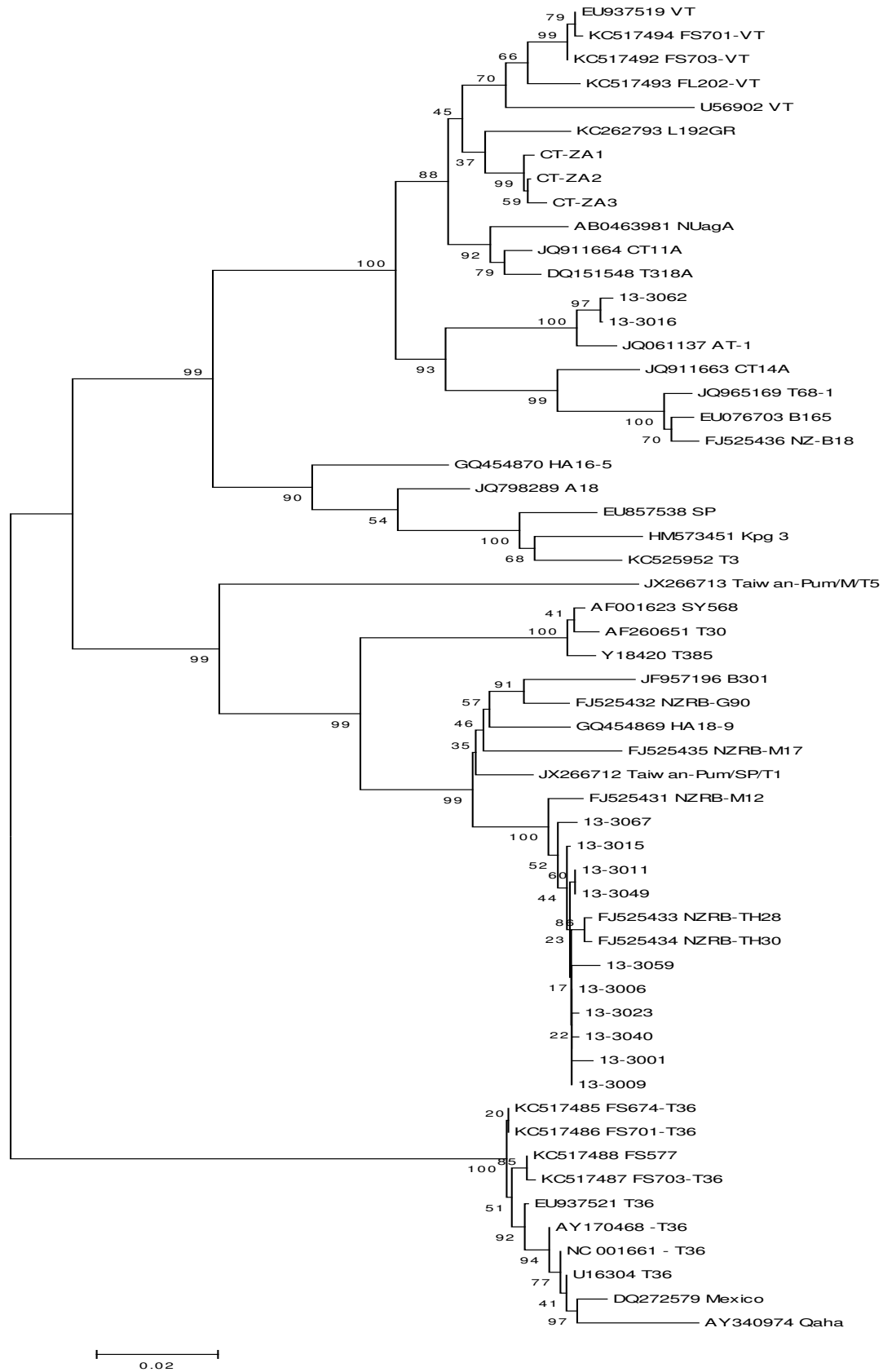


Figure 1: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Hoedspruit 1 production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

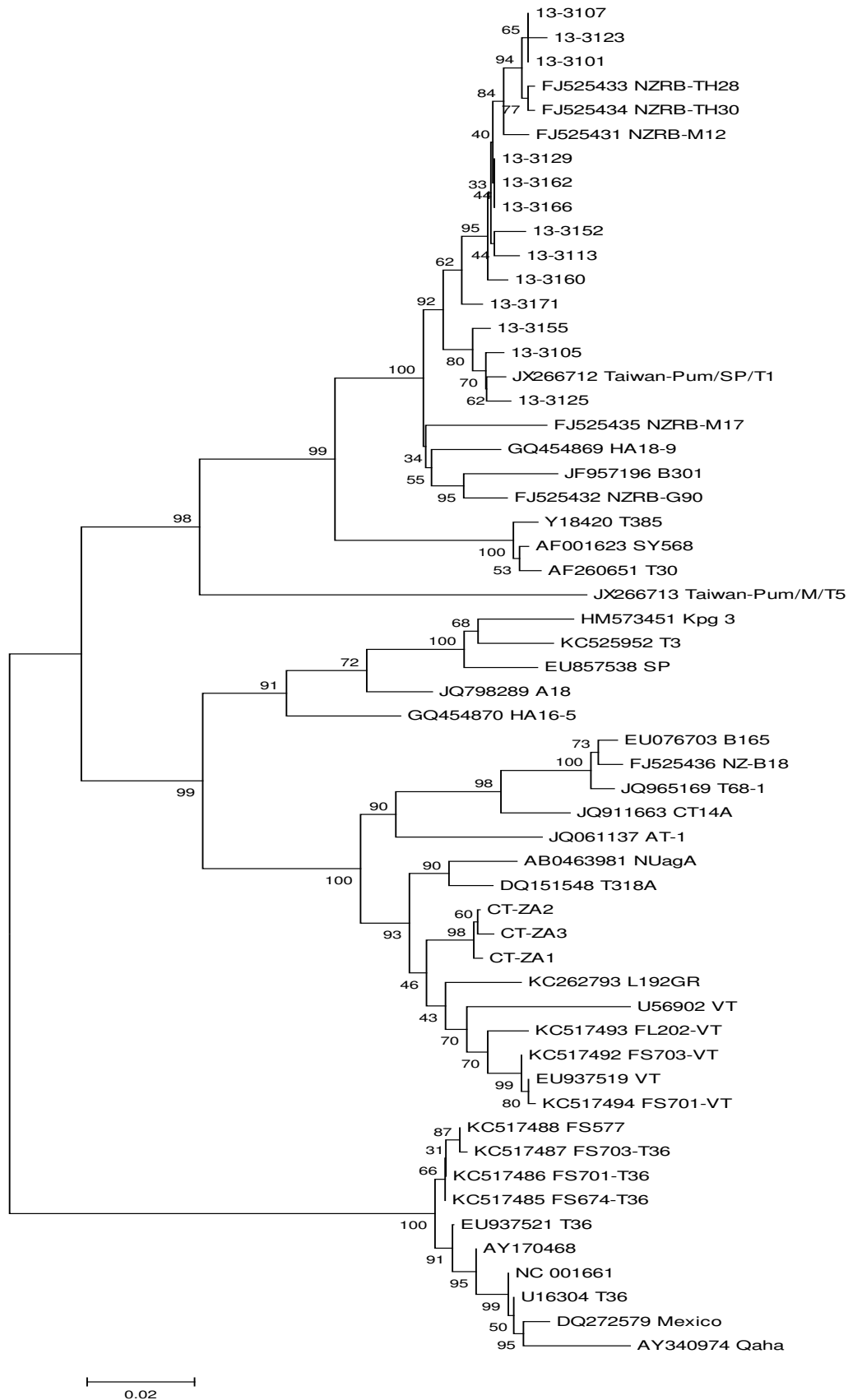


Figure 2: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Hoedspruit 2 production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

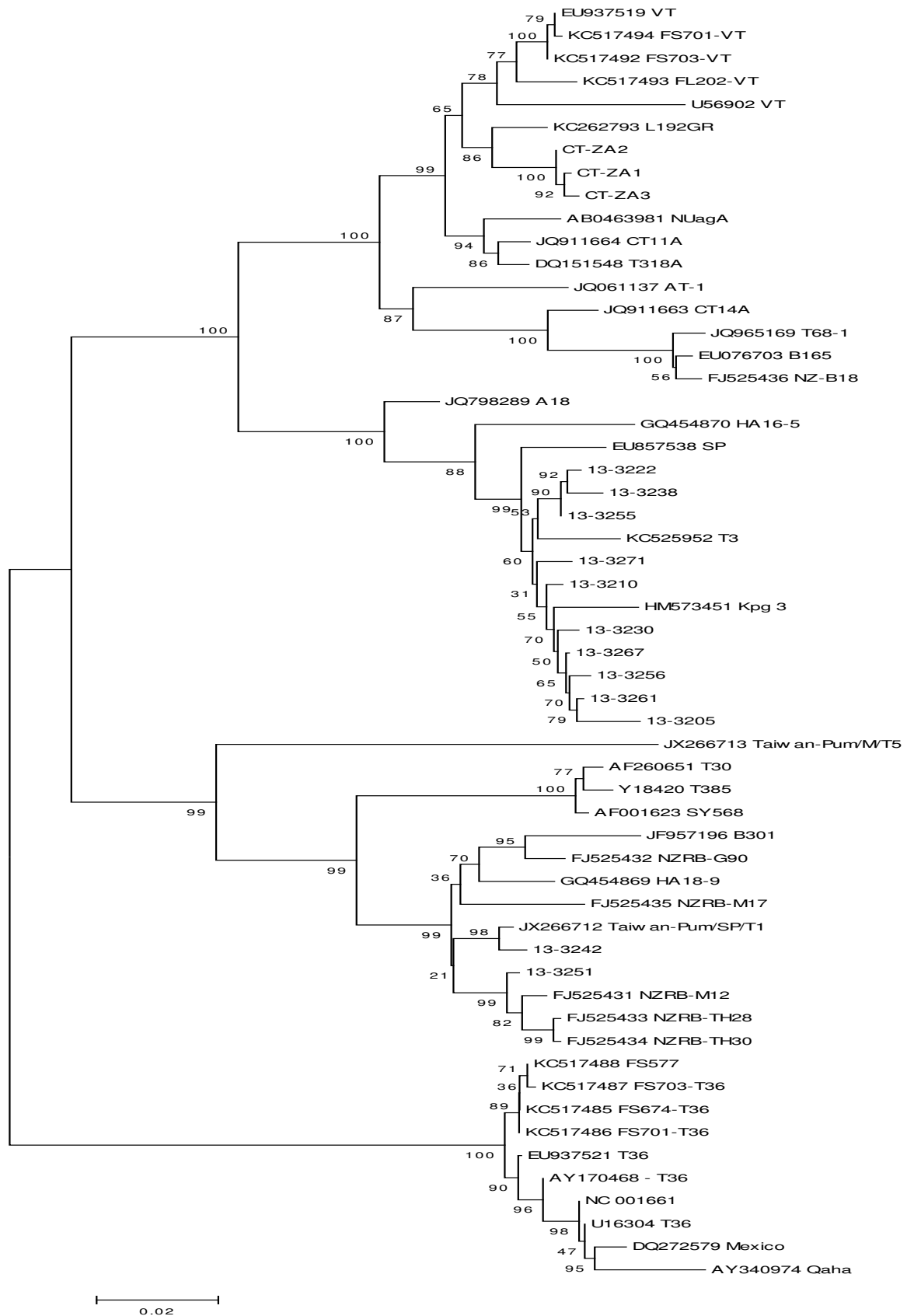


Figure 3: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Malelane 1 production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

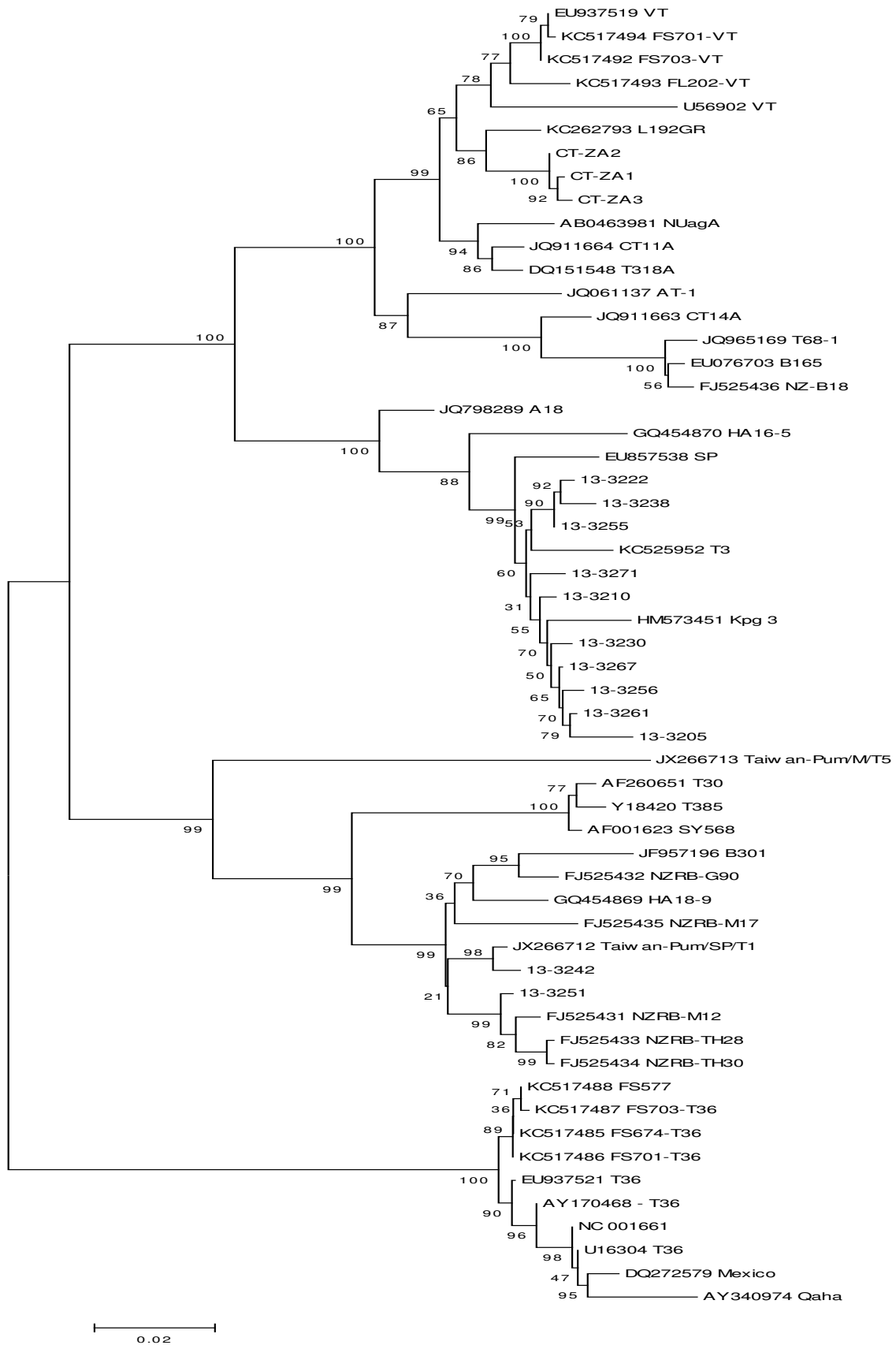


Figure 4: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Malelane 2 production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

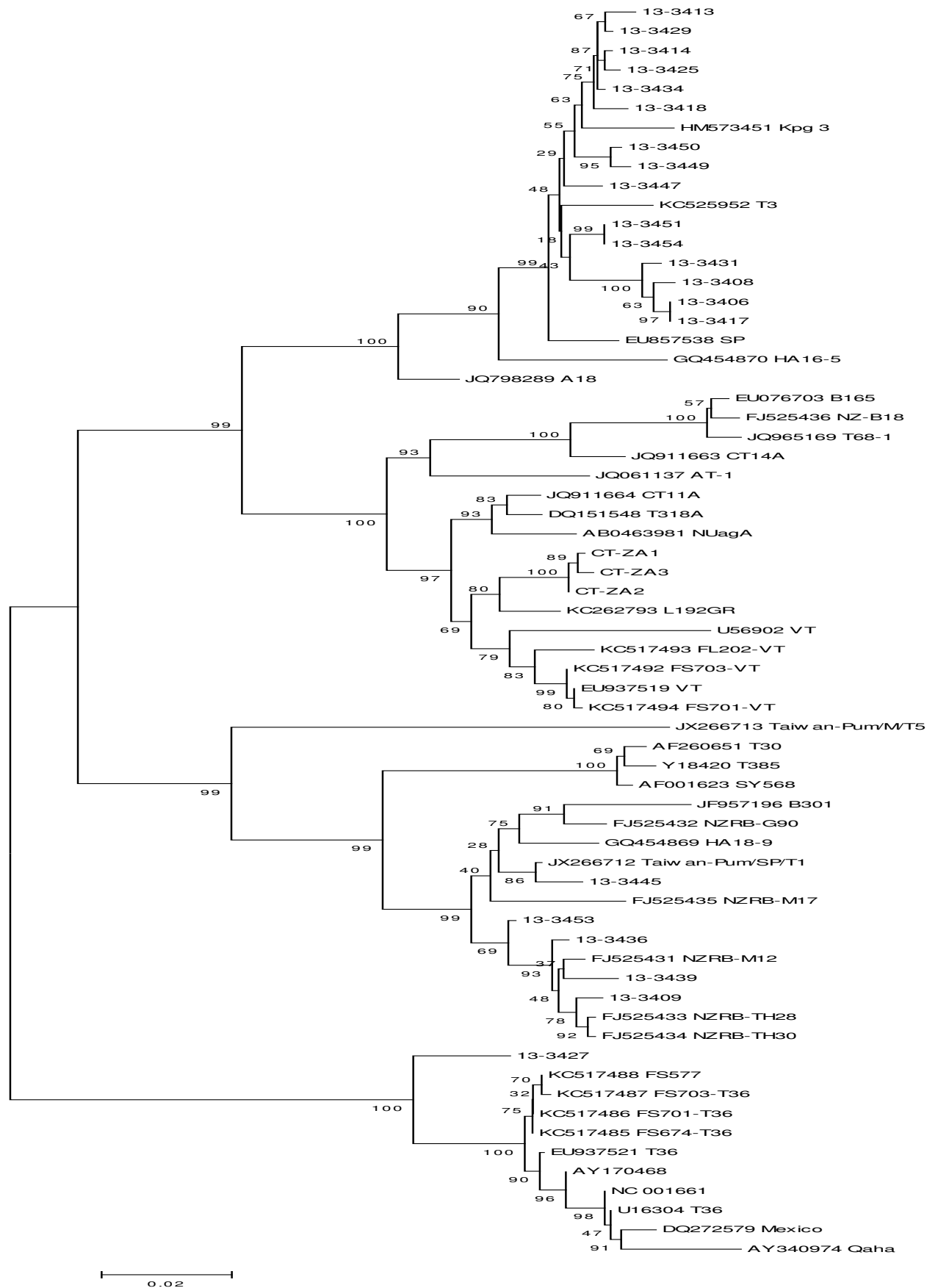


Figure 5: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Mananga, Swaziland production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

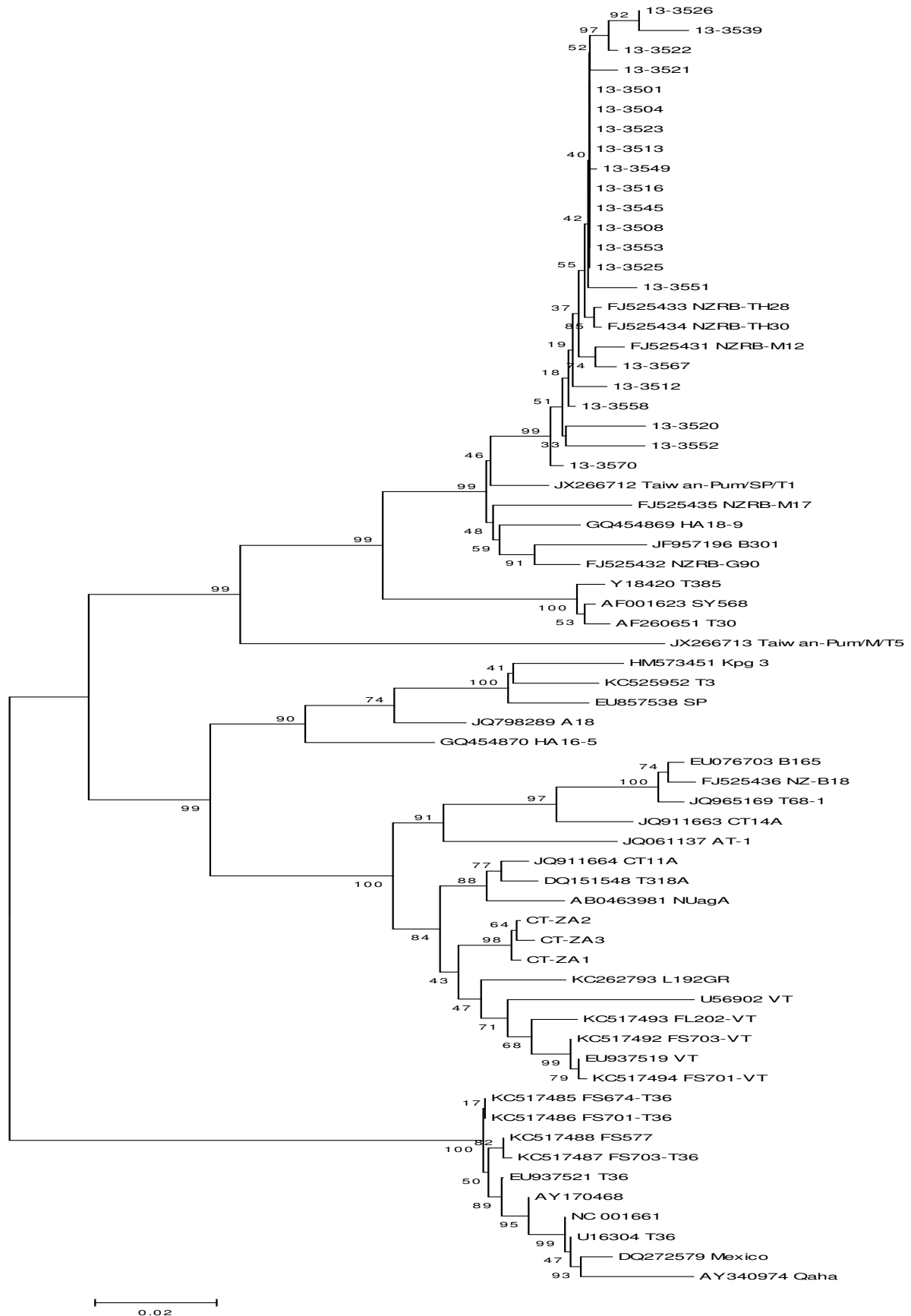


Figure 6: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Kakamas, Northern Cape production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

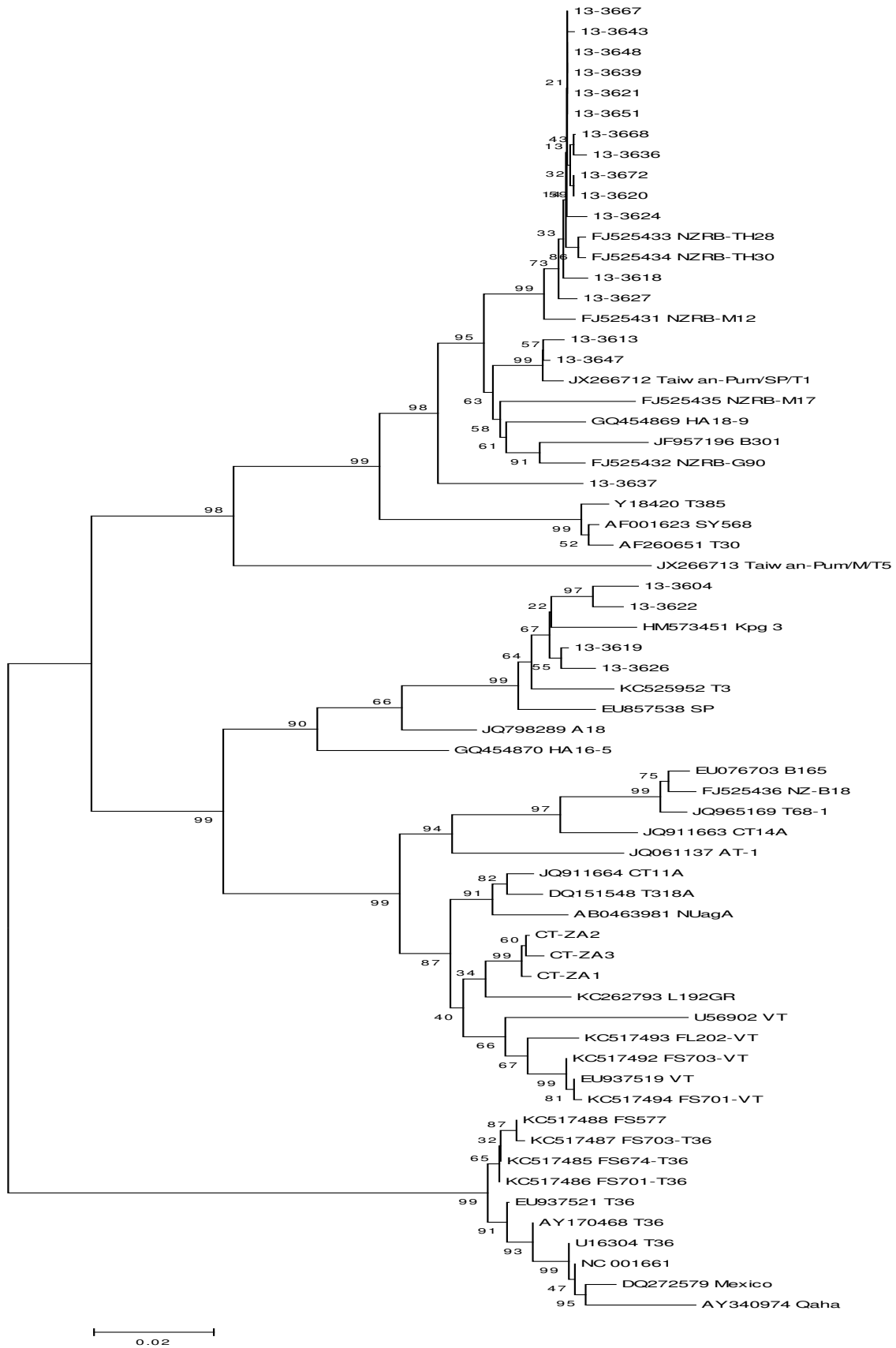


Figure 7: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Sundays River Valley production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

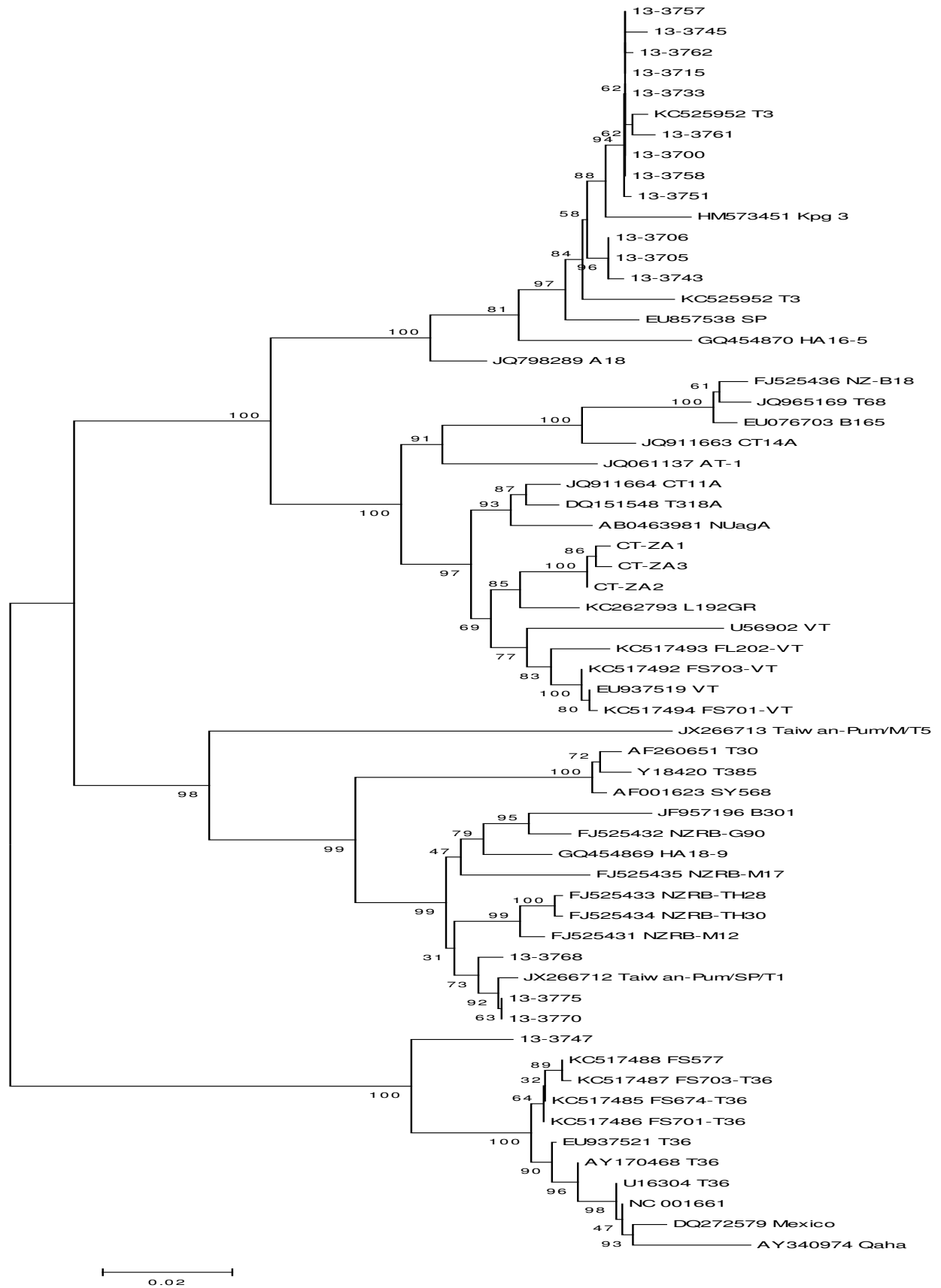


Figure 8: Neighbour-joining phylogenetic tree produced using the Maximum Composite Likelihood Model (MCL), displaying the relationships between the directly sequenced p33 gene amplicons derived from the samples collected in the Nkwalini Valley production area. The bootstrap values are shown below the branches and the scale indicates genetic distance (in nucleotide substitutions per position in the sequence).

SUPPLEMENTARY DATA

SECTION 1

Table 1.1: Number of reads mapping back to references for sample **13-3013**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	31943	55.43	15781	16162	26296	0	30552	0	14218	6135	43.81	967	0.99
JQ798289_A18	977	6746	11.71	3151	3595	5562	0	6593	0	4924	1215	43.30	926	0.95
EU857538_SP	977	5501	9.55	3235	2266	3856	0	5494	0	4428	1220	43.60	958	0.98
GQ454870_HA16-5	977	3734	6.48	1779	1955	2398	0	3704	0	2305	844	43.71	871	0.89
FJ525433_NZRB-TH28	977	2395	4.16	1280	1115	2066	0	2313	0	2368	513	43.81	663	0.68
FJ525431_NZRB-M12	977	1972	3.42	998	974	1604	0	1914	0	1350	379	43.30	932	0.95
AY170468_T36	981	1565	2.72	750	815	1168	0	1565	0	1526	419	43.63	824	0.84
KC525952_T3	977	1193	2.07	664	529	826	0	1191	0	896	214	42.89	919	0.94
JX266712_Taiwan-Pum_SP_T1	977	763	1.32	375	388	664	0	743	0	235	118	43.30	946	0.97
FJ525434_NZRB-TH30	977	543	0.94	296	247	480	0	534	0	541	113	43.81	681	0.70
FJ525432_NZRB-G90	977	241	0.42	121	120	208	0	233	0	96	28	43.40	843	0.86
FJ525435_NZRB-M17	977	219	0.38	109	110	196	0	192	0	94	17	43.71	802	0.82
EU937521_T36	981	211	0.37	119	92	144	0	211	0	183	57	43.43	905	0.92
JX266713_Taiwan-Pum_M_T5	977	194	0.34	94	100	176	0	190	0	155	17	43.30	536	0.55
JF957196_B301	977	119	0.21	58	61	102	0	118	0	85	13	42.89	566	0.58
KC262793_L192GR	977	98	0.17	48	50	88	0	98	0	55	10	43.91	833	0.85

Table 1.2: Number of reads mapping back to references for sample **13-3015**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	101298	55.74	49901	51397	90804	0	92282	0	39683	22642	43.30	970	0.99
FJ525433_NZRB-TH28	977	23997	13.20	12185	11812	21564	0	21903	0	21072	6175	43.81	960	0.98
FJ525431_NZRB-M12	977	20095	11.06	11175	8920	14316	0	19485	0	12097	4673	43.30	958	0.98
FJ525434_NZRB-TH30	977	11555	6.36	6207	5348	10462	0	11001	0	10996	2650	43.81	909	0.93
HM573451_Kpg 3	977	8770	4.83	4331	4439	8012	0	8592	0	3563	1603	43.81	963	0.99
FJ525432_NZRB-G90	977	4448	2.45	2362	2086	3300	0	4248	0	3912	855	43.40	876	0.90
FJ525435_NZRB-M17	977	3418	1.88	1768	1650	2434	0	3132	0	3082	641	43.71	867	0.89
JQ798289_A18	977	2595	1.43	1279	1316	2302	0	2577	0	2235	478	43.30	882	0.90
JF957196_B301	977	2151	1.18	1034	1117	1906	0	1455	0	2019	305	42.89	872	0.89
KC525952_T3	977	1606	0.88	822	784	1360	0	1603	0	1424	328	42.89	895	0.92
EU857538_SP	977	706	0.39	380	326	556	0	700	0	481	116	43.60	825	0.84
GQ454870_HA16-5	977	688	0.38	344	344	518	0	688	0	644	152	43.71	883	0.90
JX266713_Taiwan-Pum_M_T5	977	215	0.12	109	106	200	0	211	0	207	15	43.30	448	0.46

Table 1.3: Number of reads mapping back to references for sample 13-3023. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	51060	55.04	27724	23336	41818	0	43244	0	48701	12911	43.81	718	0.73
FJ525431_NZRB-M12	977	27733	29.90	14214	13519	18644	0	27295	0	25443	6690	43.30	963	0.99
FJ525434_NZRB-TH30	977	12826	13.83	7343	5483	10794	0	12152	0	12585	2916	43.81	715	0.73
FJ525432_NZRB-G90	977	263	0.28	134	129	214	0	259	0	220	30	43.40	832	0.85
JX266712_Taiwan-Pum_SP_T1	977	262	0.28	145	117	128	0	253	0	132	52	43.30	937	0.96
JX266713_Taiwan-Pum_M_T5	977	257	0.28	131	126	238	0	257	0	251	22	43.30	323	0.33
HM573451_Kpg3	977	137	0.15	67	70	76	0	126	0	70	30	43.81	931	0.95

Table 1.4: Number of reads mapping back to references for sample 13-3040. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	8074	65.00	4309	3765	7054	0	7991	0	6963	2410	43.81	902	0.92
FJ525431_NZRB-M12	977	2868	23.09	1634	1234	996	0	2866	0	2799	853	43.30	951	0.97
FJ525434_NZRB-TH30	977	885	7.13	571	314	574	0	883	0	753	255	43.81	781	0.80
CT-ZA1	977	200	1.61	101	99	148	0	200	0	173	59	43.81	848	0.87
JQ061137_AT-1	977	142	1.14	68	74	106	0	142	0	96	41	43.60	885	0.91
KC517493_FL202-VT	977	82	0.66	47	35	38	0	82	0	82	24	43.09	482	0.49
JQ911664_CT11A	977	39	0.31	15	24	14	0	39	0	38	11	43.60	677	0.69
JQ911663_CT14A	977	30	0.24	23	7	14	0	30	0	30	9	43.50	421	0.43
HM573451_Kpg	977	14	0.11	6	8	8	0	14	0	8	4	43.81	872	0.89
CT-ZA3	977	13	0.10	5	8	10	0	12	0	12	4	43.60	691	0.71

Table 1.5: Number of reads mapping back to references for sample **13-3042**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	37168	50.57	19319	17849	34736	0	24614	0	34914	7436	43.81	797	0.82
FJ525431_NZRB-M12	977	18936	25.77	9875	9061	15246	0	18640	0	18089	3719	43.30	927	0.95
FJ525434_NZRB-TH30	977	8165	11.11	4377	3788	7500	0	6484	0	7905	1553	43.81	791	0.81
JX266713_Taiwan-Pum_M_T5	977	1720	2.34	855	865	1578	0	1699	0	1672	123	43.30	489	0.50
KC517493_FL202-VT	977	1472	2.00	750	722	1360	0	1199	0	1356	217	43.09	798	0.82
JX266712_Taiwan-Pum_SP_T1	977	1457	1.98	728	729	1348	0	1207	0	766	154	43.30	949	0.97
FJ525432_NZRB-G90	977	1342	1.83	677	665	1182	0	1144	0	927	122	43.40	905	0.93
CT-ZA1	977	731	0.99	394	337	622	0	723	0	660	190	43.81	957	0.98
FJ525435_NZRB-M17	977	457	0.62	227	230	370	0	316	0	101	32	43.71	890	0.91
KC262793_L192GR	977	354	0.48	176	178	336	0	316	0	261	31	43.91	726	0.74
JQ061137_AT-1	977	327	0.44	160	167	310	0	306	0	113	37	43.60	755	0.77
CT-ZA3	977	301	0.41	146	155	286	0	201	0	223	47	43.60	898	0.92
HM573451_Kpg3	977	264	0.36	128	136	236	0	221	0	111	40	43.81	941	0.96
JF957196_B301	977	190	0.26	91	99	148	0	175	0	63	18	42.89	753	0.77
GQ454869_HA18-9	977	168	0.23	90	78	144	0	138	0	79	13	42.89	728	0.75

Table 1.6: Number of reads mapping back to references for sample 13-3059. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	59305	44.41	30014	29291	53744	0	54567	0	49478	15734	43.81	963	0.99
FJ525431_NZRB-M12	977	23969	17.95	12935	11034	13282	0	23814	0	21892	6124	43.30	965	0.99
CT-ZA1	977	23777	17.81	12038	11739	22720	0	23673	0	18336	6510	43.81	970	0.99
FJ525434_NZRB-TH30	977	13274	9.94	7549	5725	11368	0	12902	0	11365	3233	43.81	964	0.99
CT-ZA3	977	5054	3.78	2321	2733	4592	0	4780	0	3219	1231	43.60	969	0.99
KC262793_L192GR	977	3233	2.42	1766	1467	2544	0	3136	0	2537	567	43.91	893	0.91
KC517493_FL202-VT	977	2208	1.65	1238	970	1816	0	2128	0	1973	380	43.09	700	0.72
JX266712_Taiwan-Pum_SP_T1	977	745	0.56	413	332	540	0	713	0	306	124	43.30	951	0.97
FJ525432_NZRB-G90	977	485	0.36	248	237	404	0	402	0	256	53	43.40	893	0.91
JX266713_Taiwan-Pum_M_T5	977	233	0.17	115	118	204	0	231	0	222	18	43.30	294	0.30
HM573451_Kpg3	977	188	0.14	94	94	88	0	172	0	78	42	43.81	935	0.96
FJ525435_NZRB-M17	977	146	0.11	71	75	112	0	124	0	61	16	43.71	913	0.93

Table 1.7: Number of reads mapping back to references for sample 13-3062. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JQ061137_AT-1	977	268702	82.17	132810	135892	260512	0	242124	1	73238	33198	43.60	977	1.00
JQ911663_CT14A	977	20946	6.41	11145	9801	19358	0	14716	0	20924	3414	43.50	810	0.83
JQ911664_CT11A	977	16785	5.13	8977	7808	14492	0	13878	0	16680	2880	43.60	698	0.71
DQ151548_T318A	977	8395	2.57	4218	4177	7910	0	8113	0	8353	1075	43.30	613	0.63
FJ525431_NZRB-M12	977	3486	1.07	1752	1734	3110	0	2918	0	1896	455	43.30	899	0.92
FJ525433_NZRB-TH28	977	3296	1.01	1726	1570	3018	0	2369	0	3194	657	43.81	699	0.72
JX266712_Taiwan-Pum_SP_T1	977	1013	0.31	501	512	934	0	848	0	294	141	43.30	953	0.98
U56902_VT	977	976	0.30	491	485	912	0	941	0	907	51	43.60	498	0.51
FJ525434_NZRB-TH30	977	770	0.24	422	348	684	0	654	0	755	156	43.81	707	0.72
KC262793_L192GR	977	459	0.14	237	222	434	0	369	0	291	41	43.91	817	0.84
FJ525432_NZRB-G90	977	338	0.10	170	168	308	0	300	0	165	33	43.40	824	0.84

Table 1.8: Number of reads mapping back to references for sample **13-3067**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	114154	65.60	56932	57222	95046	0	80654	0	34283	16551	43.30	973	1.00
FJ525431_NZRB-M12	977	15990	9.19	8355	7635	11668	0	14386	0	8174	2821	43.30	927	0.95
FJ525433_NZRB-TH28	977	14435	8.30	7492	6943	10506	0	10353	0	14222	2542	43.81	588	0.60
FJ525434_NZRB-TH30	977	8182	4.70	4366	3816	6554	0	6914	0	8065	1549	43.81	604	0.62
FJ525435_NZRB-M17	977	7283	4.19	3648	3635	6910	0	4776	0	5847	723	43.71	896	0.92
FJ525432_NZRB-G90	977	5346	3.07	2680	2666	4770	0	3860	0	4170	664	43.40	944	0.97
JF957196_B301	977	4849	2.79	2406	2443	4688	0	2299	0	4107	527	42.89	820	0.84
JX266713_Taiwan-Pum_M_T5	977	1499	0.86	756	743	1392	0	1459	0	1420	105	43.30	593	0.61
KC517493_FL202-VT	977	551	0.32	278	273	466	0	505	0	520	83	43.09	657	0.67
JQ061137_AT-1	977	374	0.21	184	190	322	0	343	0	123	48	43.60	810	0.83
GQ454869_HA18-9	977	290	0.17	149	141	230	0	205	0	123	30	42.89	804	0.82
CT-ZA1	977	262	0.15	144	118	112	0	255	0	250	64	43.81	715	0.73
KC262793_L192GR	977	194	0.11	96	98	184	0	170	0	137	19	43.91	811	0.83
HM573451_Kpg3	977	182	0.10	90	92	154	0	151	0	68	26	43.81	933	0.95

Table 2.1: Number of reads mapping back to references for sample **13-3105**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	173126	90.93	86372	86754	168676	0	121382	24	49472	25445	43.30	977	1.00
FJ525432_NZRB-G90	977	7837	4.12	3913	3924	7164	0	5892	0	7688	1062	43.40	965	0.99
FJ525435_NZRB-M17	977	4470	2.35	2235	2235	4408	0	4353	0	4200	341	43.71	857	0.88
FJ525431_NZRB-M12	977	3920	2.06	2114	1806	3334	0	3280	0	3845	635	43.30	863	0.88
JF957196_B301	977	243	0.13	116	127	206	0	193	0	89	24	42.89	813	0.83
GQ454869_HA18-9	977	234	0.12	126	108	168	0	185	0	147	31	42.89	756	0.77

Table 2.2: Number of reads mapping back to references for sample **13-3113**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	86698	59.97	42386	44312	61590	0	86375	0	47209	24362	43.30	969	0.99
FJ525431_NZRB-M12	977	50204	34.73	30547	19657	32544	0	48111	0	39065	14115	43.30	963	0.99
FJ525432_NZRB-G90	977	4437	3.07	2177	2260	1002	0	4437	0	4245	1209	43.40	886	0.91
FJ525433_NZRB-TH28	977	2444	1.69	1241	1203	2300	0	2443	0	2130	705	43.81	859	0.88
FJ525435_NZRB-M17	977	203	0.14	120	83	12	0	203	0	184	50	43.71	840	0.86
FJ525434_NZRB-TH30	977	192	0.13	118	74	130	0	192	0	113	51	43.81	817	0.84
GQ454869_HA18-9	977	179	0.12	120	59	12	0	179	0	166	44	42.89	901	0.92
JF957196_B301	977	147	0.10	49	98	10	0	147	0	125	34	42.89	784	0.80

Table 2.3: Number of reads mapping back to references for sample **13-3125**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	24198	94.59	11936	12262	20958	0	23807	0	12591	7118	43.30	974	1.00
FJ525431_NZRB-M12	977	487	1.90	405	82	62	0	486	0	474	140	43.30	695	0.71
CT-ZA1	977	346	1.35	182	164	248	0	345	0	303	102	43.81	943	0.97
FJ525435_NZRB-M17	977	155	0.61	56	99	22	0	155	0	155	44	43.71	499	0.51
FJ525432_NZRB-G90	977	140	0.55	64	76	10	0	140	0	134	41	43.40	851	0.87
KC517493_FL202-VT	977	111	0.43	56	55	46	0	111	0	111	32	43.09	464	0.47
FJ525433_NZRB-TH28	977	31	0.12	19	12	24	0	31	0	27	9	43.81	612	0.63
CT-ZA3	977	30	0.12	14	16	26	0	27	0	24	9	43.60	893	0.91
GQ454869_HA18-9	977	28	0.11	17	11	4	0	28	0	21	8	42.89	911	0.93

Table 2.4: Number of reads mapping back to references for sample **13-3130**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	61977	43.08	28691	33286	38470	0	61269	0	37763	17446	43.81	969	0.99
JX266712_Taiwan-Pum_SP_T1	977	26318	18.29	12978	13340	17814	0	26256	0	14079	7338	43.30	968	0.99
FJ525431_NZRB-M12	977	20098	13.97	12247	7851	12404	0	19423	0	15381	5630	43.30	963	0.99
EU857538_SP	977	11692	8.13	8212	3480	4214	0	11692	0	10498	3281	43.60	896	0.92
JQ798289_A18	977	9378	6.52	3382	5996	2772	0	9378	0	9294	2585	43.30	902	0.92
KC525952_T3	977	6225	4.33	4159	2066	2180	0	6225	0	5874	1728	42.89	940	0.96
GQ454870_HA16-5	977	5475	3.81	2157	3318	1408	0	5475	0	4223	1514	43.71	935	0.96
FJ525432_NZRB-G90	977	1432	1.00	698	734	288	0	1432	0	1317	388	43.40	873	0.89
FJ525433_NZRB-TH28	977	873	0.61	439	434	746	0	873	0	708	247	43.81	943	0.97
FJ525435_NZRB-M17	977	167	0.12	109	58	28	0	167	0	163	44	43.71	776	0.79

Table 2.5: Number of reads mapping back to references for sample **13-3152**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	66928	61.38	32874	34054	46138	0	66720	0	35679	18623	43.30	974	1.00
FJ525431_NZRB-M12	977	36269	33.26	22134	14135	22482	0	34799	0	27896	10120	43.30	965	0.99
FJ525432_NZRB-G90	977	3521	3.23	1762	1759	840	0	3521	0	3341	960	43.40	790	0.81
FJ525433_NZRB-TH28	977	1606	1.47	808	798	1500	0	1605	0	1310	455	43.81	910	0.93
FJ525435_NZRB-M17	977	148	0.14	82	66	6	0	148	0	131	36	43.71	886	0.91
FJ525434_NZRB-TH30	977	144	0.13	81	63	106	0	143	0	95	40	43.81	816	0.84
GQ454869_HA18-9	977	144	0.13	92	52	20	0	144	0	134	36	42.89	799	0.82
JF957196_B301	977	128	0.12	46	82	4	0	128	0	110	30	42.89	837	0.86

Table 2.6: Number of reads mapping back to references for sample **13-3155**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	40063	65.78	19656	20407	32146	0	39956	0	24312	12147	43.30	974	1.00
FJ525431_NZRB-M12	977	18333	30.10	11918	6415	12076	0	17886	0	15819	5563	43.30	964	0.99
FJ525432_NZRB-G90	977	1203	1.98	490	713	264	0	1203	0	1188	364	43.40	763	0.78
FJ525433_NZRB-TH28	977	1017	1.67	518	499	990	0	1017	0	973	311	43.81	859	0.88
FJ525434_NZRB-TH30	977	72	0.12	40	32	56	0	72	0	67	22	43.81	773	0.79

Table 2.7: Number of reads mapping back to references for sample **13-3160**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	33054	57.78	15897	17157	26292	0	32972	0	20196	9735	43.30	974	1.00
FJ525431_NZRB-M12	977	21127	36.93	14237	6890	12544	0	20854	0	17145	6216	43.30	966	0.99
FJ525432_NZRB-G90	977	1402	2.45	602	800	298	0	1402	0	1361	409	43.40	914	0.94
FJ525433_NZRB-TH28	977	1313	2.30	665	648	1276	0	1313	0	1179	391	43.81	942	0.96
FJ525434_NZRB-TH30	977	102	0.18	59	43	80	0	102	0	55	28	43.81	939	0.96
FJ525435_NZRB-M17	977	88	0.15	38	50	10	0	88	0	85	25	43.71	685	0.70
GQ454869_HA18-9	977	72	0.13	41	31	2	0	72	0	68	21	42.89	841	0.86

Table 2.8: Number of reads mapping back to references for sample **13-3161**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	35531	60.47	17377	18154	22814	0	35453	0	19087	9786	43.30	971	0.99
FJ525431_NZRB-M12	977	19406	33.03	11692	7714	11516	0	18740	0	14834	5364	43.30	954	0.98
FJ525432_NZRB-G90	977	2428	4.13	1220	1208	622	0	2428	0	2277	655	43.40	825	0.84
FJ525433_NZRB-TH28	977	789	1.34	390	399	702	0	789	0	713	224	43.81	820	0.84
GQ454869_HA18-9	977	235	0.40	120	115	80	0	235	0	192	62	42.89	920	0.94
FJ525435_NZRB-M17	977	148	0.25	90	58	26	0	148	0	144	39	43.71	595	0.61
JF957196_B301	977	120	0.20	37	83	20	0	120	0	109	29	42.89	648	0.66

Table 3.1: Number of reads mapping back to references for sample **13-3205**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	44897	55.64	20337	24560	25822	0	44836	0	31377	12559	43.81	967	0.99
KC525952_T3	977	9932	12.31	6504	3428	3752	0	9932	0	9642	2750	42.89	941	0.96
EU857538_SP	977	8675	10.75	6230	2445	3142	0	8674	0	7509	2430	43.60	932	0.95
JQ798289_A18	977	6876	8.52	2748	4128	2212	0	6876	0	6807	1876	43.30	890	0.91
JX266712_Taiwan-Pum_SP_T1	977	6779	8.40	3301	3478	4524	0	6662	0	3321	1866	43.30	966	0.99
GQ454870_HA16-5	977	3035	3.76	952	2083	716	0	3034	0	2847	848	43.71	905	0.93
FJ525431_NZRB-M12	977	256	0.32	199	57	38	0	254	0	246	66	43.30	840	0.86
FJ525435_NZRB-M17	977	125	0.15	61	64	14	0	125	0	124	32	43.71	480	0.49

Table 3.2: Number of reads mapping back to references for sample **13-3242**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	76416	73.02	37167	39249	57708	0	76405	0	37683	21351	43.30	971	0.99
FJ525431_NZRB-M12	977	7796	7.45	5322	2474	2468	0	7782	0	6322	2158	43.30	963	0.99
HM573451_Kpg3	977	5636	5.39	2761	2875	3232	0	5617	0	2976	1574	43.81	968	0.99
FJ525433_NZRB-TH28	977	5495	5.25	2741	2754	4170	0	5430	0	4701	1565	43.81	946	0.97
FJ525434_NZRB-TH30	977	2534	2.42	1564	970	1492	0	2524	0	2328	683	43.81	880	0.90
GQ454870_HA16-5	977	1278	1.22	612	666	562	0	1278	0	924	357	43.71	918	0.94
JQ798289_A18	977	1210	1.16	509	701	448	0	1210	0	1206	331	43.30	748	0.77
KC525952_T3	977	1181	1.13	660	521	512	0	1181	0	1117	327	42.89	939	0.96
FJ525432_NZRB-G90	977	1071	1.02	609	462	162	0	1071	0	1036	290	43.40	861	0.88
EU857538_SP	977	884	0.84	547	337	356	0	884	0	685	248	43.60	830	0.85
FJ525435_NZRB-M17	977	882	0.84	467	415	116	0	882	0	872	231	43.71	795	0.81
JF957196_B301	977	170	0.16	46	124	32	0	169	0	155	44	42.89	695	0.71

Table 3.3: Number of reads mapping back to references for sample **13-3251**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	35214	51.01	17028	18186	26104	0	34725	0	19972	10470	43.30	974	1.00
FJ525431_NZRB-M12	977	24168	35.01	16078	8090	14662	0	23349	0	20248	7197	43.30	969	0.99
HM573451_Kpg3	977	3977	5.76	1787	2190	2454	0	3972	0	2337	1182	43.81	970	0.99
FJ525433_NZRB-TH28	977	1449	2.10	730	719	1374	0	1447	0	1330	436	43.81	866	0.89
FJ525432_NZRB-G90	977	1150	1.67	491	659	290	0	1150	0	1101	340	43.40	888	0.91
KC525952_T3	977	1140	1.65	784	356	562	0	1140	0	1120	338	42.89	895	0.92
JQ798289_A18	977	725	1.05	308	417	238	0	725	0	719	214	43.30	853	0.87
EU857538_SP	977	337	0.49	234	103	80	0	337	0	277	99	43.60	830	0.85
FJ525435_NZRB-M17	977	314	0.45	147	167	78	0	314	0	307	91	43.71	723	0.74
GQ454870_HA16-5	977	288	0.42	101	187	80	0	288	0	283	85	43.71	852	0.87
FJ525434_NZRB-TH30	977	105	0.15	67	38	70	0	105	0	72	31	43.81	897	0.92
GQ454869_HA18-9	977	72	0.10	39	33	10	0	72	0	61	20	42.89	892	0.91

Table 3.4: Number of reads mapping back to references for sample **13-3255**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	92644	29.09	34429	58215	52588	0	91535	0	70922	27380	43.81	969	0.99
EU857538_SP	977	81032	25.44	48273	32759	45606	0	80938	0	49270	24094	43.60	969	0.99
KC525952_T3	977	65465	20.55	43145	22320	28698	0	65464	0	64319	19503	42.89	957	0.98
GQ454870_HA16-5	977	54231	17.03	25019	29212	32058	0	54054	0	52243	16253	43.71	891	0.91
JX266712_Taiwan-Pum_SP_T1	977	14847	4.66	7231	7616	10768	0	14846	0	8242	4395	43.30	967	0.99
JQ798289_A18	977	4335	1.36	1316	3019	1194	0	4335	0	4309	1262	43.30	910	0.93
FJ525431_NZRB-M12	977	2433	0.76	1650	783	898	0	2430	0	2156	721	43.30	935	0.96
FJ525433_NZRB-TH28	977	2125	0.67	1180	945	1592	0	2107	0	1879	635	43.81	828	0.85
FJ525434_NZRB-TH30	977	753	0.24	512	241	450	0	752	0	733	220	43.81	764	0.78
FJ525435_NZRB-M17	977	345	0.11	144	201	96	0	345	0	343	100	43.71	547	0.56

Table 3.5: Number of reads mapping back to references for sample **13-3257**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	19119	39.53	9294	9825	14762	0	17104	0	8025	3927	43.81	969	0.99
JX266712_Taiwan-Pum_SP_T1	977	13540	28.00	6676	6864	10660	0	11385	0	4747	2804	43.30	967	0.99
KC525952_T3	977	4700	9.72	2692	2008	3496	0	4595	0	4538	972	42.89	937	0.96
FJ525431_NZRB-M12	977	2880	5.96	1721	1159	2036	0	2588	0	2096	676	43.30	934	0.96
EU857538_SP	977	2695	5.57	1531	1164	1744	0	2613	0	1522	560	43.60	900	0.92
JQ798289_A18	977	2361	4.88	990	1371	1698	0	2230	0	2176	513	43.30	856	0.88
GQ454870_HA16-5	977	1584	3.28	743	841	870	0	1519	0	1087	395	43.71	848	0.87
FJ525435_NZRB-M17	977	608	1.26	297	311	532	0	510	0	540	83	43.71	653	0.67
FJ525432_NZRB-G90	977	538	1.11	270	268	442	0	427	0	361	86	43.40	943	0.97
JF957196_B301	977	171	0.35	84	87	156	0	123	0	142	19	42.89	773	0.79
JQ061137_AT-1	977	58	0.12	29	29	56	0	58	0	28	10	43.60	860	0.88

Table 3.6: Number of reads mapping back to references for sample **13-3260**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	92597	60.79	43603	48994	61426	0	91143	0	62312	26164	43.81	974	1.00
JX266712_Taiwan-Pum_SP_T1	977	16262	10.68	7851	8411	12714	0	16235	0	7942	4553	43.30	966	0.99
KC525952_T3	977	15174	9.96	10209	4965	5148	0	15174	0	14692	4196	42.89	970	0.99
JQ798289_A18	977	11681	7.67	4296	7385	3786	0	11680	0	11596	3243	43.30	956	0.98
EU857538_SP	977	10211	6.70	7160	3051	3258	0	10211	0	8472	2870	43.60	960	0.98
GQ454870_HA16-5	977	5161	3.39	1535	3626	1194	0	5157	0	4795	1440	43.71	934	0.96
FJ525431_NZRB-M12	977	788	0.52	639	149	138	0	788	0	774	215	43.30	873	0.89
FJ525435_NZRB-M17	977	208	0.14	97	111	26	0	208	0	203	55	43.71	569	0.58
FJ525432_NZRB-G90	977	158	0.10	85	73	16	0	158	0	149	43	43.40	891	0.91

Table 3.7: Number of reads mapping back to references for sample **13-3267**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	92597	60.79	43603	48994	61426	0	91143	0	62312	26164	43.81	974	1.00
JX266712_Taiwan-Pum_SP_T1	977	16262	10.68	7851	8411	12714	0	16235	0	7942	4553	43.30	966	0.99
KC525952_T3	977	15174	9.96	10209	4965	5148	0	15174	0	14692	4196	42.89	970	0.99
JQ798289_A18	977	11681	7.67	4296	7385	3786	0	11680	0	11596	3243	43.30	956	0.98
EU857538_SP	977	10211	6.70	7160	3051	3258	0	10211	0	8472	2870	43.60	960	0.98
GQ454870_HA16-5	977	5161	3.39	1535	3626	1194	0	5157	0	4795	1440	43.71	934	0.96
FJ525431_NZRB-M12	977	788	0.52	639	149	138	0	788	0	774	215	43.30	873	0.89
FJ525435_NZRB-M17	977	208	0.14	97	111	26	0	208	0	203	55	43.71	569	0.58
FJ525432_NZRB-G90	977	158	0.10	85	73	16	0	158	0	149	43	43.40	891	0.91

*Table 3.8: Number of reads mapping back to references for sample **13-3271**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	17844	37.68	8934	8910	14104	0	17768	0	9817	5272	43.30	970	0.99
GQ454870_HA16-5	977	13164	27.80	6791	6373	9882	0	13154	0	10832	3931	43.71	929	0.95
HM573451_Kpg3	977	6204	13.10	2875	3329	3268	0	6197	1	4256	1824	43.81	977	1.00
FJ525431_NZRB-M12	977	3938	8.32	2431	1507	2564	0	3864	0	3084	1161	43.30	951	0.97
EU857538_SP	977	2246	4.74	1286	960	1096	0	2245	0	1171	665	43.60	932	0.95
JQ798289_A18	977	2129	4.50	741	1388	752	0	2129	0	2124	626	43.30	862	0.88
KC525952_T3	977	618	1.31	383	235	194	0	618	0	491	180	42.89	949	0.97
FJ525432_NZRB-G90	977	460	0.97	219	241	144	0	460	0	422	134	43.40	837	0.86
JF957196_B301	977	253	0.53	58	195	84	0	253	0	252	72	42.89	570	0.58
FJ525435_NZRB-M17	977	199	0.42	91	108	54	0	199	0	199	58	43.71	528	0.54
FJ525433_NZRB-TH28	977	174	0.37	85	89	152	0	174	0	169	52	43.81	703	0.72

*Table 4.1: Number of reads mapping back to references for sample **13-3305**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	23563	59.55	11702	11861	19028	0	23487	0	12700	6962	43.30	973	1.00
FJ525433_NZRB-TH28	977	6073	15.35	3293	2780	5114	0	6023	0	5222	1811	43.81	865	0.89
FJ525431_NZRB-M12	977	3162	7.99	2036	1126	1228	0	3154	0	2831	935	43.30	941	0.96
HM573451_Kpg3	977	2815	7.11	1319	1496	1958	0	2811	0	1677	830	43.81	969	0.99
FJ525434_NZRB-TH30	977	871	2.20	573	298	532	0	869	0	795	251	43.81	773	0.79
KC525952_T3	977	822	2.08	452	370	554	0	822	0	809	240	42.89	780	0.80
JQ798289_A18	977	624	1.58	278	346	290	0	624	0	621	183	43.30	887	0.91
EU857538_SP	977	489	1.24	304	185	228	0	489	0	429	144	43.60	810	0.83
FJ525432_NZRB-G90	977	448	1.13	221	227	108	0	448	0	440	131	43.40	714	0.73
GQ454870_HA16-5	977	279	0.71	109	170	122	0	279	0	271	83	43.71	749	0.77
JQ061137_AT-1	977	220	0.56	101	119	178	0	220	0	136	65	43.60	911	0.93
JQ911663_CT14A	977	65	0.16	40	25	46	0	65	0	65	19	43.50	416	0.43
JQ911664_CT11A	977	46	0.12	20	26	28	0	46	0	46	13	43.60	431	0.44

Table 4.2: Number of reads mapping back to references for sample **13-3309**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	33934	47.94	15456	18478	21740	0	33934	0	22011	10022	43.81	970	0.99
KC525952_T3	977	6336	8.95	4433	1903	2548	0	6335	0	6166	1855	42.89	946	0.97
EU857538_SP	977	5525	7.81	3620	1905	2310	0	5525	0	4384	1630	43.60	930	0.95
JQ061137_AT-1	977	5022	7.09	2313	2709	3816	0	5022	0	3261	1484	43.60	959	0.98
JQ798289_A18	977	4660	6.58	1684	2976	1610	0	4660	0	4617	1374	43.30	938	0.96
FJ525433_NZRB-TH28	977	4170	5.89	2313	1857	3366	0	4119	0	3580	1242	43.81	872	0.89
GQ454870_HA16-5	977	2884	4.07	991	1893	924	0	2881	0	2633	856	43.71	935	0.96
FJ525431_NZRB-M12	977	2013	2.84	1095	918	802	0	2011	0	1747	598	43.30	953	0.98
JX266712_Taiwan-Pum_SP_T1	977	1733	2.45	924	809	1164	0	1732	0	1029	514	43.30	941	0.96
JQ911663_CT14A	977	1024	1.45	747	277	450	0	1024	0	1024	300	43.50	491	0.50
JQ911664_CT11A	977	687	0.97	341	346	246	0	687	0	683	205	43.60	899	0.92
FJ525434_NZRB-TH30	977	640	0.90	439	201	388	0	640	0	560	185	43.81	769	0.79
KC262793_L192GR	977	636	0.90	317	319	394	0	626	0	594	189	43.91	936	0.96
KC517493_FL202-VT	977	469	0.66	300	169	288	0	469	0	466	138	43.09	597	0.61
AB0463981_NUagA	977	280	0.40	84	196	134	0	280	0	275	82	43.50	785	0.80
GQ454869_HA18-9	977	225	0.32	73	152	104	0	225	0	216	66	42.89	874	0.89
DQ151548_T318A	977	200	0.28	105	95	64	0	200	0	200	59	43.30	538	0.55
KC517492_FS703-VT	977	147	0.21	74	73	70	0	147	0	145	44	43.09	598	0.61
FJ525432_NZRB-G90	977	104	0.15	52	52	26	0	104	0	85	30	43.40	880	0.90

Table 4.3: Number of reads mapping back to references for sample **13-3312**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	85633	54.08	41980	43653	71154	0	80060	0	33739	19966	43.30	973	1.00
FJ525433_NZRB-TH28	977	28515	18.01	15209	13306	23630	0	26298	0	25658	7506	43.81	779	0.80
FJ525431_NZRB-M12	977	24994	15.78	14175	10819	16370	0	24545	0	16572	6127	43.30	963	0.99
FJ525434_NZRB-TH30	977	12942	8.17	7336	5606	10914	0	12507	0	12758	3017	43.81	740	0.76
FJ525435_NZRB-M17	977	1872	1.18	1001	871	1322	0	1695	0	1666	334	43.71	871	0.89
FJ525432_NZRB-G90	977	1852	1.17	934	918	1438	0	1742	0	1477	312	43.40	910	0.93
JF957196_B301	977	1619	1.02	754	865	1390	0	1241	0	1484	237	42.89	814	0.83
HM573451_Kpg3	977	180	0.11	88	92	136	0	173	0	73	34	43.81	943	0.96

Table 4.4: Number of reads mapping back to references for sample **13-3321**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	59728	46.27	29507	30221	48710	0	57253	0	22557	13868	43.30	970	0.99
FJ525433_NZRB-TH28	977	29068	22.52	16005	13063	22880	0	26385	0	27232	7593	43.81	732	0.75
FJ525431_NZRB-M12	977	22295	17.27	12475	9820	14642	0	21898	0	16836	5441	43.30	958	0.98
FJ525434_NZRB-TH30	977	13068	10.12	7490	5578	10920	0	12621	0	12921	3048	43.81	711	0.73
FJ525432_NZRB-G90	977	3176	2.46	1686	1490	2246	0	3000	0	2937	628	43.40	909	0.93
JF957196_B301	977	907	0.70	433	474	814	0	858	0	805	124	42.89	834	0.85
FJ525435_NZRB-M17	977	350	0.27	176	174	282	0	249	0	193	39	43.71	885	0.91
JX266713_Taiwan-Pum_M_T5	977	139	0.11	71	68	128	0	139	0	135	11	43.30	349	0.36
HM573451_Kpg3	977	131	0.10	64	67	94	0	129	0	51	26	43.81	920	0.94

Table 4.5: Number of reads mapping back to references for sample **13-3324**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JQ061137_AT-1	977	26559	30.50	11810	14749	20024	0	26558	0	17289	7861	43.60	968	0.99
JX266712_Taiwan-Pum_SP_T1	977	13896	15.96	7276	6620	10288	0	13892	0	8341	4129	43.30	966	0.99
HM573451_Kpg3	977	10289	11.82	4616	5673	6456	0	10289	0	5303	3037	43.81	966	0.99
EU857538_SP	977	6331	7.27	3571	2760	4204	0	6331	0	4524	1879	43.60	920	0.94
KC262793_L192GR	977	4426	5.08	1951	2475	2742	0	4372	0	4060	1317	43.91	948	0.97
JQ911663_CT14A	977	3951	4.54	2850	1101	1920	0	3951	0	3950	1160	43.50	718	0.73
JQ911664_CT11A	977	3404	3.91	2063	1341	1134	0	3403	0	3367	1016	43.60	946	0.97
GQ454869_HA18-9	977	2665	3.06	903	1762	1290	0	2665	0	2555	787	42.89	956	0.98
KC517493_FL202-VT	977	2403	2.76	1605	798	1378	0	2403	0	2396	708	43.09	622	0.64
KC525952_T3	977	2385	2.74	1549	836	1218	0	2385	0	2167	706	42.89	944	0.97
AB0463981_NUagA	977	2132	2.45	745	1387	1198	0	2132	0	2086	627	43.50	863	0.88
JQ798289_A18	977	1724	1.98	707	1017	858	0	1724	0	1674	508	43.30	941	0.96
GQ454870_HA16-5	977	1676	1.92	756	920	948	0	1676	0	1529	498	43.71	929	0.95
DQ151548_T318A	977	989	1.14	558	431	310	0	989	0	984	294	43.30	890	0.91
KC517492_FS703-VT	977	820	0.94	453	367	392	0	820	0	818	246	43.09	574	0.59
FJ525431_NZRB-M12	977	799	0.92	479	320	300	0	799	0	734	236	43.30	947	0.97
FJ525433_NZRB-TH28	977	719	0.83	410	309	494	0	713	0	672	215	43.81	813	0.83
FJ525432_NZRB-G90	977	693	0.80	314	379	226	0	693	0	644	203	43.40	773	0.79
FJ525435_NZRB-M17	977	511	0.59	247	264	158	0	511	0	490	151	43.71	814	0.83
JF957196_B301	977	434	0.50	160	274	246	0	434	0	417	123	42.89	737	0.75
FJ525434_NZRB-TH30	977	105	0.12	73	32	62	0	105	0	89	30	43.81	773	0.79

Table 4.6: Number of reads mapping back to references for sample **13-3343**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	87614	70.01	42087	45527	58956	0	86320	0	57595	24572	43.81	969	0.99
EU857538_SP	977	16252	12.99	11166	5086	6402	0	16252	0	14991	4580	43.60	902	0.92
JQ798289_A18	977	10360	8.28	3865	6495	3288	0	10360	0	10312	2843	43.30	888	0.91
GQ454870_HA16-5	977	5826	4.66	2357	3469	1254	0	5826	0	3786	1584	43.71	914	0.94
FJ525433_NZRB-TH28	977	2081	1.66	1051	1030	1632	0	2036	0	1774	589	43.81	927	0.95
KC525952_T3	977	1598	1.28	1102	496	194	0	1598	0	1239	408	42.89	950	0.97
FJ525431_NZRB-M12	977	756	0.60	418	338	256	0	756	0	669	214	43.30	938	0.96
FJ525434_NZRB-TH30	977	537	0.43	352	185	326	0	537	0	487	142	43.81	747	0.76

Table 4.7: Number of reads mapping back to references for sample **13-3347**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	140745	49.40	67818	72927	127320	0	121696	0	52769	1311	45.21	969	0.99
KC525952_T3	977	67621	23.74	36033	31588	57658	0	65566	0	65097	639	45.32	949	0.97
EU857538_SP	977	39549	13.88	21278	18271	33356	0	35262	0	22067	394	45.38	952	0.97
JQ798289_A18	977	16296	5.72	7619	8677	13856	0	15167	0	15243	152	45.59	911	0.93
GQ454870_HA16-5	977	15389	5.40	7238	8151	11358	0	14974	0	12169	198	45.27	897	0.92
FJ525433_NZRB-TH28	977	1875	0.66	948	927	1758	0	1703	0	1782	19	44.83	907	0.93
FJ525431_NZRB-M12	977	1634	0.57	832	802	1414	0	1619	0	925	15	44.79	941	0.96
FJ525434_NZRB-TH30	977	634	0.22	328	306	592	0	609	0	607	7	44.94	715	0.73
FJ525432_NZRB-G90	977	406	0.14	202	204	370	0	380	0	267	2	44.96	873	0.89

*Table 4.8: Number of reads mapping back to references for sample **13-3362**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JQ061137_AT-1	977	38160	48.40	16930	21230	28616	0	38159	0	25310	11128	43.60	970	0.99
JX266712_Taiwan-Pum_SP_T1	977	12390	15.72	6457	5933	8898	0	12388	0	7326	3622	43.30	964	0.99
JQ911663_CT14A	977	5193	6.59	3732	1461	2340	0	5191	0	5193	1486	43.50	499	0.51
JQ911664_CT11A	977	4523	5.74	2900	1623	1338	0	4523	0	4485	1330	43.60	937	0.96
KC262793_L192GR	977	3523	4.47	1512	2011	2226	0	3498	0	3204	1035	43.91	951	0.97
GQ454869_HA18-9	977	2532	3.21	892	1640	1204	0	2532	0	2379	734	42.89	954	0.98
EU857538_SP	977	2308	2.93	1230	1078	1656	0	2308	0	1586	674	43.60	846	0.87
AB0463981_NUagA	977	1749	2.22	649	1100	896	0	1749	0	1682	504	43.50	825	0.84
KC517493_FL202-VT	977	1624	2.06	1055	569	908	0	1624	0	1611	467	43.09	602	0.62
HM573451_Kpg3	977	1129	1.43	450	679	588	0	1129	0	934	324	43.81	958	0.98
DQ151548_T318A	977	937	1.19	516	421	194	0	937	0	930	273	43.30	806	0.82
KC517492_FS703-VT	977	775	0.98	457	318	340	0	775	0	768	228	43.09	637	0.65
FJ525432_NZRB-G90	977	774	0.98	345	429	168	0	774	0	697	221	43.40	946	0.97
FJ525431_NZRB-M12	977	641	0.81	384	257	190	0	641	0	589	184	43.30	883	0.90
GQ454870_HA16-5	977	589	0.75	292	297	404	0	589	0	522	173	43.71	930	0.95
FJ525433_NZRB-TH28	977	511	0.65	295	216	364	0	508	0	471	151	43.81	797	0.82
JF957196_B301	977	465	0.59	164	301	214	0	465	0	443	128	42.89	758	0.78
FJ525435_NZRB-M17	977	392	0.50	199	193	74	0	392	0	374	114	43.71	813	0.83
JQ798289_A18	977	193	0.24	86	107	78	0	193	0	183	55	43.30	889	0.91
KC525952_T3	977	120	0.15	75	45	26	0	120	0	103	35	42.89	745	0.76
AF260651_T30	977	80	0.10	74	6	8	0	80	0	80	22	43.19	316	0.32

Table 5.1: Number of reads mapping back to references for sample **13-3405**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	37379	34.65	18651	18728	21420	0	37177	0	17709	10457	43.81	969	0.99
KC262793_L192GR	977	20262	18.78	9970	10292	10644	0	19682	0	18306	5755	43.91	939	0.96
AB0463981_NUagA	977	12503	11.59	4519	7984	4944	0	12490	0	12426	3405	43.50	820	0.84
KC517493_FL202-VT	977	11350	10.52	7670	3680	5690	0	11350	0	11296	3136	43.09	937	0.96
GQ454870_HA16-5	977	6143	5.70	2844	3299	2430	0	6143	0	4651	1745	43.71	886	0.91
JQ798289_A18	977	5628	5.22	2353	3275	2348	0	5628	0	5361	1559	43.30	929	0.95
KC517492_FS703-VT	977	5033	4.67	2520	2513	1990	0	5033	0	4928	1437	43.09	659	0.67
EU857538_SP	977	3586	3.32	2203	1383	1266	0	3586	0	2916	1009	43.60	961	0.98
KC525952_T3	977	1739	1.61	1023	716	380	0	1739	0	1531	468	42.89	948	0.97
DQ151548_T318A	977	1589	1.47	895	694	418	0	1589	0	1520	445	43.30	918	0.94
JX266712_Taiwan-Pum_SP_T1	977	1266	1.17	602	664	792	0	1265	0	733	350	43.30	955	0.98
JQ911663_CT14A	977	279	0.26	219	60	24	0	279	0	279	69	43.50	387	0.40
CT-ZA1	977	229	0.21	127	102	78	0	229	0	218	65	43.81	870	0.89
JQ911664_CT11A	977	225	0.21	61	164	16	0	225	0	189	53	43.60	874	0.89
FJ525433_NZRB-TH28	977	182	0.17	92	90	130	0	181	0	173	51	43.81	699	0.72
FJ525431_NZRB-M12	977	168	0.16	100	68	50	0	168	0	118	46	43.30	917	0.94

Table 5.2: Number of reads mapping back to references for sample **13-3406**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
EU857538_SP	977	48414	46.91	27662	20752	37140	0	48414	0	37794	14381	43.60	965	0.99
HM573451_Kpg3	977	16914	16.39	7721	9193	7122	0	16911	0	14078	4665	43.81	975	1.00
JX266712_Taiwan-Pum_SP_T1	977	10686	10.35	5221	5465	9084	0	10601	0	5728	3145	43.30	966	0.99
JQ798289_A18	977	10208	9.89	3402	6806	3810	0	10208	0	10145	2987	43.30	850	0.87
GQ454870_HA16-5	977	10200	9.88	4324	5876	5420	0	10200	0	10040	3033	43.71	913	0.93
AY170468_T36	981	4104	3.98	1820	2284	3150	0	4104	0	4054	1219	43.63	912	0.93
KC525952_T3	977	1569	1.52	873	696	376	0	1569	0	1219	456	42.89	937	0.96
FJ525431_NZRB-M12	977	263	0.25	199	64	60	0	263	0	259	76	43.30	599	0.61
EU937521_T36	981	225	0.22	139	86	108	0	225	0	192	66	43.43	944	0.96
KC262793_L192GR	977	123	0.12	68	55	82	0	122	0	117	36	43.91	840	0.86
FJ525435_NZRB-M17	977	118	0.11	63	55	22	0	118	0	116	34	43.71	704	0.72
KC517493_FL202-VT	977	104	0.10	61	43	64	0	104	0	103	30	43.09	570	0.58

Table 5.3: Number of reads mapping back to references for sample **13-3410**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	23829	46.16	11709	12120	18948	0	23753	0	13011	7054	43.30	971	0.99
FJ525433_NZRB-TH28	977	16374	31.72	8980	7394	13638	0	16261	0	13882	4882	43.81	906	0.93
FJ525431_NZRB-M12	977	6607	12.80	4022	2585	2466	0	6601	0	6028	1964	43.30	963	0.99
FJ525434_NZRB-TH30	977	1531	2.97	989	542	982	0	1530	0	1332	442	43.81	904	0.93
AY170468_T36	981	1178	2.28	566	612	1026	0	1114	0	584	347	43.63	951	0.97
HM573451_Kpg3	977	701	1.36	328	373	474	0	700	0	439	207	43.81	938	0.96
FJ525432_NZRB-G90	977	667	1.29	326	341	128	0	667	0	652	195	43.40	887	0.91
KC525952_T3	977	201	0.39	110	91	130	0	201	0	182	59	42.89	823	0.84
JQ798289_A18	977	136	0.26	57	79	46	0	136	0	135	40	43.30	790	0.81
GQ454870_HA16-5	977	102	0.20	54	48	58	0	102	0	85	30	43.71	764	0.78
FJ525435_NZRB-M17	977	83	0.16	36	47	6	0	83	0	82	24	43.71	736	0.75
EU857538_SP	977	81	0.16	50	31	42	0	81	0	62	24	43.60	795	0.81

Table 5.4: Number of reads mapping back to references for sample **13-3416**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Standard deviation	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	59027	56.17	28957	30070	43070	95.0618	0	58567	0	28882	16560	43.30	974	1.00
FJ525433_NZRB-TH28	977	11835	11.26	6074	5761	9670	98.97245	0	11638	0	9458	3375	43.81	953	0.98
HM573451_Kpg	977	8147	7.75	3771	4376	4422	88.36484	0	8142	0	4659	2279	43.81	965	0.99
FJ525431_NZRB-M12	977	8060	7.67	5482	2578	2492	75.10433	0	8029	0	6092	2255	43.30	965	0.99
FJ525434_NZRB-TH30	977	3910	3.72	2434	1476	2554	136.4062	0	3904	0	3268	1047	43.81	920	0.94
CT-ZA1	977	3309	3.15	1795	1514	2294	105.6101	0	3289	0	2737	932	43.81	971	0.99
JQ798289_A18	977	2024	1.93	876	1148	756	35.70152	0	2024	0	1992	561	43.30	896	0.92
KC525952_T3	977	1972	1.88	1169	803	826	37.89403	0	1972	0	1863	553	42.89	917	0.94
GQ454870_HA16-5	977	1498	1.43	684	814	500	70.43395	0	1498	0	1298	423	43.71	919	0.94
KC517493_FL202-VT	977	1399	1.33	813	586	482	26.84184	0	1399	0	1397	391	43.09	697	0.71
FJ525432_NZRB-G90	977	1155	1.10	618	537	184	25.31104	0	1155	0	1121	317	43.40	859	0.88
EU857538_SP	977	1101	1.05	755	346	340	74.12611	0	1101	0	901	309	43.60	859	0.88
FJ525435_NZRB-M17	977	742	0.71	355	387	94	22.61883	0	742	0	727	196	43.71	778	0.80
CT-ZA3	977	282	0.27	122	160	198	100.0324	0	264	0	138	80	43.60	918	0.94
KC262793_L192GR	977	199	0.19	111	88	58	32.38891	0	199	0	152	55	43.91	849	0.87
JF957196_B301	977	169	0.16	54	115	34	32.99219	0	169	0	156	42	42.89	780	0.80

Table 5.5: Number of reads mapping back to references for sample **13-3417**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
EU857538_SP	977	111004	48.00	64348	46656	84146	0	110442	0	83719	29732	43.60	930	0.95
HM573451_Kpg3	977	58358	25.24	23364	34994	37286	0	57951	0	41032	13614	43.81	967	0.99
JQ798289_A18	977	28220	12.20	10718	17502	13694	0	28068	0	28022	7351	43.30	909	0.93
GQ454870_HA16-5	977	22913	9.91	10553	12360	13496	0	22908	0	21835	6294	43.71	925	0.95
KC525952_T3	977	5213	2.25	2899	2314	2878	0	5197	0	3496	1164	42.89	954	0.98
JX266712_Taiwan-Pum_SP_T1	977	2883	1.25	1417	1466	2404	0	2772	0	1224	649	43.30	958	0.98
FJ525433_NZRB-TH28	977	1077	0.47	573	504	884	0	1024	0	994	276	43.81	761	0.78
FJ525431_NZRB-M12	977	642	0.28	350	292	382	0	638	0	474	154	43.30	931	0.95
FJ525434_NZRB-TH30	977	441	0.19	254	187	366	0	440	0	439	103	43.81	579	0.59

Table 5.6: Number of reads mapping back to references for sample **13-3418**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	312536	57.67	148950	163586	232912	0	276774	0	161787	72521	43.81	967	0.99
EU857538_SP	977	122110	22.53	72621	49489	92386	0	121859	0	100128	26153	43.60	950	0.97
JQ798289_A18	977	37667	6.95	16772	20895	27158	0	36043	0	36736	8194	43.30	897	0.92
JX266712_Taiwan-Pum_SP_T1	977	37040	6.83	18313	18727	32064	0	35178	0	15032	7457	43.30	969	0.99
KC525952_T3	977	13376	2.47	7364	6012	9840	0	13353	0	11731	2194	42.89	954	0.98
GQ454870_HA16-5	977	10697	1.97	4002	6695	5866	0	10662	0	10235	2552	43.71	918	0.94
FJ525431_NZRB-M12	977	2388	0.44	1318	1070	1736	0	2353	0	1611	496	43.30	928	0.95
FJ525432_NZRB-G90	977	2009	0.37	1034	975	1708	0	1896	0	1357	307	43.40	847	0.87
FJ525433_NZRB-TH28	977	1204	0.22	629	575	992	0	1177	0	1172	261	43.81	639	0.65
FJ525435_NZRB-M17	977	1112	0.21	564	548	932	0	1064	0	877	157	43.71	904	0.93
FJ525434_NZRB-TH30	977	698	0.13	375	323	612	0	675	0	687	151	43.81	571	0.58
JF957196_B301	977	569	0.10	281	288	544	0	535	0	547	62	42.89	804	0.82

*Table 5.7: Number of reads mapping back to references for sample **13-3421**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	19967	39.90	9700	10267	15654	0	19641	0	11069	5919	43.30	967	0.99
FJ525431_NZRB-M12	977	14638	29.25	9084	5554	9970	0	14178	0	12033	4343	43.30	964	0.99
HM573451_Kpg3	977	4598	9.19	2079	2519	3036	0	4594	0	2915	1360	43.81	969	0.99
AY170468_T36	981	4595	9.18	2059	2536	3718	0	4595	0	4544	1369	43.63	883	0.90
EU857538_SP	977	1295	2.59	771	524	834	0	1295	0	1023	385	43.60	844	0.86
FJ525432_NZRB-G90	977	1085	2.17	520	565	394	0	1085	0	1047	319	43.40	846	0.86
KC525952_T3	977	997	1.99	615	382	564	0	997	0	962	294	42.89	895	0.92
JQ798289_A18	977	971	1.94	384	587	388	0	971	0	951	287	43.30	876	0.90
FJ525433_NZRB-TH28	977	678	1.35	328	350	616	0	676	0	608	203	43.81	859	0.88
GQ454870_HA16-5	977	411	0.82	191	220	224	0	411	0	381	122	43.71	874	0.89
FJ525435_NZRB-M17	977	193	0.39	99	94	26	0	193	0	191	56	43.71	701	0.72
EU937521_T36	981	147	0.29	97	50	54	0	147	0	118	43	43.43	935	0.95
KC517493_FL202-VT	977	144	0.29	75	69	80	0	144	0	144	42	43.09	508	0.52
CT-ZA1	977	120	0.24	81	39	58	0	120	0	120	35	43.81	551	0.56
FJ525434_NZRB-TH30	977	84	0.17	53	31	60	0	84	0	67	25	43.81	863	0.88

Table 5.8: Number of reads mapping back to references for sample **13-3422**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	86402	33.43	38666	47736	69120	0	84478	0	47478	19819	43.81	969	0.99
GQ454870_HA16-5	977	33190	12.84	16476	16714	23956	0	33019	0	25876	8923	43.71	931	0.95
EU857538_SP	977	32448	12.55	18340	14108	22762	0	32192	0	20605	8233	43.60	946	0.97
KC525952_T3	977	28212	10.91	16463	11749	18742	0	28077	0	26678	6705	42.89	949	0.97
FJ525433_NZRB-TH28	977	20784	8.04	10482	10302	18736	0	20522	0	17564	5418	43.81	960	0.98
JX266712_Taiwan-Pum_SP_T1	977	18382	7.11	9061	9321	15408	0	18225	0	7067	4272	43.30	967	0.99
JQ798289_A18	977	14878	5.76	6779	8099	11568	0	14828	0	13381	3001	43.30	937	0.96
FJ525431_NZRB-M12	977	12125	4.69	6740	5385	7834	0	12083	0	9431	2946	43.30	960	0.98
FJ525434_NZRB-TH30	977	8413	3.25	4616	3797	7474	0	8335	0	7567	2028	43.81	937	0.96
FJ525432_NZRB-G90	977	2047	0.79	1099	948	1546	0	2008	0	1201	357	43.40	941	0.96
FJ525435_NZRB-M17	977	424	0.16	215	209	328	0	375	0	253	58	43.71	802	0.82
JX266713_Taiwan-Pum_M_T5	977	349	0.14	176	173	328	0	349	0	342	31	43.30	461	0.47
JF957196_B301	977	307	0.12	138	169	254	0	299	0	277	43	42.89	718	0.73

Table 5.9: Number of reads mapping back to references for sample **13-3425**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	46272	53.28	21438	24834	27702	0	46263	0	31572	13670	43.81	969	0.99
KC525952_T3	977	9953	11.46	6898	3055	3998	0	9953	0	9656	2909	42.89	924	0.95
GQ454870_HA16-5	977	8805	10.14	3468	5337	3398	0	8805	0	8718	2626	43.71	871	0.89
EU857538_SP	977	7949	9.15	5453	2496	3094	0	7949	0	6337	2348	43.60	878	0.90
JQ798289_A18	977	6804	7.83	2402	4402	2634	0	6804	0	6775	2005	43.30	944	0.97
FJ525433_NZRB-TH28	977	2707	3.12	1488	1219	2226	0	2680	0	2367	807	43.81	848	0.87
JX266712_Taiwan-Pum_SP_T1	977	2527	2.91	1260	1267	1896	0	2524	0	1564	747	43.30	973	0.99
FJ525431_NZRB-M12	977	1273	1.47	751	522	410	0	1272	0	1118	377	43.30	949	0.97
FJ525434_NZRB-TH30	977	455	0.52	304	151	268	0	455	0	429	131	43.81	822	0.84

Table 5.10: Number of reads mapping back to references for sample **13-3427**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	21227	37.13	10440	10787	16554	0	20463	0	7957	4969	43.30	963	0.99
AY170468_T36	981	15593	27.28	7073	8520	11888	0	15593	0	15059	4365	43.63	947	0.97
FJ525431_NZRB-M12	977	12491	21.85	7191	5300	9740	0	11974	0	9169	3099	43.30	965	0.99
FJ525432_NZRB-G90	977	2857	5.00	1381	1476	1760	0	2823	0	2344	674	43.40	918	0.94
EU937521_T36	981	1523	2.66	832	691	1134	0	1523	0	1318	420	43.43	952	0.97
FJ525435_NZRB-M17	977	1000	1.75	491	509	768	0	972	0	794	147	43.71	917	0.94
JQ798289_A18	977	751	1.31	364	387	650	0	739	0	657	94	43.30	742	0.76
HM573451_Kpg3	977	415	0.73	209	206	330	0	414	0	160	84	43.81	945	0.97
JF957196_B301	977	370	0.65	166	204	284	0	348	0	311	61	42.89	918	0.94
FJ525433_NZRB-TH28	977	253	0.44	133	120	200	0	245	0	204	69	43.81	739	0.76
KC517487_FS703-T36	981	101	0.18	60	41	50	0	101	0	96	25	43.63	628	0.64
KC525952_T3	977	81	0.14	45	36	58	0	81	0	72	18	42.89	646	0.66
JX266713_Taiwan-Pum_M_T5	977	76	0.13	38	38	68	0	68	0	35	5	43.30	601	0.62
EU857538_SP	977	63	0.11	33	30	48	0	62	0	37	13	43.60	740	0.76
GQ454869_HA18-9	977	60	0.10	29	31	38	0	60	0	29	11	42.89	899	0.92

*Table 5.11: Number of reads mapping back to references for sample **13-3436**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	101881	48.34	50075	51806	86216	0	98659	0	47207	25546	43.30	968	0.99
FJ525431_NZRB-M12	977	51459	24.41	30057	21402	37240	0	49218	0	31673	13360	43.30	965	0.99
HM573451_Kpg	977	23156	10.99	11376	11780	17954	0	22906	0	10872	5378	43.81	963	0.99
JQ798289_A18	977	6169	2.93	2721	3448	4316	0	6120	0	5837	1420	43.30	909	0.93
FJ525432_NZRB-G90	977	5192	2.46	2319	2873	2580	0	4938	0	3848	1175	43.40	911	0.93
FJ525433_NZRB-TH28	977	5080	2.41	2594	2486	4426	0	5046	0	3613	1393	43.81	950	0.97
FJ525434_NZRB-TH30	977	4399	2.09	2449	1950	3822	0	4351	0	3726	1131	43.81	962	0.98
KC525952_T3	977	4152	1.97	2441	1711	2722	0	4148	0	3947	1017	42.89	915	0.94
EU857538_SP	977	2191	1.04	1374	817	1202	0	2182	0	1824	529	43.60	914	0.94
GQ454870_HA16-5	977	2126	1.01	929	1197	1222	0	2124	0	1944	537	43.71	888	0.91
FJ525435_NZRB-M17	977	2079	0.99	1113	966	1238	0	1966	0	1751	391	43.71	926	0.95
JF957196_B301	977	1051	0.50	487	564	868	0	919	0	972	163	42.89	932	0.95
CT-ZA1	977	667	0.32	349	318	520	0	665	0	582	179	43.81	931	0.95
KC517493_FL202-VT	977	436	0.21	238	198	310	0	428	0	416	100	43.09	834	0.85
JQ061137_AT-1	977	213	0.10	92	121	122	0	211	0	98	47	43.60	895	0.92

Table 5.12: Number of reads mapping back to references for sample **13-3439**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	77297	29.87	38305	38992	63570	0	70917	0	28423	16023	43.30	973	1.00
FJ525431_NZRB-M12	977	64273	24.84	35496	28777	50756	0	58423	0	45468	14770	43.30	951	0.97
HM573451_Kpg3	977	61499	23.77	30342	31157	49918	0	59118	0	20042	12366	43.81	973	1.00
KC525952_T3	977	11599	4.48	6373	5226	8756	0	11381	0	10696	2434	42.89	938	0.96
JQ798289_A18	977	10287	3.98	4797	5490	8432	0	10126	0	8554	1952	43.30	938	0.96
FJ525432_NZRB-G90	977	9943	3.84	4825	5118	7458	0	9347	0	7062	1956	43.40	927	0.95
EU857538_SP	977	8322	3.22	4506	3816	5492	0	8195	0	5134	1793	43.60	925	0.95
GQ454870_HA16-5	977	7280	2.81	3528	3752	5010	0	7246	0	4967	1787	43.71	911	0.93
FJ525435_NZRB-M17	977	3728	1.44	1909	1819	3142	0	3376	0	3062	524	43.71	943	0.97
JF957196_B301	977	2663	1.03	1273	1390	2434	0	2125	0	2500	380	42.89	869	0.89
FJ525433_NZRB-TH28	977	859	0.33	461	398	692	0	815	0	634	221	43.81	658	0.67
FJ525434_NZRB-TH30	977	322	0.12	182	140	254	0	268	0	206	69	43.81	618	0.63

Table 5.13: Number of reads mapping back to references for sample **13-3445**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	37711	60.61	18348	19363	29200	0	37709	1	22183	11176	43.30	977	1
FJ525433_NZRB-TH28	977	8852	14.23	4839	4013	7368	0	8749	0	7180	2645	43.81	936	0.96
FJ525431_NZRB-M12	977	6031	9.69	4240	1791	1800	0	6026	0	5353	1791	43.30	970	0.99
HM573451_Kpg3	977	3737	6.01	1476	2261	2282	0	3733	0	2375	1109	43.81	958	0.98
FJ525434_NZRB-TH30	977	1798	2.89	1203	595	1114	0	1794	0	1591	519	43.81	864	0.88
CT-ZA1	977	967	1.55	537	430	674	0	959	0	858	288	43.81	804	0.82
KC525952_T3	977	783	1.26	535	248	410	0	783	0	773	232	42.89	917	0.94
JQ798289_A18	977	544	0.87	225	319	184	0	544	0	539	162	43.30	883	0.90
FJ525432_NZRB-G90	977	510	0.82	282	228	86	0	510	0	491	150	43.40	837	0.86
KC517493_FL202-VT	977	385	0.62	220	165	130	0	385	0	385	113	43.09	488	0.50
FJ525435_NZRB-M17	977	300	0.48	160	140	48	0	300	0	295	87	43.71	791	0.81
EU857538_SP	977	259	0.42	172	87	70	0	258	0	222	75	43.60	820	0.84
GQ454870_HA16-5	977	193	0.31	85	108	74	0	193	0	178	57	43.71	926	0.95

Table 5.14: Number of reads mapping back to references for sample **13-3447**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
GQ454870_HA16-5	977	36400	37.93	19720	16680	26920	0	36400	0	29294	10888	43.71	891	0.91
HM573451_Kpg3	977	16650	17.35	7756	8894	7266	0	16642	1	10969	4910	43.81	977	1.00
JX266712_Taiwan-Pum_SP_T1	977	15673	16.33	7599	8074	12360	0	15568	0	8285	4638	43.30	973	1.00
FJ525431_NZRB-M12	977	13647	14.22	8781	4866	8824	0	13186	0	11271	4052	43.30	947	0.97
JQ798289_A18	977	7168	7.47	2004	5164	2108	0	7168	0	7155	2115	43.30	875	0.90
EU857538_SP	977	4007	4.18	2242	1765	1756	0	4007	0	3011	1194	43.60	883	0.90
FJ525432_NZRB-G90	977	843	0.88	352	491	208	0	843	0	811	248	43.40	858	0.88
KC525952_T3	977	768	0.80	491	277	96	0	768	0	555	220	42.89	955	0.98
FJ525433_NZRB-TH28	977	644	0.67	323	321	594	0	644	0	564	193	43.81	926	0.95

Table 5.15: Number of reads mapping back to references for sample **13-3449**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	58105	62.58	25746	32359	36680	0	57131	7	31946	17265	43.81	977	1
EU857538_SP	977	15446	16.64	9868	5578	6992	0	15446	0	13554	4622	43.60	932	0.95
KC525952_T3	977	8202	8.83	5382	2820	3144	0	8202	0	8078	2407	42.89	946	0.97
JQ798289_A18	977	4679	5.04	1185	3494	1376	0	4679	0	4630	1390	43.30	926	0.95
GQ454870_HA16-5	977	4635	4.99	2425	2210	1198	0	4631	0	4433	1393	43.71	936	0.96
JX266712_Taiwan-Pum_SP_T1	977	988	1.06	468	520	732	0	988	0	555	293	43.30	958	0.98
FJ525433_NZRB-TH28	977	446	0.48	233	213	370	0	442	0	374	133	43.81	822	0.84
FJ525431_NZRB-M12	977	229	0.25	144	85	82	0	229	0	199	68	43.30	942	0.96

Table 5.16: Number of reads mapping back to references for sample 13-3454. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
EU857538_SP	977	10399	29.04	6096	4303	6310	0	10385	0	6375	3074	43.60	844	0.86
HM573451_Kpg3	977	9413	26.29	3475	5938	6212	0	9384	0	9210	2759	43.81	953	0.98
GQ454870_HA16-5	977	8847	24.71	4325	4522	6224	0	8816	0	8772	2643	43.71	817	0.84
KC525952_T3	977	5762	16.09	3650	2112	2572	0	5759	0	5638	1708	42.89	925	0.95
JX266712_Taiwan-Pum_SP_T1	977	469	1.31	229	240	370	0	465	0	251	138	43.30	936	0.96
FJ525433_NZRB-TH28	977	425	1.19	240	185	336	0	417	0	389	127	43.81	774	0.79
FJ525431_NZRB-M12	977	253	0.71	149	104	112	0	253	0	224	75	43.30	899	0.92
JQ798289_A18	977	131	0.37	52	79	38	0	131	0	129	35	43.30	886	0.91
FJ525434_NZRB-TH30	977	95	0.27	58	37	58	0	95	0	93	27	43.81	636	0.65

Table 6.1: Number of reads mapping back to references for sample 13-3508. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	55972	51.00	30379	25593	45942	0	47642	0	53312	13818	43.81	726	0.74
FJ525431_NZRB-M12	977	29448	26.83	15458	13990	19992	0	29319	0	28125	6817	43.30	958	0.98
FJ525434_NZRB-TH30	977	15960	14.54	9170	6790	13412	0	14950	0	15627	3519	43.81	726	0.74
KC517493_FL202-VT	977	2202	2.01	1157	1045	1902	0	2095	0	2078	394	43.09	835	0.85
CT-ZA1	977	1850	1.69	995	855	1412	0	1838	0	1611	465	43.81	957	0.98
JX266712_Taiwan-Pum_SP_T1	977	1169	1.07	597	572	930	0	1087	0	338	182	43.30	930	0.95
FJ525432_NZRB-G90	977	783	0.71	398	385	682	0	700	0	438	84	43.40	903	0.92
CT-ZA3	977	515	0.47	238	277	460	0	431	0	320	104	43.60	850	0.87
KC262793_L192GR	977	443	0.40	225	218	402	0	437	0	225	67	43.91	846	0.87
JX266713_Taiwan-Pum_M_T5	977	388	0.35	195	193	350	0	385	0	366	33	43.30	340	0.35
JF957196_B301	977	207	0.19	102	105	176	0	197	0	101	25	42.89	812	0.83
FJ525435_NZRB-M17	977	178	0.16	91	87	154	0	156	0	58	17	43.71	898	0.92

Table 6.2: Number of reads mapping back to references for sample **13-3512**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	120475	56.63	58585	61890	104196	0	112039	0	48835	28044	43.30	970	0.99
FJ525431_NZRB-M12	977	31576	14.84	17869	13707	22108	0	30878	0	19140	7692	43.30	963	0.99
FJ525433_NZRB-TH28	977	29720	13.97	15073	14647	26730	0	27509	0	25295	7772	43.81	958	0.98
FJ525434_NZRB-TH30	977	16211	7.62	8698	7513	14658	0	15650	0	15253	3823	43.81	941	0.96
FJ525432_NZRB-G90	977	7730	3.63	4322	3408	5130	0	7402	0	6770	1678	43.40	929	0.95
FJ525435_NZRB-M17	977	4574	2.15	2538	2036	2820	0	4387	0	4396	1022	43.71	895	0.92
JF957196_B301	977	1852	0.87	867	985	1616	0	1318	0	1711	276	42.89	888	0.91

Table 6.3: Number of reads mapping back to references for sample **13-3520**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	59974	56.87	32409	27565	48014	0	54042	0	55474	15500	43.81	746	0.76
FJ525431_NZRB-M12	977	26584	25.21	14123	12461	14960	0	26517	0	25425	6565	43.30	958	0.98
FJ525434_NZRB-TH30	977	17385	16.48	10440	6945	13626	0	16832	0	17016	4052	43.81	736	0.75
JF957196_B301	977	407	0.39	165	242	188	0	358	0	388	75	42.89	813	0.83
JX266712_Taiwan-Pum_SP_T1	977	392	0.37	235	157	178	0	386	0	201	79	43.30	941	0.96
FJ525432_NZRB-G90	977	274	0.26	137	137	194	0	268	0	220	39	43.40	827	0.85
JX266713_Taiwan-Pum_M_T5	977	211	0.20	103	108	180	0	209	0	202	18	43.30	379	0.39

Table 6.4: Number of reads mapping back to references for sample **13-3523**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	75221	40.27	37433	37788	64712	0	61851	0	23978	14103	43.30	967	0.99
FJ525433_NZRB-TH28	977	24868	13.31	13663	11205	19582	0	19773	0	24209	5604	43.81	683	0.70
FJ525431_NZRB-M12	977	20101	10.76	10738	9363	15142	0	19230	0	14338	4261	43.30	958	0.98
HM573451_Kpg3	977	18160	9.72	8842	9318	16006	0	17794	0	10010	3220	43.81	966	0.99
KC525952_T3	977	10568	5.66	5837	4731	8294	0	10247	0	9505	2147	42.89	930	0.95
FJ525434_NZRB-TH30	977	9263	4.96	5099	4164	8210	0	8150	0	9146	1945	43.81	672	0.69
EU857538_SP	977	8091	4.33	4063	4028	5526	0	7737	0	3999	1808	43.60	926	0.95
FJ525435_NZRB-M17	977	4252	2.28	2131	2121	3880	0	3176	0	3140	488	43.71	944	0.97
JQ798289_A18	977	4138	2.22	2073	2065	3760	0	4114	0	2670	603	43.30	954	0.98
GQ454870_HA16-5	977	3706	1.98	1822	1884	2632	0	3684	0	2679	881	43.71	918	0.94
FJ525432_NZRB-G90	977	3565	1.91	1771	1794	3116	0	2996	0	2676	516	43.40	958	0.98
JF957196_B301	977	2170	1.16	1061	1109	2060	0	1199	0	1828	235	42.89	875	0.90
JX266713_Taiwan-Pum_M_T5	977	939	0.50	469	470	858	0	918	0	884	67	43.30	634	0.65
KC517493_FL202-VT	977	662	0.35	342	320	588	0	642	0	653	110	43.09	575	0.59
CT-ZA1	977	603	0.32	324	279	470	0	601	0	571	153	43.81	941	0.96

Table 6.5: Number of reads mapping back to references for sample **13-3526**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	17178	37.67	9033	8145	14556	0	17027	0	14913	5121	43.81	888	0.91
JX266712_Taiwan-Pum_SP_T1	977	12032	26.39	5946	6086	8846	0	11924	54	6395	3546	43.30	977	1.00
FJ525431_NZRB-M12	977	7307	16.02	4213	3094	2970	0	7303	6	6583	2167	43.30	977	1.00
FJ525434_NZRB-TH30	977	3200	7.02	2044	1156	2038	0	3196	0	2972	926	43.81	879	0.90
HM573451_Kpg3	977	2914	6.39	1498	1416	1818	0	2914	15	1913	858	43.81	977	1.00
GQ454870_HA16-5	977	722	1.58	320	402	312	0	722	0	688	214	43.71	863	0.88
JQ798289_A18	977	608	1.33	237	371	272	0	608	0	595	179	43.30	887	0.91
CT-ZA1	977	469	1.03	259	210	340	0	468	2	407	139	43.81	977	1.00
KC525952_T3	977	354	0.78	218	136	112	0	354	0	324	103	42.89	819	0.84
FJ525432_NZRB-G90	977	281	0.62	129	152	72	0	281	0	271	82	43.40	832	0.85
KC517493_FL202-VT	977	176	0.39	89	87	66	0	176	0	174	52	43.09	602	0.62
EU857538_SP	977	155	0.34	96	59	42	0	155	0	92	45	43.60	851	0.87
FJ525435_NZRB-M17	977	110	0.24	50	60	28	0	110	0	107	31	43.71	788	0.81

Table 6.6: Number of reads mapping back to references for sample **13-3534**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	116600	53.54	57775	58825	110420	0	94617	0	37142	16748	43.81	968	0.99
FJ525433_NZRB-TH28	977	20651	9.48	10734	9917	19162	0	14056	0	19943	3877	43.81	749	0.77
KC525952_T3	977	19728	9.06	10328	9400	17760	0	17443	0	19042	3146	42.89	917	0.94
EU857538_SP	977	17655	8.11	9259	8396	16008	0	14412	0	10463	2416	43.60	935	0.96
JQ798289_A18	977	14830	6.81	7113	7717	13246	0	13683	0	11616	2503	43.30	916	0.94
FJ525431_NZRB-M12	977	12292	5.64	6214	6078	10564	0	12190	0	10583	2180	43.30	952	0.97
FJ525434_NZRB-TH30	977	5068	2.33	2678	2390	4740	0	4169	0	4951	962	43.81	712	0.73
GQ454870_HA16-5	977	4825	2.22	2206	2619	3956	0	3967	0	3075	778	43.71	899	0.92
FJ525432_NZRB-G90	977	1583	0.73	800	783	1440	0	1475	0	798	146	43.40	890	0.91
JX266713_Taiwan-Pum_M_T5	977	1429	0.66	713	716	1292	0	1398	0	1316	107	43.30	603	0.62
KC517493_FL202-VT	977	724	0.33	363	361	684	0	712	0	682	106	43.09	750	0.77
JX266712_Taiwan-Pum_SP_T1	977	603	0.28	301	302	500	0	532	0	178	69	43.30	953	0.98
FJ525435_NZRB-M17	977	589	0.27	287	302	534	0	400	0	278	36	43.71	853	0.87
CT-ZA1	977	452	0.21	236	216	402	0	449	0	392	108	43.81	951	0.97
KC262793_L192GR	977	296	0.14	149	147	278	0	290	0	123	26	43.91	830	0.85

Table 6.7: Number of reads mapping back to references for sample **13-3536**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	56800	55.02	31219	25581	46718	0	51458	0	51168	14976	43.81	782	0.80
FJ525431_NZRB-M12	977	28711	27.81	15772	12939	17214	0	28538	0	27234	7232	43.30	951	0.97
FJ525434_NZRB-TH30	977	14824	14.36	8997	5827	11502	0	14373	0	14463	3471	43.81	780	0.80
CT-ZA1	977	731	0.71	381	350	574	0	727	0	633	186	43.81	946	0.97
KC517493_FL202-VT	977	681	0.66	359	322	556	0	667	0	661	139	43.09	633	0.65
JX266712_Taiwan-Pum_SP_T1	977	492	0.48	264	228	336	0	480	0	195	100	43.30	943	0.97
FJ525432_NZRB-G90	977	237	0.23	108	129	168	0	223	0	178	33	43.40	864	0.88
JX266713_Taiwan-Pum_M_T5	977	158	0.15	81	77	140	0	154	0	150	14	43.30	338	0.35
CT-ZA3	977	132	0.13	63	69	120	0	126	0	67	28	43.60	783	0.80
KC262793_L192GR	977	110	0.11	57	53	94	0	108	0	68	20	43.91	795	0.81

Table 6.8: Number of reads mapping back to references for sample **13-3539**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	30482	55.10	16203	14279	26558	0	30191	0	26336	9085	43.81	899	0.92
FJ525431_NZRB-M12	977	10420	18.83	5710	4710	4164	0	10415	0	10097	3100	43.30	956	0.98
HM573451_Kpg3	977	6766	12.23	2982	3784	4302	0	6706	0	4903	2003	43.81	953	0.97
FJ525434_NZRB-TH30	977	3307	5.98	2138	1169	2196	0	3295	0	2841	955	43.81	906	0.93
KC525952_T3	977	1380	2.49	1016	364	514	0	1380	0	1350	404	42.89	909	0.93
EU857538_SP	977	1107	2.00	787	320	416	0	1107	0	976	326	43.60	961	0.98
JQ798289_A18	977	1023	1.85	372	651	364	0	1023	0	1019	302	43.30	810	0.83
GQ454870_HA16-5	977	468	0.85	128	340	120	0	468	0	455	139	43.71	926	0.95
CT-ZA1	977	173	0.31	90	83	132	0	172	0	145	51	43.81	942	0.97
KC517493_FL202-VT	977	67	0.12	40	27	32	0	67	0	67	19	43.09	477	0.49

Table 6.9: Number of reads mapping back to references for sample **13-3545**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	232775	66.11	115560	117215	218386	0	211688	0	87416	43954	43.30	969	0.99
FJ525431_NZRB-M12	977	73120	20.77	39006	34114	65264	0	65325	0	49852	15881	43.30	960	0.98
FJ525432_NZRB-G90	977	32038	9.10	16031	16007	27668	0	29204	0	22316	5462	43.40	895	0.92
JF957196_B301	977	6831	1.94	3338	3493	6460	0	6450	0	5711	831	42.89	882	0.90
FJ525435_NZRB-M17	977	5169	1.47	2596	2573	4888	0	4262	0	4131	432	43.71	930	0.95
FJ525433_NZRB-TH28	977	1015	0.29	524	491	932	0	975	0	746	269	43.81	843	0.86
GQ454869_HA18-9	977	632	0.18	336	296	464	0	604	0	363	100	42.89	904	0.93

Table 6.10: Number of reads mapping back to references for sample **13-3552**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	33966	69.33	16715	17251	27384	0	33875	0	18424	10045	43.30	967	0.99
FJ525433_NZRB-TH28	977	7087	14.47	3795	3292	5954	0	7029	0	6133	2119	43.81	871	0.89
FJ525431_NZRB-M12	977	4723	9.64	3081	1642	1834	0	4720	0	4002	1395	43.30	893	0.91
FJ525434_NZRB-TH30	977	1564	3.19	1005	559	1034	0	1562	0	1455	451	43.81	818	0.84
FJ525432_NZRB-G90	977	537	1.10	279	258	82	0	537	0	526	156	43.40	825	0.84
HM573451_Kpg3	977	445	0.91	223	222	276	0	444	0	350	131	43.81	911	0.93
KC525952_T3	977	187	0.38	106	81	142	0	187	0	186	55	42.89	689	0.71
GQ454870_HA16-5	977	136	0.28	52	84	54	0	136	0	133	40	43.71	850	0.87
JQ798289_A18	977	123	0.25	48	75	62	0	123	0	123	36	43.30	511	0.52
EU857538_SP	977	111	0.23	69	42	60	0	111	0	89	33	43.60	830	0.85
GQ454869_HA18-9	977	56	0.11	36	20	8	0	56	0	53	16	42.89	798	0.82

Table 6.11: Number of reads mapping back to references for sample **13-3553**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	17687	62.68	9031	8656	15216	0	17629	0	14556	5098	43.81	956	0.98
FJ525431_NZRB-M12	977	5075	17.98	3062	2013	1498	0	5073	0	4695	1468	43.30	943	0.97
FJ525434_NZRB-TH30	977	3174	11.25	1946	1228	2234	0	3173	0	2520	881	43.81	960	0.98
JX266712_Taiwan-Pum_SP_T1	977	2193	7.77	1069	1124	1484	0	2189	0	1252	619	43.30	953	0.97
FJ525435_NZRB-M17	977	33	0.12	15	18	2	0	33	0	31	9	43.71	696	0.71
FJ525432_NZRB-G90	977	30	0.11	16	14	0	0	30	0	25	8	43.40	708	0.72

Table 6.12: Number of reads mapping back to references for sample **13-3556**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	82339	70.50	40178	42161	70814	0	82190	0	48396	24948	43.30	969	0.99
FJ525431_NZRB-M12	977	29237	25.03	18395	10842	20164	0	28884	0	25001	8857	43.30	965	0.99
FJ525432_NZRB-G90	977	2842	2.43	1291	1551	774	0	2842	0	2793	859	43.40	896	0.92
FJ525433_NZRB-TH28	977	1716	1.47	862	854	1654	0	1715	0	1569	525	43.81	949	0.97
GQ454869_HA18-9	977	146	0.13	87	59	12	0	146	0	139	44	42.89	934	0.96
FJ525435_NZRB-M17	977	124	0.11	63	61	16	0	124	0	118	37	43.71	883	0.90

Table 6.13: Number of reads mapping back to references for sample **13-3570**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	52743	64.43	26173	26570	41036	0	52586	3	28631	15595	43.30	977	1
FJ525433_NZRB-TH28	977	15206	18.57	8143	7063	12692	0	15110	0	12887	4532	43.81	893	0.91
FJ525431_NZRB-M12	977	8620	10.53	5498	3122	3296	0	8609	0	7479	2550	43.30	963	0.99
FJ525434_NZRB-TH30	977	2892	3.53	1917	975	1776	0	2890	0	2631	835	43.81	874	0.89
FJ525432_NZRB-G90	977	1118	1.37	537	581	232	0	1118	0	1080	326	43.40	892	0.91
HM573451_Kpg3	977	346	0.42	155	191	224	0	345	0	184	102	43.81	939	0.96
CT-ZA1	977	172	0.21	94	78	134	0	172	0	151	51	43.81	941	0.96
AY170468_T36	981	157	0.19	81	76	134	0	138	0	92	46	43.63	907	0.92
GQ454870_HA16-5	977	124	0.15	67	57	76	0	124	0	99	37	43.71	894	0.92
GQ454869_HA18-9	977	99	0.12	58	41	16	0	99	0	90	29	42.89	916	0.94
FJ525435_NZRB-M17	977	86	0.11	40	46	16	0	86	0	79	24	43.71	773	0.79

Table 7.1: Number of reads mapping back to references for sample **13-3604**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped Reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	64925	58.84	29850	35075	44742	0	64626	0	36049	19215	43.81	972	0.99
GQ454870_HA16-5	977	18197	16.49	9315	8882	9880	0	18193	0	17010	5437	43.71	861	0.88
EU857538_SP	977	9973	9.04	6246	3727	4674	0	9972	0	8594	2962	43.60	928	0.95
KC525952_T3	977	9162	8.30	5723	3439	4274	0	9160	0	9047	2669	42.89	956	0.98
JQ798289_A18	977	3540	3.21	1343	2197	1466	0	3540	0	3468	1046	43.30	919	0.94
JX266712_Taiwan-Pum_SP_T1	977	2169	1.97	1055	1114	1730	0	2127	0	1201	643	43.30	949	0.97
FJ525431_NZRB-M12	977	2050	1.86	1249	801	1412	0	2047	0	1533	607	43.30	952	0.97
FJ525432_NZRB-G90	977	155	0.14	60	95	40	0	155	0	151	46	43.40	674	0.69
FJ525433_NZRB-TH28	977	120	0.11	61	59	106	0	119	0	90	36	43.81	874	0.89

Table 7.2: Number of reads mapping back to references for sample **13-3608**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	27483	40.64	13431	14052	21192	0	27406	0	15061	8129	43.30	973	1.00
FJ525433_NZRB-TH28	977	23127	34.20	12350	10777	19732	0	22837	0	19661	6908	43.81	910	0.93
FJ525431_NZRB-M12	977	10932	16.16	6555	4377	4166	0	10910	0	10047	3247	43.30	943	0.97
FJ525434_NZRB-TH30	977	3448	5.10	2269	1179	2188	0	3444	0	3012	997	43.81	904	0.93
CT-ZA1	977	863	1.28	460	403	660	0	855	0	720	256	43.81	953	0.97
FJ525432_NZRB-G90	977	427	0.63	236	191	88	0	427	0	413	125	43.40	899	0.92
KC517493_FL202-VT	977	324	0.48	199	125	122	0	324	0	324	95	43.09	477	0.49
FJ525435_NZRB-M17	977	272	0.40	126	146	52	0	272	0	267	79	43.71	805	0.82
HM573451_Kpg3	977	231	0.34	108	123	164	0	230	0	127	68	43.81	942	0.96
CT-ZA3	977	209	0.31	94	115	170	0	193	0	153	61	43.60	951	0.97

Table 7.3: Number of reads mapping back to references for sample **13-3612**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of total reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	126566	53.93	63091	63475	104182	0	107490	69	37291	21598	43.2958	977	1.00
FJ525433_NZRB-TH28	977	27152	11.57	14439	12713	19406	0	19367	0	26794	5441	43.81	605	0.62
FJ525431_NZRB-M12	977	25779	10.99	13444	12335	20020	0	24166	1	16541	4977	43.30	977	1.00
FJ525434_NZRB-TH30	977	11106	4.73	5946	5160	9916	0	9363	0	10977	2156	43.81	601	0.62
HM573451_Kpg3	977	7192	3.06	3551	3641	6436	0	6953	0	3244	1142	43.81	965	0.99
FJ525432_NZRB-G90	977	5969	2.54	3037	2932	5412	0	4582	0	4620	788	43.40	965	0.99
FJ525435_NZRB-M17	977	5855	2.50	2916	2939	5430	0	4310	0	4734	662	43.71	943	0.97
JF957196_B301	977	4644	1.98	2286	2358	4436	0	4267	0	4132	541	42.89	915	0.94
EU857538_SP	977	4178	1.78	2068	2110	3066	0	3911	3	1798	704	43.60	977	1.00
GQ454870_HA16-5	977	3581	1.53	1839	1742	2272	0	3561	0	2201	795	43.71	902	0.92
JQ061137_AT-1	977	3395	1.45	1678	1717	3046	0	3335	0	1381	520	43.60	916	0.94
KC525952_T3	977	3007	1.28	1566	1441	2440	0	2949	0	2573	529	42.89	913	0.93
JQ798289_A18	977	2032	0.87	988	1044	1786	0	2008	0	1362	314	43.30	936	0.96
JX266713_Taiwan-Pum_M_T5	977	1105	0.47	551	554	992	0	1082	0	1047	81	43.30	552	0.56
KC517493_FL202-VT	977	751	0.32	383	368	674	0	751	0	731	123	43.09	628	0.64
CT-ZA1	977	628	0.27	334	294	354	0	628	0	606	163	43.81	841	0.86
JQ911664_CT11A	977	464	0.20	229	235	378	0	452	0	448	87	43.60	727	0.74
JQ911663_CT14A	977	324	0.14	175	149	268	0	313	0	322	68	43.50	410	0.42
GQ454869_HA18-9	977	287	0.12	145	142	210	0	249	0	135	38	42.89	829	0.85
KC262793_L192GR	977	276	0.12	141	135	250	0	245	0	114	36	43.91	868	0.89

Table 7.4: Number of reads mapping back to references for sample **13-3619**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	48632	43.87	20401	28231	31244	0	48617	0	35988	14409	43.81	966	0.99
JX266712_Taiwan-Pum_SP_T1	977	23927	21.59	11505	12422	19750	0	23720	0	13234	7083	43.30	971	0.99
KC525952_T3	977	10780	9.72	7768	3012	4084	0	10780	0	10552	3170	42.89	945	0.97
EU857538_SP	977	10648	9.61	7777	2871	3958	0	10648	0	9696	3160	43.60	965	0.99
JQ798289_A18	977	6249	5.64	2023	4226	1968	0	6249	0	6220	1844	43.30	923	0.94
FJ525431_NZRB-M12	977	5856	5.28	3847	2009	3446	0	5840	0	4641	1730	43.30	948	0.97
GQ454870_HA16-5	977	3436	3.10	1028	2408	804	0	3435	0	3202	1017	43.71	929	0.95
FJ525432_NZRB-G90	977	476	0.43	232	244	128	0	476	0	464	140	43.40	859	0.88
FJ525435_NZRB-M17	977	368	0.33	181	187	86	0	367	0	359	106	43.71	885	0.91
FJ525433_NZRB-TH28	977	230	0.21	111	119	198	0	227	0	183	69	43.81	909	0.93

Table 7.5: Number of reads mapping back to references for sample **13-3621**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	52453	41.21	26184	26269	47528	0	47465	0	17929	10992	43.30	974	1.00
FJ525433_NZRB-TH28	977	32361	25.42	16289	16072	29778	0	28480	0	28095	8095	43.81	967	0.99
FJ525431_NZRB-M12	977	21908	17.21	11836	10072	15800	0	21311	0	16227	4975	43.30	965	0.99
FJ525434_NZRB-TH30	977	14455	11.36	7820	6635	13134	0	13559	0	13252	3247	43.81	933	0.95
FJ525432_NZRB-G90	977	2554	2.01	1316	1238	2122	0	2260	0	2277	415	43.40	896	0.92
JF957196_B301	977	1664	1.31	795	869	1510	0	1051	0	1561	231	42.89	749	0.77
FJ525435_NZRB-M17	977	1307	1.03	677	630	1020	0	1108	0	1110	208	43.71	866	0.89
JX266713_Taiwan-Pum_M_T5	977	269	0.21	137	132	250	0	269	0	264	23	43.30	409	0.42

Table 7.6: Number of reads mapping back to references for sample **13-3626**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	63640	72.57	28146	35494	48972	0	63135	0	43026	18801	43.81	968	0.99
KC525952_T3	977	8734	9.96	5876	2858	3896	0	8733	0	8680	2555	42.89	951	0.97
EU857538_SP	977	7231	8.25	5036	2195	3028	0	7231	0	6683	2134	43.60	863	0.88
JX266712_Taiwan-Pum_SP_T1	977	3260	3.72	1586	1674	2540	0	3256	0	1776	964	43.30	953	0.98
FJ525433_NZRB-TH28	977	2615	2.98	1447	1168	2178	0	2590	0	2325	781	43.81	827	0.85
FJ525431_NZRB-M12	977	1131	1.29	650	481	426	0	1128	0	1038	336	43.30	702	0.72
GQ454870_HA16-5	977	338	0.39	172	166	112	0	338	0	254	97	43.71	836	0.86
JQ798289_A18	977	303	0.35	133	170	70	0	303	0	260	80	43.30	920	0.94
FJ525434_NZRB-TH30	977	297	0.34	204	93	166	0	296	0	276	85	43.81	761	0.78

Table 7.7: Number of reads mapping back to references for sample **13-3627**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	69973	38.73	34428	35545	56028	0	66118	0	28128	16532	43.30	974	1.00
HM573451_Kpg3	977	24145	13.36	11415	12730	19598	0	24112	0	15583	5158	43.81	966	0.99
FJ525433_NZRB-TH28	977	16235	8.99	8915	7320	12428	0	15092	0	15298	4269	43.81	725	0.74
FJ525431_NZRB-M12	977	15311	8.47	8614	6697	9962	0	15043	0	10046	3714	43.30	965	0.99
KC525952_T3	977	11185	6.19	6333	4852	7732	0	11155	0	10133	2652	42.89	943	0.97
GQ454870_HA16-5	977	10668	5.90	5263	5405	7482	0	10649	0	8501	2851	43.71	918	0.94
EU857538_SP	977	10607	5.87	5778	4829	6694	0	10578	0	6491	2789	43.60	932	0.95
FJ525434_NZRB-TH30	977	7893	4.37	4523	3370	6606	0	7651	0	7818	1840	43.81	717	0.73
JQ798289_A18	977	5565	3.08	2768	2797	4646	0	5550	0	4391	1046	43.30	938	0.96
FJ525432_NZRB-G90	977	2046	1.13	1101	945	1468	0	1876	0	1618	384	43.40	920	0.94
FJ525435_NZRB-M17	977	2007	1.11	1031	976	1510	0	1797	0	1504	340	43.71	930	0.95
CT-ZA1	977	1783	0.99	937	846	1470	0	1776	0	1706	441	43.81	929	0.95
KC517493_FL202-VT	977	1252	0.69	647	605	1040	0	1240	0	1225	260	43.09	663	0.68
JF957196_B301	977	944	0.52	450	494	856	0	735	0	883	138	42.89	899	0.92
KC262793_L192GR	977	332	0.18	180	152	280	0	332	0	287	71	43.91	798	0.82

Table 7.8: Number of reads mapping back to references for sample **13-3637**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	103212	52.53	53066	50146	84348	0	101163	0	52294	24017	43.30	973	1.00
FJ525431_NZRB-M12	977	31891	16.23	17050	14841	23262	0	31761	0	25434	8064	43.30	964	0.99
FJ525433_NZRB-TH28	977	16345	8.32	8530	7815	14170	0	15780	0	15825	4003	43.81	740	0.76
FJ525435_NZRB-M17	977	13287	6.76	7031	6256	9184	0	13222	0	11414	3027	43.71	874	0.89
FJ525432_NZRB-G90	977	12569	6.40	5916	6653	10074	0	12510	0	11083	2590	43.40	941	0.96
GQ454869_HA18-9	977	9263	4.71	5196	4067	6332	0	9202	0	8533	2014	42.89	936	0.96
FJ525434_NZRB-TH30	977	4787	2.44	2647	2140	4180	0	4686	0	4676	1074	43.81	727	0.74
JF957196_B301	977	2895	1.47	1254	1641	2214	0	2621	0	2559	594	42.89	825	0.84
AF260651_T30	977	813	0.41	588	225	430	0	813	0	811	202	43.19	378	0.39
JX266713_Taiwan-Pum_M_T5	977	575	0.29	283	292	542	0	574	0	555	56	43.30	450	0.46
KC517493_FL202-VT	977	198	0.10	104	94	172	0	194	0	196	39	43.09	532	0.54

Table 7.9: Number of reads mapping back to references for sample **13-3642**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	19760	56.61	9732	10028	15494	0	19701	0	11067	5813	43.30	976	1.00
FJ525433_NZRB-TH28	977	8081	23.15	4301	3780	6798	0	8004	0	6825	2401	43.81	867	0.89
FJ525431_NZRB-M12	977	4181	11.98	2692	1489	1480	0	4180	0	3691	1235	43.30	948	0.97
FJ525434_NZRB-TH30	977	1444	4.14	954	490	886	0	1444	0	1329	415	43.81	754	0.77
CT-ZA1	977	313	0.90	150	163	224	0	309	0	265	93	43.81	943	0.97
FJ525432_NZRB-G90	977	305	0.87	148	157	66	0	305	0	300	88	43.40	763	0.78
KC517493_FL202-VT	977	257	0.74	170	87	140	0	257	0	257	75	43.09	507	0.52
FJ525435_NZRB-M17	977	191	0.55	87	104	40	0	191	0	188	55	43.71	688	0.70
HM573451_Kpg3	977	152	0.44	69	83	112	0	152	0	85	44	43.81	915	0.94
CT-ZA3	977	45	0.13	20	25	38	0	40	0	30	13	43.60	838	0.86
KC525952_T3	977	42	0.12	25	17	20	0	42	0	41	12	42.89	671	0.69
GQ454869_HA18-9	977	35	0.10	15	20	6	0	35	0	33	10	42.89	741	0.76

Table 7.10: Number of reads mapping back to references for sample **13-3643**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	12174	46.45	6713	5461	9930	0	12009	0	10706	3633	43.81	871	0.89
FJ525431_NZRB-M12	977	8165	31.15	4125	4040	3992	0	8163	0	7856	2425	43.30	933	0.95
FJ525434_NZRB-TH30	977	1832	6.99	1208	624	1126	0	1830	0	1743	532	43.81	804	0.82
JX266712_Taiwan-Pum_SP_T1	977	1771	6.76	910	861	1076	0	1767	0	1248	520	43.30	960	0.98
FJ525435_NZRB-M17	977	594	2.27	246	348	334	0	594	0	590	177	43.71	653	0.67
JF957196_B301	977	560	2.14	219	341	324	0	560	0	559	163	42.89	616	0.63
FJ525432_NZRB-G90	977	297	1.13	137	160	116	0	297	0	227	87	43.40	898	0.92
GQ454869_HA18-9	977	231	0.88	166	65	62	0	231	0	211	67	42.89	917	0.94
HM573451_Kpg3	977	138	0.53	60	78	86	0	138	0	103	40	43.81	942	0.96
GQ454870_HA16-5	977	137	0.52	65	72	100	0	137	0	126	41	43.71	790	0.81
EU857538_SP	977	120	0.46	67	53	76	0	120	0	105	36	43.60	826	0.85
KC525952_T3	977	76	0.29	44	32	48	0	76	0	74	22	42.89	663	0.68
CT-ZA1	977	44	0.17	22	22	30	0	44	0	38	13	43.81	927	0.95
KC517493_FL202-VT	977	29	0.11	14	15	18	0	29	0	29	9	43.09	435	0.45

Table 7.11: Number of reads mapping back to references for sample **13-3647**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	42382	73.60	20766	21616	35242	0	41689	0	21464	12474	43.30	972	0.99
GQ454870_HA16-5	977	4682	8.13	2378	2304	3422	0	4682	0	4411	1400	43.71	866	0.89
EU857538_SP	977	2785	4.84	1438	1347	1652	0	2785	0	2105	828	43.60	866	0.89
HM573451_Kpg3	977	2216	3.85	802	1414	1260	0	2216	0	2082	648	43.81	950	0.97
KC525952_T3	977	1796	3.12	1190	606	780	0	1796	0	1735	525	42.89	931	0.95
FJ525431_NZRB-M12	977	1061	1.84	841	220	190	0	1061	0	1044	305	43.30	900	0.92
FJ525435_NZRB-M17	977	862	1.50	418	444	206	0	862	0	860	251	43.71	709	0.72
FJ525432_NZRB-G90	977	811	1.41	417	394	206	0	811	0	804	236	43.40	841	0.86
CT-ZA1	977	414	0.72	227	187	282	0	411	0	365	123	43.81	929	0.95
JQ798289_A18	977	247	0.43	128	119	80	0	247	0	192	70	43.30	887	0.91
KC517493_FL202-VT	977	148	0.26	98	50	48	0	148	0	148	43	43.09	475	0.49
CT-ZA3	977	70	0.12	26	44	52	0	70	0	44	21	43.60	872	0.89

Table 7.12: Number of reads mapping back to references for sample **13-3651**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	69535	57.7351	37465	32070	57990	0	62491	0	63174	18253.15	43.81	782	0.80
FJ525431_NZRB-M12	977	32766	27.205699	17708	15058	19202	0	32695	0	31487	8248.723	43.30	949	0.97
FJ525434_NZRB-TH30	977	16939	14.064498	10253	6686	13210	0	16310	0	16553	3971.46	43.81	779	0.80
JX266712_Taiwan-Pum_SP_T1	977	308	0.2557332	183	125	130	0	296	0	189	60.55271	43.30	958	0.98
FJ525432_NZRB-G90	977	251	0.208406	126	125	176	0	243	0	189	34.87206	43.40	841	0.86
JX266713_Taiwan-Pum_M_T5	977	251	0.208406	127	124	232	0	245	0	246	21.73081	43.30	393	0.40
HM573451_Kpg 3	977	130	0.1079394	66	64	56	0	118	0	53	28.75333	43.81	911	0.93

Table 7.13: Number of reads mapping back to references for sample **13-3655**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	58604	55.22	32068	26536	47152	0	53354	0	52839	15751	43.81	770	0.79
FJ525431_NZRB-M12	977	27695	26.10	15104	12591	14986	0	27615	0	26420	7116	43.30	950	0.97
FJ525434_NZRB-TH30	977	13694	12.90	8595	5099	10050	0	13257	0	13339	3287	43.81	766	0.78
JQ061137_AT-1	977	4125	3.89	1971	2154	3630	0	4101	0	1827	843	43.60	943	0.97
JQ911664_CT11A	977	507	0.48	278	229	360	0	501	0	505	116	43.60	548	0.56
JQ911663_CT14A	977	282	0.27	175	107	210	0	279	0	282	63	43.50	434	0.44
JX266712_Taiwan-Pum_SP_T1	977	244	0.23	137	107	122	0	226	0	107	44	43.30	918	0.94
FJ525432_NZRB-G90	977	243	0.23	127	116	178	0	231	0	183	31	43.40	794	0.81
JX266713_Taiwan-Pum_M_T5	977	220	0.21	111	109	194	0	214	0	214	19	43.30	356	0.36
DQ151548_T318A	977	149	0.14	72	77	108	0	149	0	145	28	43.30	560	0.57

Table 7.14: Number of reads mapping back to references for sample **13-3658**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	48769	32.64	26648	22121	39864	0	44199	0	45155	12850	43.81	763	0.78
HM573451_Kpg3	977	26944	18.03	12934	14010	22942	0	26900	0	17857	5546	43.81	966	0.99
FJ525431_NZRB-M12	977	22681	15.18	11995	10686	13506	0	22629	0	20661	5673	43.30	960	0.98
FJ525434_NZRB-TH30	977	12822	8.58	7789	5033	9942	0	12496	0	12574	3041	43.81	759	0.78
KC525952_T3	977	10166	6.80	5656	4510	7272	0	10139	0	8891	2385	42.89	953	0.98
GQ454870_HA16-5	977	9583	6.41	4716	4867	6852	0	9563	0	6953	2521	43.71	899	0.92
EU857538_SP	977	9036	6.05	4892	4144	6042	0	9005	0	5384	2309	43.60	930	0.95
JQ798289_A18	977	6517	4.36	3230	3287	5618	0	6505	0	5102	1141	43.30	918	0.94
JX266712_Taiwan-Pum_SP_T1	977	966	0.65	509	457	772	0	963	0	605	157	43.30	957	0.98
FJ525432_NZRB-G90	977	674	0.45	347	327	578	0	614	0	388	89	43.40	872	0.89
FJ525435_NZRB-M17	977	643	0.43	320	323	608	0	433	0	468	48	43.71	867	0.89
JX266713_Taiwan-Pum_M_T5	977	297	0.20	147	150	268	0	268	0	285	26	43.30	541	0.55

Table 7.15: Number of reads mapping back to references for sample **13-3661**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JQ061137_AT-1	977	257825	78.29	117659	140166	216364	0	255559	0	124189	63401	43.60	967	0.99
JQ911663_CT14A	977	31269	9.49	20654	10615	20662	0	30033	0	31260	7779	43.50	690	0.71
JQ911664_CT11A	977	26521	8.05	16727	9794	12786	0	26328	0	26497	6883	43.60	904	0.93
DQ151548_T318A	977	4601	1.40	2582	2019	2506	0	4582	0	4592	947	43.30	604	0.62
FJ525433_NZRB-TH28	977	4248	1.29	2303	1945	3524	0	3915	0	3857	1112	43.81	785	0.80
FJ525431_NZRB-M12	977	2492	0.76	1330	1162	1498	0	2443	0	2029	602	43.30	955	0.98
FJ525434_NZRB-TH30	977	975	0.30	599	376	736	0	955	0	951	231	43.81	642	0.66
KC517493_FL202-VT	977	451	0.14	286	165	220	0	448	0	407	109	43.09	606	0.62

Table 7.16: Number of reads mapping back to references for sample **13-3674**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	59110	50.38	31687	27423	49222	0	46044	0	57312	13437	43.81	702	0.72
FJ525431_NZRB-M12	977	34801	29.66	17822	16979	25842	0	34650	0	33965	7391	43.30	949	0.97
FJ525434_NZRB-TH30	977	19728	16.81	10951	8777	17330	0	17519	0	19037	4048	43.81	700	0.72
JX266713_Taiwan-Pum_M_T5	977	1533	1.31	778	755	1446	0	1531	0	1508	131	43.30	704	0.72
FJ525432_NZRB-G90	977	1130	0.96	576	554	1020	0	1041	0	891	118	43.40	902	0.92
JX266712_Taiwan-Pum_SP_T1	977	370	0.32	205	165	230	0	343	0	133	55	43.30	946	0.97
FJ525435_NZRB-M17	977	224	0.19	109	115	194	0	183	0	64	21	43.71	883	0.90
JF957196_B301	977	152	0.13	73	79	122	0	134	0	73	18	42.89	848	0.87

*Table 8.1: Number of reads mapping back to references for sample **13-3705**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
GQ454870_HA16-5	977	50041	45.71	25708	24333	40672	0	50009	0	40546	14965	43.71	945	0.97
HM573451_Kpg3	977	17559	16.04	8270	9289	9216	0	17526	0	11583	5177	43.81	974	1.00
JX266712_Taiwan-Pum_SP_T1	977	14784	13.50	7155	7629	12344	0	14733	0	8271	4377	43.30	965	0.99
FJ525431_NZRB-M12	977	10963	10.01	6710	4253	7714	0	10679	0	9233	3249	43.30	964	0.99
JQ798289_A18	977	6334	5.79	2173	4161	2350	0	6334	0	6319	1869	43.30	909	0.93
EU857538_SP	977	5751	5.25	3107	2644	2954	0	5749	0	3885	1712	43.60	862	0.88
KC525952_T3	977	2561	2.34	1672	889	1042	0	2561	0	2326	748	42.89	953	0.98
FJ525433_NZRB-TH28	977	579	0.53	278	301	526	0	579	0	513	174	43.81	923	0.94
FJ525432_NZRB-G90	977	453	0.41	206	247	126	0	453	0	424	133	43.40	776	0.79
CT-ZA1	977	188	0.17	105	83	122	0	188	0	183	55	43.81	654	0.67
KC517493_FL202-VT	977	118	0.11	66	52	60	0	118	0	118	35	43.09	498	0.51

*Table 8.2: Number of reads mapping back to references for sample **13-3708**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.*

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	132345	62.04	65183	67162	108394	0	112740	0	56335	26923	43.81	968	0.99
KC525952_T3	977	23607	11.07	13085	10522	18248	0	22741	0	22597	4708	42.89	968	0.99
JQ798289_A18	977	17311	8.12	7760	9551	13004	0	15924	0	16421	3602	43.30	909	0.93
JX266712_Taiwan-Pum_SP_T1	977	13949	6.54	6901	7048	12436	0	13065	0	5342	2617	43.30	966	0.99
EU857538_SP	977	13080	6.13	7643	5437	8698	0	12842	0	10161	2709	43.60	930	0.95
GQ454870_HA16-5	977	8890	4.17	3921	4969	6042	0	8753	0	4967	1940	43.71	916	0.94
FJ525431_NZRB-M12	977	1025	0.48	570	455	814	0	971	0	735	202	43.30	926	0.95
FJ525435_NZRB-M17	977	834	0.39	422	412	740	0	778	0	693	114	43.71	642	0.66
AY170468_T36	981	833	0.39	396	437	640	0	833	0	805	225	43.63	842	0.86
FJ525432_NZRB-G90	977	678	0.32	338	340	570	0	636	0	433	111	43.40	835	0.85
JF957196_B301	977	250	0.12	121	129	234	0	235	0	200	28	42.89	680	0.70

Table 8.3: Number of reads mapping back to references for sample **13-3713**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
GQ454870_HA16-5	977	84525	31.11	44094	40431	57152	0	84166	0	59252	22199	43.71	942	0.96
HM573451_Kpg3	977	84060	30.94	40596	43464	64564	0	79158	0	50583	18525	43.81	969	0.99
JQ798289_A18	977	27152	9.99	12049	15103	18838	0	26585	0	26791	6287	43.30	953	0.98
JX266712_Taiwan-Pum_SP_T1	977	22452	8.26	11132	11320	18786	0	21975	0	9831	4817	43.30	967	0.99
FJ525431_NZRB-M12	977	20867	7.68	11127	9740	17268	0	20116	0	14035	4711	43.30	962	0.98
EU857538_SP	977	16860	6.21	9948	6912	9976	0	16705	0	11061	3916	43.60	930	0.95
KC525952_T3	977	9523	3.51	5221	4302	6478	0	9423	0	7507	1986	42.89	954	0.98
FJ525432_NZRB-G90	977	4215	1.55	2104	2111	3248	0	4168	0	3171	810	43.40	868	0.89
FJ525435_NZRB-M17	977	845	0.31	426	419	730	0	812	0	597	112	43.71	792	0.81
JF957196_B301	977	400	0.15	197	203	356	0	371	0	378	51	42.89	646	0.66

Table 8.4: Number of reads mapping back to references for sample **13-3714**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	70826	52.00	32238	38588	44682	0	70739	0	48524	20975	43.81	968	0.99
KC525952_T3	977	17373	12.75	11869	5504	7970	0	17373	0	17022	5118	42.89	953	0.98
JX266712_Taiwan-Pum_SP_T1	977	14874	10.92	7396	7478	10152	0	14866	0	6907	4402	43.30	973	1.00
GQ454870_HA16-5	977	10006	7.35	3708	6298	3580	0	10006	0	9851	2988	43.71	929	0.95
JQ798289_A18	977	9964	7.32	3695	6269	4112	0	9964	0	9840	2945	43.30	936	0.96
EU857538_SP	977	9728	7.14	6701	3027	3408	0	9728	0	8471	2884	43.60	930	0.95
FJ525432_NZRB-G90	977	2658	1.95	1232	1426	1000	0	2658	0	2656	795	43.40	625	0.64
FJ525431_NZRB-M12	977	343	0.25	270	73	62	0	342	0	329	99	43.30	884	0.90
FJ525435_NZRB-M17	977	332	0.24	150	182	140	0	332	0	332	95	43.71	543	0.56

Table 8.5: Number of reads mapping back to references for sample **13-3719**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	16218	42.96	7826	8392	12540	0	16216	0	8989	4784.72	43.30	964	0.99
AY170468_T36	981	16138	42.75	6849	9289	12156	0	16138	0	15987	4782.18	43.63	958	0.98
FJ525431_NZRB-M12	977	1569	4.16	1151	418	414	0	1567	0	1358	461.43	43.30	933	0.95
EU937521_T36	981	981	2.60	594	387	546	0	981	0	881	289.18	43.43	945	0.96
FJ525433_NZRB-TH28	977	967	2.56	520	447	772	0	957	0	841	288.65	43.81	804	0.82
HM573451_Kpg3	977	676	1.79	309	367	436	0	675	0	377	198.99	43.81	953	0.98
FJ525434_NZRB-TH30	977	306	0.81	201	105	192	0	306	0	285	88.51	43.81	703	0.72
FJ525435_NZRB-M17	977	194	0.51	85	109	38	0	194	0	192	56.06	43.71	597	0.61
KC525952_T3	977	189	0.50	123	66	94	0	189	0	186	55.08	42.89	723	0.74
FJ525432_NZRB-G90	977	145	0.38	75	70	26	0	145	0	138	42.43	43.40	822	0.90
KC517487_FS703-T36	981	85	0.23	28	57	32	0	85	0	84	24.81	43.63	601	0.61
JQ798289_A18	977	83	0.22	33	50	34	0	83	0	70	23.95	43.30	706	0.72
EU857538_SP	977	71	0.19	46	25	18	0	71	0	58	20.71	43.60	910	0.93
GQ454870_HA16-5	977	60	0.16	27	33	34	0	60	0	41	17.84	43.71	826	0.85

Table 8.6: Number of reads mapping back to references for sample **13-3727**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	192028	65.74	91240	100788	158172	0	177052	0	102799	46652	43.81	967	0.99
KC525952_T3	977	33240	11.38	20495	12745	20852	0	32958	0	32363	7974	42.89	965	0.99
JQ798289_A18	977	26830	9.18	10670	16160	16400	0	26026	0	26462	6565	43.30	881	0.90
EU857538_SP	977	19667	6.73	13142	6525	10396	0	19423	0	16075	4864	43.60	930	0.95
JX266712_Taiwan-Pum_SP_T1	977	11244	3.85	5584	5660	10378	0	10750	0	4645	2514	43.30	974	1.00
GQ454870_HA16-5	977	7392	2.53	2548	4844	3248	0	7358	0	6966	1868	43.71	900	0.92
FJ525435_NZRB-M17	977	612	0.21	320	292	482	0	610	0	571	119	43.71	726	0.74
FJ525432_NZRB-G90	977	585	0.20	282	303	338	0	562	0	515	131	43.40	729	0.75

Table 8.7: Number of reads mapping back to references for sample **13-3729**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	139495	40.34	67627	71868	106512	0	128432	0	62700	30753	43.81	969	0.99
GQ454870_HA16-5	977	85985	24.86	44120	41865	60378	0	85456	0	57481	22011	43.71	937	0.96
EU857538_SP	977	32732	9.47	19141	13591	21302	0	32201	0	21739	7741	43.60	925	0.95
JQ798289_A18	977	29357	8.49	12474	16883	19130	0	28356	0	28759	6753	43.30	900	0.92
KC525952_T3	977	25510	7.38	13679	11831	18666	0	25170	0	22745	5297	42.89	961	0.98
JX266712_Taiwan-Pum_SP_T1	977	20525	5.94	10105	10420	17668	0	20056	0	8210	4214	43.30	967	0.99
FJ525431_NZRB-M12	977	8538	2.47	4510	4028	7088	0	8237	0	5797	1875	43.30	958	0.98
FJ525432_NZRB-G90	977	1965	0.57	989	976	1516	0	1927	0	1107	365	43.40	896	0.92
FJ525435_NZRB-M17	977	893	0.26	470	423	752	0	857	0	745	140	43.71	711	0.73

Table 8.8: Number of reads mapping back to references for sample **13-3732**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	183607	52.75	88420	95187	135948	0	168392	0	87907	41756	43.81	971	0.99
JX266712_Taiwan-Pum_SP_T1	977	57328	16.47	28219	29109	50444	0	56523	0	25378	11772	43.30	969	0.99
KC525952_T3	977	34073	9.79	19964	14109	23328	0	33732	0	32804	7862	42.89	963	0.99
JQ798289_A18	977	30980	8.90	13627	17353	22134	0	29949	0	29628	6916	43.30	897	0.92
EU857538_SP	977	20339	5.84	12993	7346	11698	0	20037	0	16308	4597	43.60	930	0.95
GQ454870_HA16-5	977	11617	3.34	4650	6967	6788	0	11557	0	10969	2784	43.71	935	0.96
FJ525435_NZRB-M17	977	3730	1.07	1955	1775	3160	0	3695	0	3466	676	43.71	676	0.69
FJ525432_NZRB-G90	977	2766	0.79	1448	1318	2016	0	2685	0	2171	546	43.40	885	0.91
FJ525431_NZRB-M12	977	2512	0.72	1402	1110	1910	0	2497	0	1932	487	43.30	862	0.88
JF957196_B301	977	630	0.18	315	315	600	0	630	0	595	68	42.89	813	0.83

Table 8.9: Number of reads mapping back to references for sample **13-3743**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
GQ454870_HA16-5	977	32922	40.74	17125	15797	25266	0	32916	0	26600	9838	43.71	926	0.95
HM573451_Kpg3	977	16887	20.89	7933	8954	9766	0	16884	0	12097	4990	43.81	975	1.00
JX266712_Taiwan-Pum_SP_T1	977	14065	17.40	6865	7200	10728	0	13966	0	7915	4161	43.30	975	1.00
FJ525431_NZRB-M12	977	6921	8.56	4263	2658	4558	0	6734	0	5738	2048	43.30	951	0.97
JQ798289_A18	977	4497	5.56	1725	2772	1550	0	4497	0	4489	1324	43.30	827	0.85
EU857538_SP	977	3982	4.93	2258	1724	1938	0	3980	0	2403	1182	43.60	869	0.89
KC525952_T3	977	762	0.94	483	279	166	0	762	0	577	220	42.89	937	0.96
FJ525432_NZRB-G90	977	381	0.47	163	218	98	0	381	0	350	112	43.40	671	0.69
FJ525433_NZRB-TH28	977	262	0.32	131	131	232	0	261	0	242	79	43.81	747	0.76

Table 8.10: Number of reads mapping back to references for sample **13-3745**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	46276	49.71	21424	24852	29058	0	46167	0	32769	13689	43.81	966	0.99
JX266712_Taiwan-Pum_SP_T1	977	14639	15.72	7184	7455	10934	0	14497	0	7568	4327	43.30	964	0.99
KC525952_T3	977	8463	9.09	5935	2528	3266	0	8463	0	8255	2477	42.89	940	0.97
JQ798289_A18	977	8145	8.75	3039	5106	3340	0	8145	0	8121	2408	43.30	909	0.87
EU857538_SP	977	6452	6.93	4635	1817	2146	0	6451	0	5559	1905	43.60	935	0.96
GQ454870_HA16-5	977	5072	5.45	1671	3401	1428	0	5072	0	4908	1505	43.71	915	0.94
FJ525431_NZRB-M12	977	3162	3.40	2056	1106	1864	0	3087	0	2556	933	43.30	938	0.96
FJ525432_NZRB-G90	977	488	0.52	231	257	142	0	488	0	476	145	43.40	853	0.75
FJ525435_NZRB-M17	977	176	0.19	90	86	56	0	176	0	175	50	43.71	883	0.55
FJ525433_NZRB-TH28	977	150	0.16	76	74	132	0	150	0	129	45	43.81	774	0.84

Table 8.11: Number of reads mapping back to references for sample **13-3747**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	60824	69.66	29868	30956	52720	59581	0	23750	13937	43.30	974	1.00
AY170468_T36	981	15368	17.60	7099	8269	12118	15368	0	14861	4245	43.63	941	0.96
FJ525431_NZRB-M12	977	2421	2.77	1526	895	1464	2386	0	2271	542	43.30	899	0.92
FJ525435_NZRB-M17	977	1847	2.12	925	922	1446	1816	0	1622	301	43.71	763	0.78
FJ525432_NZRB-G90	977	1562	1.79	813	749	1174	1457	0	1122	292	43.40	921	0.94
EU937521_T36	981	1407	1.61	763	644	1030	1407	0	1076	378	43.43	955	0.97
JQ798289_A18	977	1300	1.49	630	670	1112	1294	0	932	189	43.30	798	0.82
HM573451_Kpg	977	1131	1.30	551	580	952	1090	0	437	233	43.81	959	0.98
JF957196_B301	977	316	0.36	158	158	276	312	0	256	46	42.89	805	0.82
KC525952_T3	977	206	0.24	108	98	164	205	0	187	45	42.89	742	0.76
EU857538_SP	977	133	0.15	77	56	100	133	0	95	24	43.60	844	0.86
GQ454869_HA18-9	977	116	0.13	65	51	82	116	0	52	19	42.89	779	0.80
KC517487_FS703-T36	981	111	0.13	55	56	62	111	0	107	27	43.63	653	0.67
GQ454870_HA16-5	977	108	0.12	48	60	66	106	0	98	24	43.71	831	0.85

Table 8.12: Number of reads mapping back to references for sample **13-3748**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
HM573451_Kpg3	977	155164	53.41	74222	80942	105858	0	144465	0	79758	37343	43.81	968	0.99
JX266712_Taiwan-Pum_SP_T1	977	41381	14.25	20268	21113	35066	0	40547	0	17919	8930	43.30	968	0.99
KC525952_T3	977	30131	10.37	18109	12022	19220	0	29875	0	29003	7166	42.89	948	0.97
JQ798289_A18	977	25375	8.74	10684	14691	16212	0	24820	0	24472	5953	43.30	895	0.92
EU857538_SP	977	18952	6.52	12387	6565	9946	0	18752	0	15022	4566	43.60	930	0.95
GQ454870_HA16-5	977	11403	3.93	4528	6875	6050	0	11328	0	9816	2881	43.71	921	0.94
FJ525431_NZRB-M12	977	2662	0.92	1487	1175	1964	0	2628	0	2087	551	43.30	935	0.96
FJ525435_NZRB-M17	977	2445	0.84	1291	1154	1996	0	2426	0	2229	450	43.71	727	0.74
FJ525432_NZRB-G90	977	2058	0.71	1108	950	1444	0	2011	0	1609	424	43.40	829	0.85
JF957196_B301	977	428	0.15	214	214	406	0	422	0	401	47	42.89	787	0.80

Table 8.13: Number of reads mapping back to references for sample **13-3768**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	40616	60.71	19757	20859	31516	0	40604	0	24034	12035	43.30	972	0.99
FJ525431_NZRB-M12	977	18319	27.38	11500	6819	11866	0	18207	0	15982	5420	43.30	969	0.99
HM573451_Kpg3	977	2472	3.70	1126	1346	1532	0	2472	0	1406	733	43.81	965	0.99
FJ525433_NZRB-TH28	977	1511	2.26	759	752	1444	0	1511	0	1449	456	43.81	819	0.84
FJ525432_NZRB-G90	977	1282	1.92	564	718	452	0	1282	0	1217	380	43.40	868	0.89
KC525952_T3	977	797	1.19	479	318	478	0	797	0	773	235	42.89	910	0.93
FJ525435_NZRB-M17	977	523	0.78	302	221	138	0	523	0	504	152	43.71	713	0.73
JQ798289_A18	977	510	0.76	217	293	170	0	510	0	509	151	43.30	800	0.82
GQ454870_HA16-5	977	356	0.53	171	185	144	0	356	0	339	106	43.71	871	0.89
EU857538_SP	977	309	0.46	174	135	110	0	309	0	258	91	43.60	830	0.85
JF957196_B301	977	77	0.12	20	57	26	0	77	0	69	22	42.89	685	0.70
GQ454869_HA18-9	977	70	0.10	41	29	6	0	70	0	60	20	42.89	909	0.93

Table 8.14: Number of reads mapping back to references for sample **13-3770**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	79139	79.94	38855	40284	71430	0	77602	0	40591	23317	43.30	971	0.99
HM573451_Kpg3	977	8322	8.41	3882	4440	5672	0	8321	0	4973	2447	43.81	963	0.99
KC525952_T3	977	2565	2.59	1556	1009	1386	0	2565	0	2496	755	42.89	920	0.94
EU857538_SP	977	2184	2.21	1145	1039	922	0	2184	0	1124	644	43.60	852	0.87
FJ525431_NZRB-M12	977	1634	1.65	1337	297	260	0	1634	0	1599	469	43.30	833	0.85
GQ454870_HA16-5	977	1399	1.41	662	737	846	0	1399	0	1238	415	43.71	906	0.93
CT-ZA1	977	1217	1.23	613	604	970	0	1212	0	1069	360	43.81	944	0.97
FJ525435_NZRB-M17	977	944	0.95	462	482	186	0	944	0	939	274	43.71	833	0.85
FJ525432_NZRB-G90	977	671	0.68	346	325	126	0	671	0	667	194	43.40	830	0.85
KC517493_FL202-VT	977	406	0.41	234	172	168	0	406	0	406	118	43.09	497	0.51
JQ798289_A18	977	239	0.24	116	123	82	0	239	0	203	69	43.30	877	0.90

Table 8.15: Number of reads mapping back to references for sample **13-3776**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JX266712_Taiwan-Pum_SP_T1	977	33707	59.15	16431	17276	26254	0	33446	0	18463	9941	43.30	970	0.99
FJ525431_NZRB-M12	977	13140	23.06	8204	4936	8604	0	13069	0	11474	3866	43.30	936	0.96
GQ454870_HA16-5	977	2411	4.23	1139	1272	1650	0	2411	0	2257	718	43.71	914	0.94
HM573451_Kpg3	977	2116	3.71	1123	993	1148	0	2116	0	1025	619	43.81	961	0.98
EU857538_SP	977	1568	2.75	785	783	934	0	1568	0	1226	465	43.60	852	0.87
FJ525432_NZRB-G90	977	1192	2.09	548	644	386	0	1192	0	1089	350	43.40	743	0.76
KC525952_T3	977	1132	1.99	680	452	444	0	1132	0	1105	334	42.89	772	0.79
FJ525433_NZRB-TH28	977	916	1.61	456	460	880	0	916	0	876	276	43.81	818	0.84
FJ525435_NZRB-M17	977	432	0.76	247	185	114	0	432	0	421	124	43.71	695	0.71
JQ798289_A18	977	177	0.31	81	96	70	0	177	0	146	50	43.30	887	0.91
GQ454869_HA18-9	977	103	0.18	55	48	26	0	103	0	91	29	42.89	846	0.87

Table 9: Number of reads mapping back to references for sample **Laf control**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
EU921620 Laf 16S	1432	369533	99.75	184363	185170	332640	0	342399	0	81016	46662	53.00	1174	0.82
JQ061137_AT-1	977	396	0.11	197	199	388	0	378	0	174	50	43.60	907	0.93
JX266712_Taiwan-Pum_SP_T1	977	94	0.03	51	43	44	0	81	0	38	16	43.30	938	0.96
KC262793_L192GR	977	47	0.01	24	23	46	0	47	0	26	6	43.91	604	0.62
FJ525431_NZRB-M12	977	45	0.01	24	21	8	0	43	0	34	10	43.30	680	0.70
HM573451_Kpg	977	44	0.01	17	27	10	0	36	0	16	8	43.81	937	0.96
JQ911664_CT11A	977	39	0.01	20	19	38	0	33	0	29	5	43.60	475	0.49
KC517493_FL202-VT	977	37	0.01	19	18	36	0	37	0	34	5	43.09	379	0.39
CT-ZA1	977	34	0.01	18	16	32	0	34	0	34	8	43.81	456	0.47
JQ911663_CT14A	977	31	0.01	15	16	30	0	29	0	31	5	43.50	359	0.37
CT-ZA3	977	30	0.01	14	16	28	0	24	0	25	6	43.60	589	0.60
DQ151548_T318A	977	26	0.01	13	13	26	0	26	0	26	3	43.30	183	0.19
KC525952_T3	977	20	0.01	11	9	6	0	20	0	19	4	42.89	615	0.63
FJ525433_NZRB-TH28	977	17	0.00	9	8	4	0	10	0	16	4	43.81	475	0.49
FJ525432_NZRB-G90	977	15	0.00	6	9	4	0	15	0	12	3	43.40	469	0.48
FJ525434_NZRB-TH30	977	11	0.00	5	6	6	0	10	0	10	2	43.81	425	0.44
EU857538_SP	977	8	0.00	6	2	0	0	8	0	8	2	43.60	294	0.30
AB0463981_NUagA	977	7	0.00	3	4	6	0	7	0	6	1	43.50	156	0.16
GQ454870_HA16-5	977	6	0.00	3	3	2	0	5	0	3	1	43.71	536	0.54
JQ798289_A18	977	6	0.00	2	4	0	0	6	0	4	1	43.30	368	0.38
FJ525435_NZRB-M17	977	5	0.00	3	2	2	0	4	0	3	0	43.71	173	0.18
U56902_VT	977	5	0.00	2	3	4	0	5	0	4	0	43.60	179	0.18
JX266713_Taiwan-Pum_M_T5	977	3	0.00	1	2	0	0	1	0	1	0	43.30	78	0.08
EU076703_B165	978	2	0.00	1	1	2	0	2	0	2	0	44.48	152	0.16
EU937521_T36	981	1	0.00	1	0	0	0	1	0	1	0	43.43	283	0.29
JQ965169_T68	977	1	0.00	0	1	0	0	0	0	1	0	44.63	8	0.01
KC517492_FS703-VT	977	1	0.00	1	0	0	0	1	0	1	0	43.09	117	0.12

Table 10.1: Number of reads mapping back to references for sample **14-6000**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JQ061137_AT-1	977	76997	49.38	37604	39393	69916	0	76337	0	34055	13519	43.60	955	0.98
KC517493_FL202-VT	977	20193	12.95	11190	9003	15478	0	18641	0	19984	4309	43.09	963	0.99
CT-ZA1	977	16941	10.86	9839	7102	10590	0	16776	0	16108	4482	43.81	969	0.99
JQ911664_CT11A	977	9872	6.33	5006	4866	8098	0	9741	0	8354	1894	43.60	912	0.93
KC262793_L192GR	977	8536	5.47	4294	4242	7758	0	8475	0	5453	1385	43.91	946	0.97
CT-ZA3	977	8260	5.30	3795	4465	7254	0	6659	0	6522	1765	43.60	970	0.99
JQ911663_CT14A	977	6308	4.05	3727	2581	4808	0	5961	0	6259	1373	43.50	815	0.83
DQ151548_T318A	977	4113	2.64	2042	2071	3532	0	4085	0	4024	606	43.30	797	0.82
JX266712_Taiwan-Pum_SP_T1	977	1720	1.10	848	872	1578	0	1696	0	594	291	43.30	943	0.97
AB0463981_NUagA	977	1558	1.00	778	780	1492	0	1552	0	1438	171	43.50	856	0.88
U56902_VT	977	272	0.17	129	143	242	0	268	0	162	22	43.60	741	0.76
KC517492_FS703-VT	977	243	0.16	149	94	158	0	233	0	241	38	43.09	486	0.50
JQ798289_A18	977	238	0.15	118	120	218	0	229	0	167	18	43.30	838	0.86

Table 10.2: Number of reads mapping back to references for sample **14-6001**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
FJ525433_NZRB-TH28	977	47499	40.67	24047	23452	42388	0	44826	0	39080	12519	43.81	970	0.99
FJ525431_NZRB-M12	977	28169	24.12	15434	12735	17118	0	27770	0	23167	7053	43.30	971	0.99
JX266712_Taiwan-Pum_SP_T1	977	19869	17.01	10136	9733	15946	0	18867	0	6742	4608	43.30	970	0.99
FJ525434_NZRB-TH30	977	16055	13.75	8874	7181	14108	0	15373	0	14169	3956	43.81	965	0.99
JQ061137_AT-1	977	1787	1.53	852	935	1584	0	1783	0	850	360	43.60	937	0.96
FJ525432_NZRB-G90	977	891	0.76	465	426	624	0	798	0	767	168	43.40	820	0.84
FJ525435_NZRB-M17	977	480	0.41	244	236	336	0	429	0	391	86	43.71	846	0.87
JF957196_B301	977	413	0.35	185	228	324	0	312	0	358	65	42.89	791	0.81
KC517493_FL202-VT	977	341	0.29	177	164	268	0	337	0	338	76	43.09	781	0.80
JQ911664_CT11A	977	270	0.23	138	132	196	0	268	0	262	62	43.60	871	0.89
CT-ZA1	977	251	0.21	133	118	190	0	251	0	219	65	43.81	950	0.97
JQ911663_CT14A	977	149	0.13	89	60	114	0	149	0	149	35	43.50	446	0.46
GQ454869_HA18-9	977	133	0.11	68	65	96	0	111	0	73	17	42.89	915	0.94
HM573451_Kpg3	977	120	0.10	56	64	42	0	109	0	55	26	43.81	899	0.92

Table 10.3: Number of reads mapping back to references for sample **14-6002**. The percentages of the total reads that mapped back to specific strains is listed in column 4. This provides an indication of the relative abundance of specific strains with the population. Additional data relating to the coverage of the reference sequence is listed in columns 10, 11 and 12.

Reference name	Reference length	Mapped reads	Percentage of mapped reads (%)	Forward reads	Reversed reads	Reads in aligned pairs	Non-specific matches	Non-perfect matches	Minimum coverage	Maximum coverage	Average coverage	% GC	Consensus length	Fraction of reference covered
JQ061137_AT-1	977	69804	48.69	32834	36970	59424	0	69610	1	33834	15409	43.60	977	1.00
CT-ZA1	977	27400	19.11	13854	13546	22152	0	27217	0	22728	7151	43.81	969	0.99
KC517493_FL202-VT	977	10299	7.18	6223	4076	6032	0	10299	0	10113	2330	43.09	956	0.98
JQ911664_CT11A	977	9148	6.38	4862	4286	5246	0	9123	0	8495	2097	43.60	945	0.97
KC262793_L192GR	977	8084	5.64	4239	3845	5610	0	8068	0	4775	1579	43.91	897	0.92
JQ911663_CT14A	977	7334	5.12	4642	2692	4766	0	7236	0	7301	1655	43.50	497	0.51
CT-ZA3	977	6508	4.54	2929	3579	5680	0	6218	0	4574	1541	43.60	955	0.98
DQ151548_T318A	977	3281	2.29	1611	1670	1794	0	3269	0	3162	665	43.30	883	0.90
AB0463981_NUagA	977	754	0.53	353	401	574	0	752	0	679	120	43.50	856	0.88
KC517492_FS703-VT	977	235	0.16	148	87	54	0	233	0	232	52	43.09	509	0.52

SECTION 2

Figure 1.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3013**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

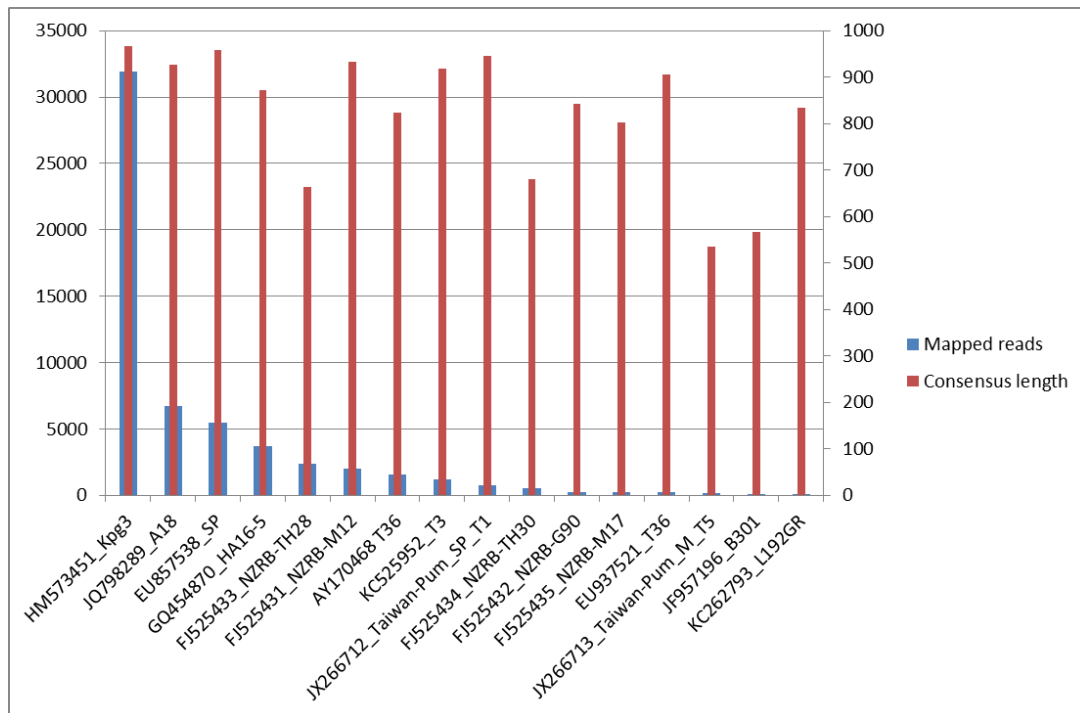


Figure 1.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3015**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

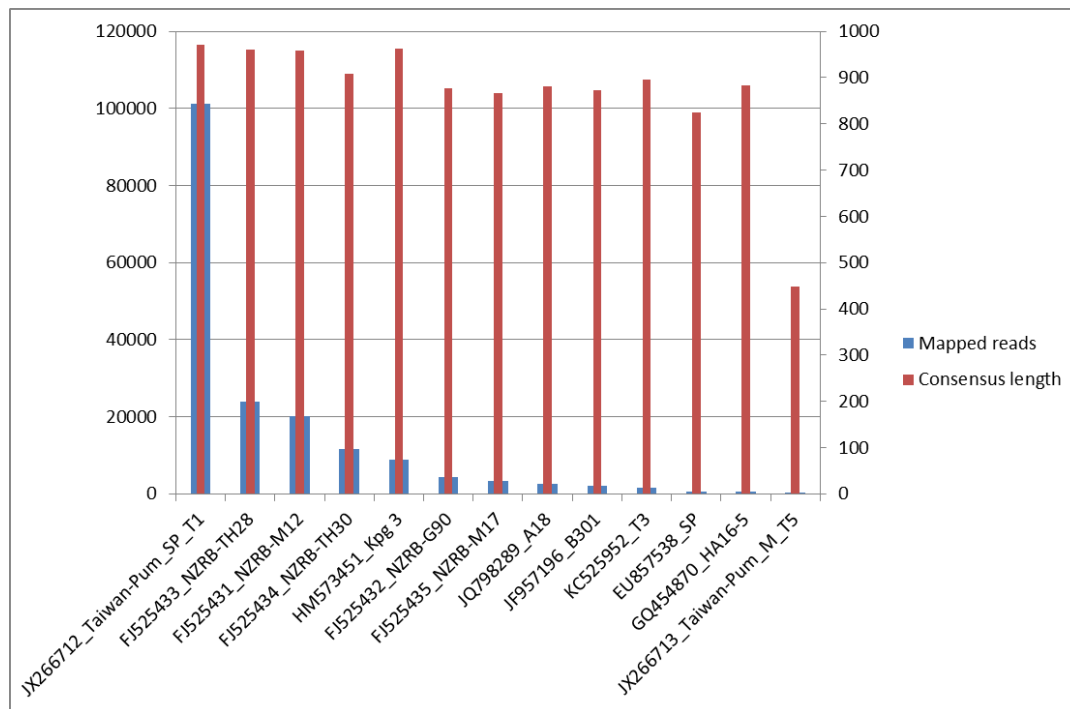


Figure 1.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3023**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

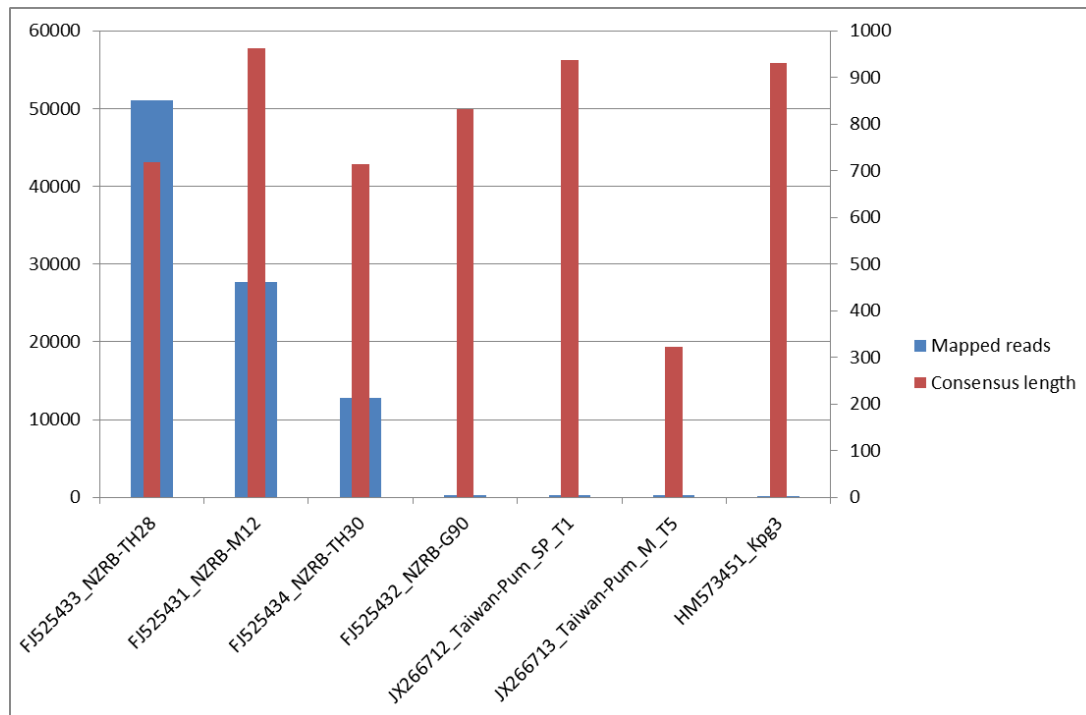


Figure 1.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3040**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

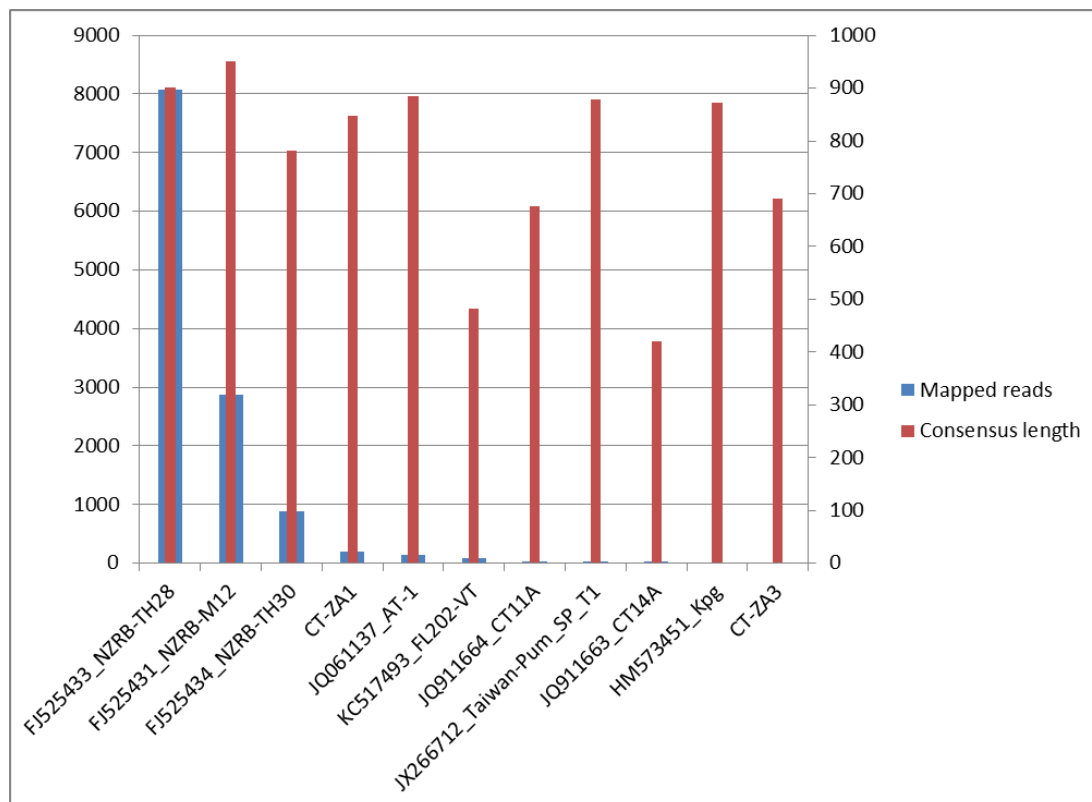


Figure 1.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3042**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

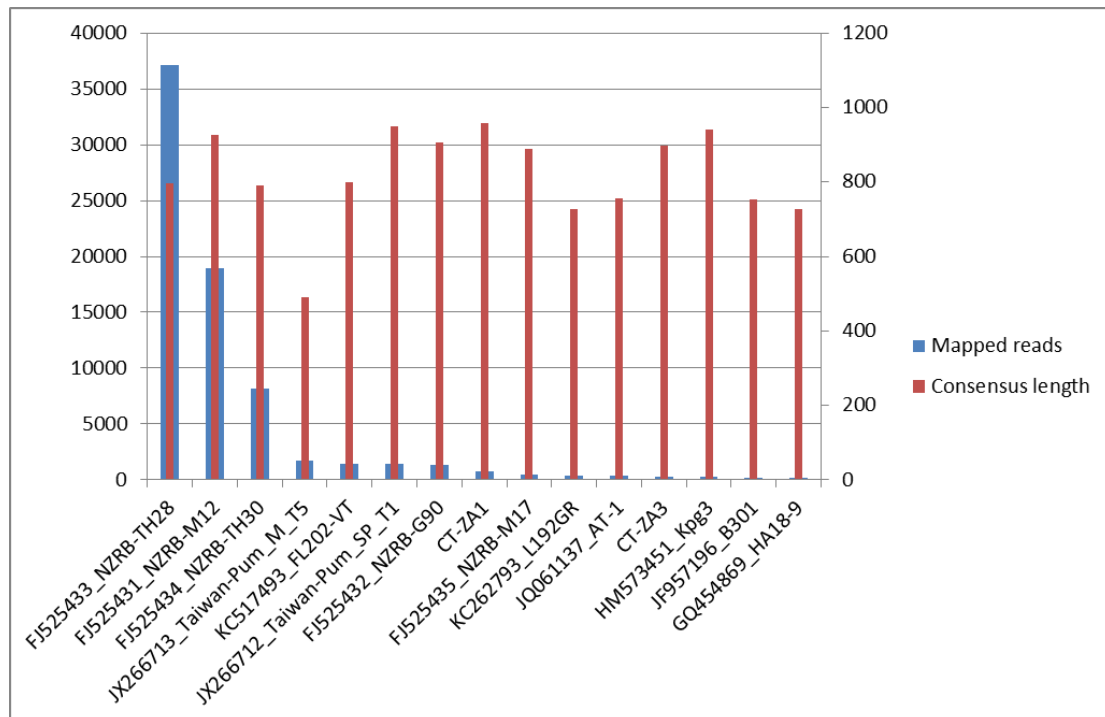


Figure 1.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3059**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

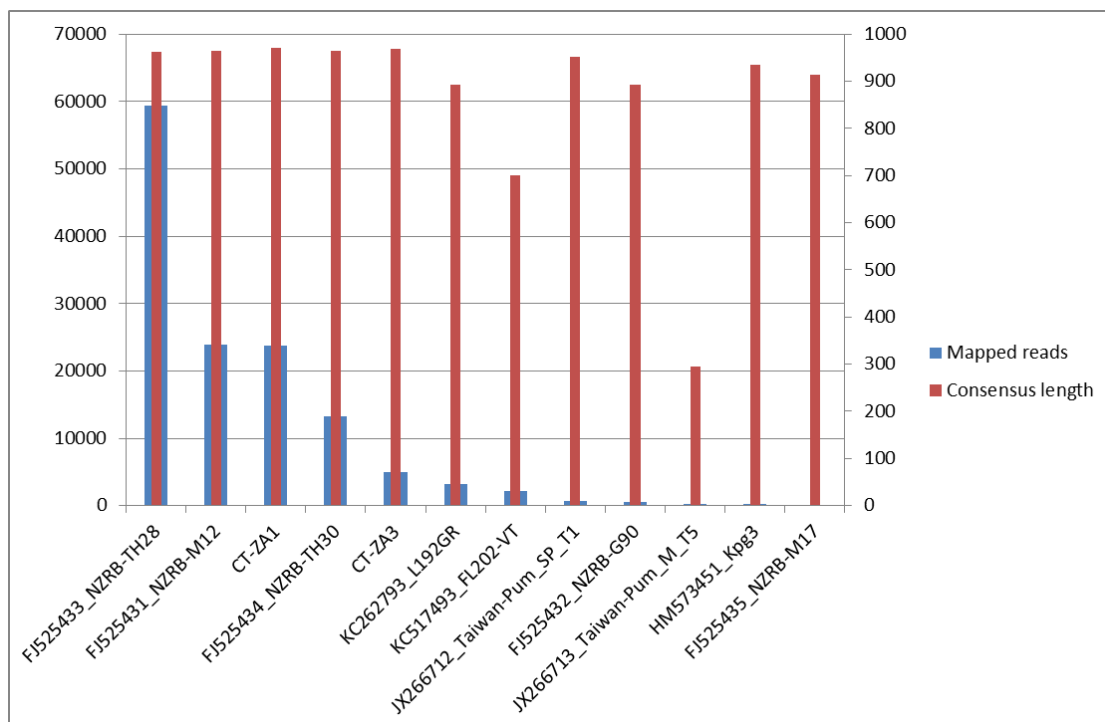


Figure 1.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3062**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

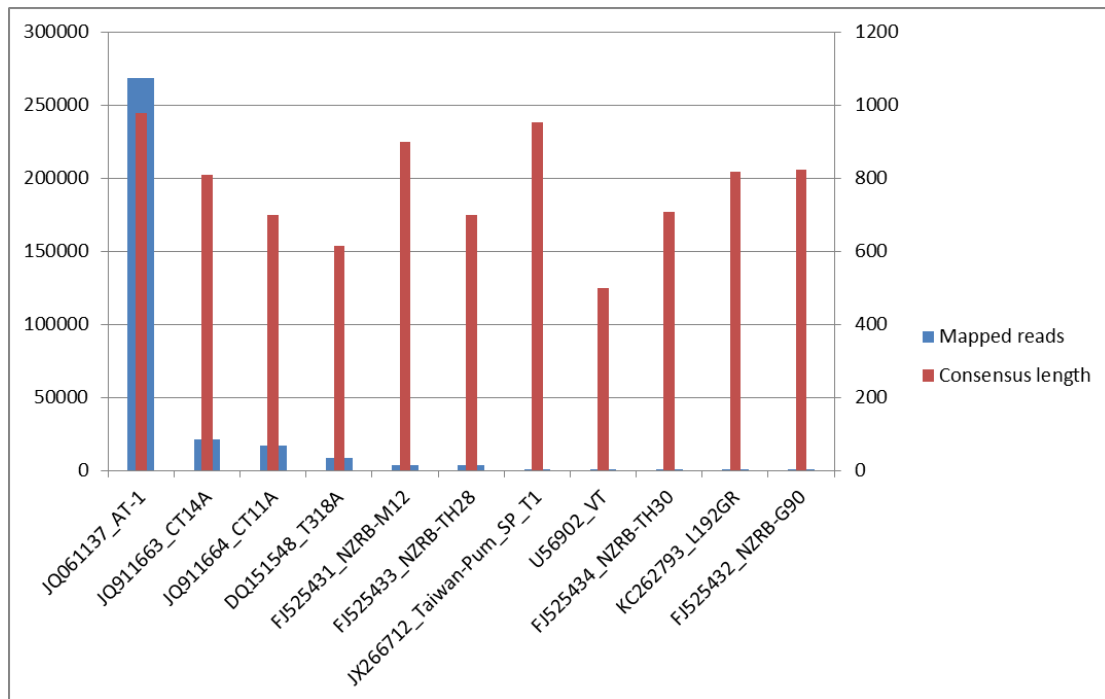


Figure 1.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3067**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

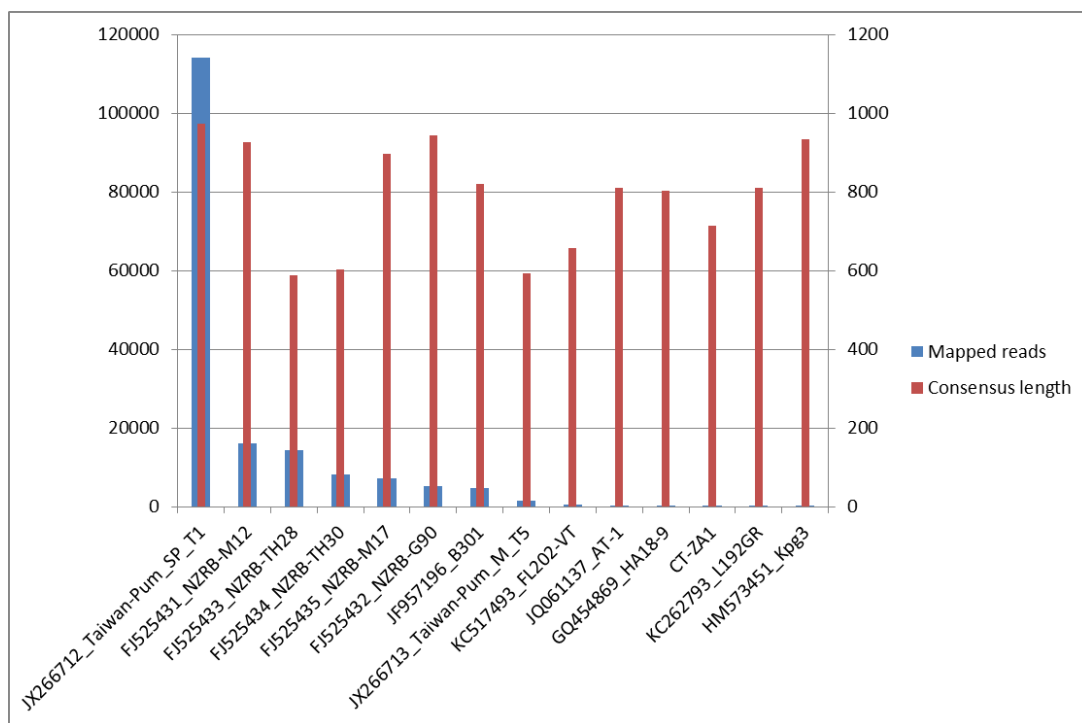


Figure 2.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3105**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

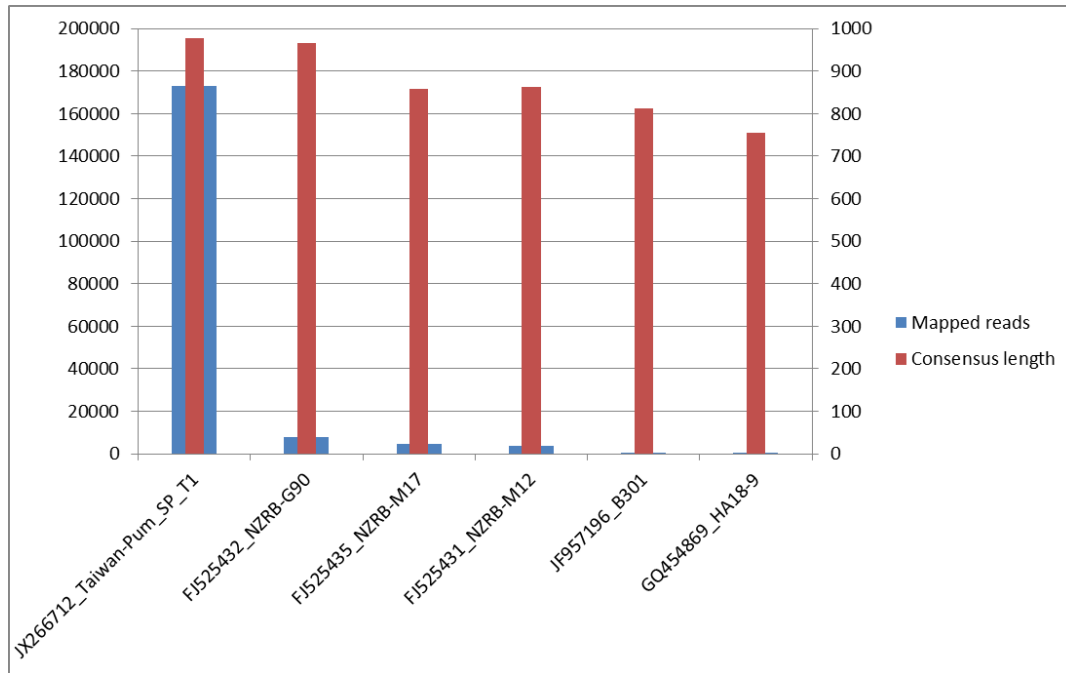


Figure 2.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3113**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

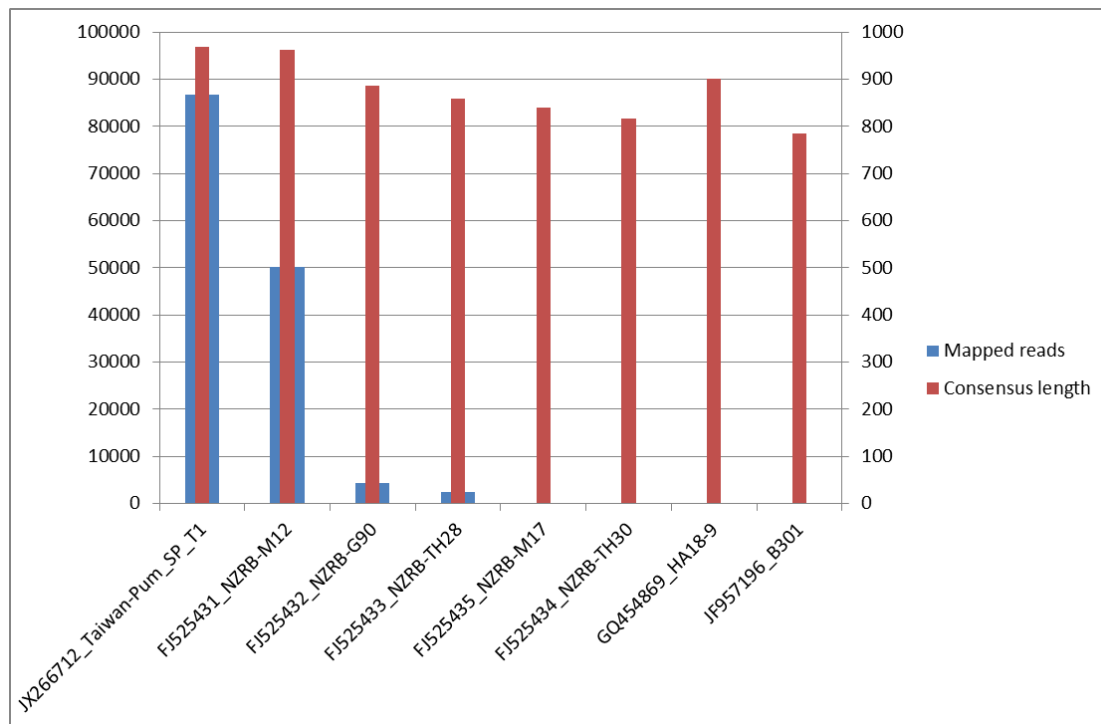


Figure 2.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3125**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

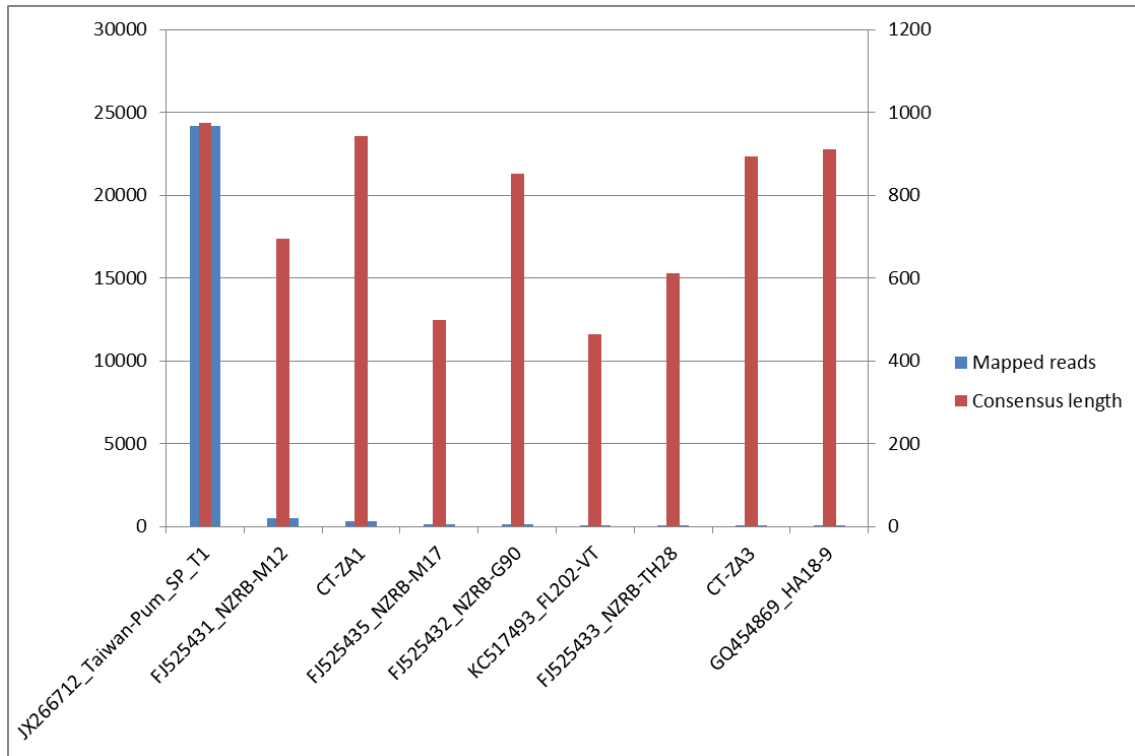


Figure 2.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3130**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

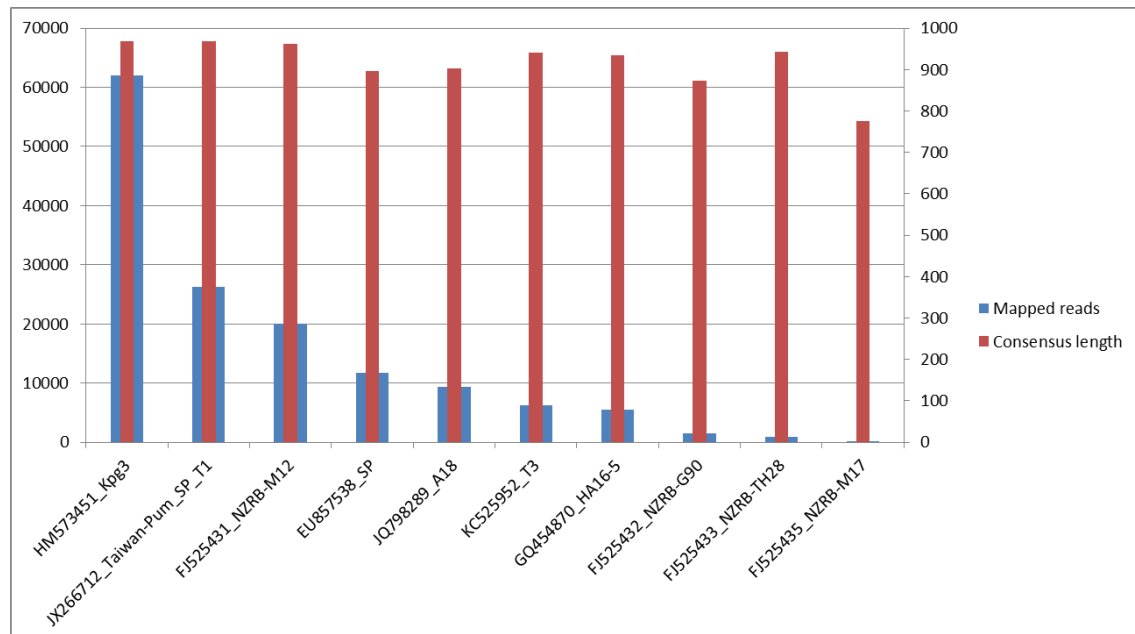


Figure 2.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3152**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

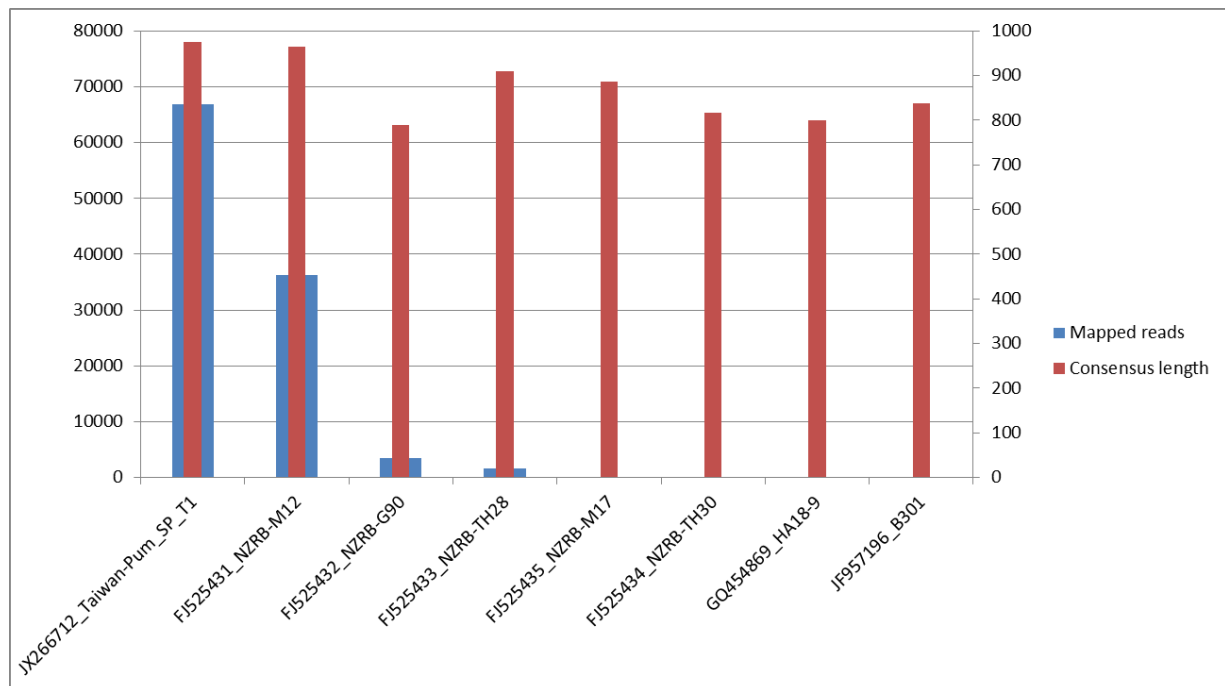


Figure 2.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3155**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

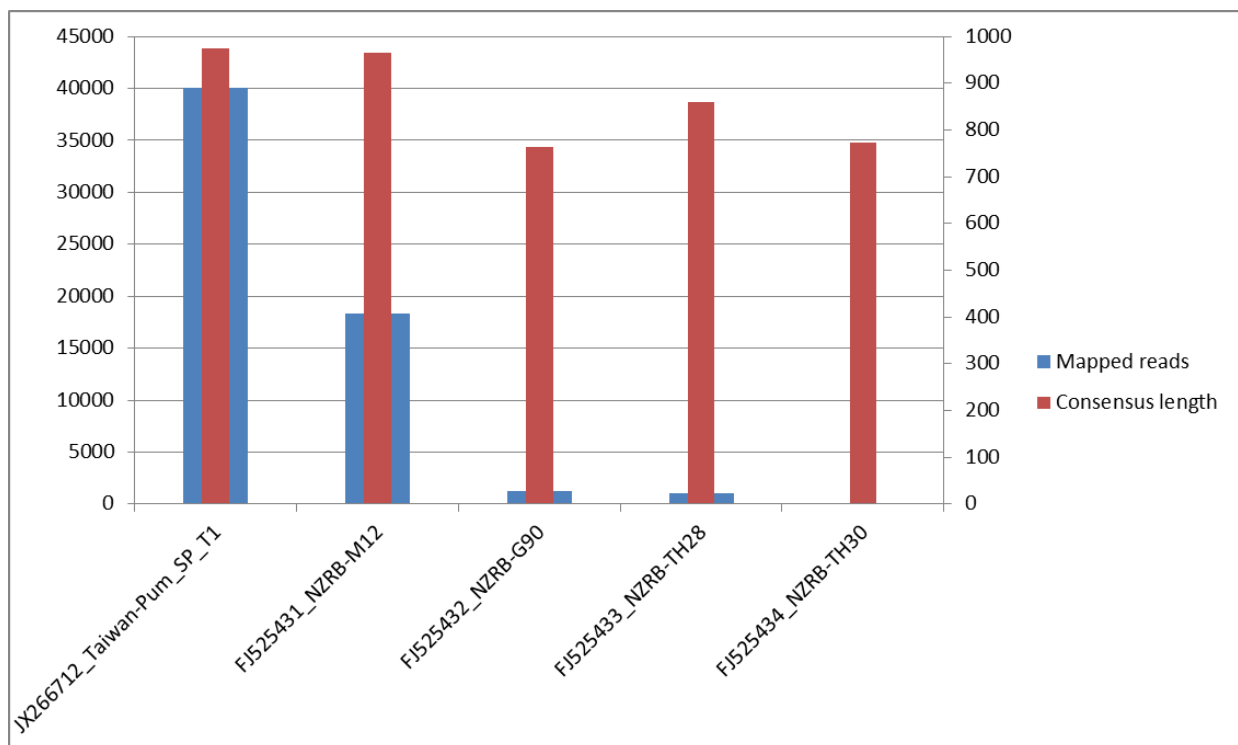


Figure 2.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3160**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

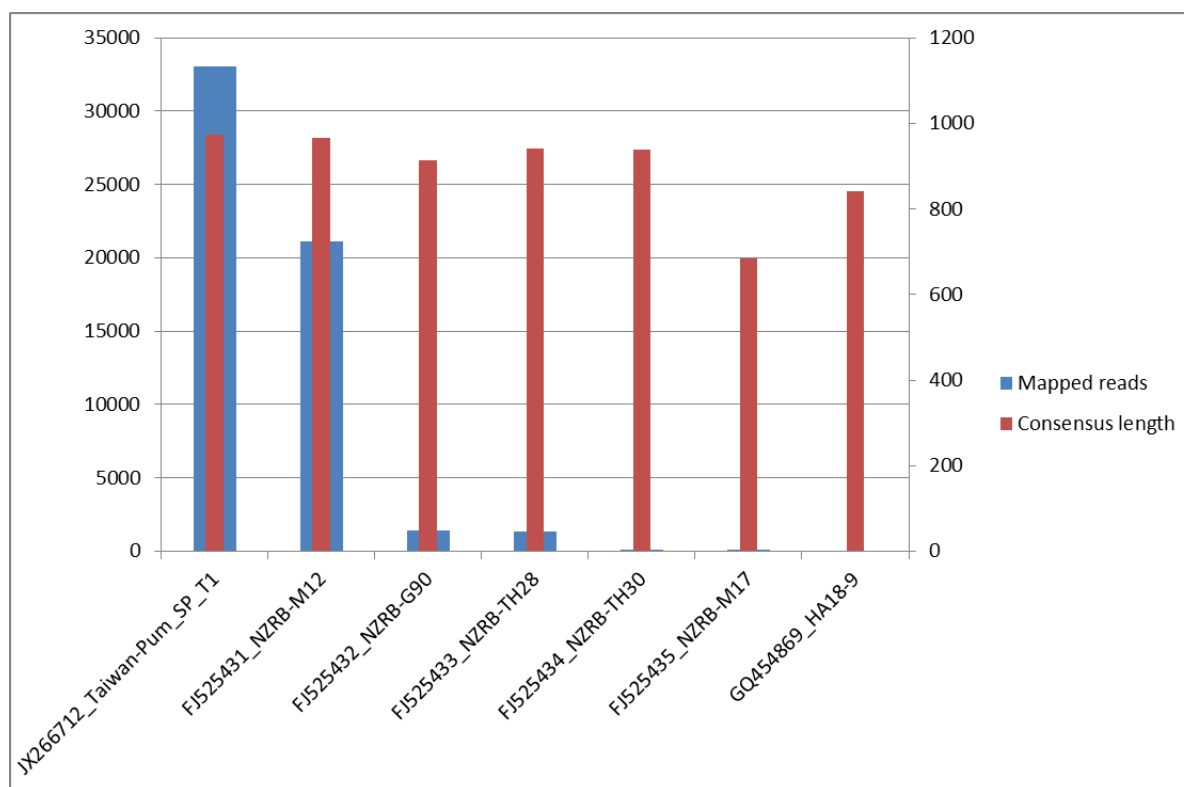


Figure 2.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3161**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

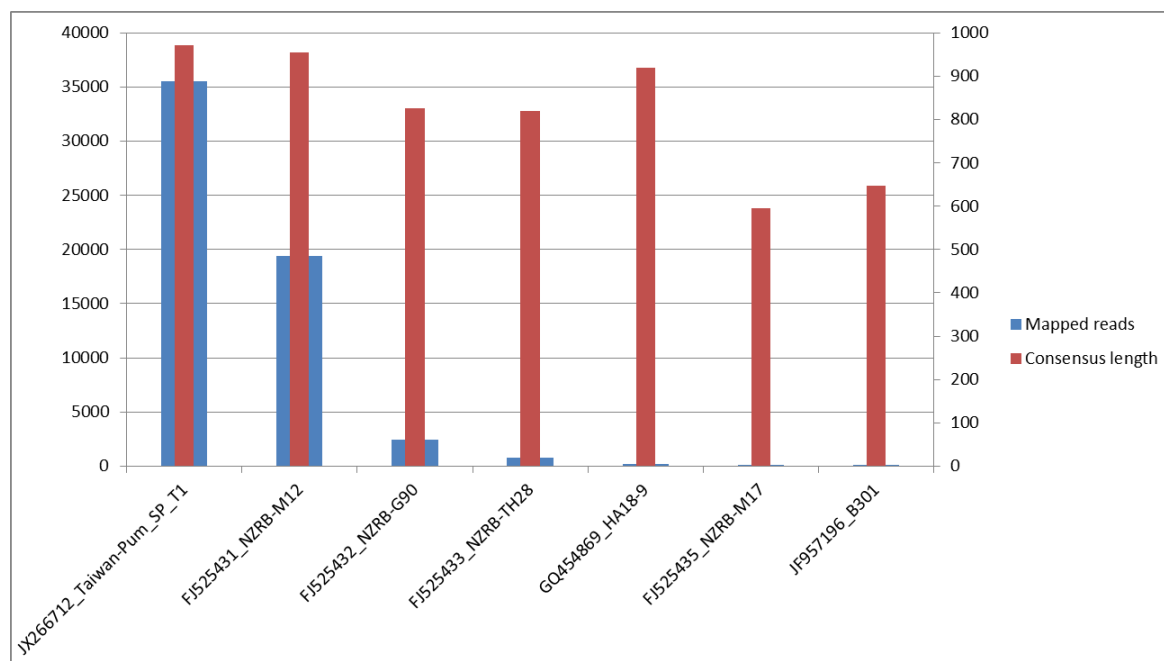


Figure 3.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3205**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

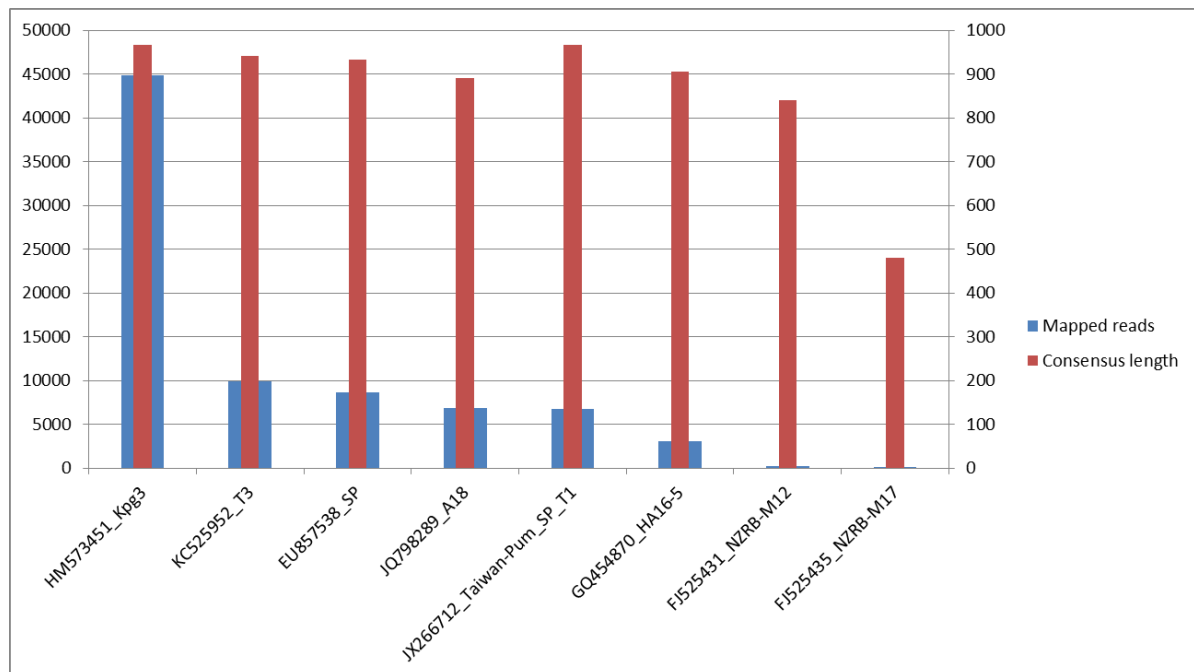


Figure 3.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3242**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

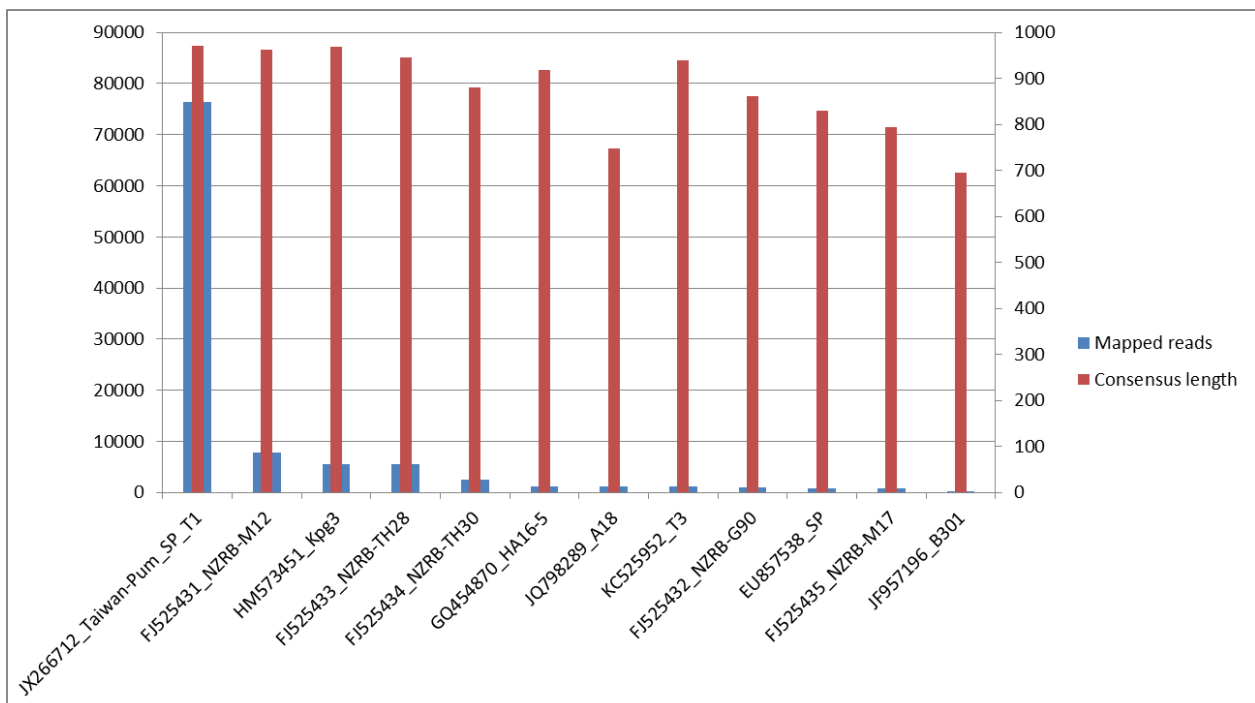


Figure 3.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3251**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

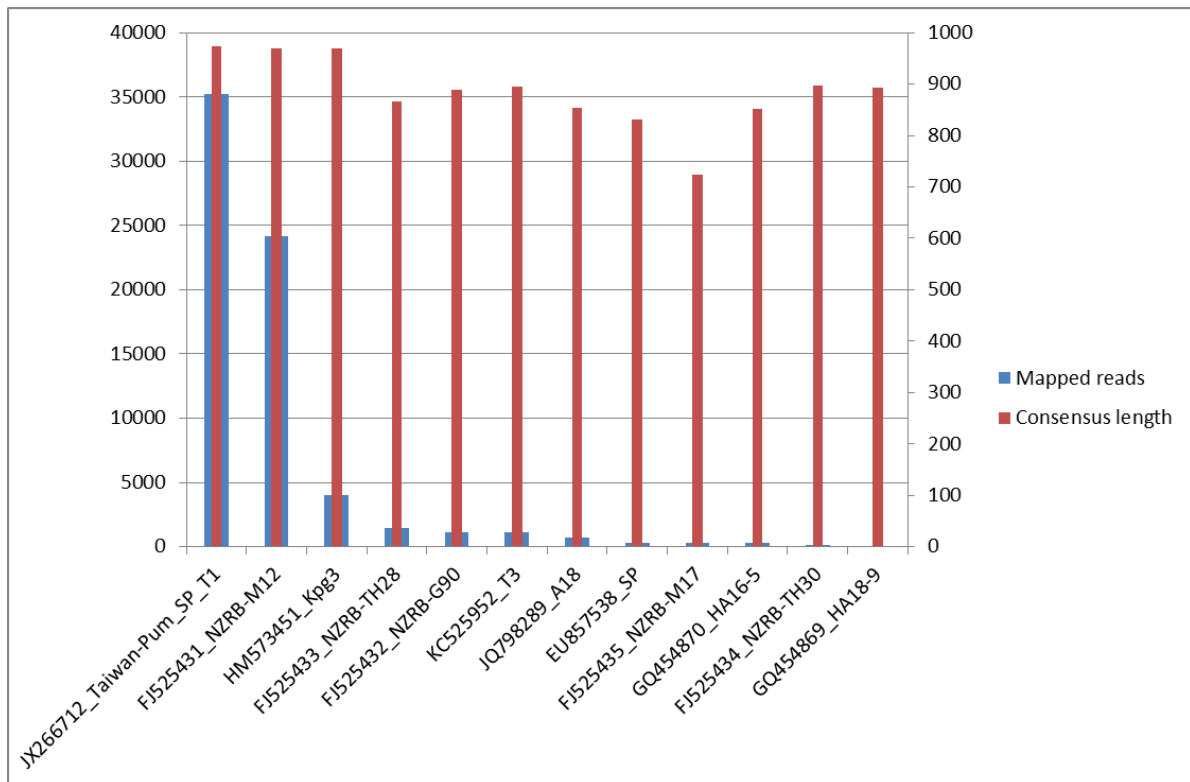


Figure 3.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3255**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

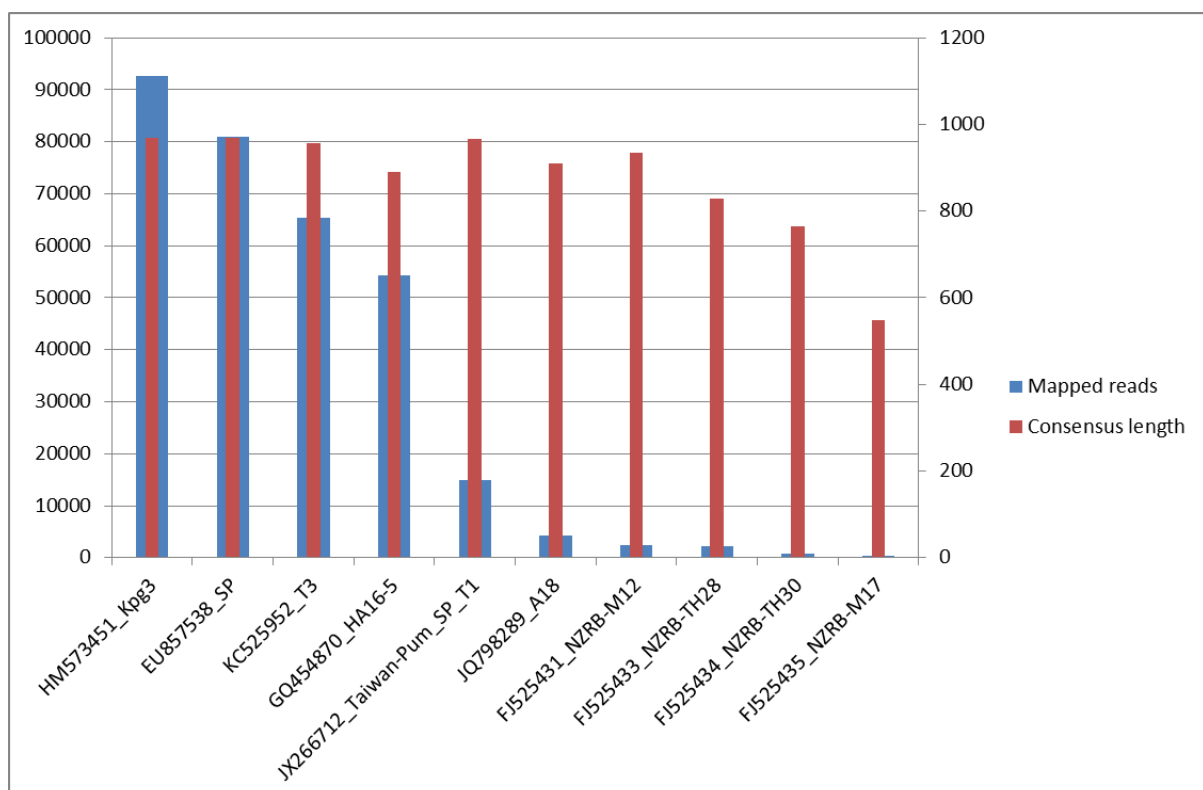


Figure 3.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3257**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

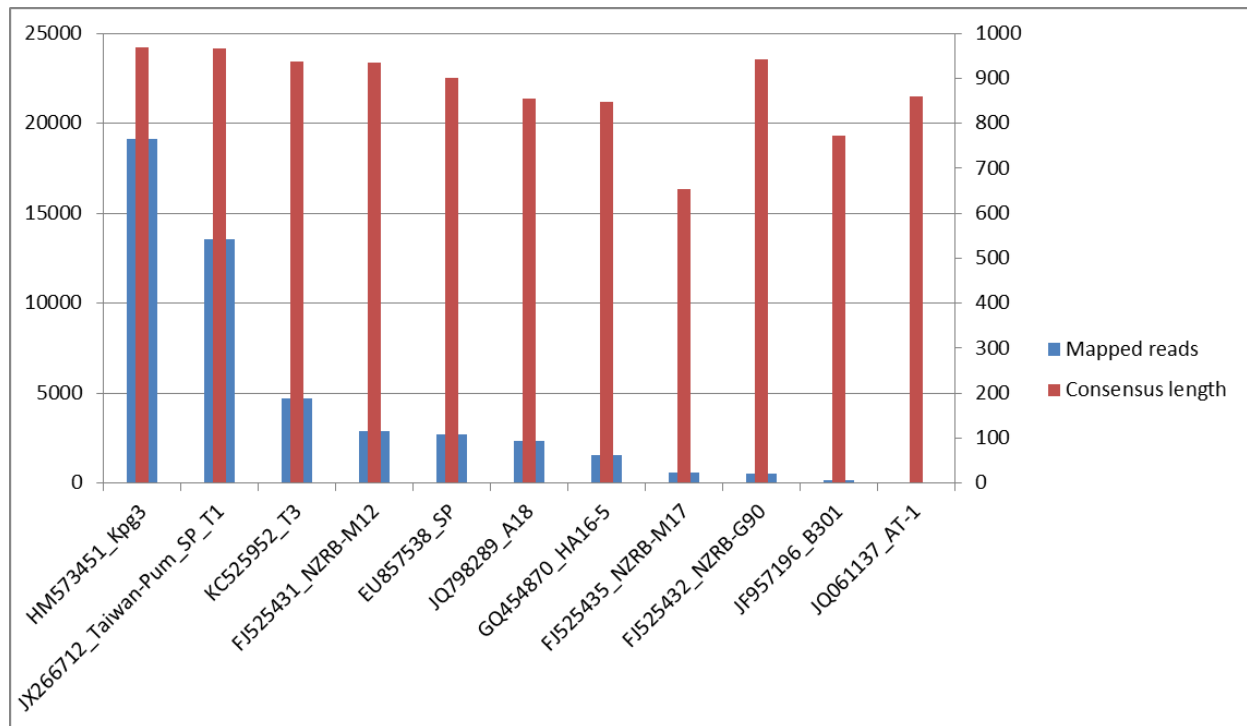


Figure 3.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3260**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

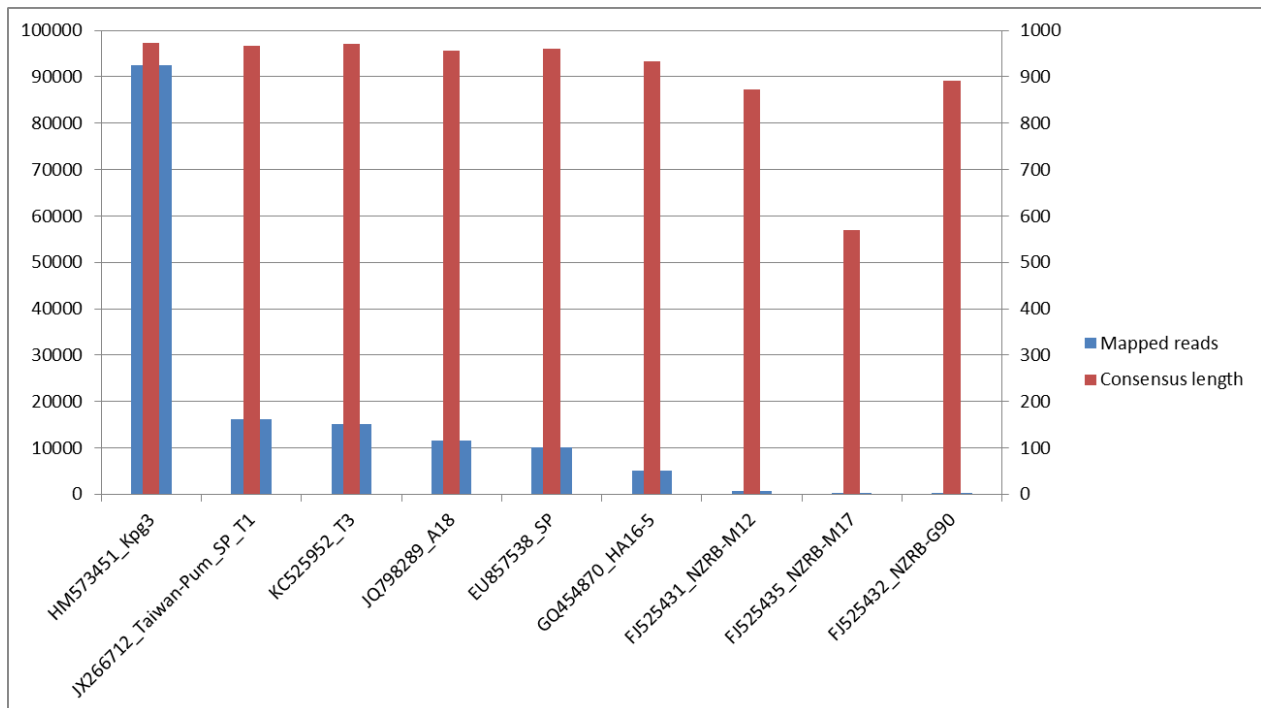


Figure 3.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3267**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

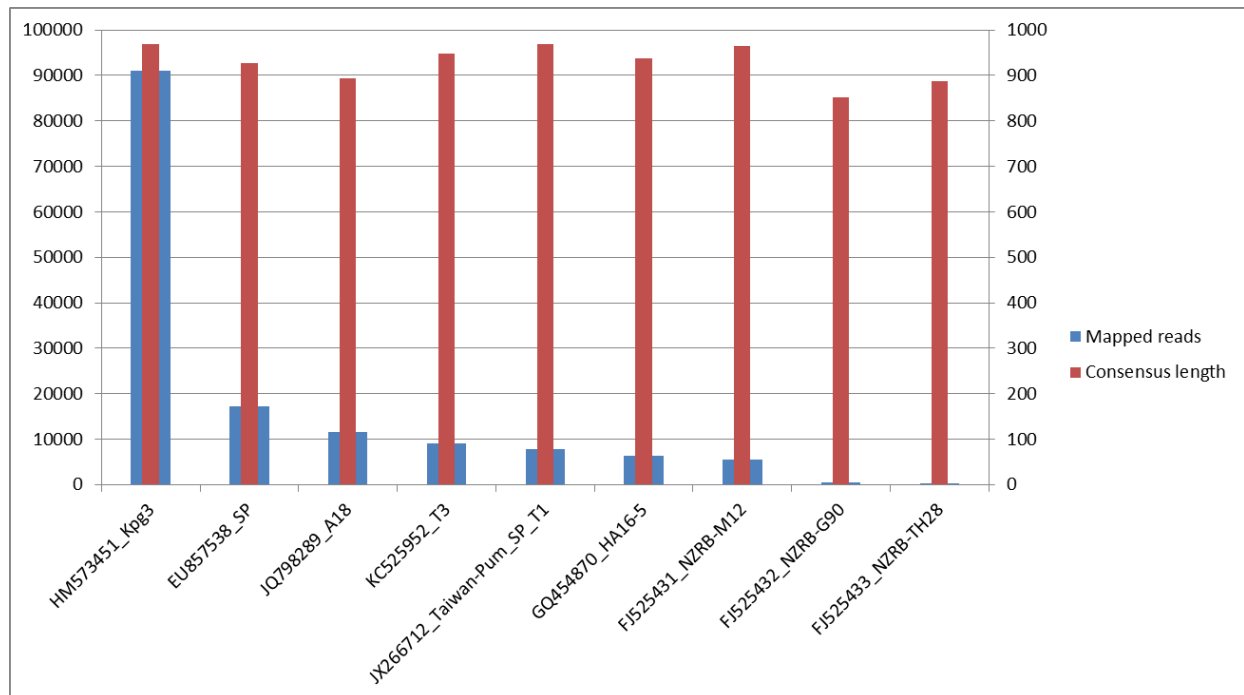


Figure 3.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3271**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

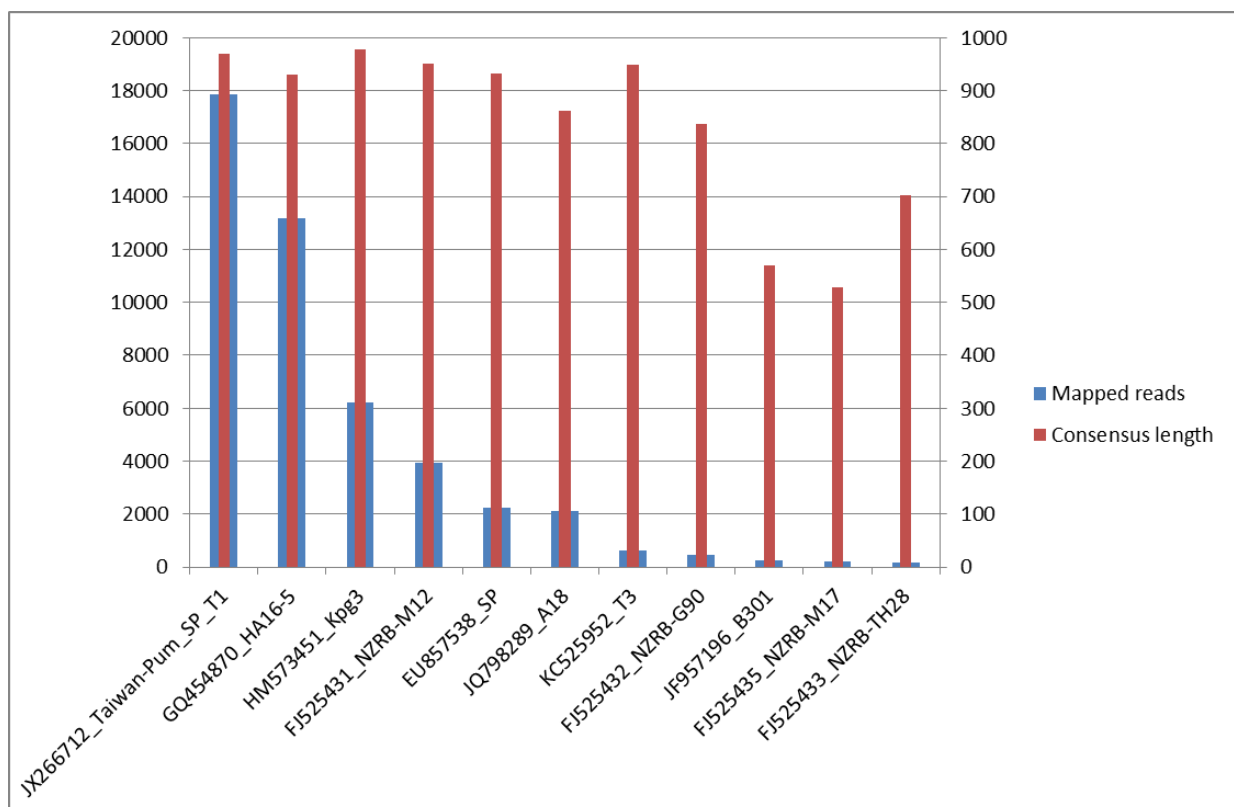


Figure 4.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3305**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

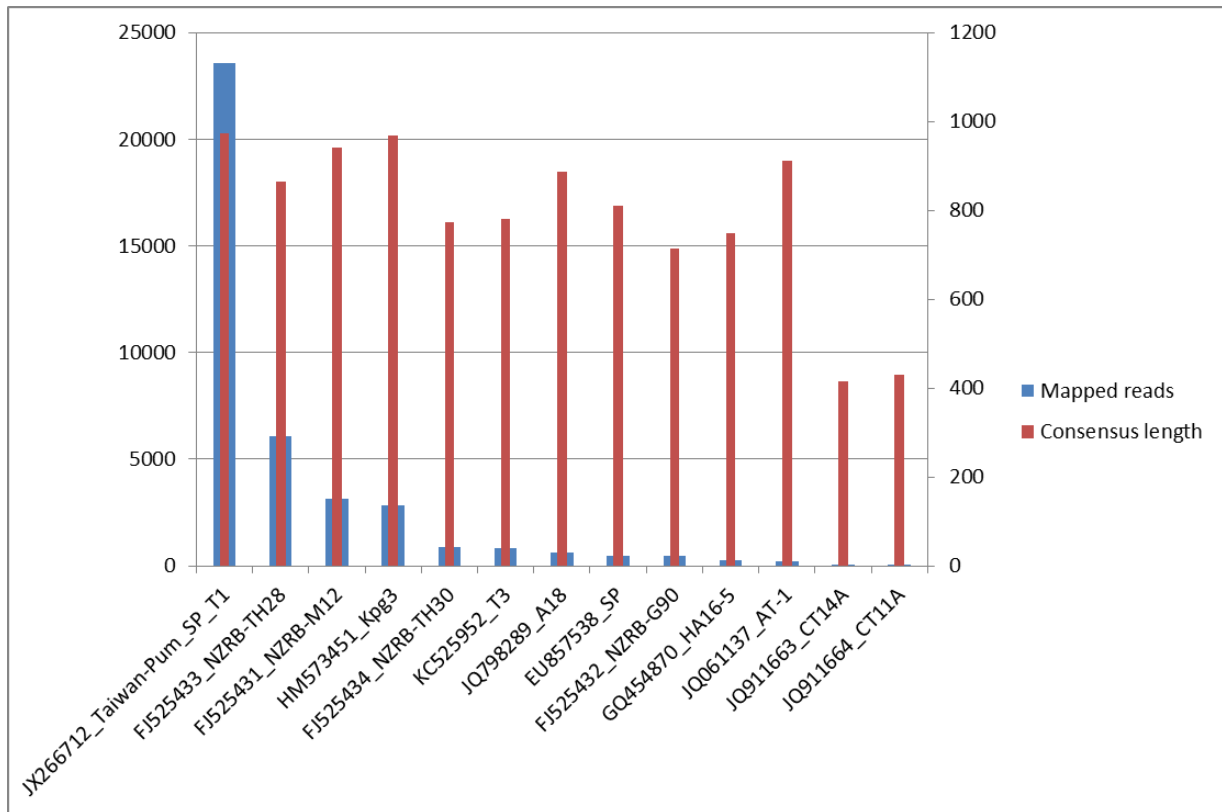


Figure 4.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3309**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

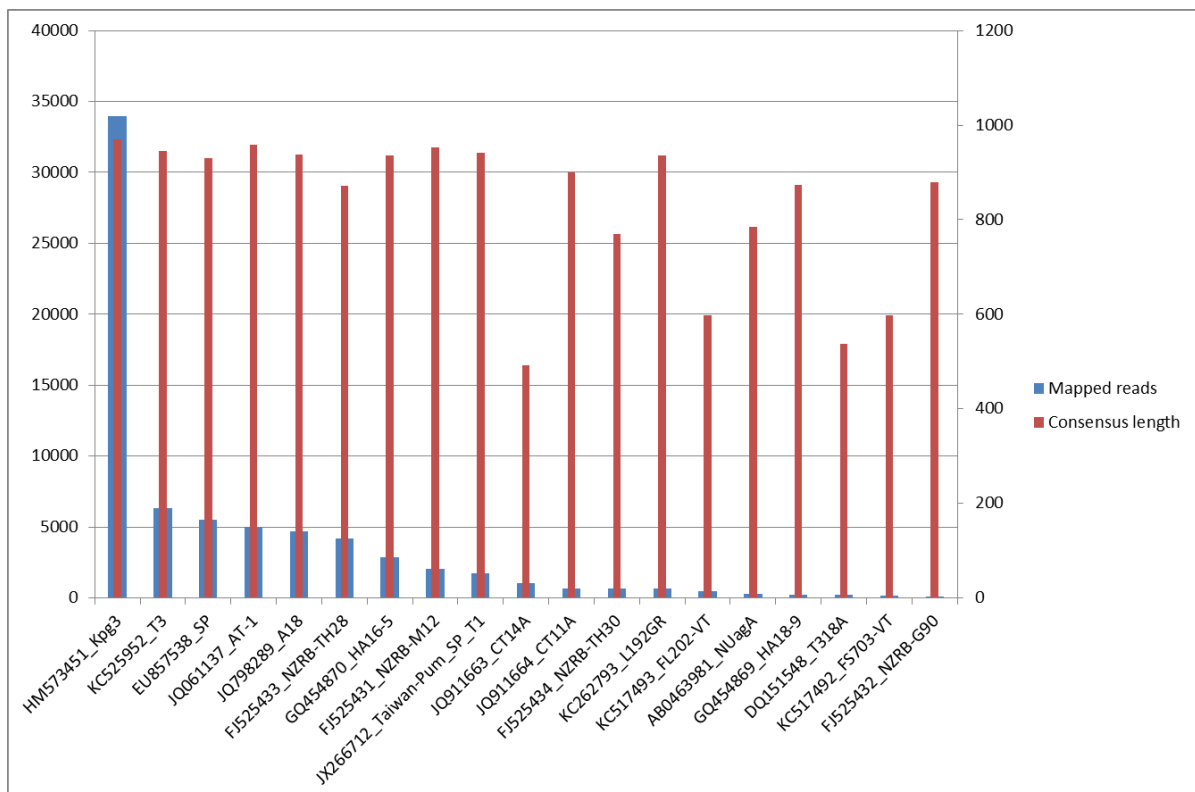


Figure 4.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3312**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

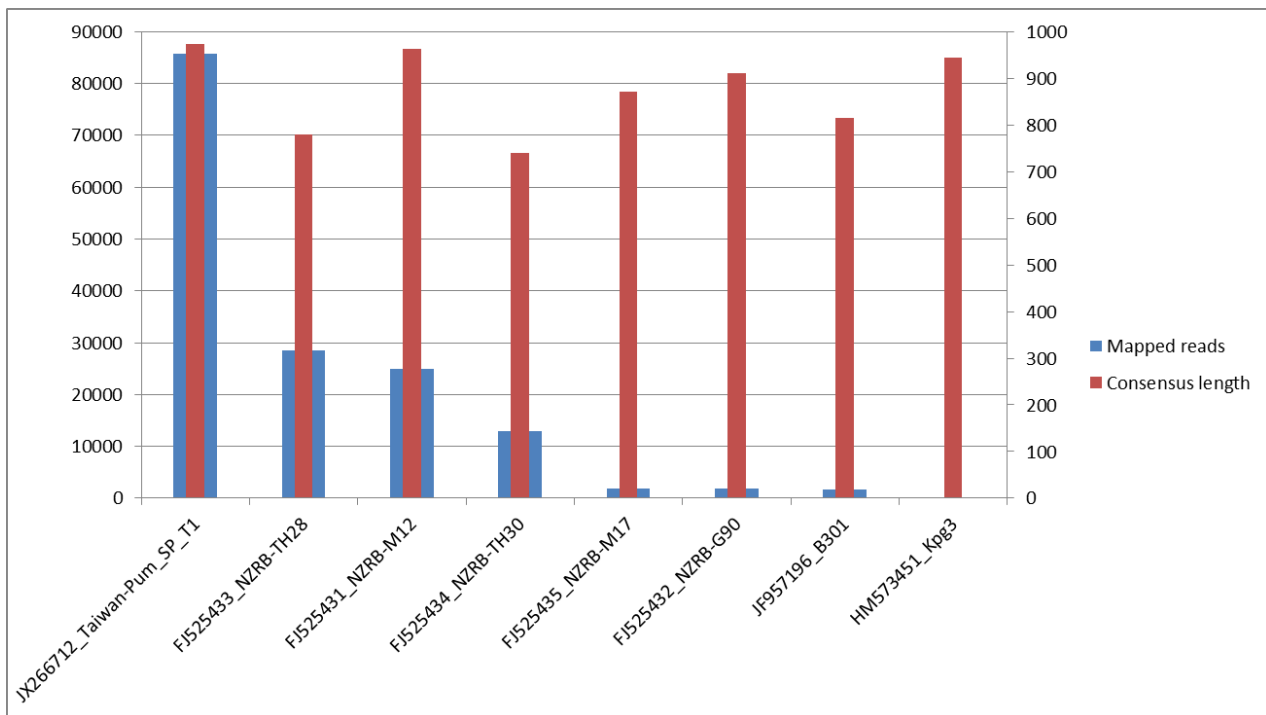


Figure 4.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3321**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

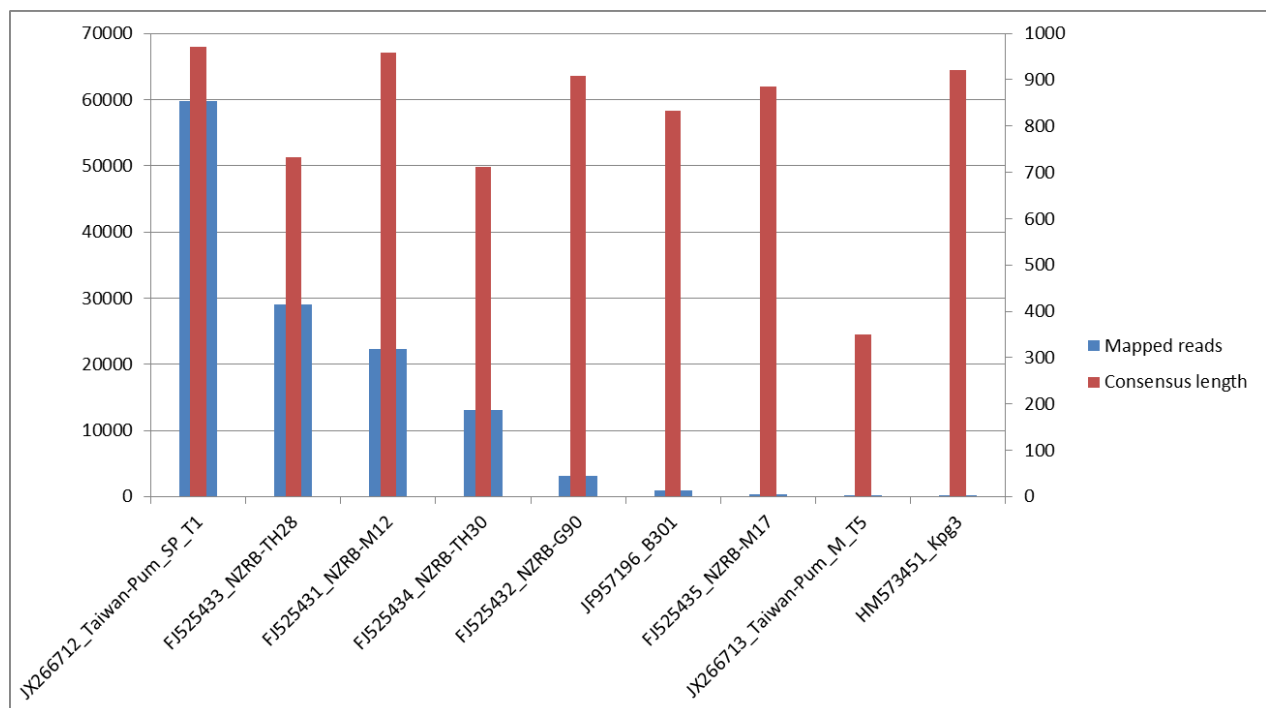


Figure 4.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3324**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

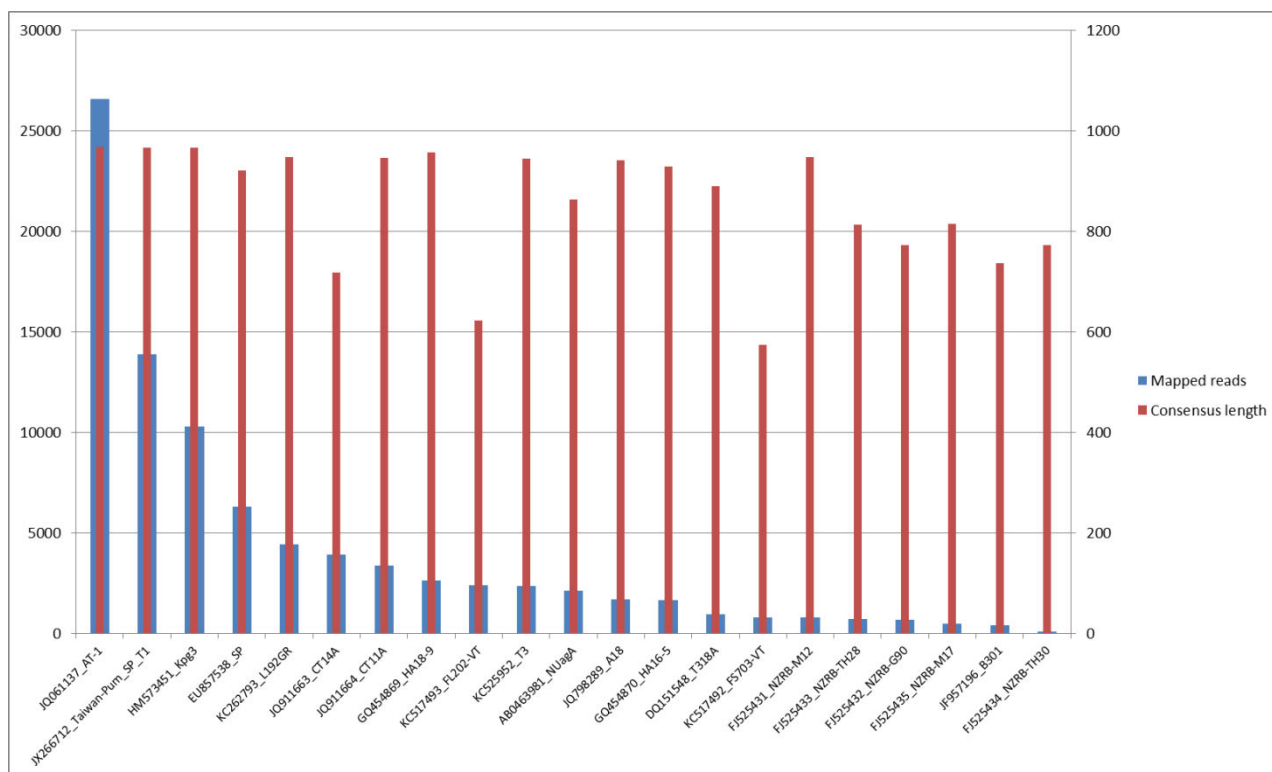


Figure 4.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3343**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

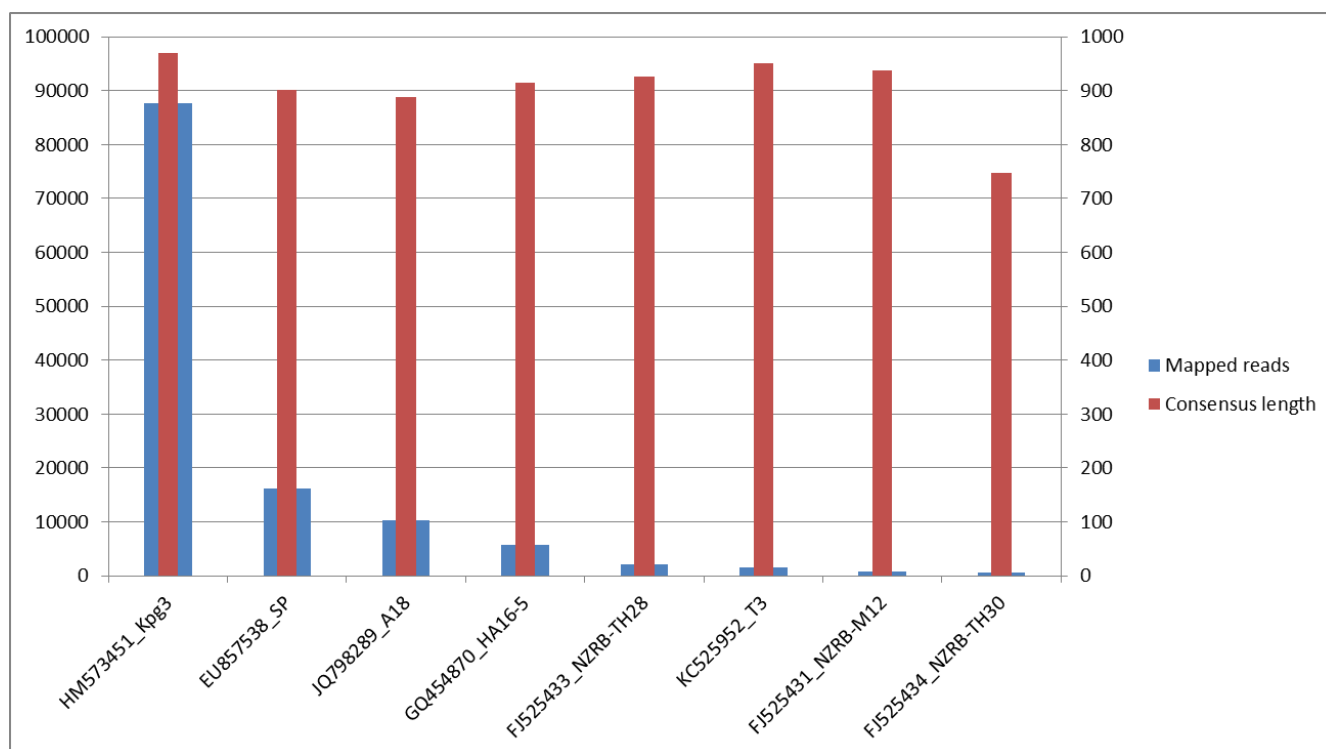


Figure 4.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3347**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

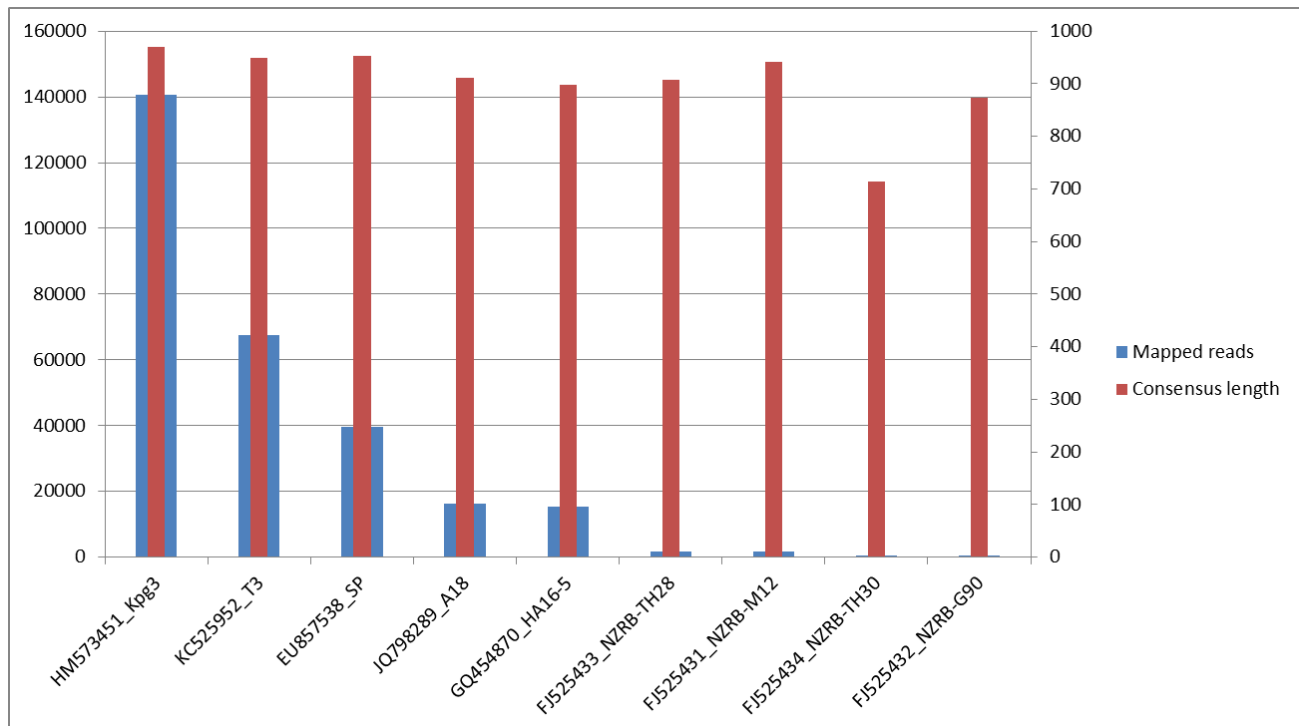


Figure 4.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3362**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

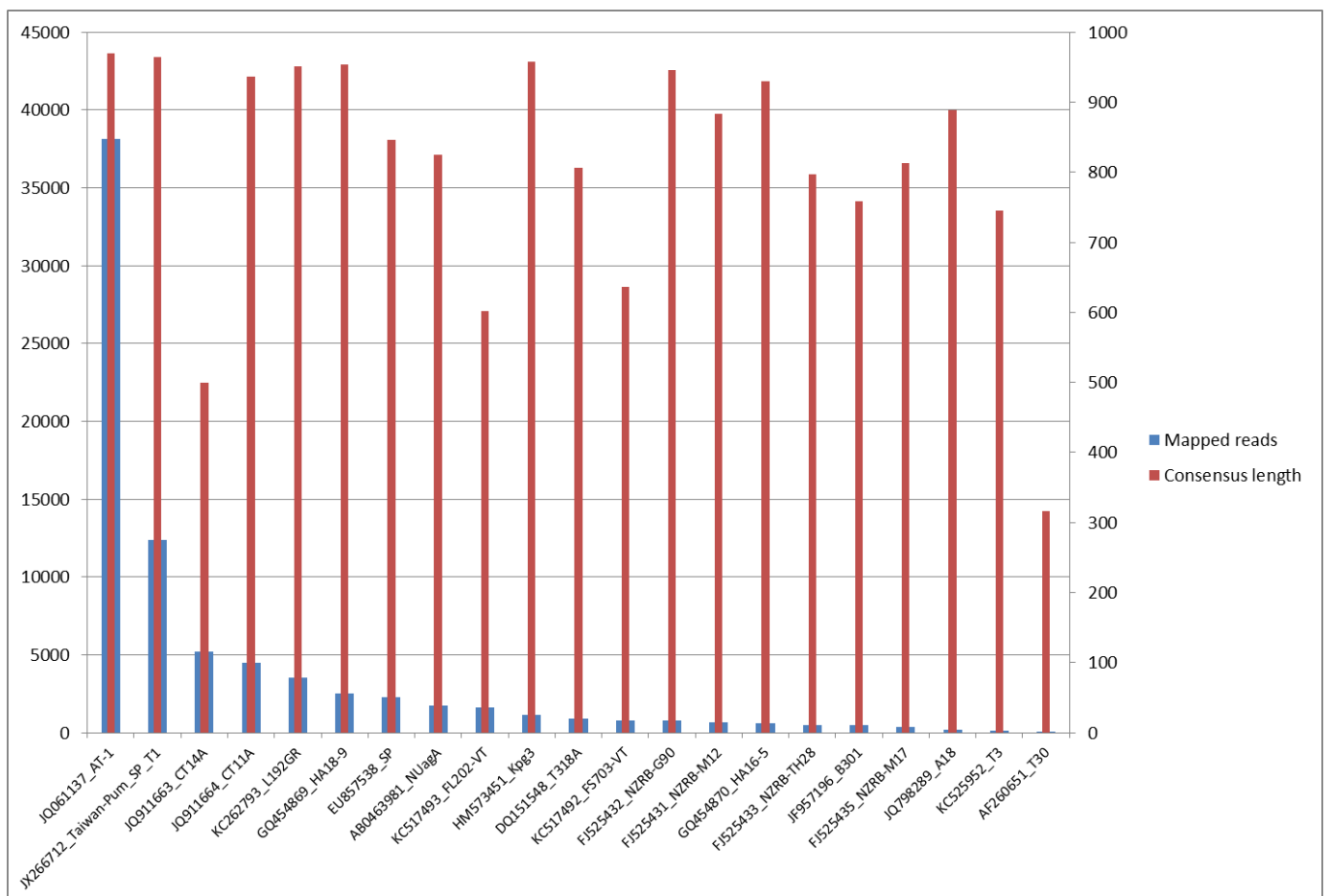


Figure 5.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3405**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

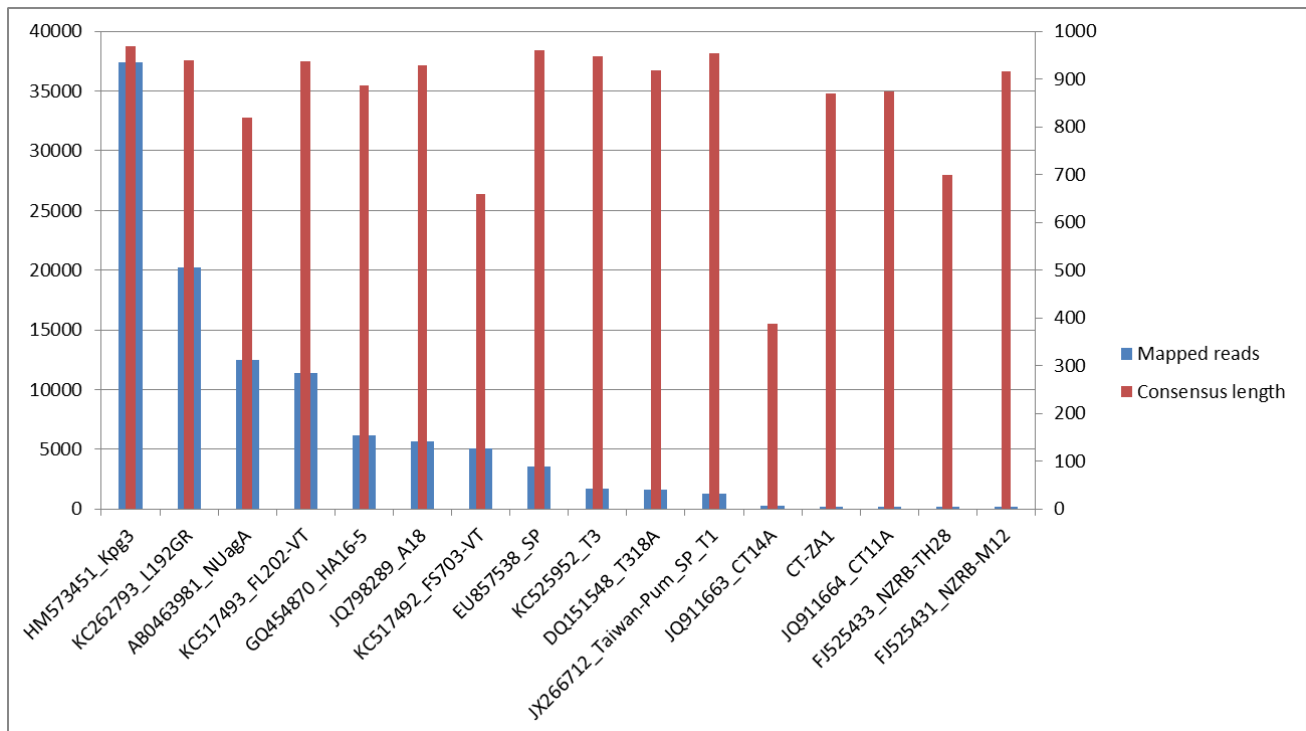


Figure 5.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3406**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

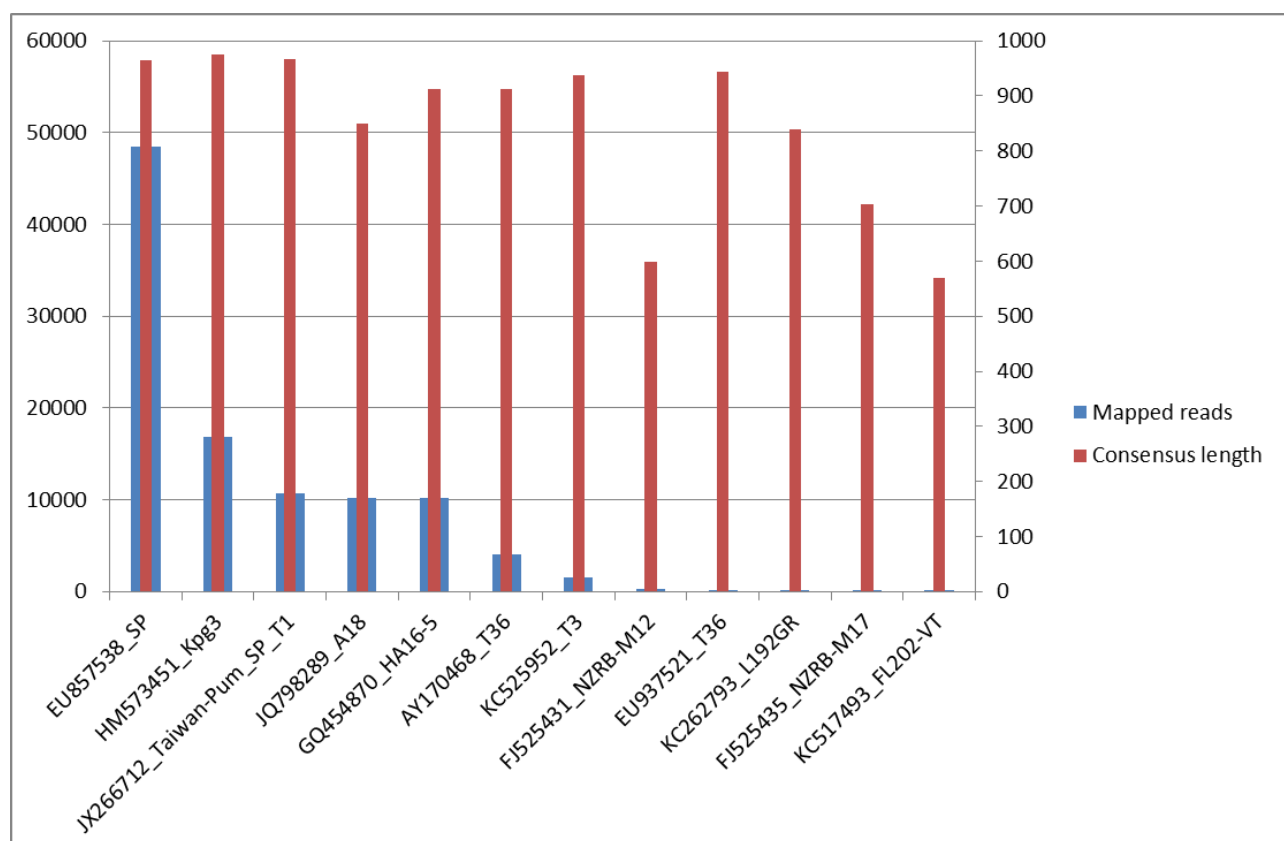


Figure 5.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3410**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

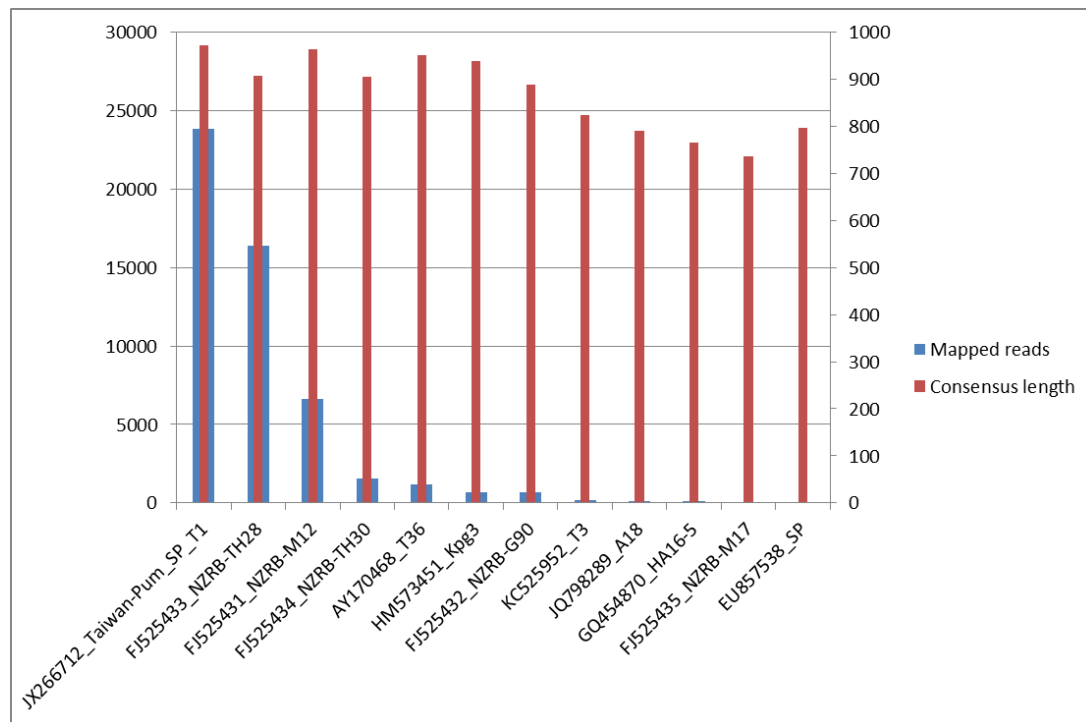


Figure 5.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3416**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

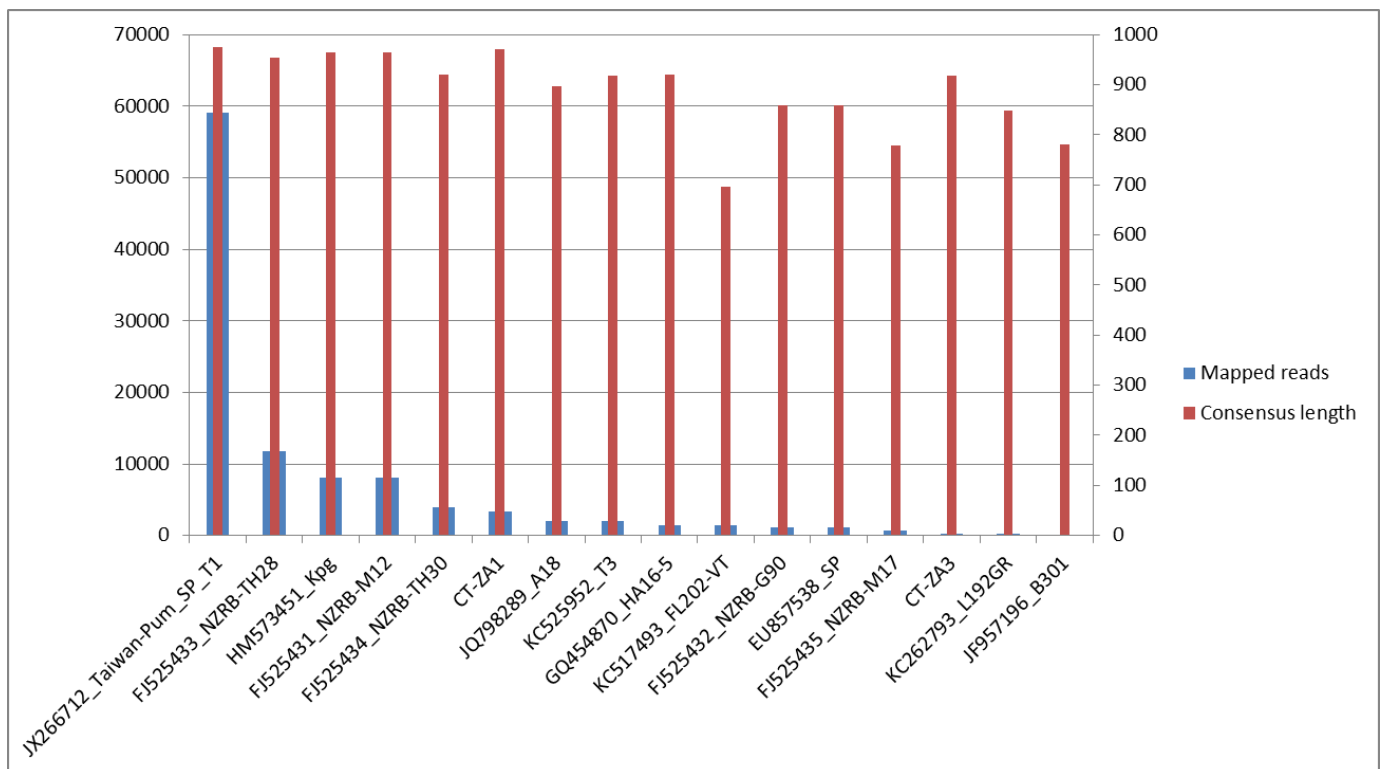


Figure 5.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3417**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

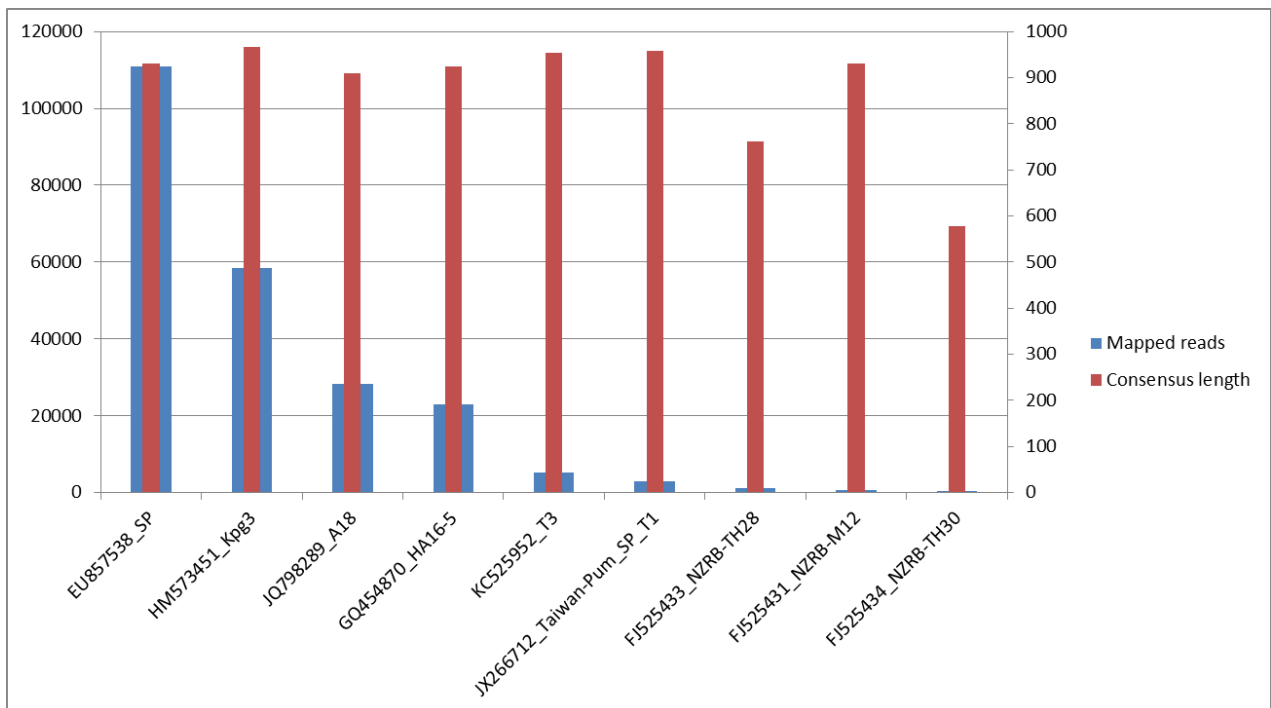


Figure 5.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3418**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

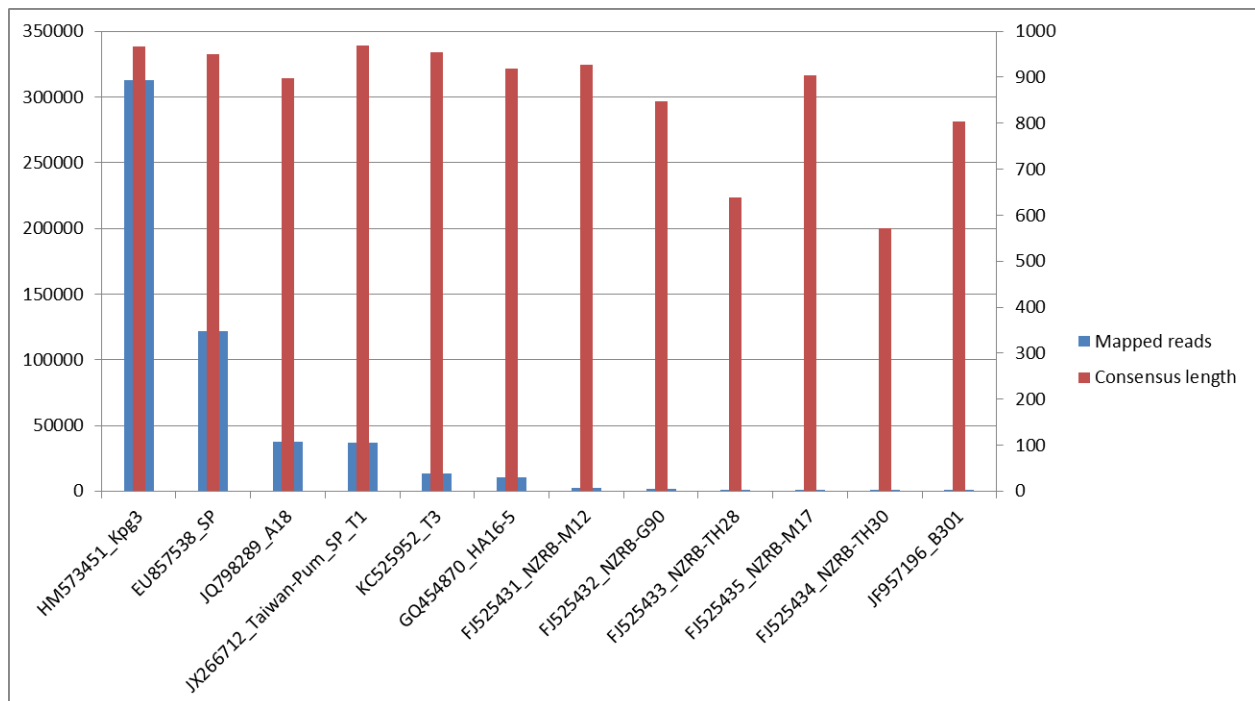


Figure 5.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3421**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

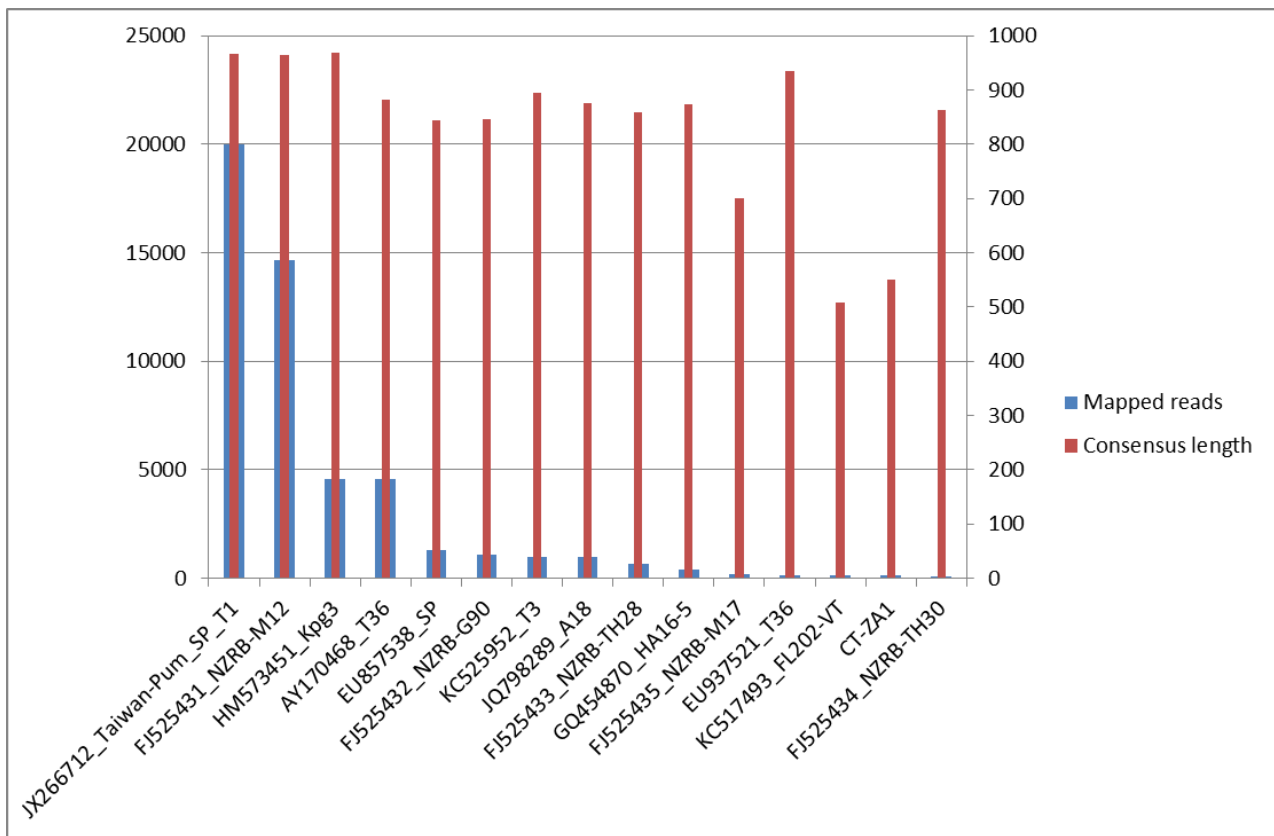


Figure 5.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3422**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

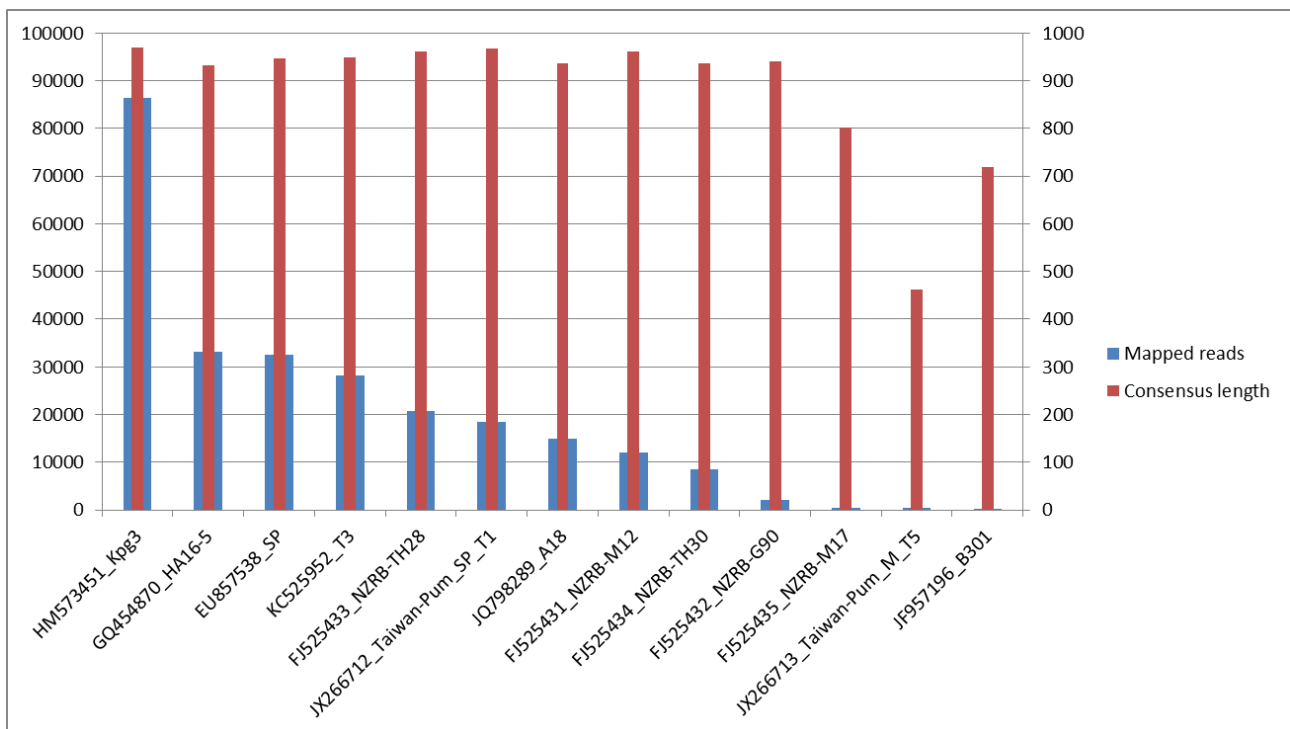


Figure 5.9: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3425**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

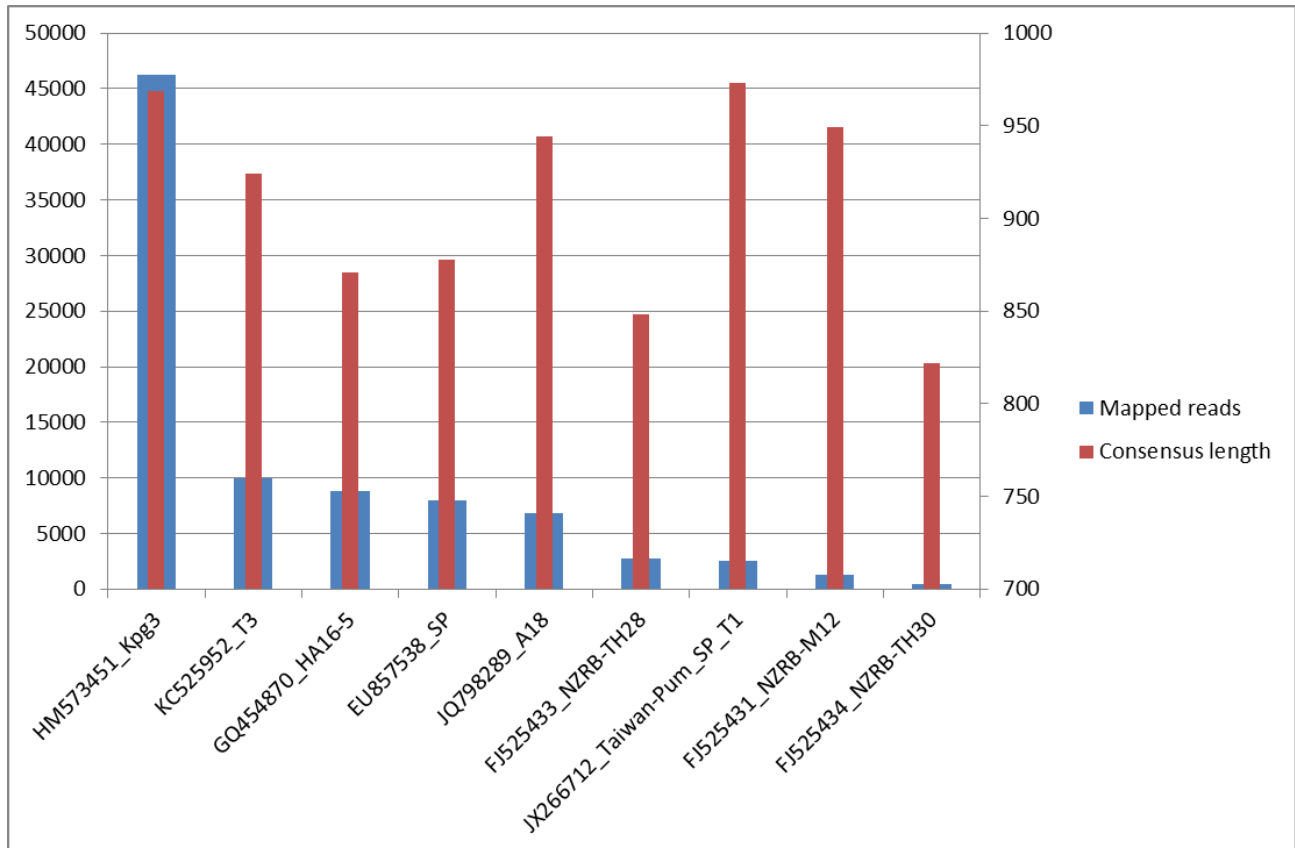


Figure 5.10: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3427**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

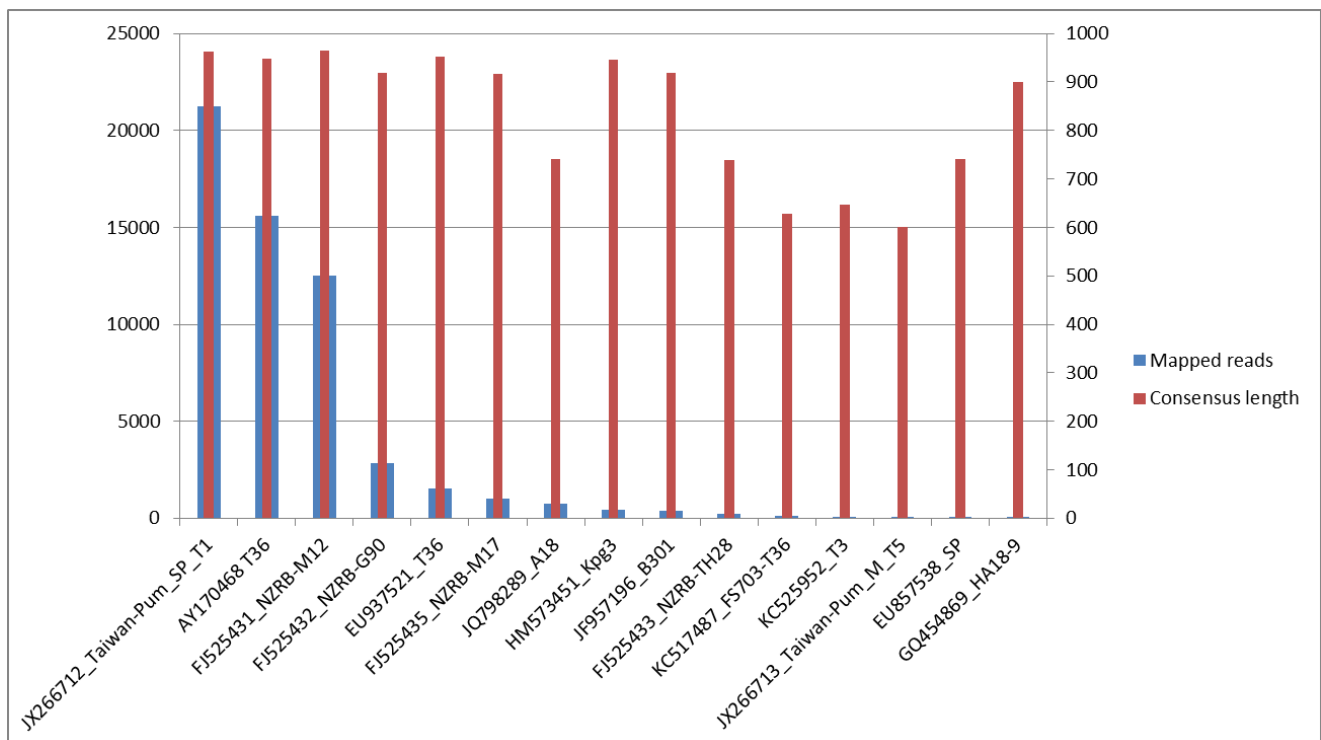


Figure 5.11: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3436**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

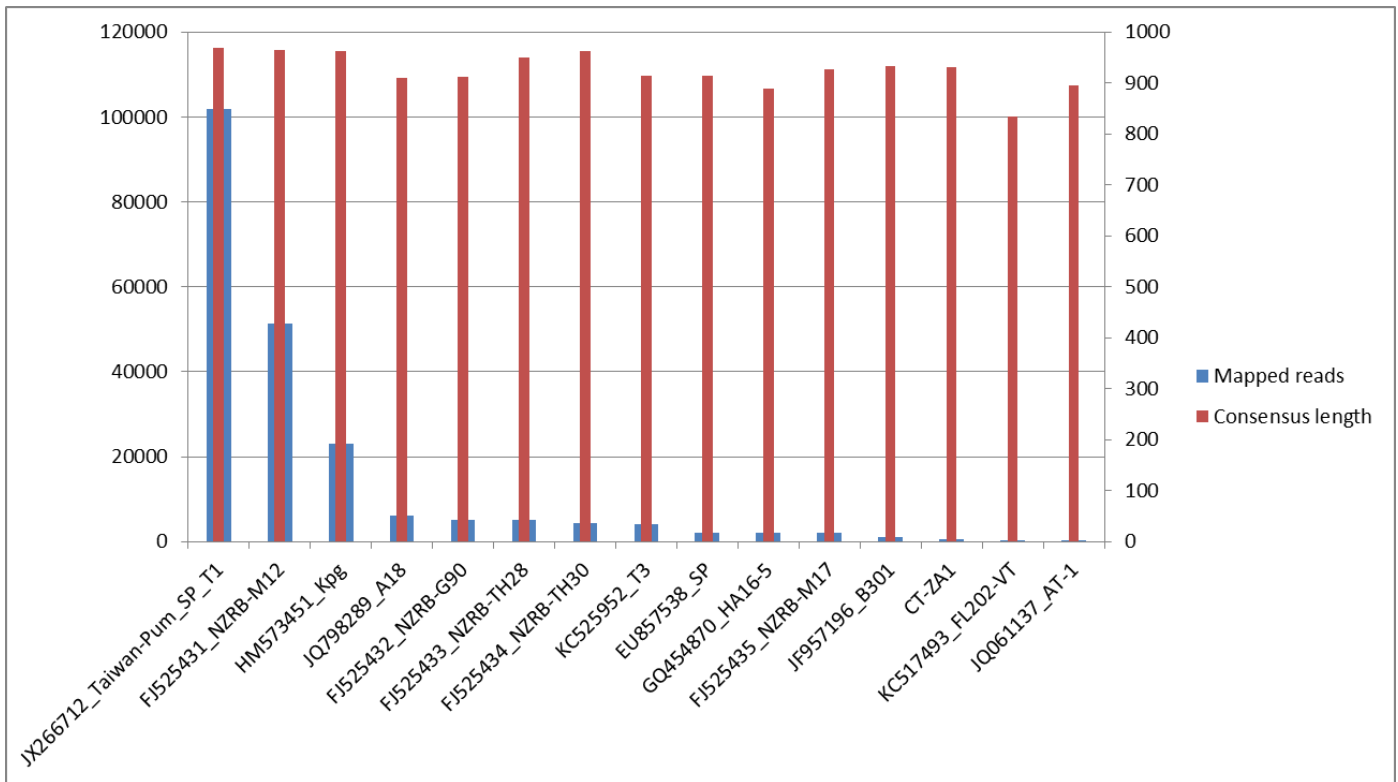


Figure 5.12: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3439**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

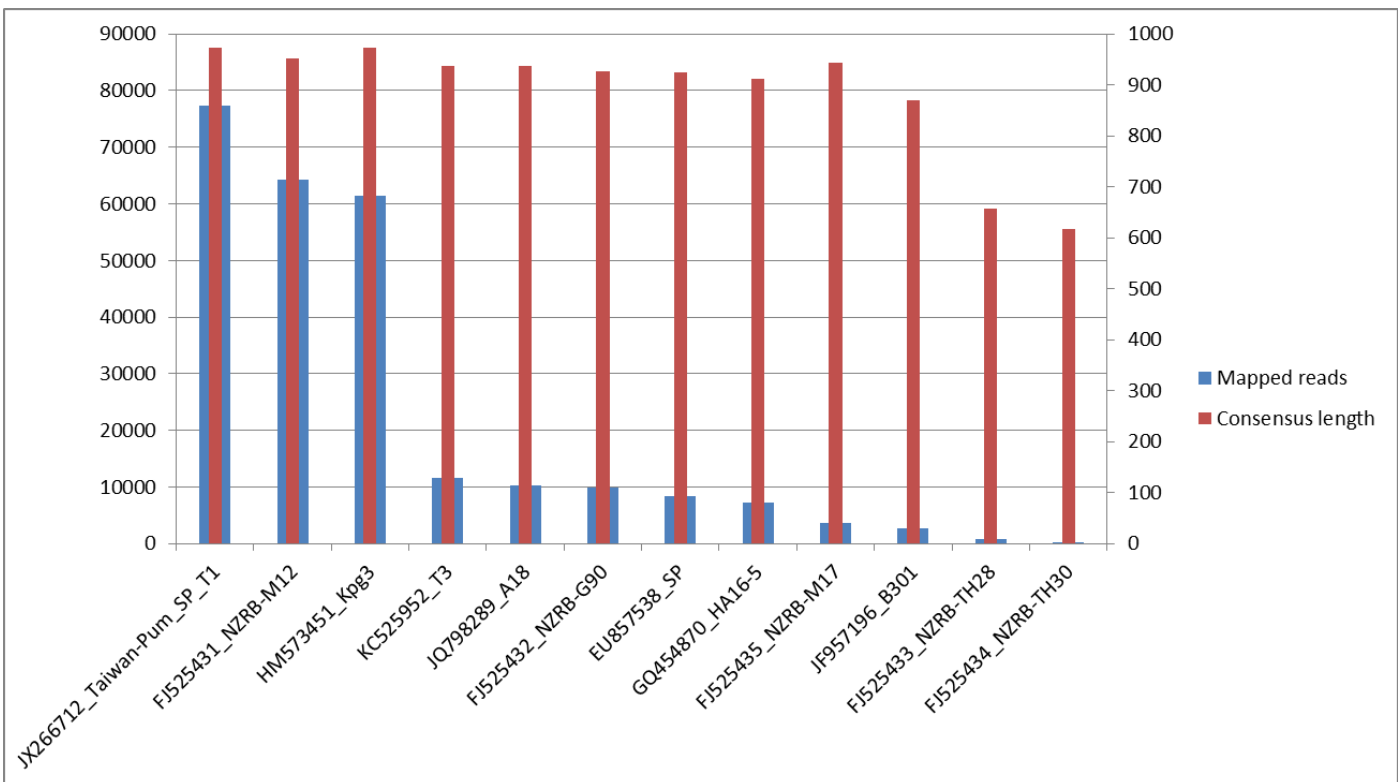


Figure 5.13: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3445**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

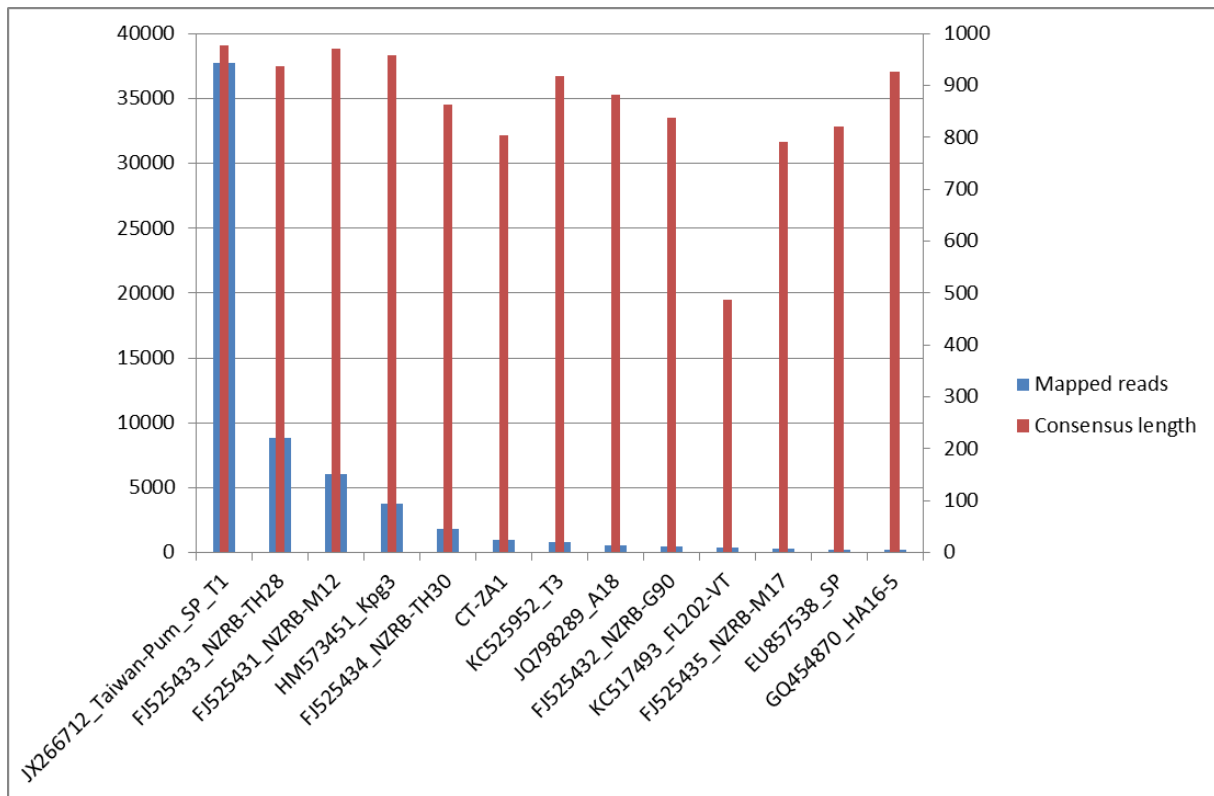


Figure 5.14: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3447**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

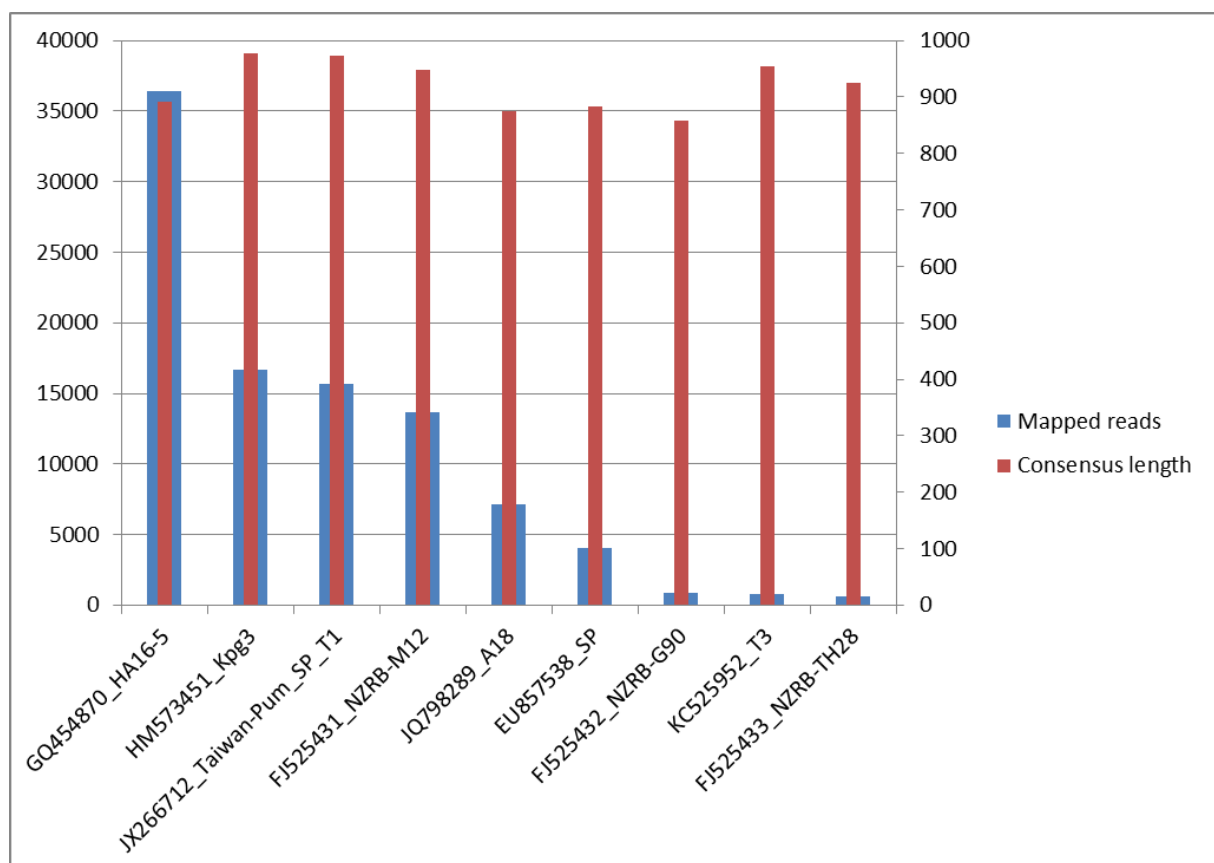


Figure 5.15: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3449**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

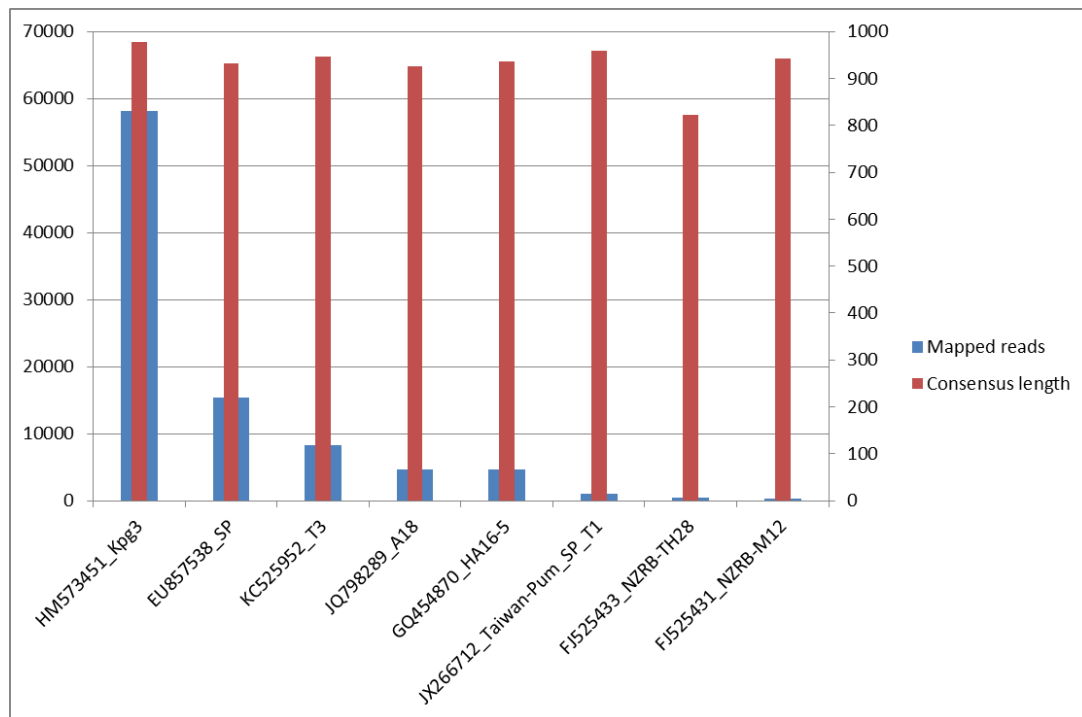


Figure 5.16: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3454**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

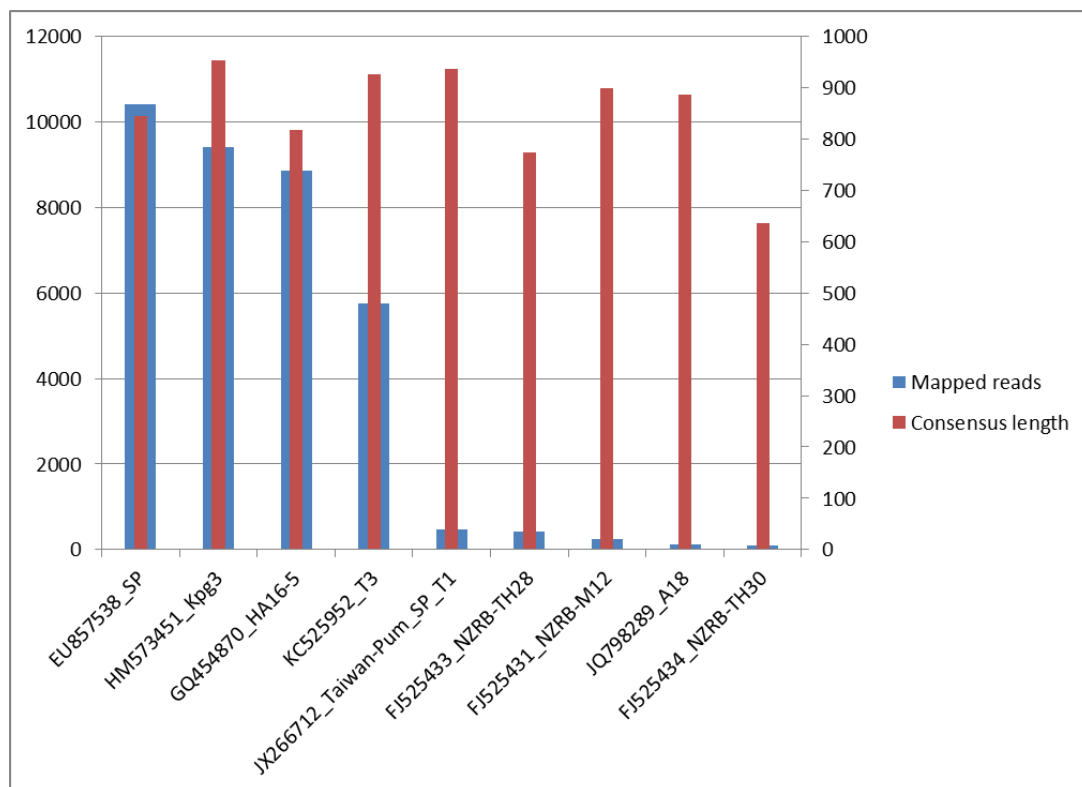


Figure 6.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3508**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

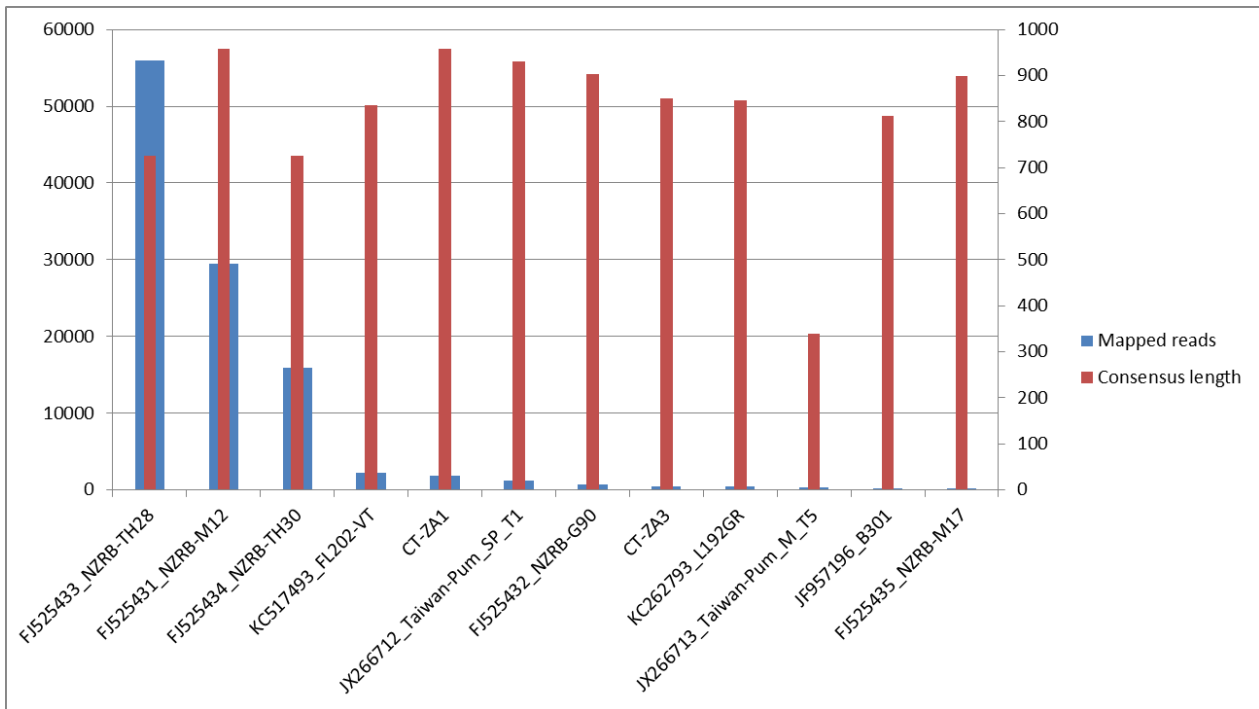


Figure 6.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3512**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

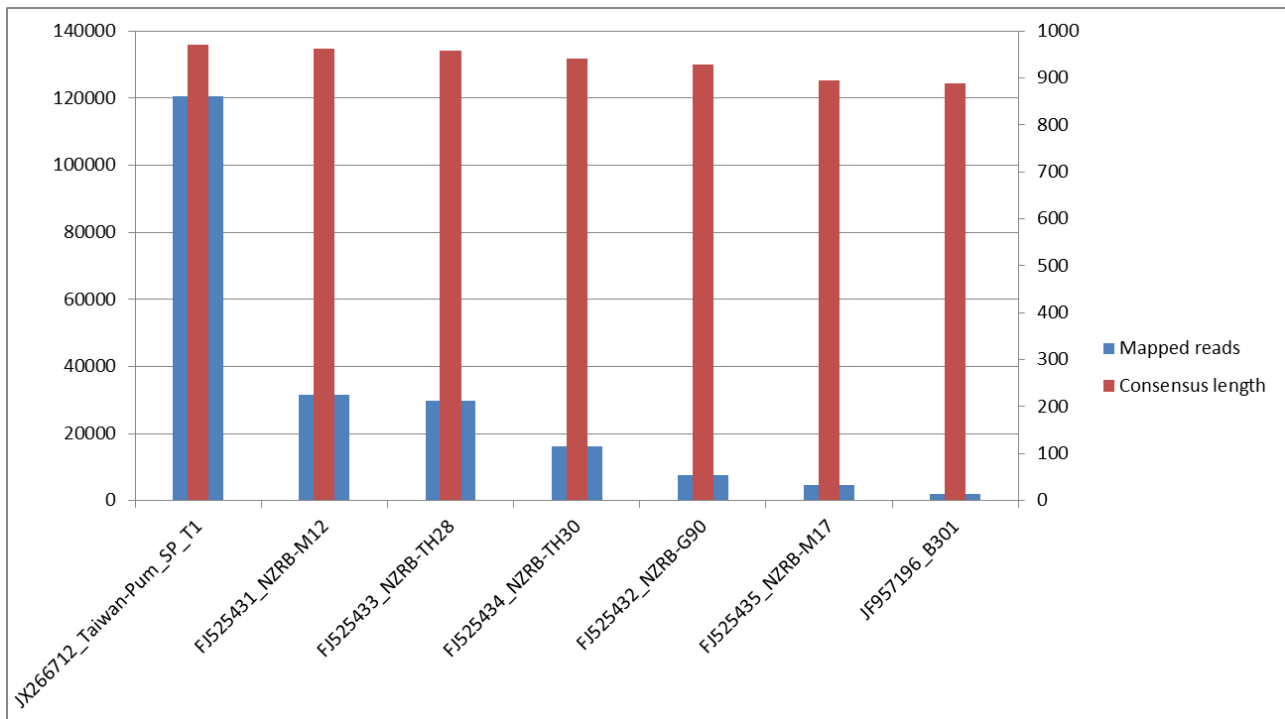


Figure 6.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3520**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

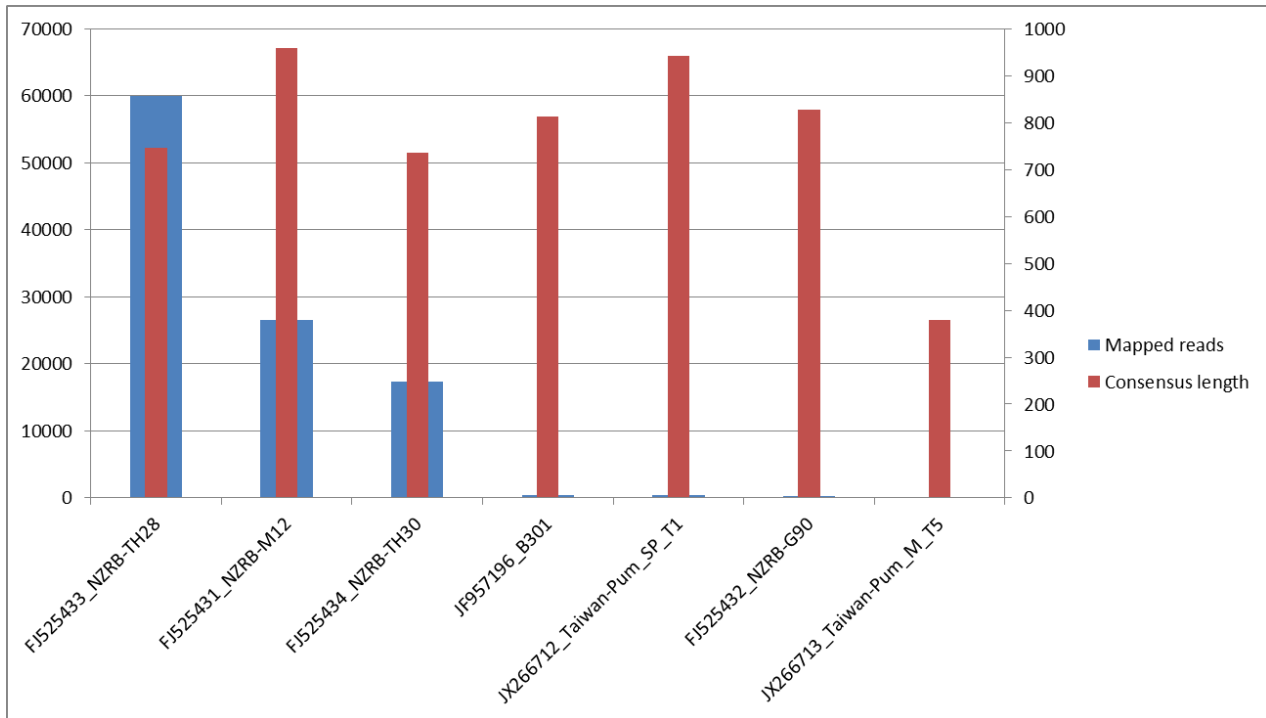


Figure 6.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3523**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

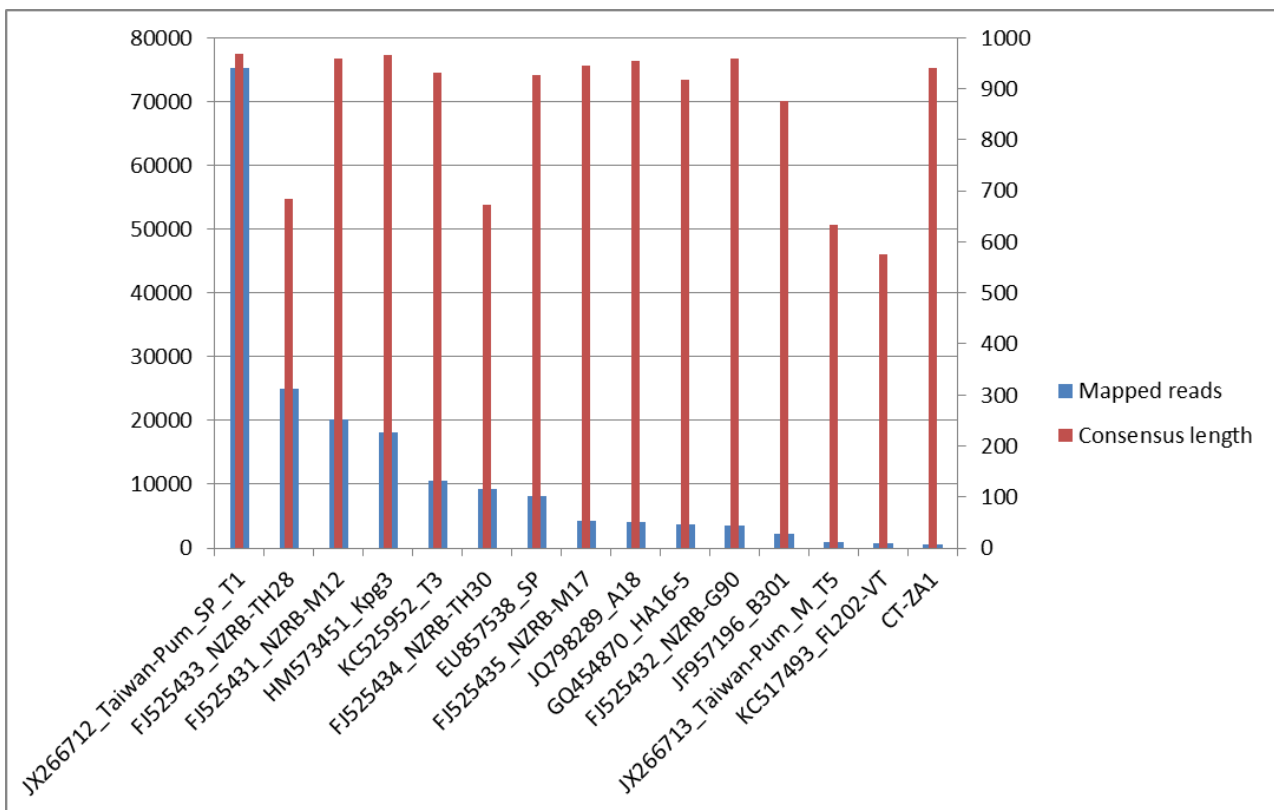


Figure 6.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3526**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

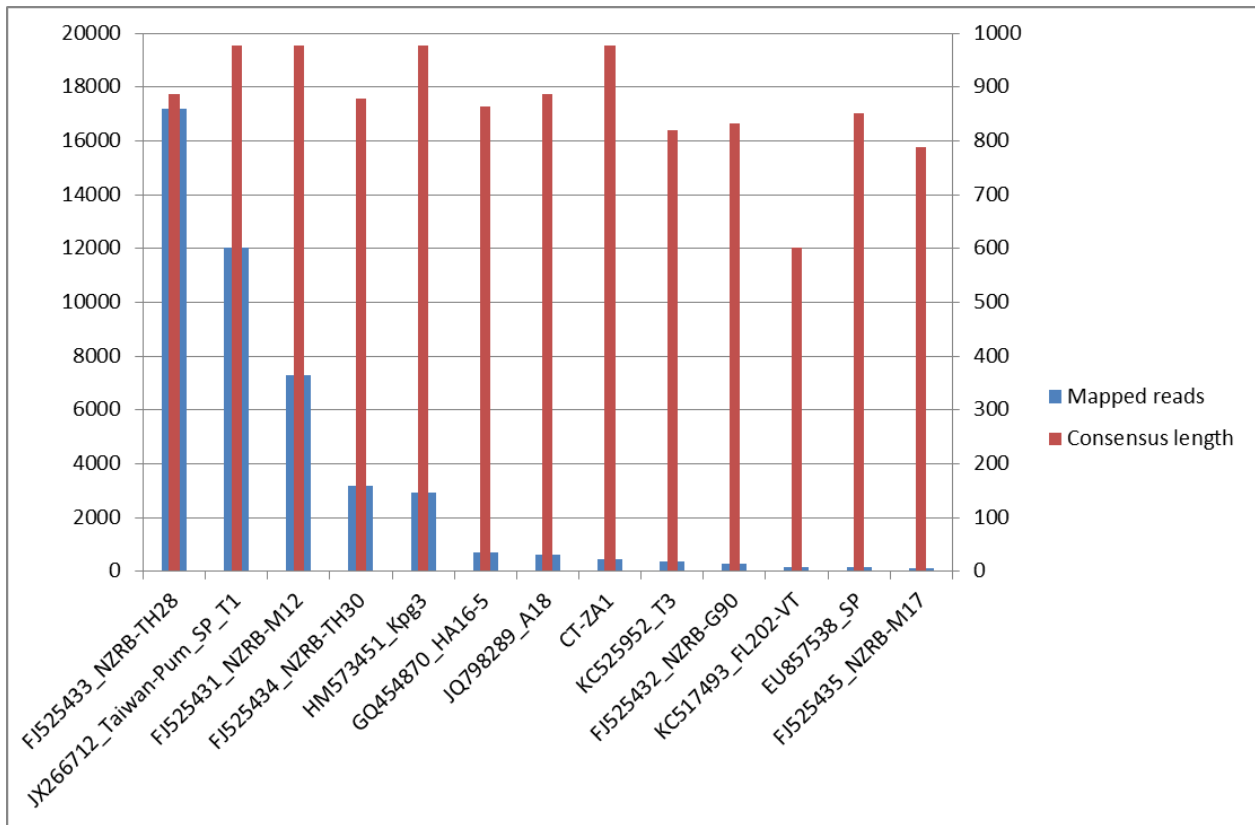


Figure 6.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3534**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

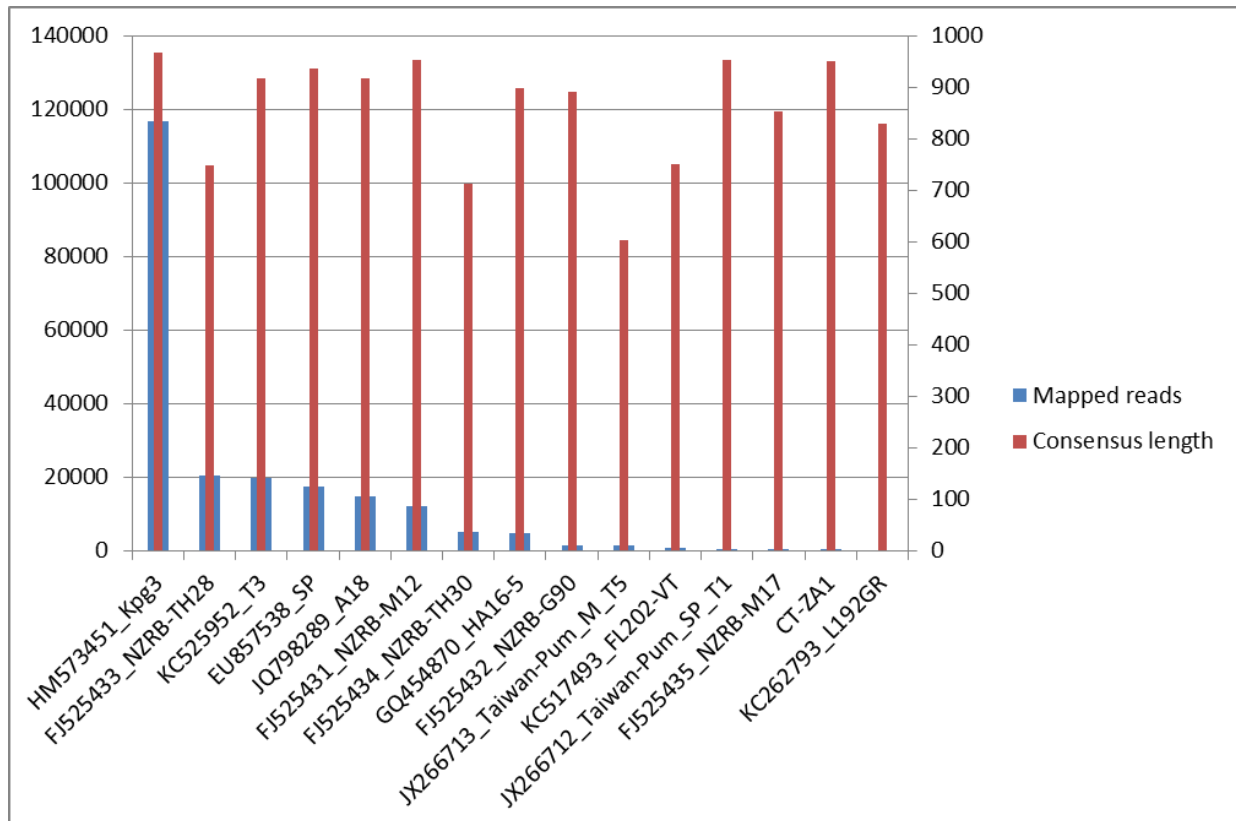


Figure 6.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3536**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

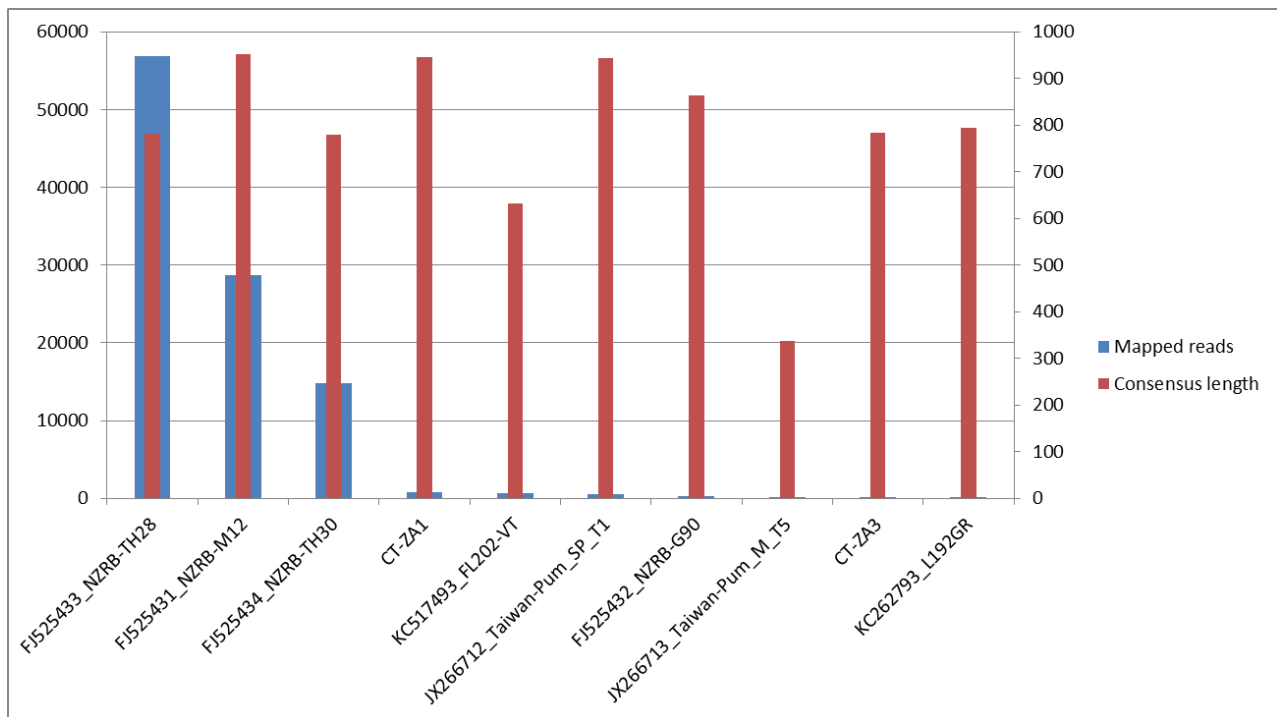


Figure 6.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3539**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

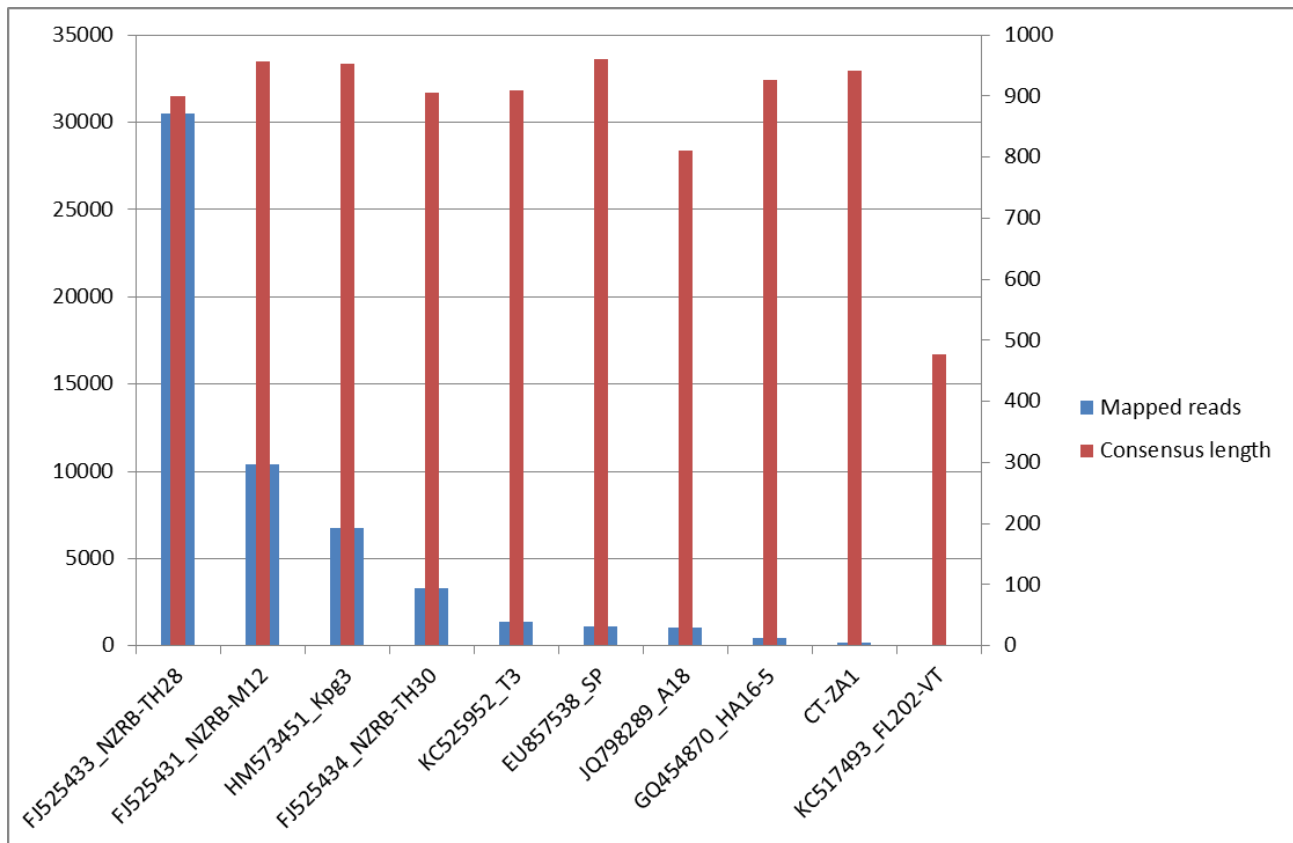


Figure 6.9: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3545**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

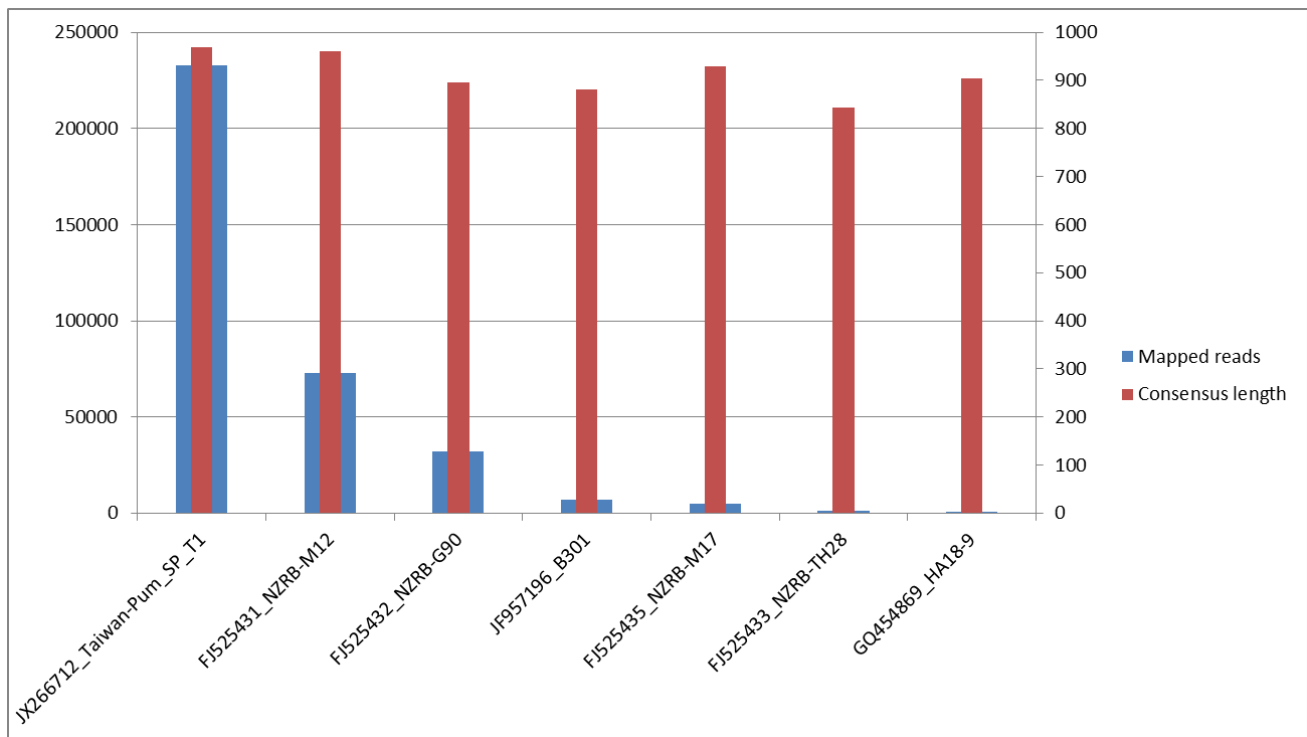


Figure 6.10: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3552**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

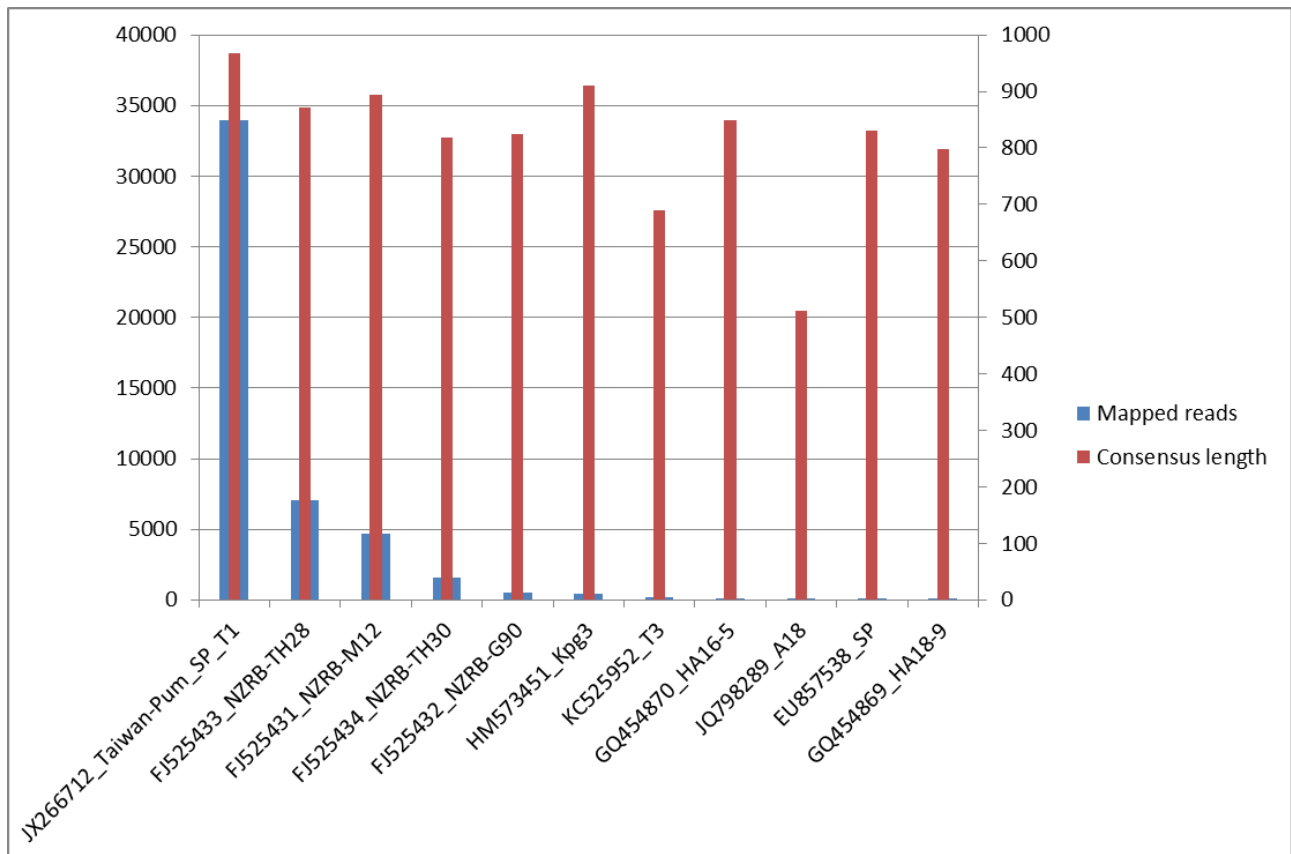


Figure 6.11: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3553**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

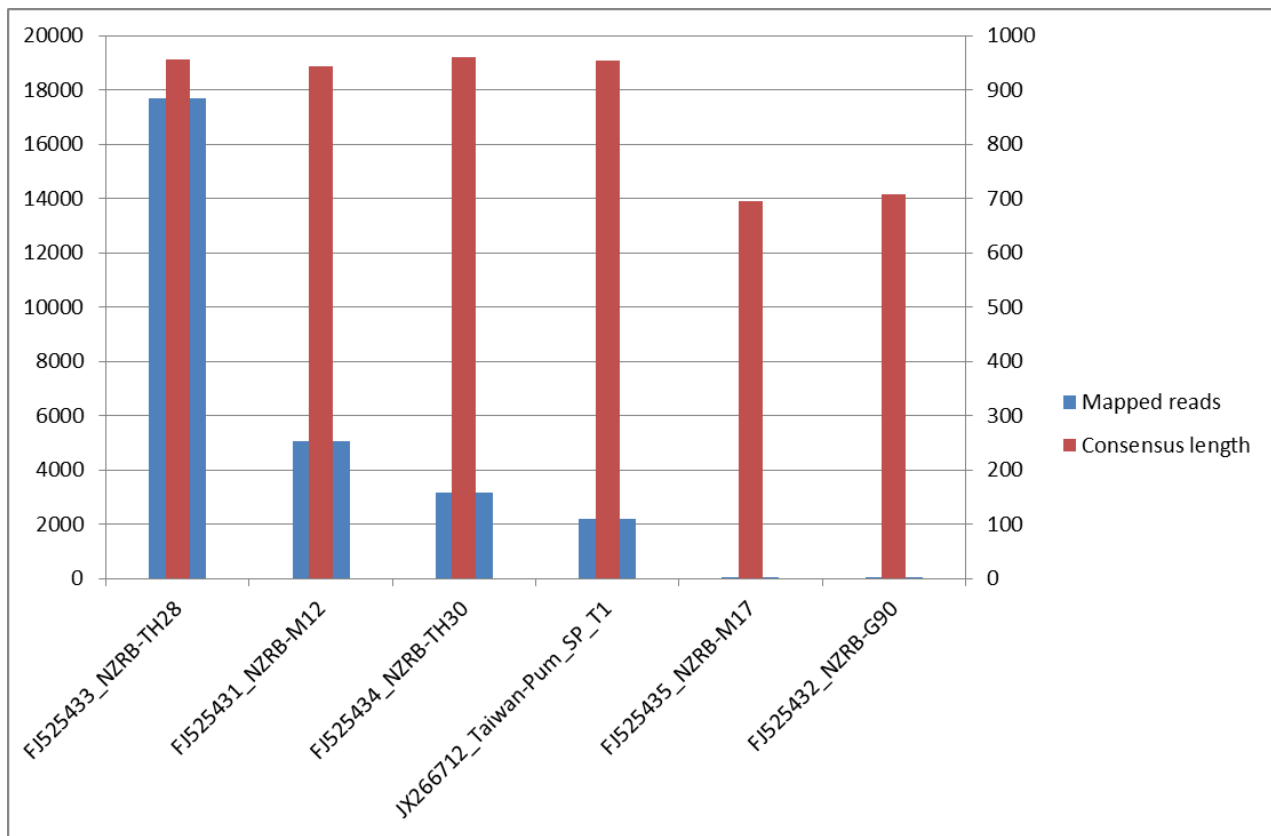


Figure 6.12: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3556**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

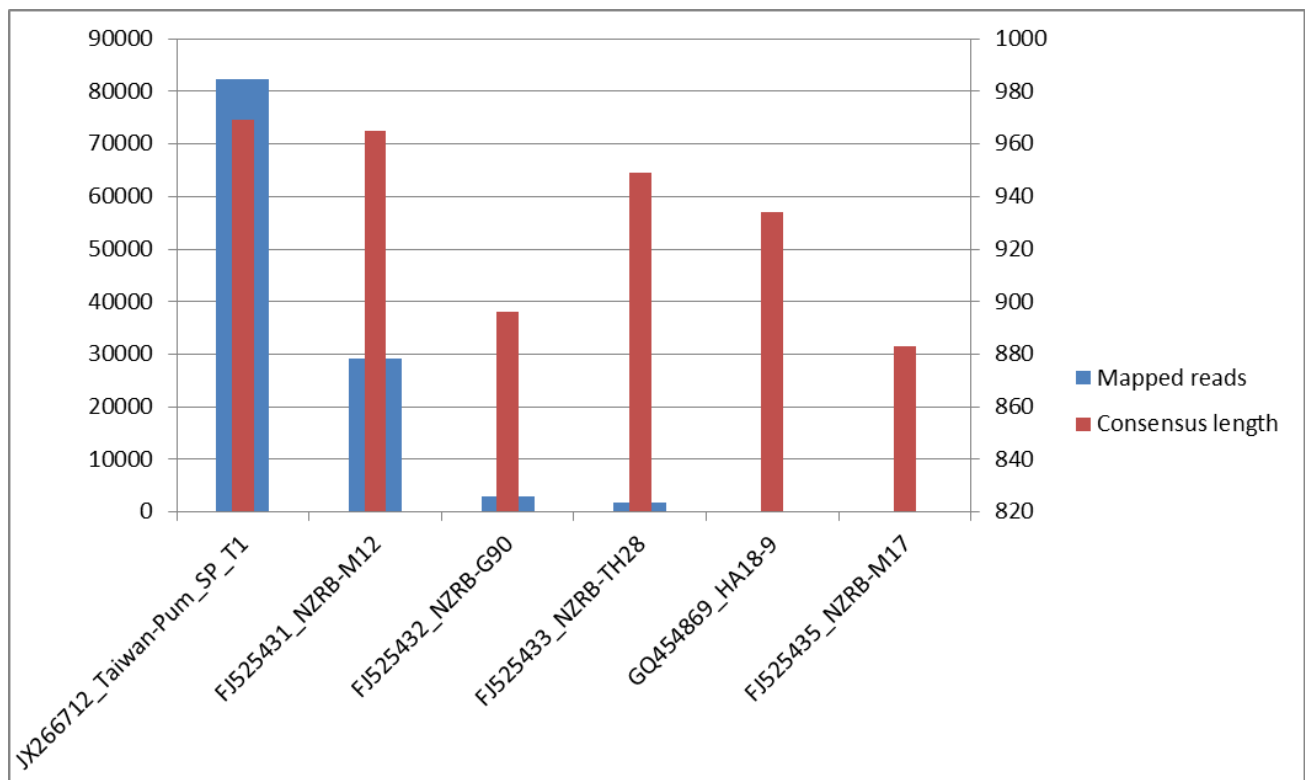


Figure 6.13: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3570**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

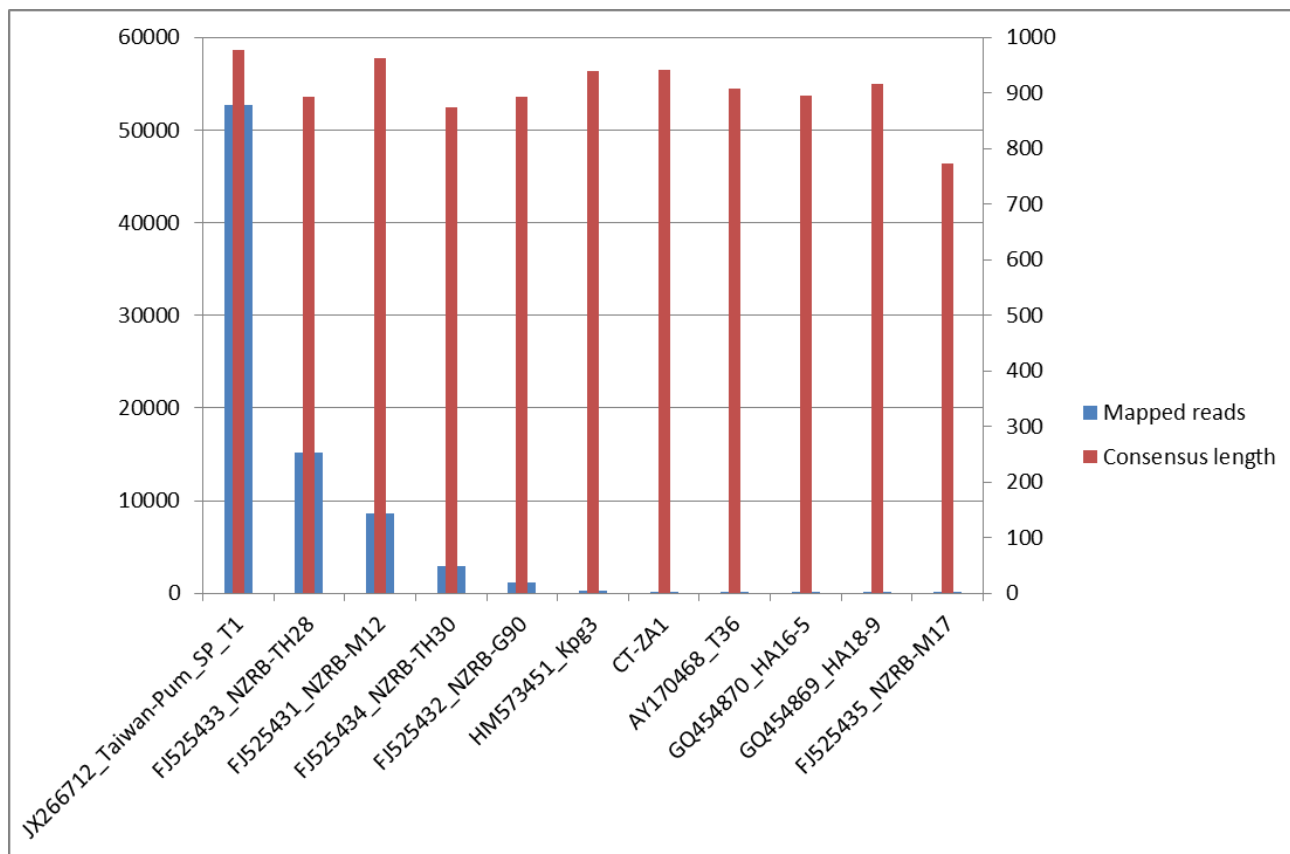


Figure 7.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3604**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

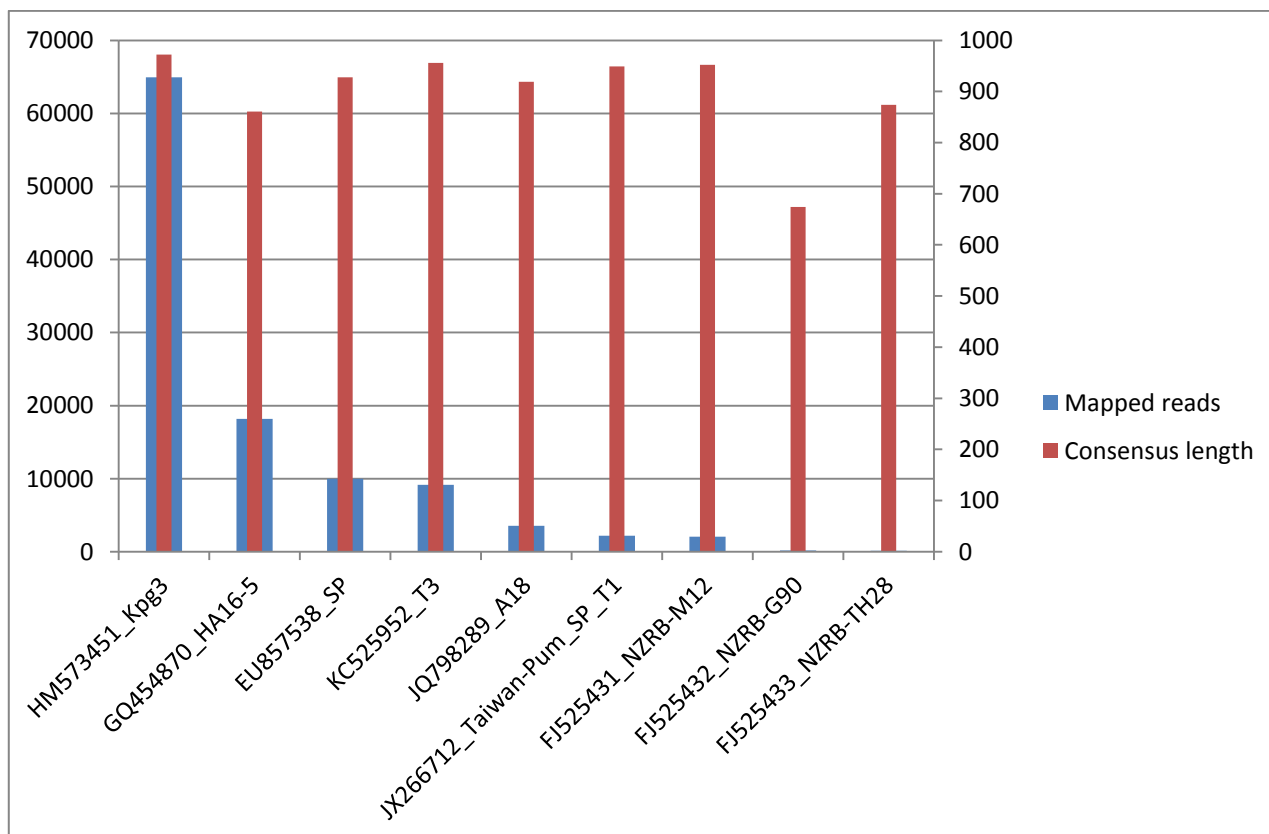


Figure 7.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3608**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

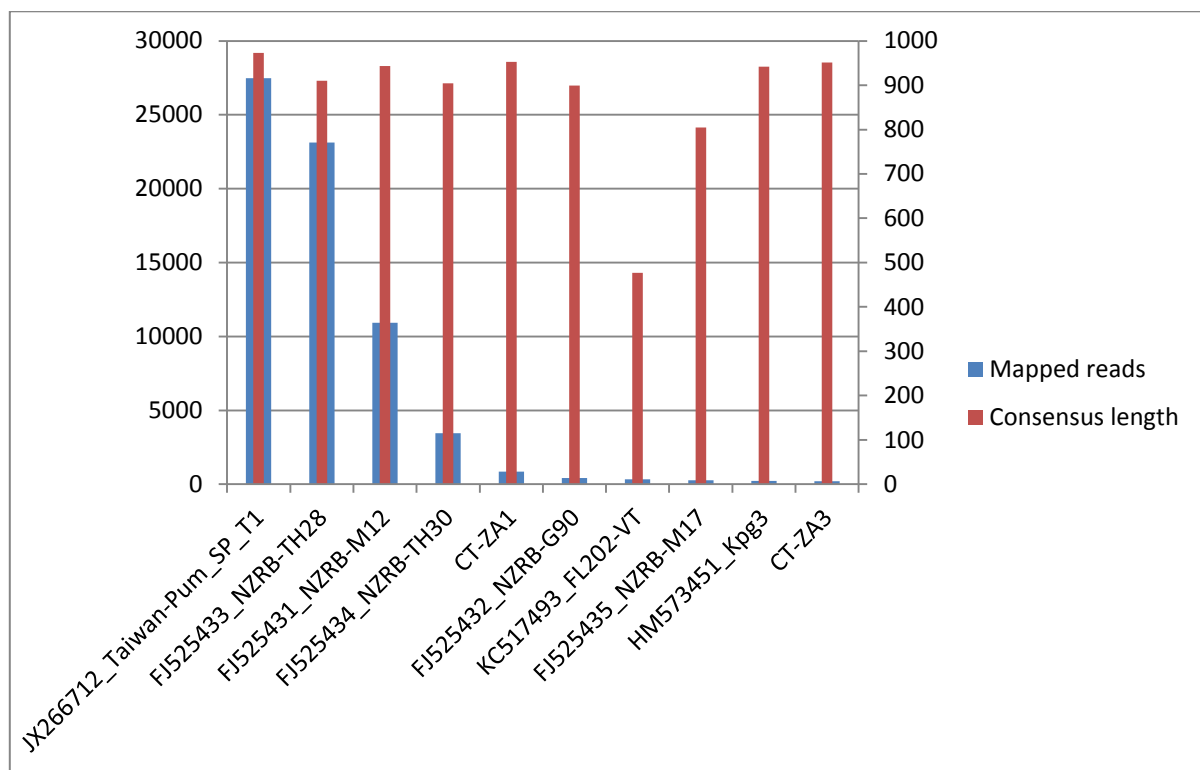


Figure 7.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3612**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

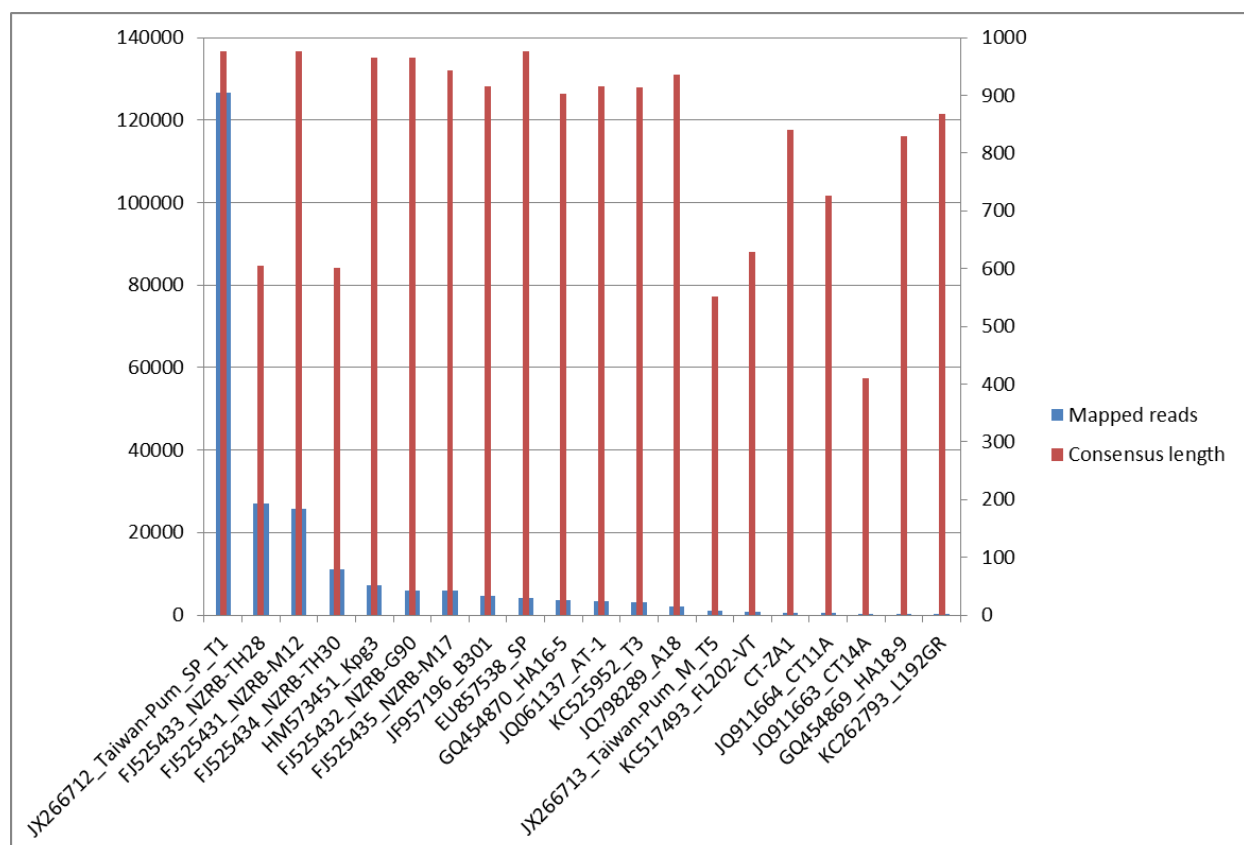


Figure 7.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3619**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

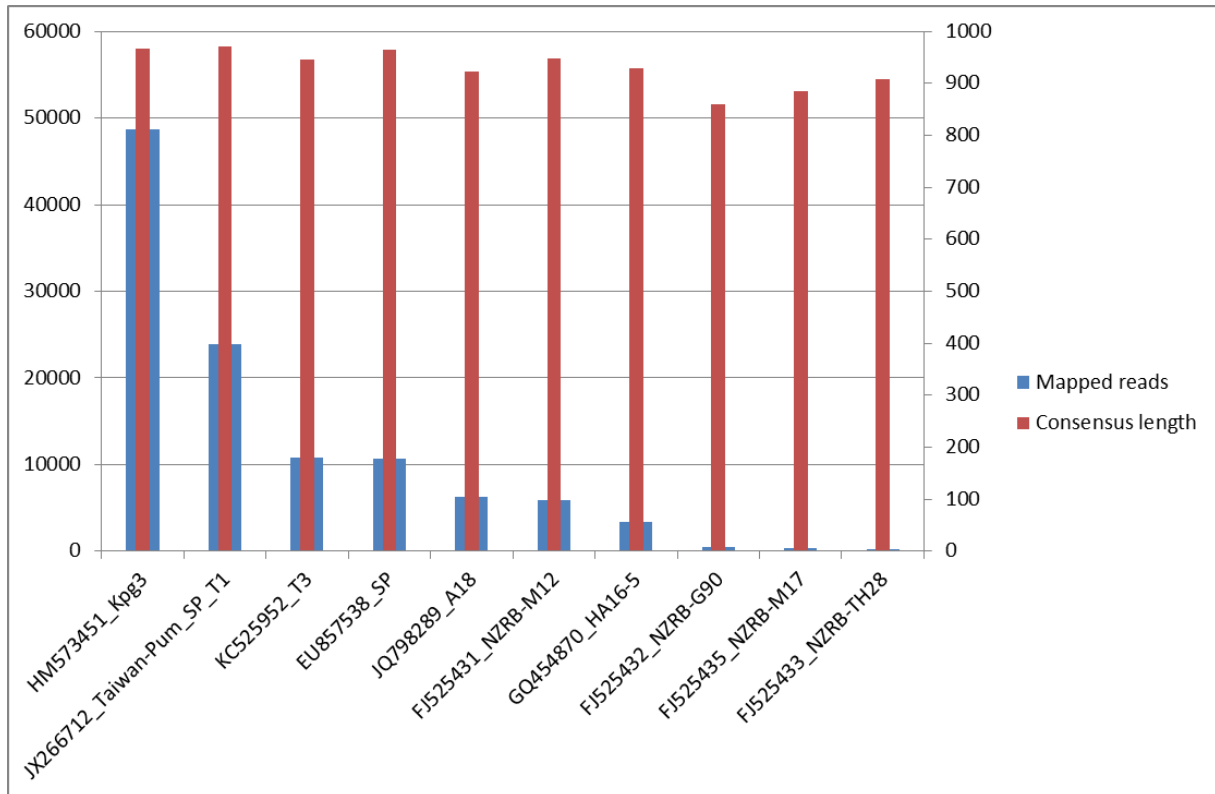


Figure 7.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3621**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

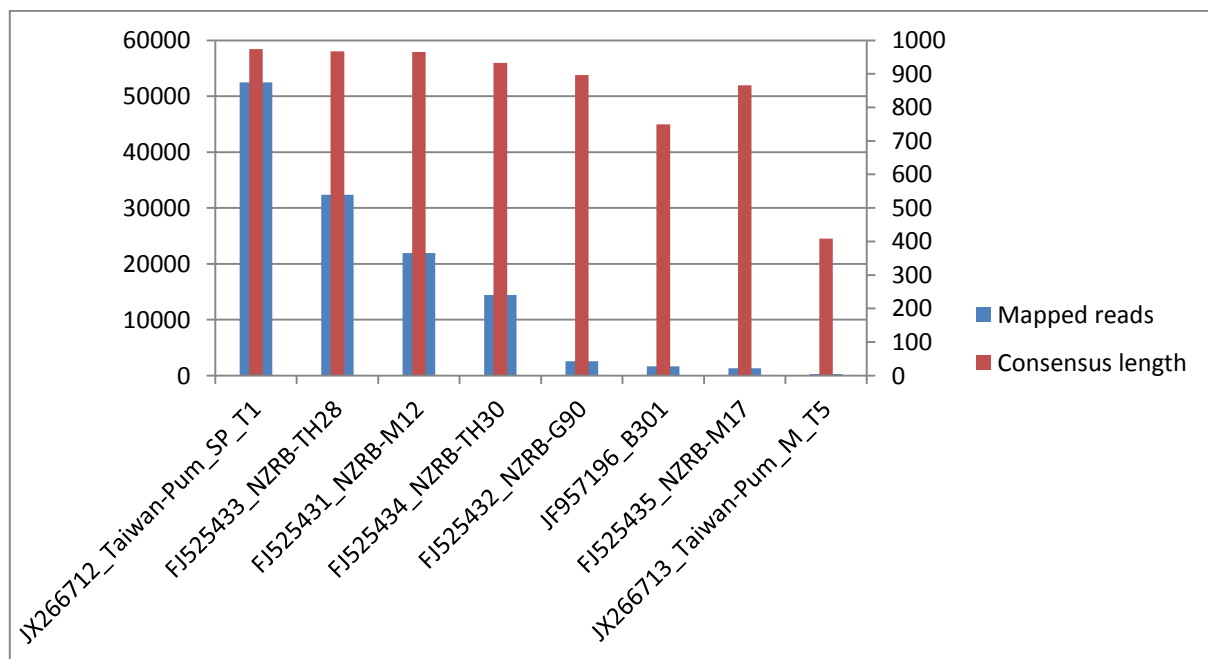


Figure 7.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3626**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

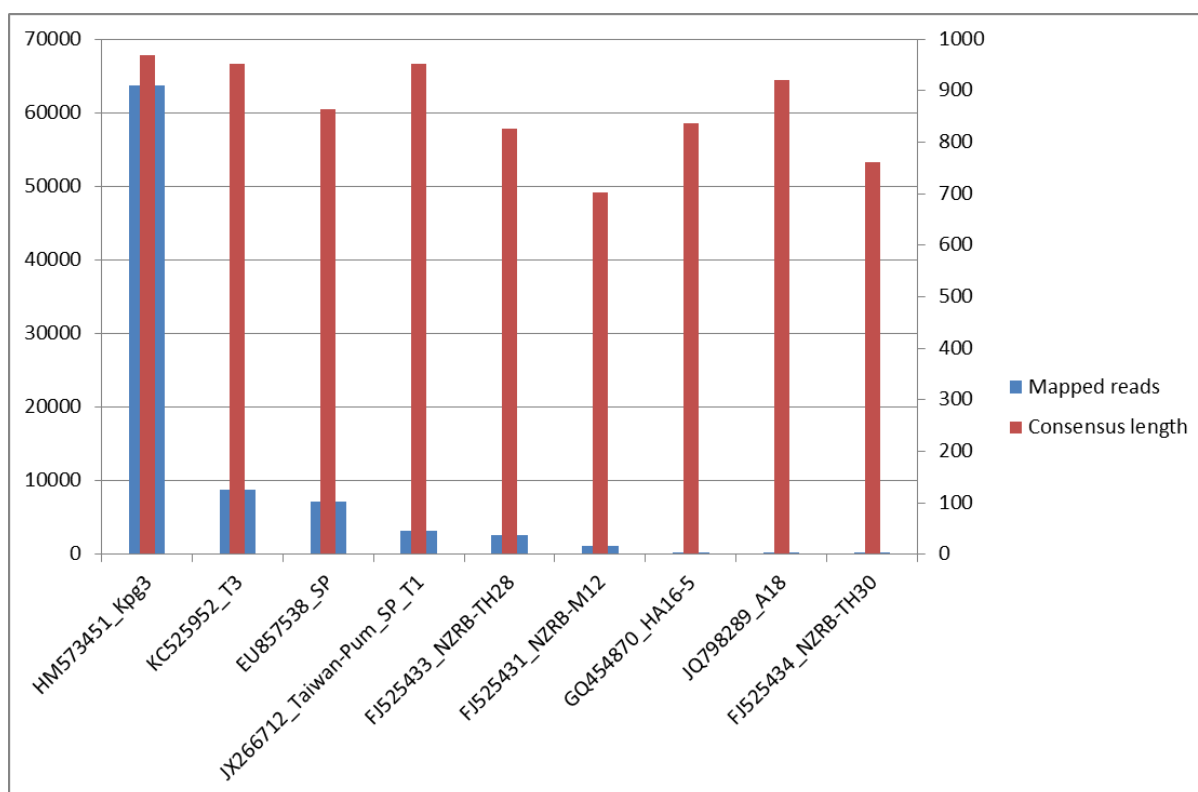


Figure 7.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3627**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

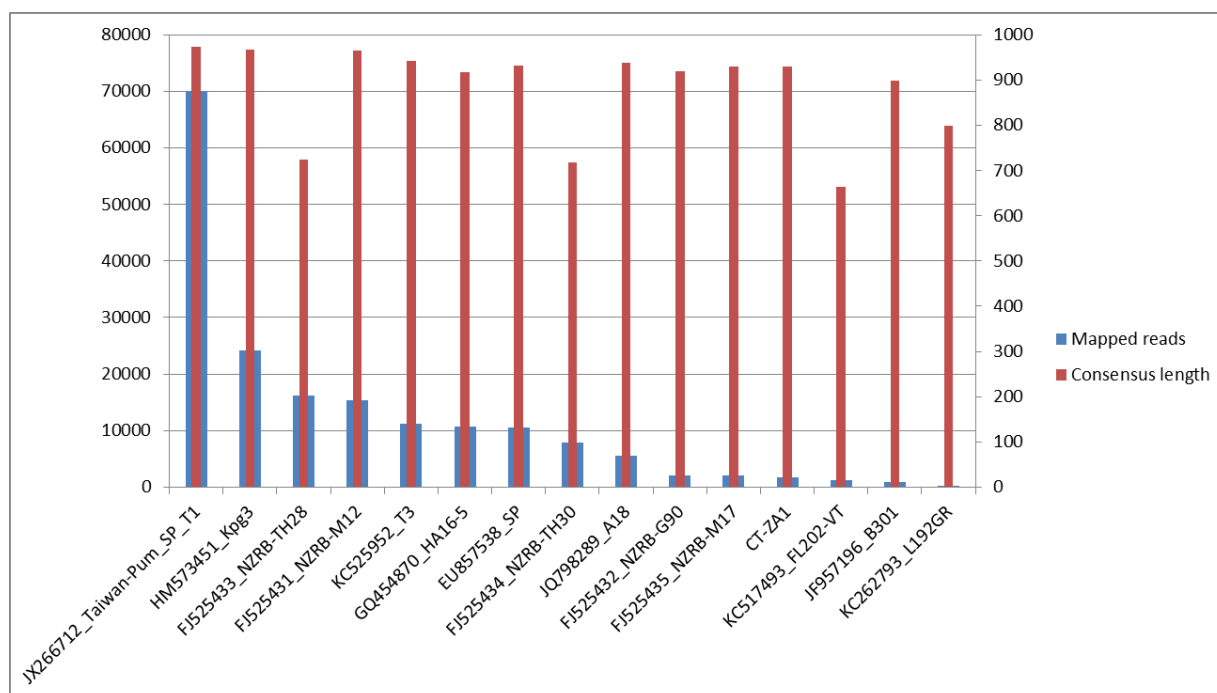


Figure 7.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3637**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

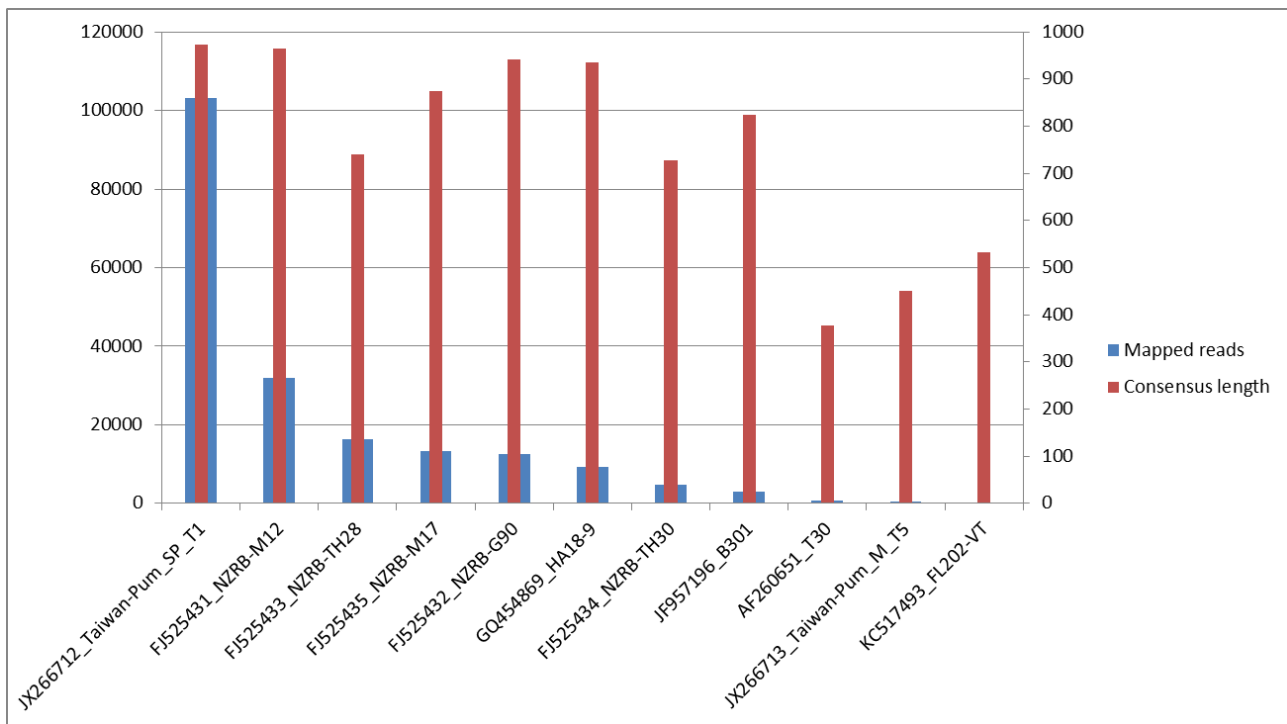


Figure 7.9: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3642**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

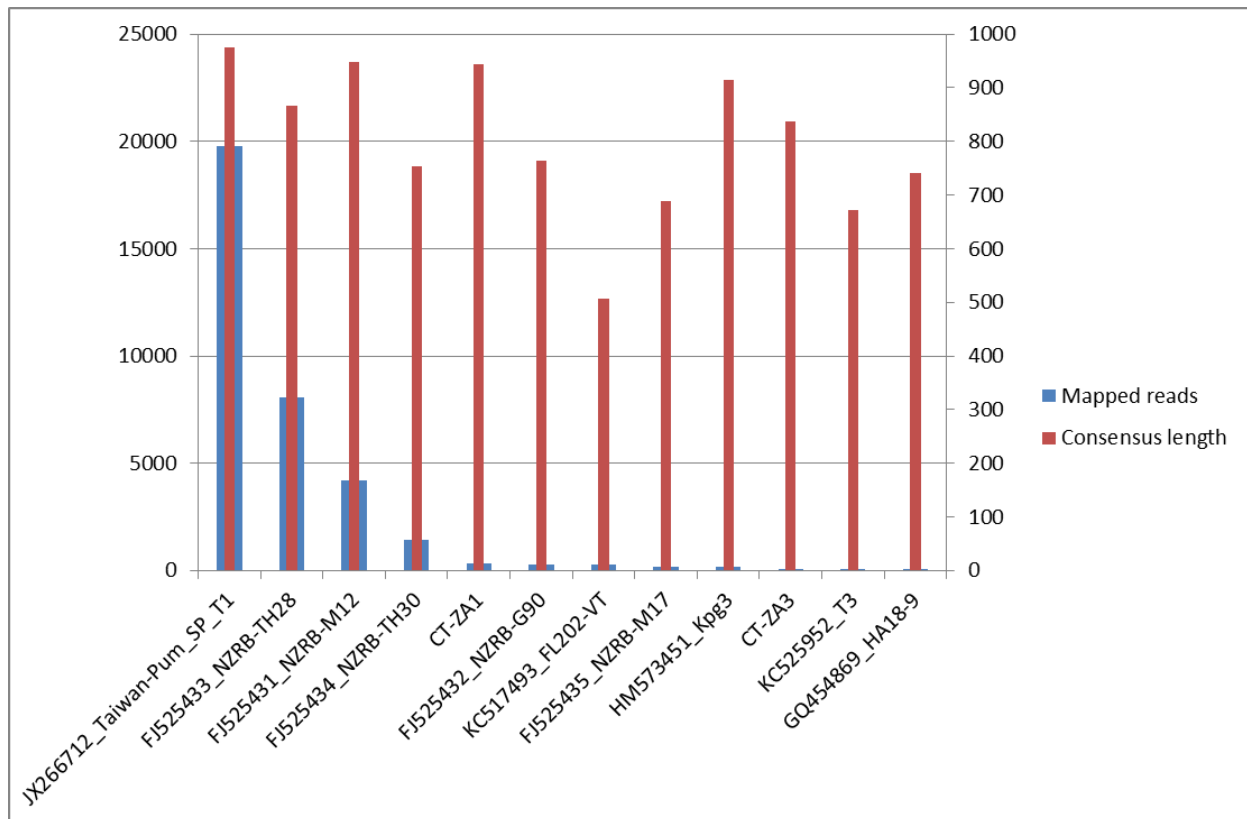


Figure 7.10: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3643**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

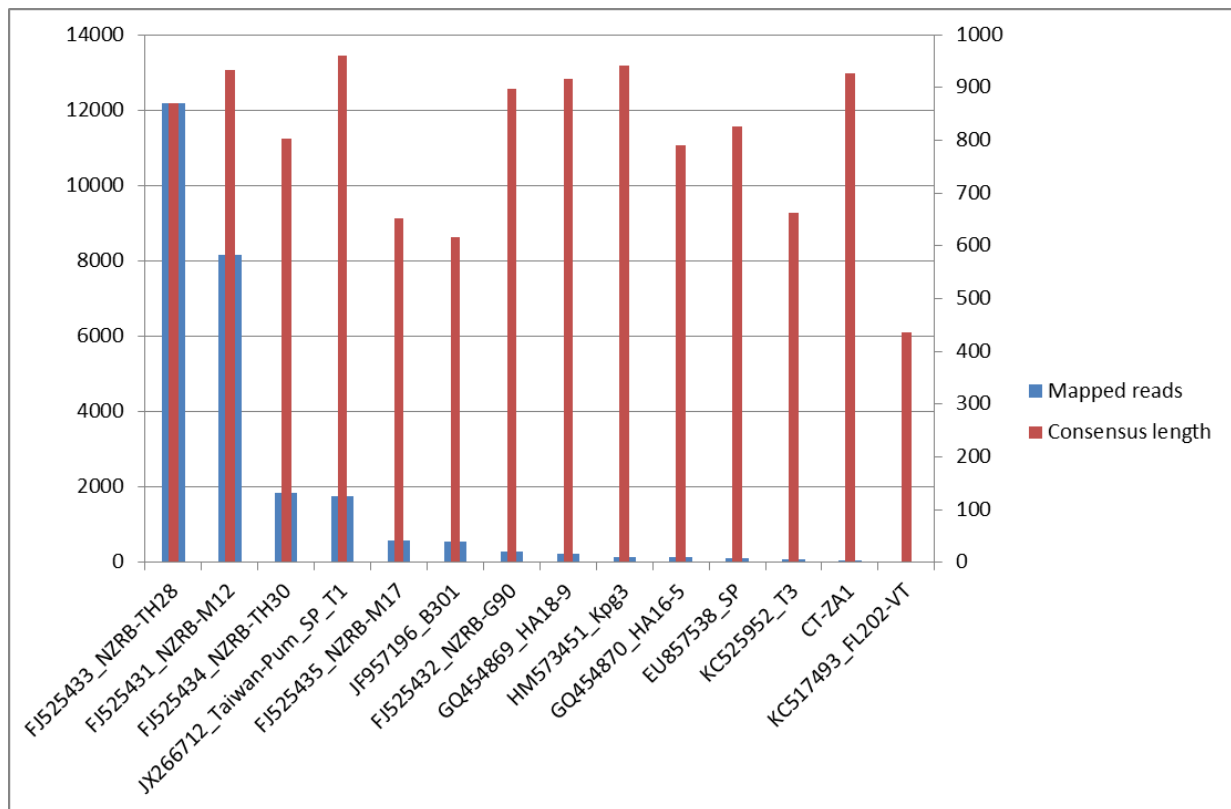


Figure 7.11: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3647**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

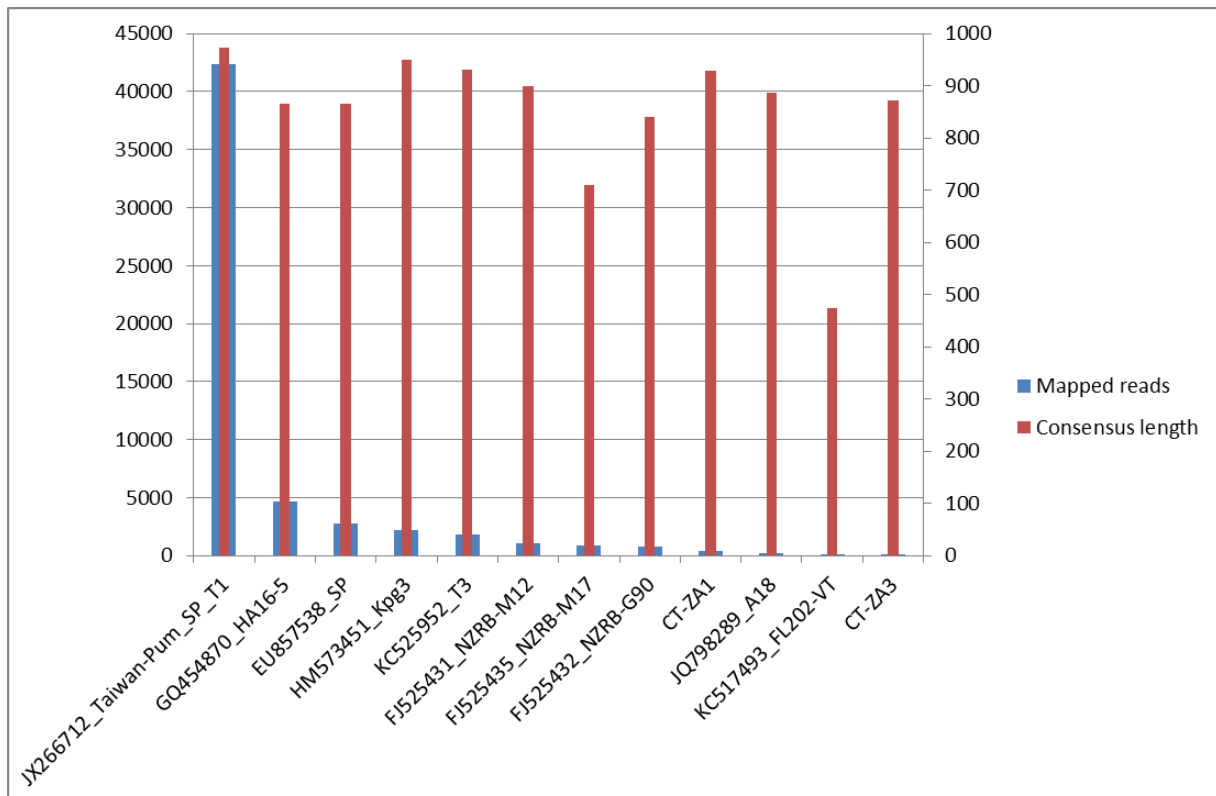


Figure 7.12: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3651**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

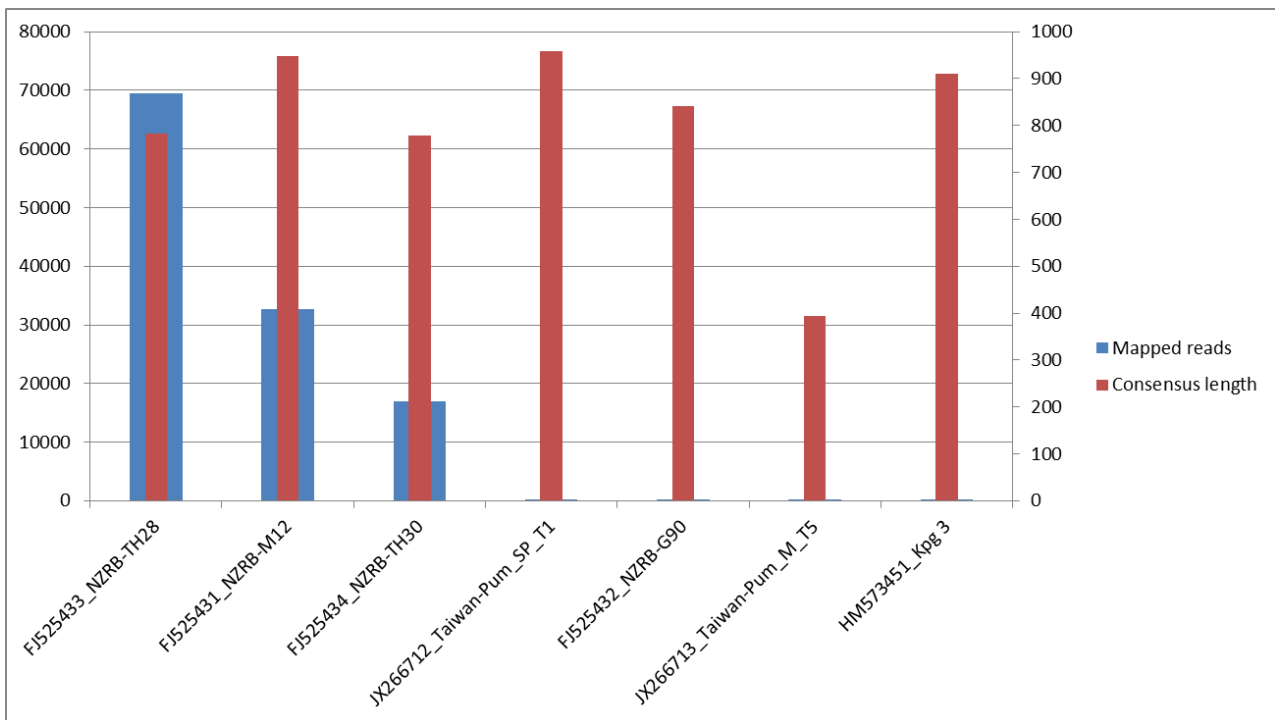


Figure 7.13: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3655**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

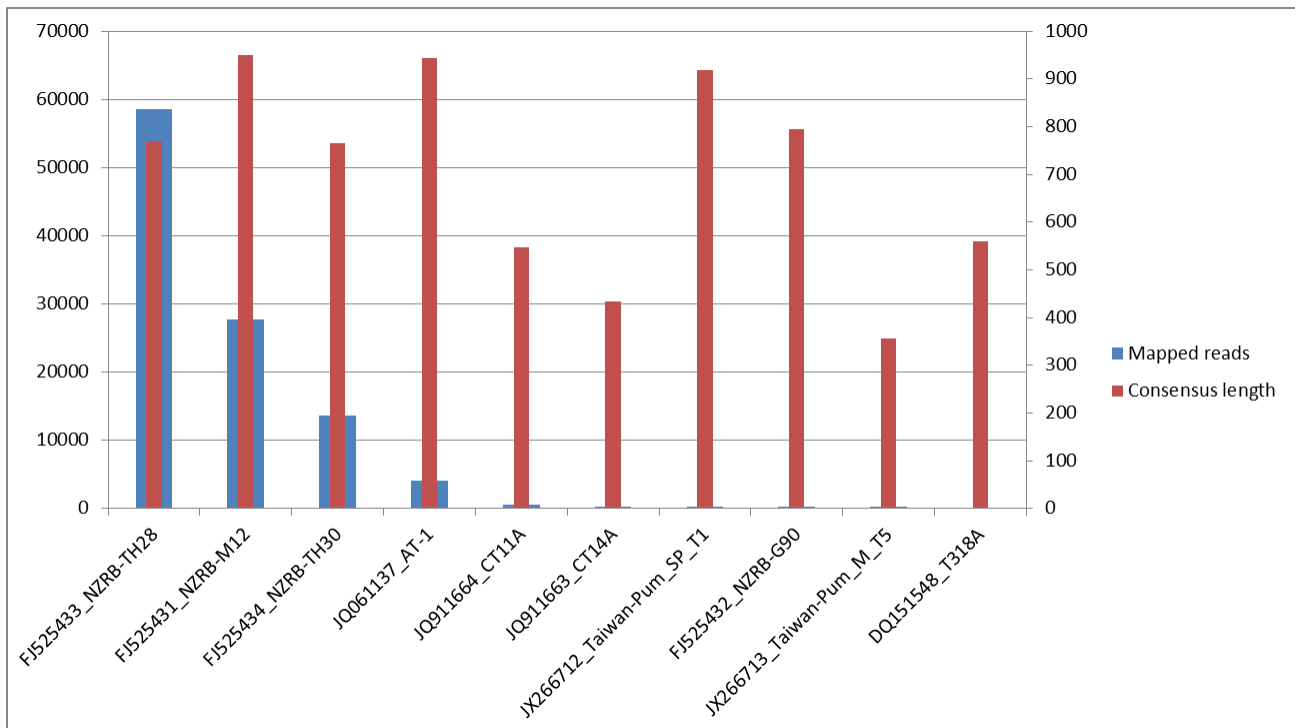


Figure 7.14: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3658**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

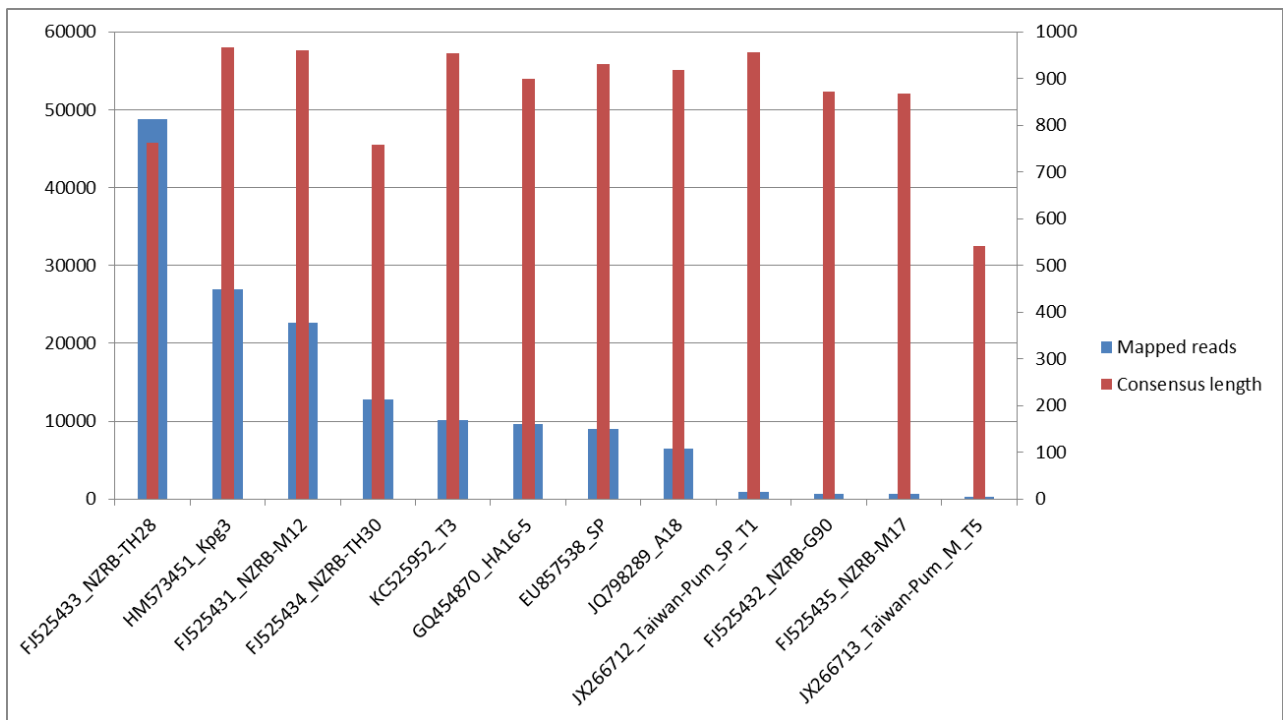


Figure 7.15: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3661**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

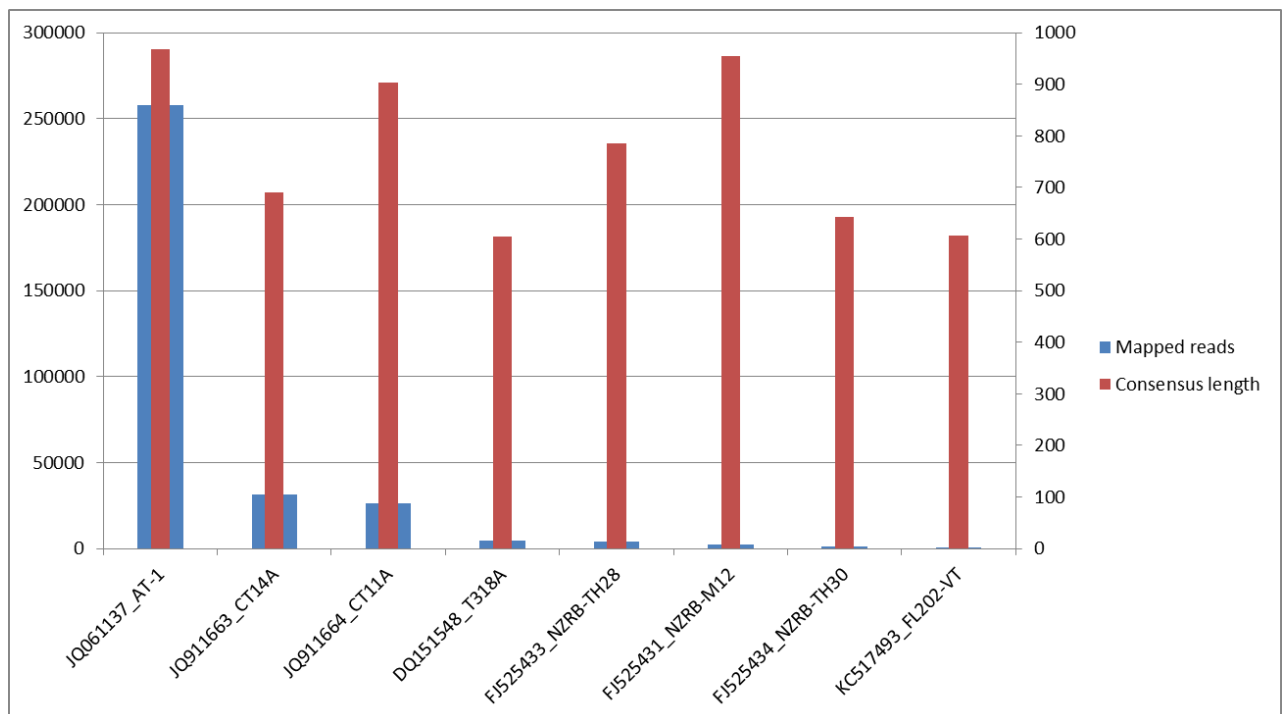


Figure 7.16: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3674**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

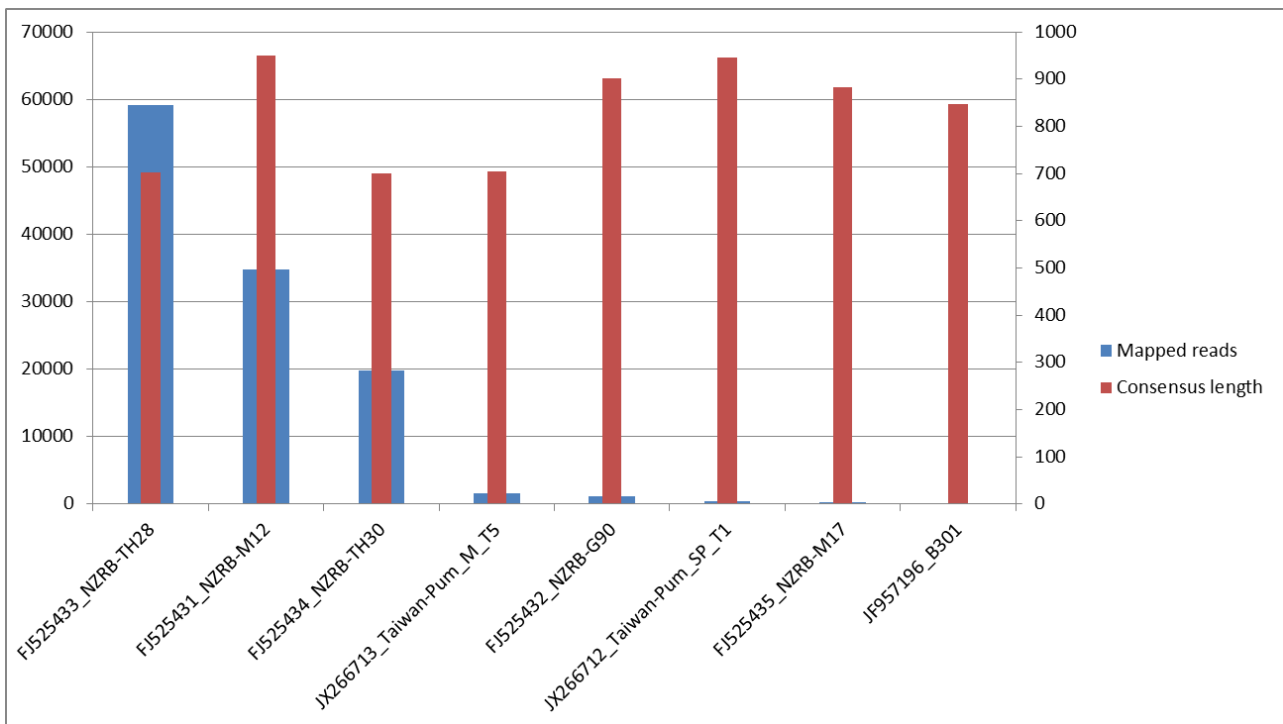


Figure 8.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3705**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

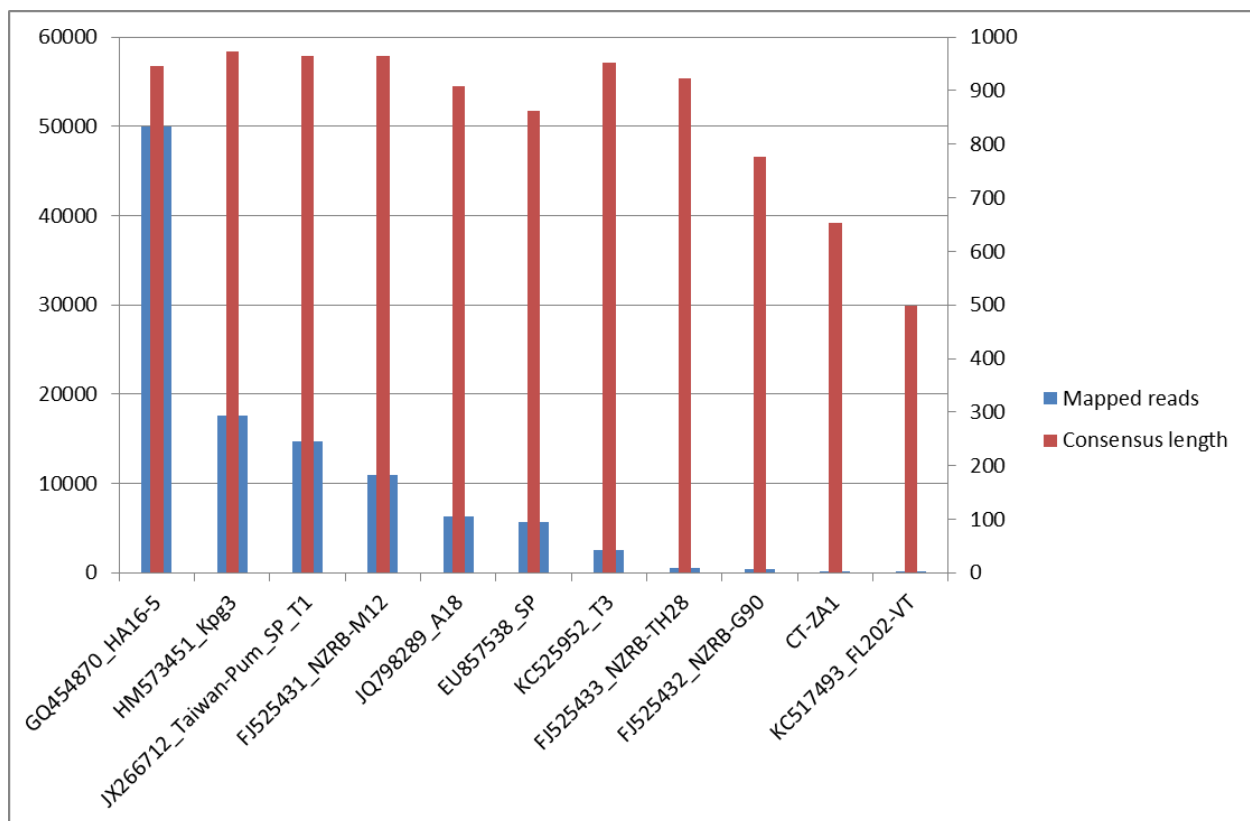


Figure 8.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3708**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

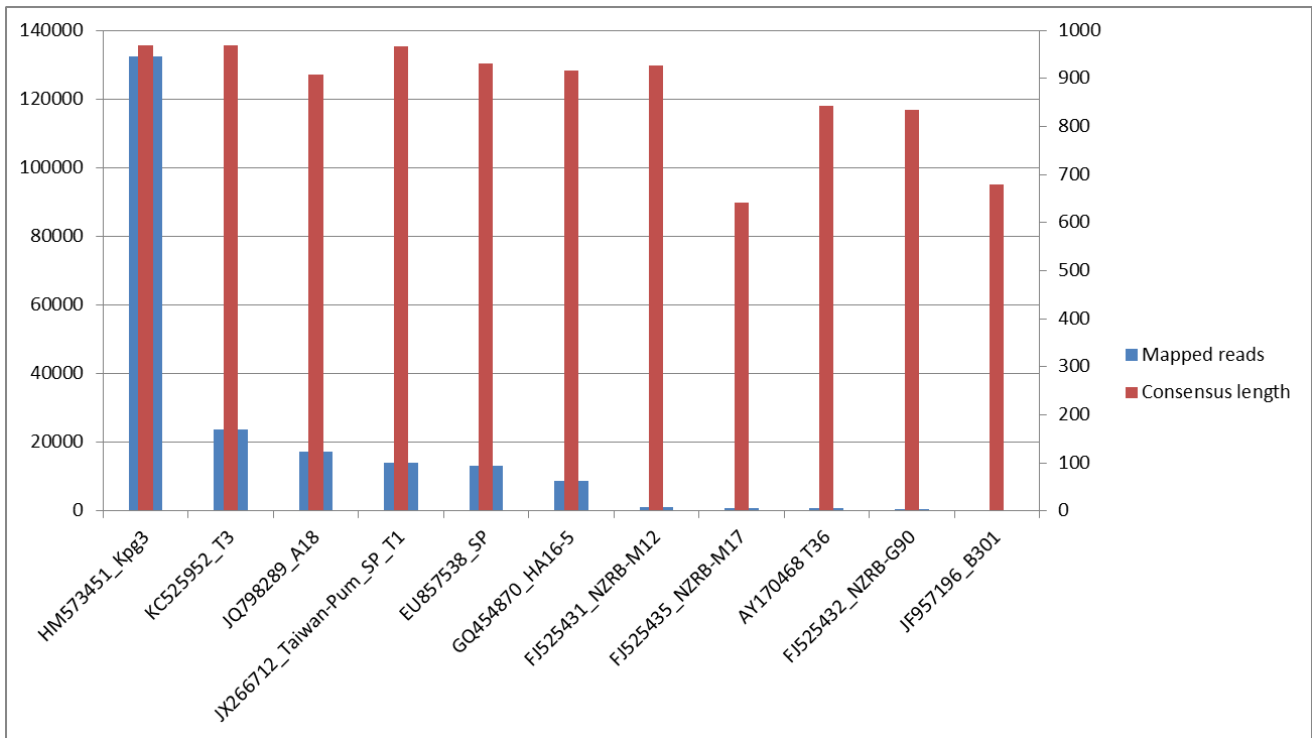


Figure 8.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3713**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

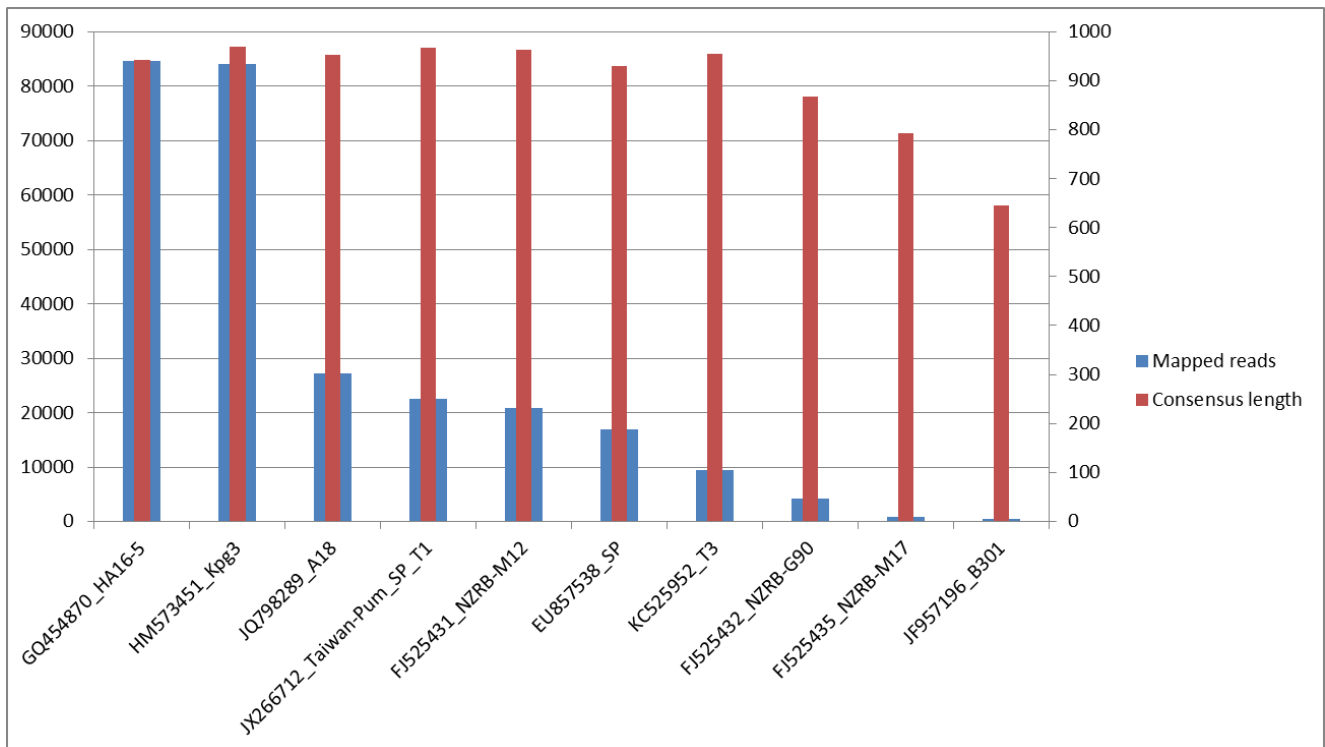


Figure 8.4: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3714**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

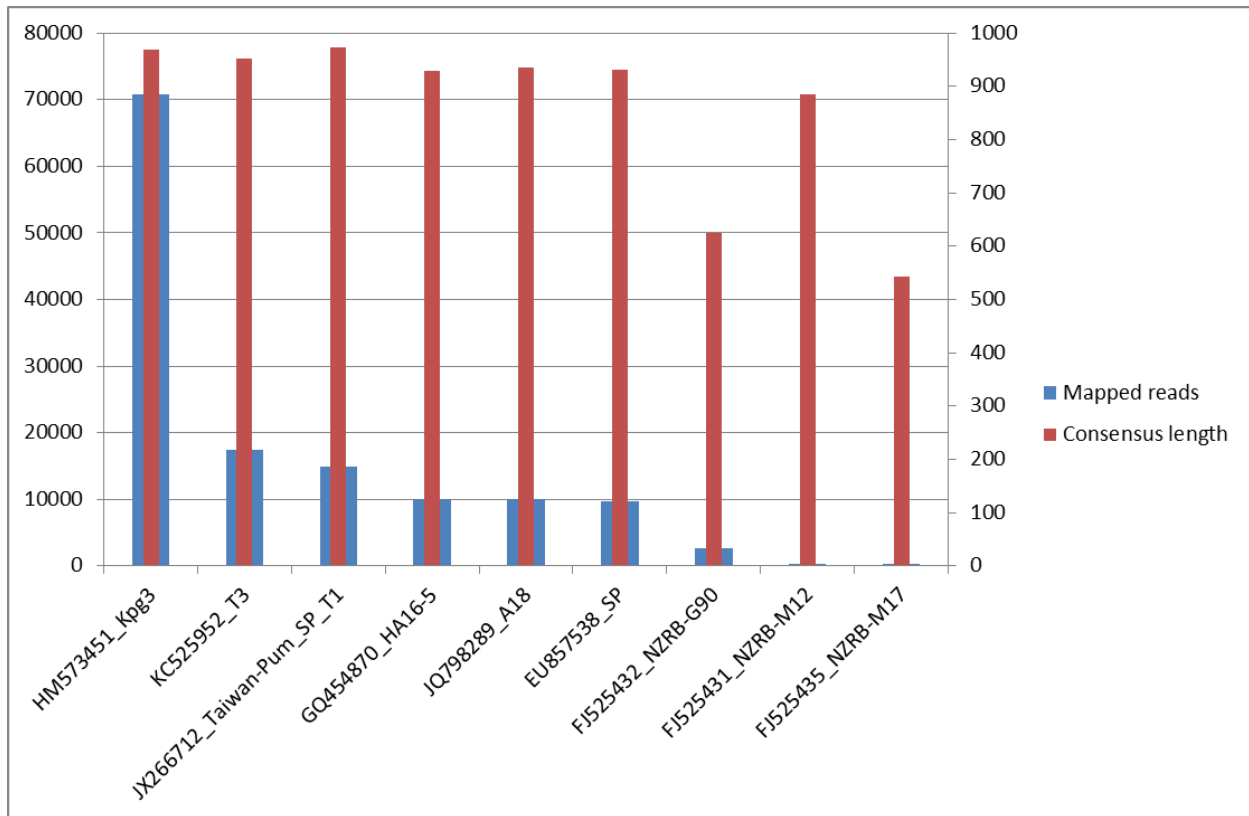


Figure 8.5: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3727**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

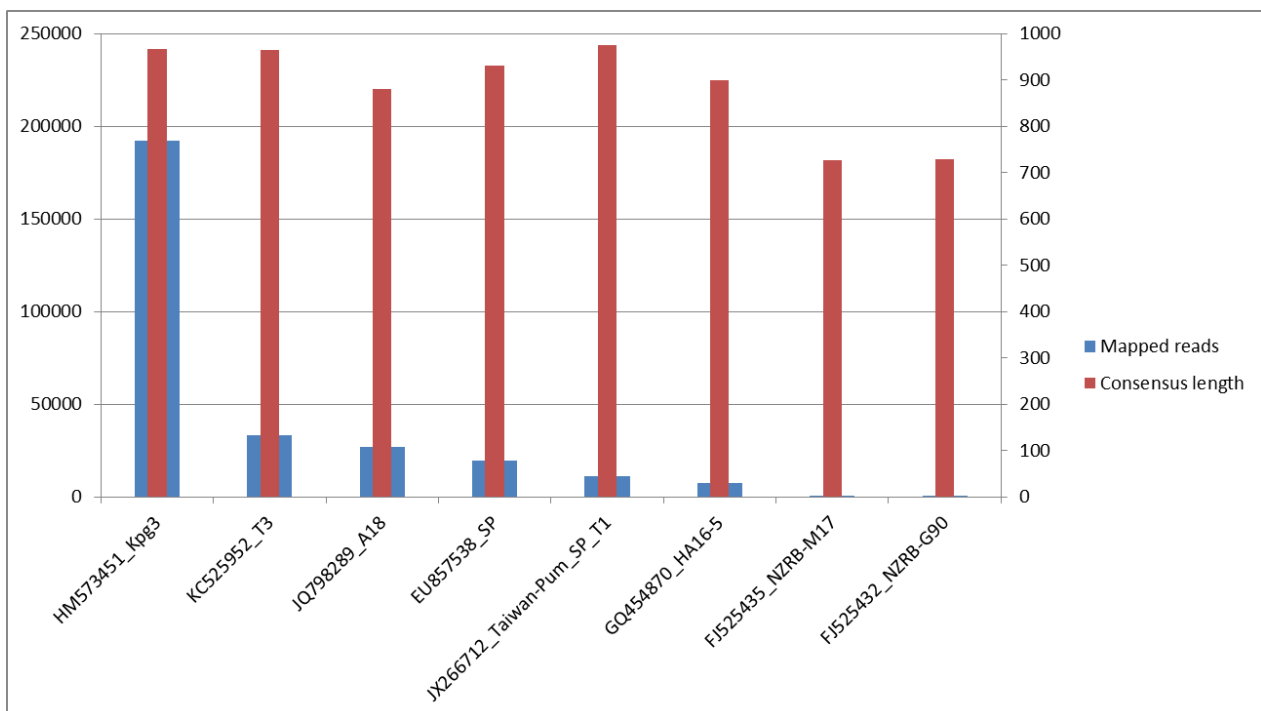


Figure 8.6: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3729**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

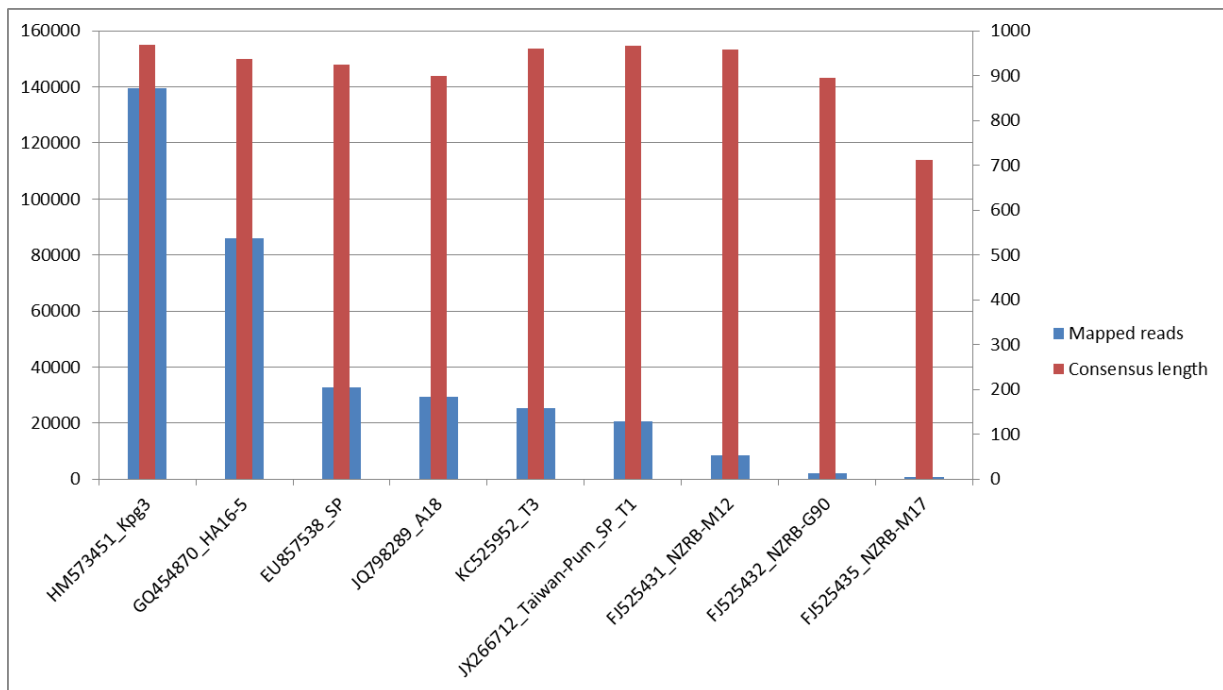


Figure 8.7: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3732**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

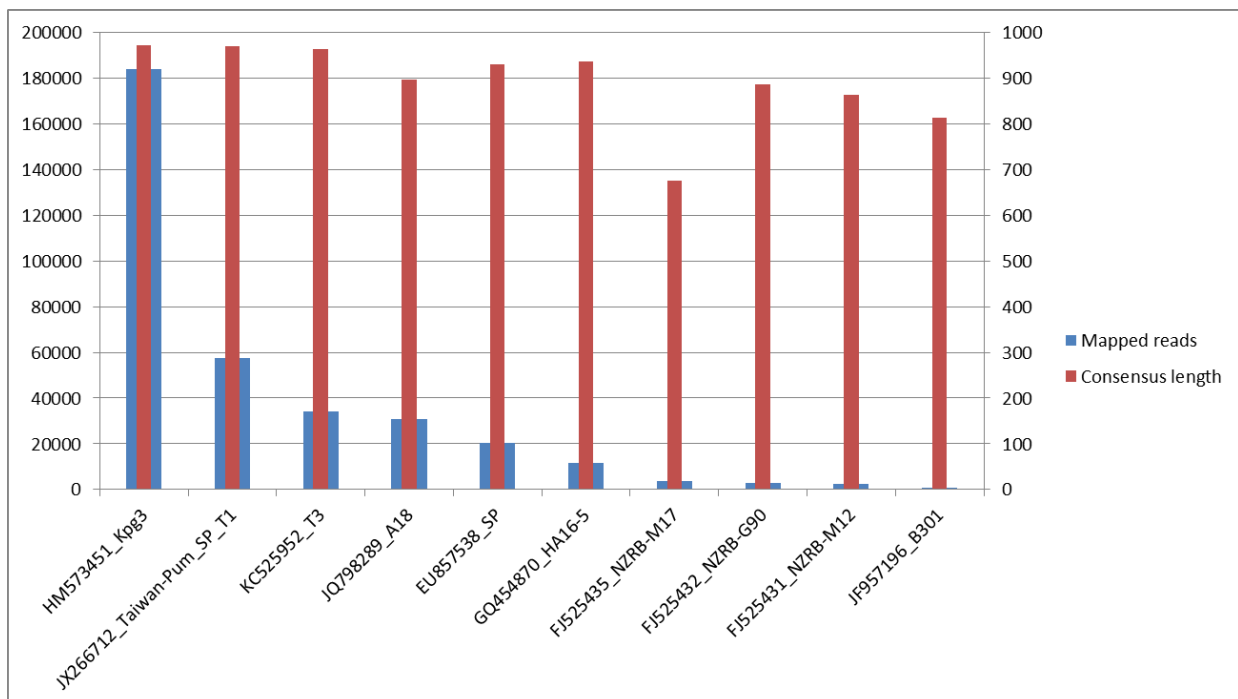


Figure 8.8: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3743**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

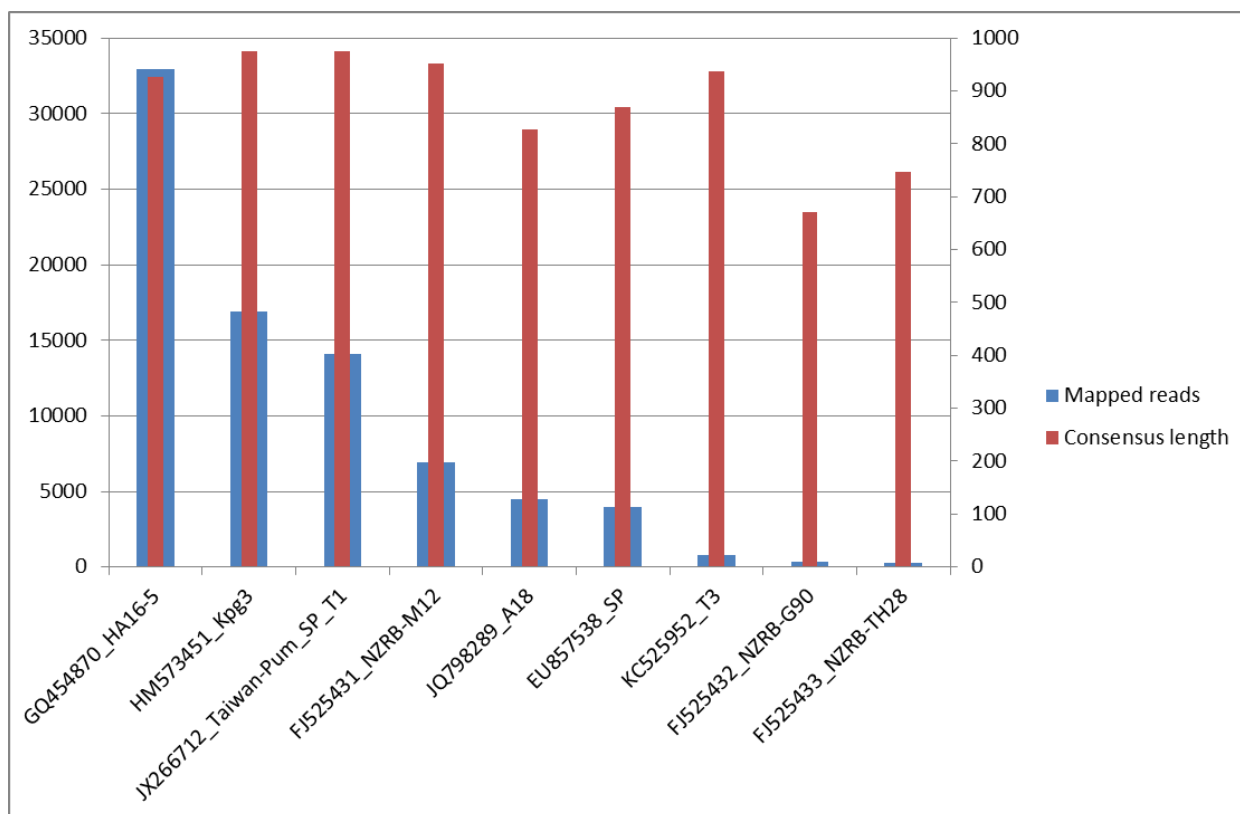


Figure 8.9: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3745**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

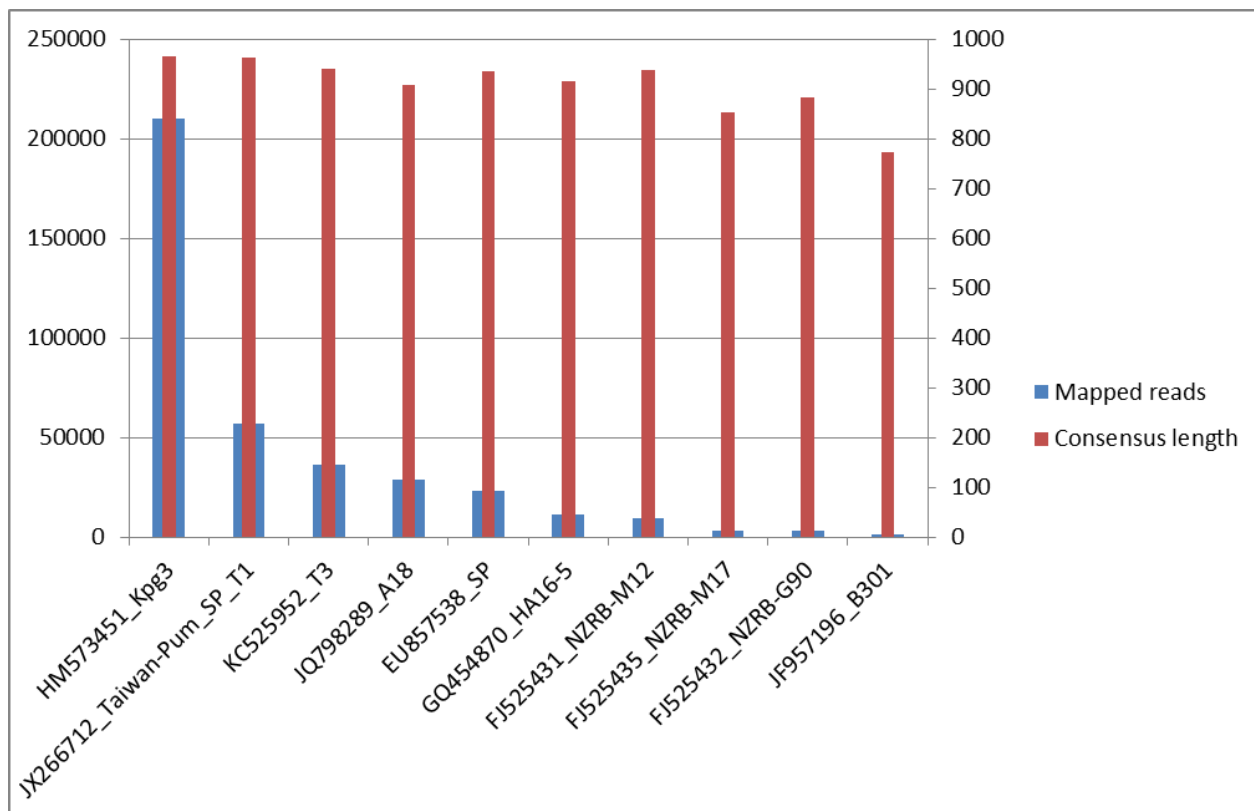


Figure 8.10: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3747**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

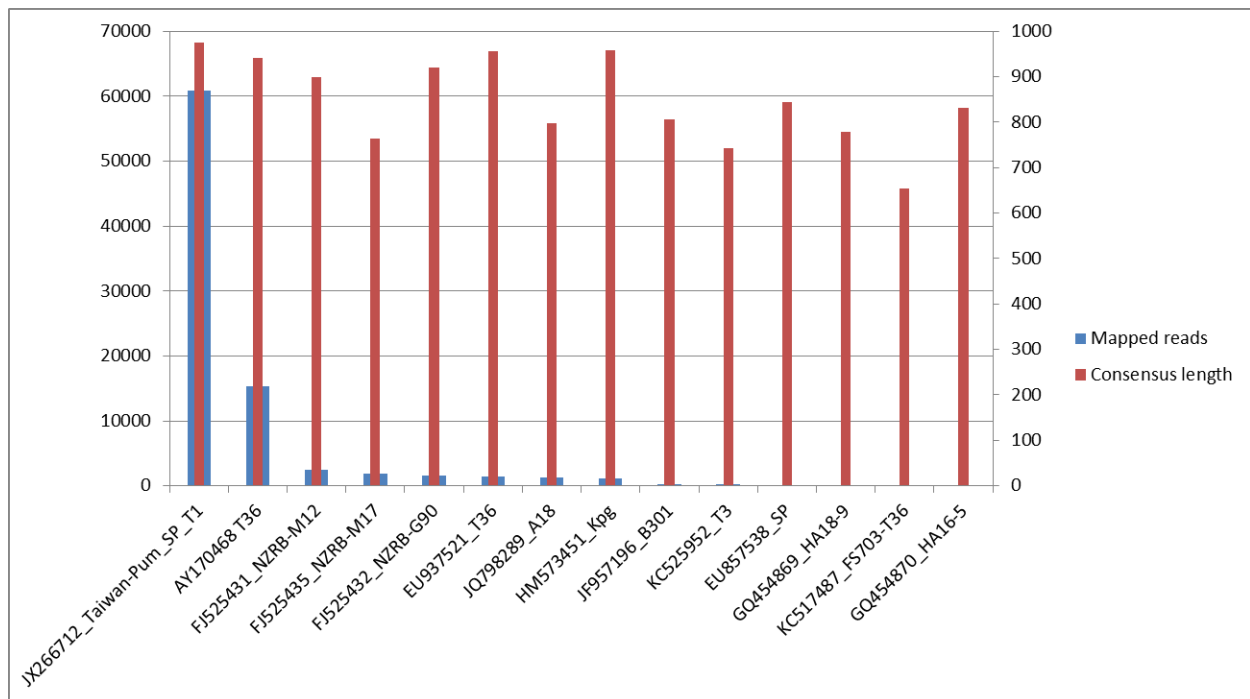


Figure 8.11: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3748**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

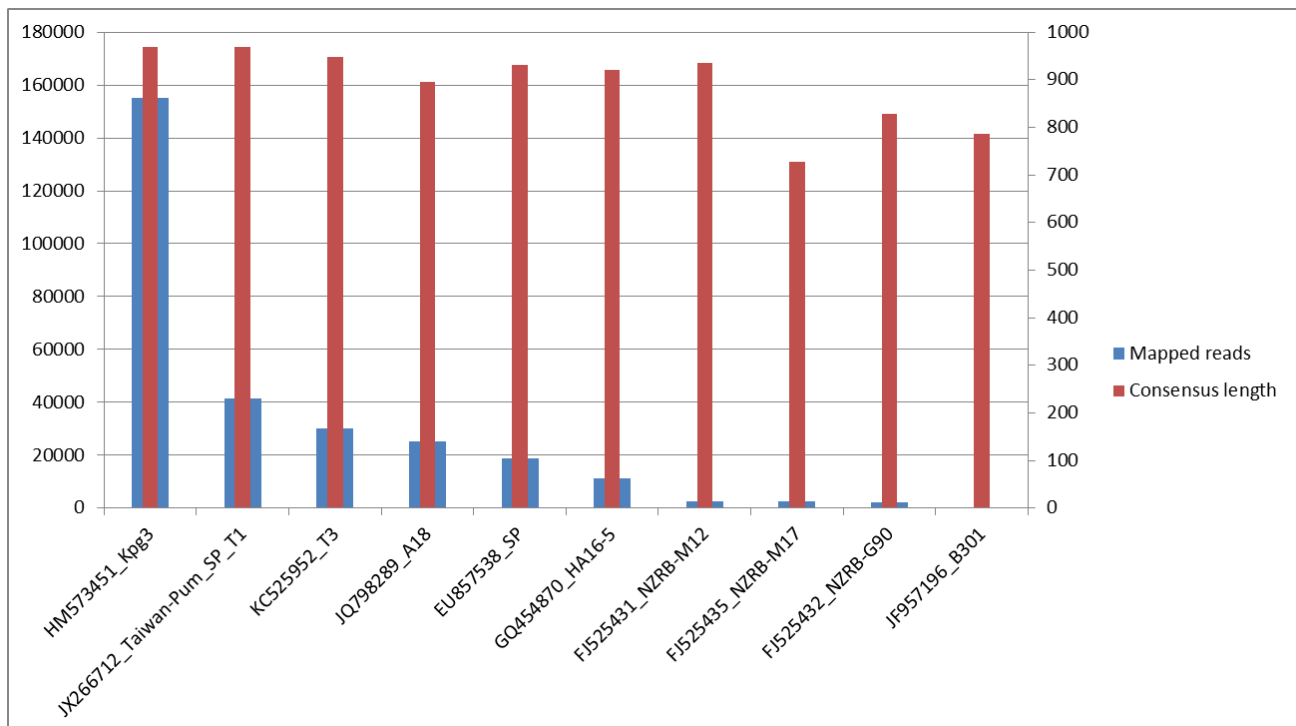


Figure 8.12: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3768**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

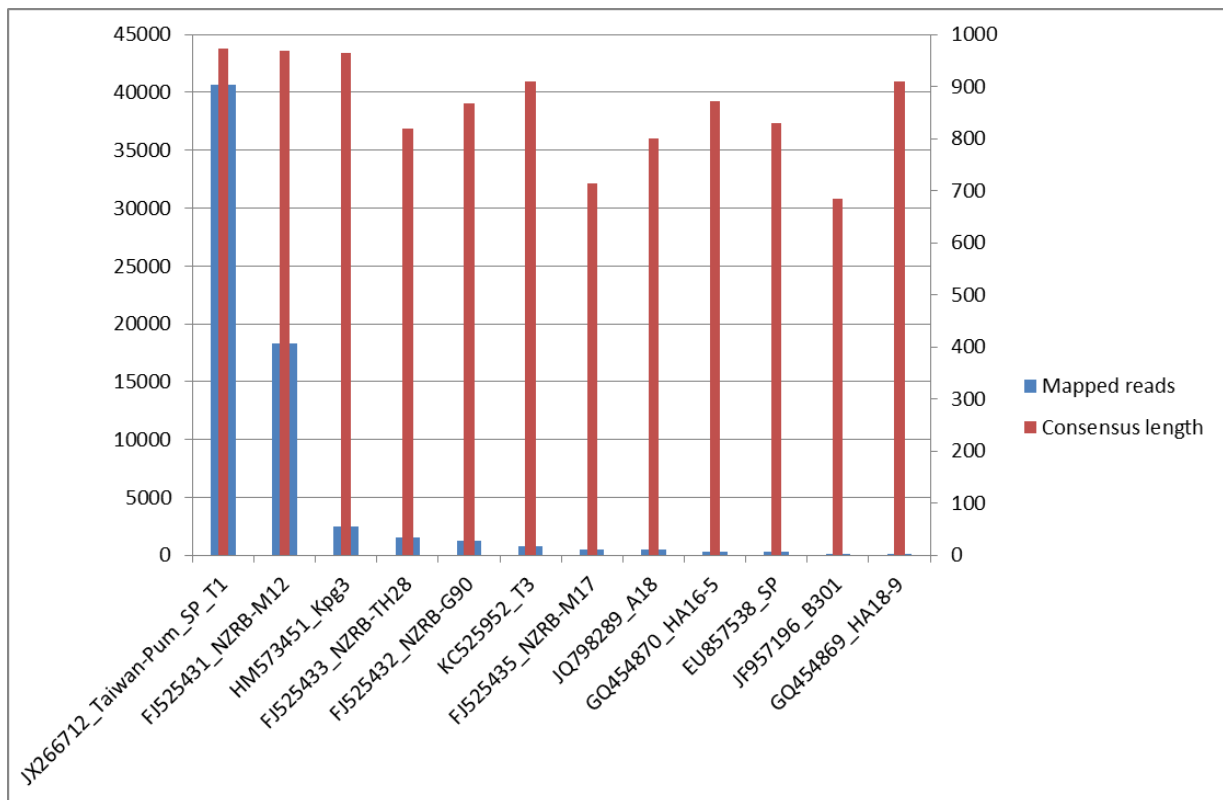


Figure 8.13: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3770**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

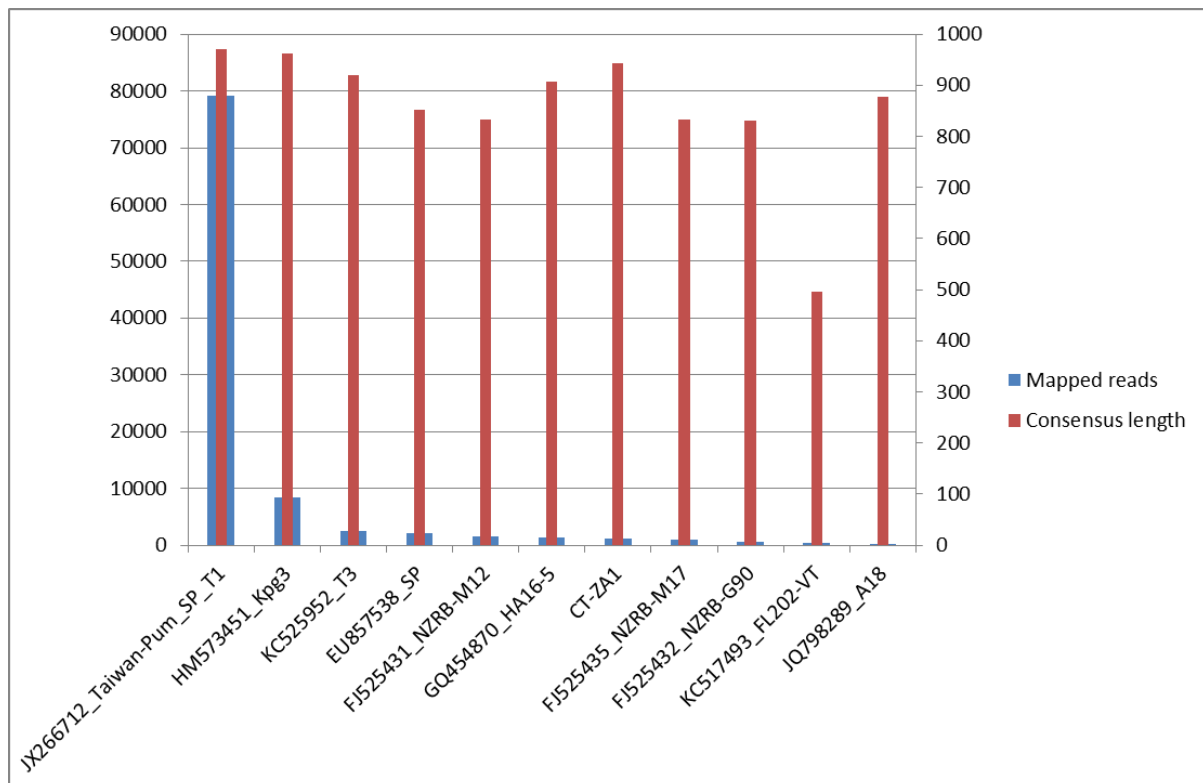


Figure 8.14: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **13-3776**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

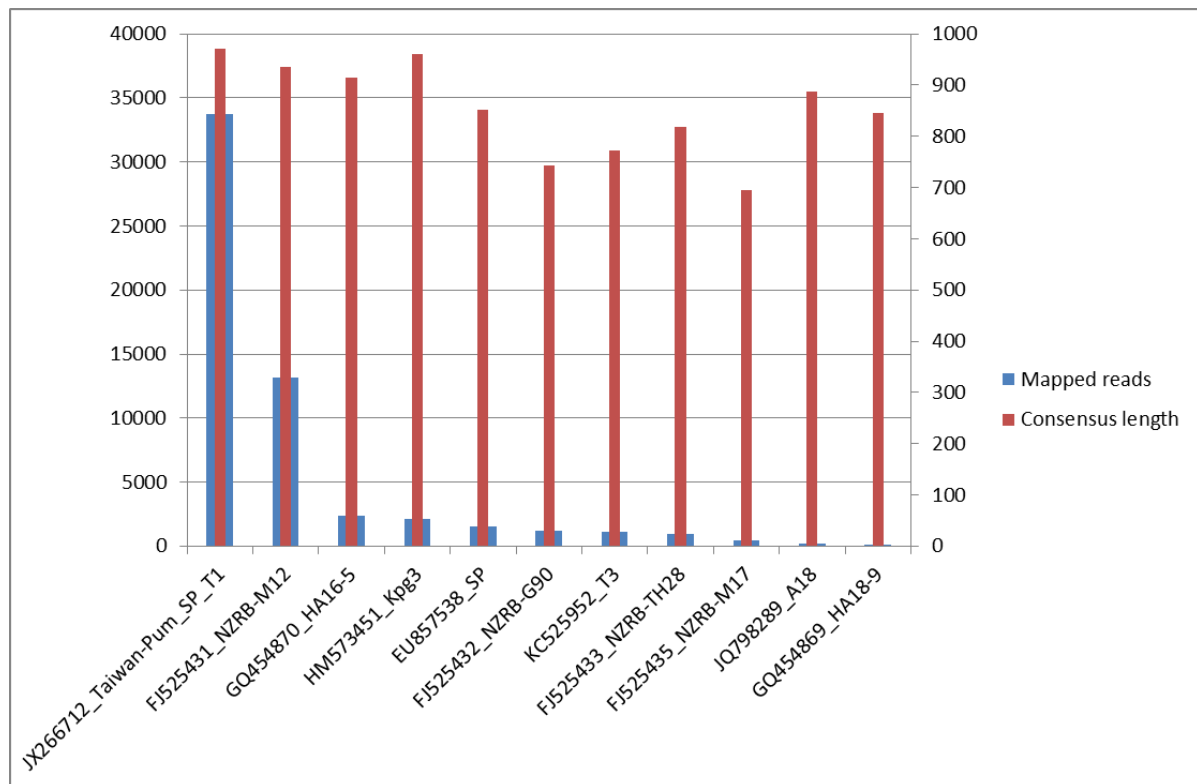


Figure 9.1: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **14-6000**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

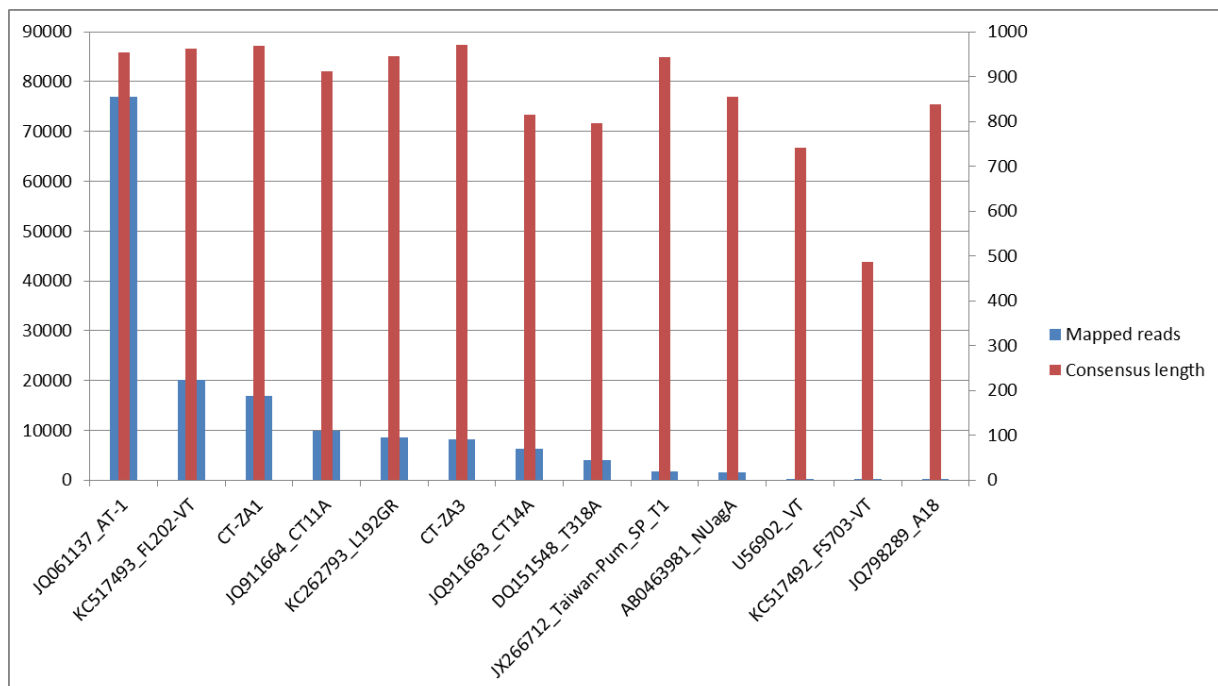


Figure 9.2: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **14-6001**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

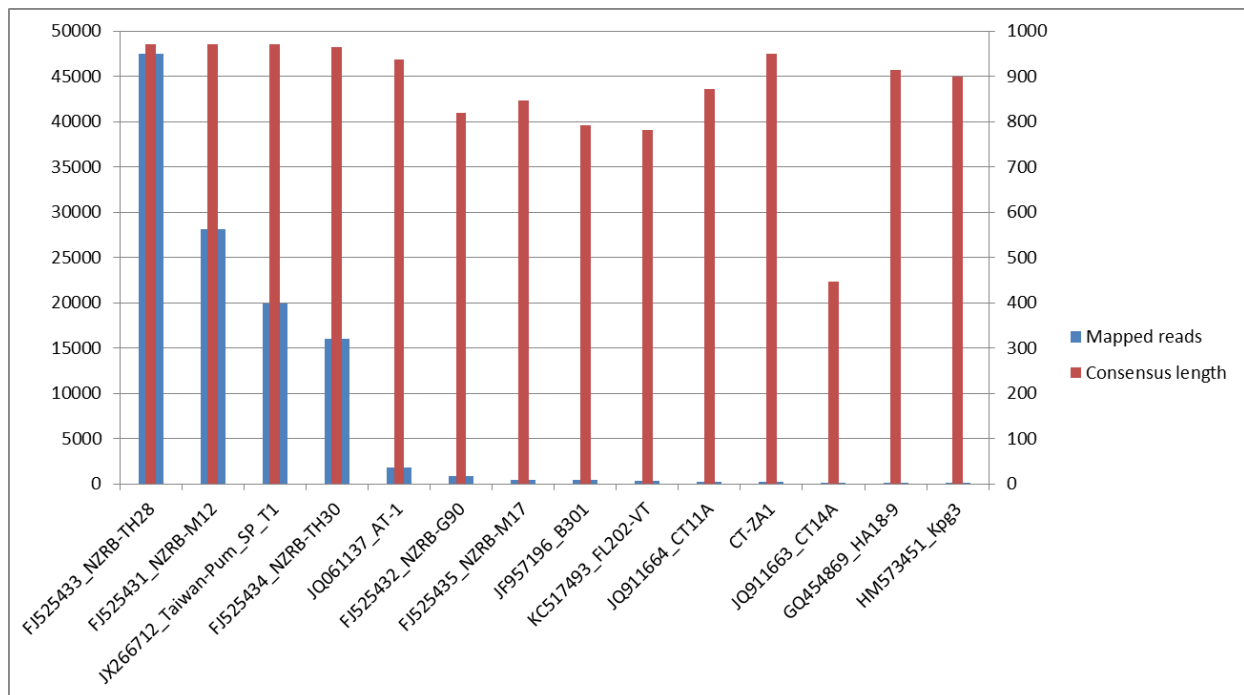


Figure 9.3: Graph representing the number of reads for each of the reference sequences used for sample (y-axis left) versus the consensus length (y-axis right), which is related to the length of the reference that was covered by the Illumina sequence reads for sample **14-6002**. Consensus lengths of >900bp indicate that these reads mapped covered close to 100% of the reference sequence.

