Australian Government

Cotton Research and Development Corporation

DYNAMICS, DIVERSITY AND EVOLUTION OF BACULOVIRUSES

Christopher Noune

BSc (Microbiology) – Queensland University of Technology, 2012

BSc (Ecology) (Honours) – Queensland University of Technology, 2013

Principal Supervisor: Associate Professor Caroline Hauxwell

Associate Supervisor: Associate Professor Jim Hogan

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Science and Engineering Faculty

Queensland University of Technology

2017

Page intentionally left blank

Keywords

Selection pressure; virus evolution; Baculovirus; HaSNPV-AC53; performance metrics; viral dynamics; virulence-transmission; transmission bottlenecks; bioinformatics; metapopulation; meta-barcoding; MetaGaAP; MetaGaAP-Py; community analysis; *Helicoverpa armigera*; Next Generation Sequencing; High Throughput Sequencing; strain selection; *Alphabaculovirus*; Nucleopolyhedrovirus; $HzAM_1$; tissue culture derivatisation; AC53-derived; phylogenetics; median lethal dose; median survival time; occlusion body; HzSNPV; maximum-likelihood estimation; quasispecies; time-course; performance metrics; biopesticides; biological control; *in vivo*; *in vitro*; competitive exclusion principal; niche differentiation.

Abstract

The family *Baculoviridae* (Baculovirus) are dsDNA invertebrate-specific obligate pathogens of the insect orders Lepidoptera, Hymenopteran and Dipteran. An NPV consists of a genome with lengths between 80 and 180 kbp and a community of phenotypically and genetically diverse virus strains that are co-occluded within a protein body. NPVs are widely used in biological control of Lepidopteran pests, and understanding isolate dynamics, diversity and evolution is important in resistance management strategies and developing next generation biopesticides with desired phenotypic traits.

The aim of this study was to apply next generation sequencing and develop bioinformatic techniques to expand and accelerate current knowledge of baculoviruses by studying the dynamics, diversity and evolution. This included the development of a new bioinformatic pipeline to analyse the within-isolate and within-strain diversity, applying this pipeline to monitor the change in genotype abundance during the infection cycle and derivatisation of *in vitro* and *in vivo* selected strains from a wild type isolate of commercial importance, Helicoverpa armigera single nucleopolyhedrovirus isolate AC53 (HaSNPV-AC53). Derived genomes were analysed to identify trait or isolation technique specific mutations, and the global relationships of these strains to all known HaSNPV isolates.

Phylogenetic analysis of all known HaSNPV and *H. zea* SNPV isolates with the addition of AC53 and some of its derivatives supported the claim that these viruses are the same viral species, and suggests that the HaSNPV species may have originated in Australia. The use of whole genomes in phylogenetic analysis gave greater resolution than the more commonly used analysis using selected open reading frames.

Five strain derivatisation approaches were applied: two *in vitro* (in tissue culture) and three *in vivo*. Analysis of both *in vitro* and *in vivo* derived strains' genomes identified selection specific mutations, with fast speed of kill, slow speed of kill and maximum virus production strains containing trait specific mutations. Biological characterisation of these trait-specific strains identified significant virulence-transmission trade-offs such as enhanced speed of kill but reduced efficacy which implicates commercial optimisation of strains.

A new software pipeline called the 'Meta-barcoding Genotyping and Abundance Pipeline' (MetaGaAP) was developed to identify genotypes and their relative abundance within the AC53 isolate. This was validated by Sanger sequencing and comparison to the AC53-T2 strain. The pipeline was applied to monitor AC53 during the infection cycle and identified two evolutionary effects occurring within the population; weak-negative selection with mutation bias and a 'drift barrier' to limit the effects of genetic drift. Furthermore, time-course assays revealed

a significant reduction in dominant genotype abundance with an increase in minor genotype abundance when the inoculum is compared to the final viral product. This implicates commercial production as the starting material and the produced material contain different genotype abundance profiles, however, both products contain the same genotype composition.

In addition, results presented throughout this study suggested that NPVs fit the viral quasispecies model as mutations that arose were the result of mutational robustness and genotype cooperation. Limitations observed with current NGS and bioinformatic techniques partially impacted the described results but may eventually resolve with advent of third-generation sequencing.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Table 1-1: Abbreviations and meanings used throughout the thesis

Abbreviation	Meaning or description
ORF(s)	Open Reading Frame(s)
NGS	Next Generation Sequencing
(S or M) NPV	(Single or Multiple) Nucleopolyhedrovirus
HaSNPV	Helicoverpa armigera single nucleopolyhedrovirus
HzSNPV	Helicoverpa zea single nucleopolyhedrovirus
OB(s)	Occlusion Body(ies)
GV	Granulovirus
AC53	Helicoverpa armigera single nucleopolyhedrovirus isolate AC53
H25EA1	Helicoverpa armigera single Nucleopolyhedrovirus isolate H25EA1
AC53-C1	Helicoverpa armigera Single nucleopolyhedrovirus strain AC53-C1
AC53-C3	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-C3
AC53-C5	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-C5
AC53-C6	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-C6
AC53-C9	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-C9
AC53-T2	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-T2
AC53-T4.1	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-T4.1
AC53-T4.2	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-T4.2
AC53-T5	Helicoverpa armigera single nucleopolyhedrovirus strain AC53-T5
MaxOB	Maximum Occlusion Body production
egt	*ecdysteroid UDP-glucosyltransferase*
MLE	Maximum-Likelihood Estimation
ST_{50}	Median Survival Time
LC_{50}	Median Lethal Concentration
PCA	Principal Component Analysis
SNP(s)	Single Nucleotide Polymorphism(s)
BV	Budded Virus
ODV	Occlusion Derived Virus
MetaGaAP(-Py)	Meta-barcoding Genotyping and Abundance Pipeline (-Python)
GATK	Genome Analysis Toolkit
BWA	Burrows-Wheeler Transform Aligner
(q)PCR	(quantitative)Polymerase Chain Reaction
DGGE	Denaturing Gel Gradient Electrophoresis
BRO-(A or B)	Baculovirus Repeated Open Reading Frame – (A or B)
RFLP	Restriction Fragment Length Polymorphism
P.I.	Post Infection
Hr	Homologous Repeat
mM	Millimole
mL	Milliliter
OB/µg	Occlusion bodies per microgram
θ	Tajima's D and Fay and Wu's H statistic
PxGV	*Plutella xylostella* Granulovirus
Bash	Bourne again shell
RAM	Random Access Memory
GLM	General linear model
GTR	General time reversible
AcMNPV	*Autographica californica* Multiple Nucleopolyhedrovirus
TAE	Tris-acetate- Ethylenediaminetetraacetic acid
SDS	Sodium dodecyl sulfate
ENT	Effective number treated

Publications Incorporated into the Thesis

1. Noune, C., & Hauxwell, C. (2015). Complete Genome Sequences of *Helicoverpa armigera* Single Nucleopolyhedrovirus Strains AC53 and H25EA1 from Australia. *Genome announcements, 3*(5). doi:10.1128/genomeA.01083-15

2. Noune, C., & Hauxwell, C. (2016). Complete Genome Sequences of Seven *Helicoverpa armigera* SNPV-AC53-Derived Strains. *Genome announcements, 4*(3). doi:10.1128/genomeA.00260-16

3. Noune, C., & Hauxwell, C. (2016). Comparative Analysis of HaSNPV-AC53 and Derived Strains. *Viruses, 8*(11), 280.

4. Noune, C., & Hauxwell, C. (2017). MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data. *Biology, 6*(1), 14.

5. *Noune, C., & Hauxwell, C. (2017). Enhanced Pipeline 'MetaGaAP-Py' for the Analysis of Quasispecies and Non-Model Microbial Populations using Ultra-Deep 'Meta-barcode' Sequencing. *bioRxiv*. doi:10.1101/171520

6. *Spence, R. J., Noune, C., & Hauxwell, C. (2016). Complete Genome Sequences of Four Isolates of *Plutella xylostella* Granulovirus. *Genome announcements, 4*(3). doi:10.1128/genomeA.00633-16

*Publication 5 & 6 are incorporated into the appendix.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: QUT Verified Signature

Date: October 2017

Acknowledgments

I wish to thank my principle supervisor **Associate Professor Caroline Hauxwell** for all her hard work in bestowing me with the knowledge to improve my academic writing, designing successful experiments and leading me into the world of baculoviruses. All my academic and employment opportunities have stemmed from your guidance and teachings. I could not ask for a better mentor. I would also like to thank my associate supervisor **Associate Professor Jim Hogan** for his guidance and support. In addition, I would like to thank **Associate Professor James McGree** for all his last-minute contributions and recommendations regarding the statistical analysis of the thesis. I look forward to working with James in the future.

I would to thank my research group the **Invertebrates & Microbiology Group** for all their support and help, especially **Mr Robert Spence** and **Mr Andrew Dickson**, and thank all the **Earth, Environment and Biological Sciences** and the **Molecular Genetics Research Facility** technical staff, especially **Ms Anne-Marie McKinnon**, **Mr Vincent Chand** and **Dr Kevin Dudley**.

I would also like to acknowledge and thank my project manager **Ms. Susan Maas** from the **Cotton Research Development Corporation** for all her help, and the organisations which funded this project, the **Cotton Research Development Corporation**, the **Australian Government Research Training Program** and **QUT** for their support and financial contribution. If not for their financial backing and trusting in my abilities, I could not have been able to complete this PhD. In addition, I would like to thank **AgBiTech Pty Ltd** for providing the baculovirus isolate and insects used in this study, and thank **Dr Steve Reid** for providing the tissue-culture cells and training in the maintenance of these cells.

Finally, I would not be here writing this thesis without the love and support from **my parents, my brother and my partner Steph**. Through-out my life and this academic journey you have picked me up when I have fallen and pushed me to always be the best I possibly can. You've supported and encouraged me during the hard times and celebrated with me during the happy times. No one in this world is more important to me than my family and I love them dearly. I dedicate this thesis to them.

Chapter 1: Introduction

1.1 BACKGROUND AND SIGNIFICANCE

The baculoviruses (*Baculoviridae*) are a family of diverse double-stranded DNA viruses that are obligate pathogens of the insect orders *Lepidoptera*, *Hymenoptera* and *Diptera* (G. F. Rohrmann, 2013c). These viruses are increasingly used as biopesticides against agro-economically important pests (Marie Berling et al., 2009; B. C. Black, L. A. Brennan, P. M. Dierks, & I. E. Gard, 1997; Buerger, Hauxwell, & Murray, 2007; Hauxwell, 2008a; G. Zhang, 1989). The development of technology in bioinformatics and 'next generation sequencing' (NGS) has created new opportunities to increase our understanding of fundamental questions in the dynamics, diversity and evolution of these virus (Chateigner et al., 2015; Fleming-Davies, Dwyer, Rohani, & Kalisz, 2015; McElroy, Thomas, & Luciani, 2014; Noune & Hauxwell, 2016a; White, Burden, Maini, & Hails, 2012).

Previous studies have relied on culture-based techniques to isolate and identify strain variants within isolates but have recently begun including bioinformatic and NGS approaches (V.L. Baillie & Bouwer, 2011; Vicky Lynne Baillie & Bouwer, 2012a, 2012b; Chateigner et al., 2015; Noune & Hauxwell, 2016a). However, the baculoviruses are both double-stranded DNA viruses and pathogens of invertebrates, and thus are a 'non-model' organism, and although they may provide insights into some aspects of virology, there has been comparatively little investment to date to analyse strain variation within isolates, and the technology available may be either impractical, have high error rates or are not yet available (V.L. Baillie & Bouwer, 2011; Vicky Lynne Baillie & Bouwer, 2012a; Chateigner et al., 2015; Lueders & Friedrich, 2003; McElroy et al., 2014; Neilson, Jordan, & Maier, 2013; Noune & Hauxwell, 2016a; Schloss, Gevers, & Westcott, 2011; Sipos, Székely, Révész, & Márialigeti, 2010).

The biology of baculoviruses has been well described, and the presence of multiple strains within isolates is a well-documented feature (Blissard & Rohrmann, 1990; G. F. Rohrmann, 2013b, 2013c, 2013d). Studies on the diversity of strains within isolates and their biology, ecology and interactions with the host have used strains derived from isolates by *in vitro* and *in vivo* selection, and include modelling of trade-offs between virulence and transmission, but work on the effects of selection pressure, such as selection for increased speed of kill, and the resultant changes in phenotypic and genotypic traits within isolates have been relatively little investigated (Marie Berling et al., 2009; D.J Hodgson et al., 2004; Elizabeth M Redman, Wilson, & Cory, 2016; White et al., 2012).

The aim of the research presented in this thesis was to develop and apply new NGS and bioinformatics approaches to the study of the dynamics, diversity and evolution of baculoviruses during the infection cycle in both insects and tissue culture and in response to selection pressure during *in vitro* and *in vivo* passage. From a commercial perspective, the inadvertent or deliberate application of selection pressure to isolates can lead to amplification or selection of strains and production of isolates with different traits from the parent isolate. From a research perspective, these studies develop new approaches to the use of shotgun and ultra-deep viral sequencing and apply these to the understanding of fundamental questions in virology, including trade-offs between virulence and transmission, and evolution of strains dynamics of strains during infection or baculovirus production.

To achieve this aim, a commercialised biological control against the polyphagous insect pests of *Helicoverpa* spp. (Lepidoptera: Noctuidae), a group II singly-enveloped baculovirus isolate Helicoverpa armigera single nucleopolyhedrovirus isolate AC53 (AC53) (also known as A44WT) was used as a model system (Christian, Gibb, Kasprzak, & Richards, 2001; Richards & Christian, 1999; Rowley, Popham, & Harrison, 2011). The virus isolate belongs to the *Helicopverpa armigera Nucleopolyhedrovirus* species, and throughout the thesis, the species will be referred to as HaSNPV ("Virus Taxonomy: 2016 Release," 2016). AC53 is manufactured in Australia and included in the commercial biopesticides “Vivus” and “Vivus Max” (AgBiTech Pty Ltd., Brisbane, Queensland, Australia). The evolutionary relationships of this isolate on a global and local level, the development of new analytical approaches for NGS data and phenotypic and genotypic analysis of trait-specific strains are discussed in this thesis.

1.2 RESEARCH OBJECTIVES

Three research objectives were identified to achieve the research aim and described as follows:

1. Apply NGS and develop bioinformatic techniques to assemble whole-genome sequences and develop and analyse custom meta-barcodes to quantify and describe the strain diversity and abundance within isolates.
    a. Identify suitable bioinformatic software and extend existing techniques for non-model systems.
    b. Develop custom scripts which automate the whole-genome assembly, within-isolate and within-strain community composition process.
2. Apply *in vivo* and *in vitro* selection to derive strains from the wild-type isolate and identify, characterise and quantify strain variation resulting from:
    a. *In vitro* culture and plaque purification.
    b. *In vivo* infection
    c. Repeated *in vivo* passage under selection pressure for three characteristics:

i. Fast speed of kill
ii. Slow speed of kill
iii. Maximum virus occlusion body (OB) yield

d. Apply NGS and bioinformatics (from objective 1) to construct, annotate and analyse the derived strains genomes.
e. Describe evolutionary relationships between isolates and strains, including virulence-transmission trade-offs and if the viral quasispecies model could be extended to include baculoviruses.

3. Characterise the biology of the strains selected by *in vivo* passage.
   i. Virus OB yield
   ii. Speed of kill
   iii. Percentage kill

1.3 THESIS OUTLINE

The thesis is divided into a literature review providing context, background research and knowledge gaps as the basis for the research completed and experimental chapters that address each of the objectives outlined above. The concluding chapter summarises, links and discusses the findings of the study conducted.

1.3.1 Chapter 1: General Introduction

This chapter provides a brief background into the intended research and outlines research objectives, with a brief overview of the other chapters described in this thesis.

1.3.2 Chapter 2: Literature Review

The literature review is divided into three main parts; an introduction to the baculoviruses and commercial use, the dynamics, diversity and evolution of these viruses, and the introduction and limitations of NGS and bioinformatics. This chapter identifies the knowledge gaps that will be addressed in this thesis.

1.3.3 Chapter 3: Complete Genome Sequences of *Helicoverpa armigera* Single Nucleopolyhedrovirus Strains AC53 and H25EA1

Publication 1 describes the NGS and computation method used for the construction of the virus genome consensus sequence, and highlights the main features identified within the viral genomes.

1.3.4 Chapter 4: Complete Genome Sequences of Seven *Helicoverpa armigera* SNPV-AC53-Derived Strains

Publication 2 describes the assembly technique used to generate the consensus sequence of the genomes of seven strains derived from the AC53 parent isolate by passage and plaque

selection in tissue culture, describing the main features identified in the genomes including unique open reading frames (ORFs) and a new hypothetical ORF.

1.3.5 Chapter 5: Comparative analysis of AC53 and its Tissue Culture Derived Strains

Publication 3 is a detailed comparative analysis of the whole genomes of the parent isolate and derived strains identifying key ORFs that are associated with *in vitro* passage, and application of maximum likelihood estimates (MLE) to describe the systematic and geographic relationships of SNPVs isolated from *Helicoverpa* spp.

1.3.6 Chapter 6: MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data

Publication 4 describes the development and validation of a new bioinformatics pipeline to analyse and determine the relative abundance and community composition of strains within the HaSNPV-AC53 isolate and the derived strain AC53-T2. This approach identifies and applies custom DNA 'barcodes' that amplify regions within the BRO-A and DNA polymerase ORFs and use ultra-deep sequencing data to create a custom library of all possible combinations of polymorphisms within the barcode region with which to align and quantify the sequence read copy number. The chapter also describes the further development of the computational scripts and the transfer to Python.

1.3.7 Chapter 7: Time-Course Analysis of Strains with Polymorphisms in the *BRO-A* ORF During *In Vivo* Infection by the HaSNPV-AC53 isolate.

This chapter applies the MetaGaAP pipeline to quantify the relative abundance of strains by inference from variance and abundance of reads produced by amplification and ultra-deep sequencing of the BRO-A ORF 'barcode' during infection *in vivo* by the HaSNPV-AC53 isolate. Hierarchical clustering, heat maps, and linear regression were used to determine the statistical significance of the change in BRO-A sequence read abundance during the infection cycle. Evolutionary statistics (Tajima's D and Fay and Wu's H) and mean relative evolutionary rate were used to describe the evolutionary relationships of strains within the AC53 isolate population based on sequence polymorphisms in the BRO-A ORF.

1.3.8 Chapter 8: Strain Selection & Trade Offs Between Virulence and Transmission Under Selection Pressure during *In Vivo* passage of the HaSNPV-AC53 isolate.

This chapter applies selection for specific traits in the HaSNPV-AC53 isolate during serial passage *in vivo*. The derived strains were characterised, quantifying traits including viral yield and speed of kill to evaluate the trade-off between virulence and virus transmission. General linear models and estimation of relative potency using Fieller's theorem were applied to determine the mean or median values and their statistical significance.

1.3.9 Chapter 9: Genetic Analysis of Trait-Specific *In Vivo* Derived Strains from HaSNPV-AC53

This chapter further develops the results from chapter 8 and applies shotgun sequencing to sequence the genomes of these derived viruses. The derived strains were analysed to identify trait-specific polymorphisms, mutations within ORFs, estimation of divergence time and clustering of polymorphisms to identify any ecological effects occurring. In addition, the concept of 'viral quasispecies' to describe the evolutionary and ecological effects and virulence-transmission trade-offs which have occurred within these viruses is discussed.

1.3.10 Chapter 10: Conclusions and Future Work

This chapter concludes the thesis and ties all experimental chapters together and discusses the biological and commercial implications of the study taken and future work.

Chapter 2: Literature Review

2.1 ABSTRACT

Baculoviruses (family: *Baculoviridae*) are double-stranded DNA viruses specific to invertebrates that have been commercialised as biopesticides against Lepidopteran pests. They contain a population of genotypes with differing phenotypic properties within a single isolate. Studies on the diversity of strains within isolates and their biology, ecology and interactions with the host have used strains derived from isolates by *in vitro* and *in vivo* selection, and include modelling of trade-offs between virulence and transmission, but work on the effects of selection pressure, such as selection for increased speed of kill, and the resultant changes in phenotypic and genotypic traits within isolates have been relatively little investigated The development of technology in bioinformatics and 'next generation sequencing' (NGS) has created new opportunities to increase our understanding of fundamental questions in the dynamics, diversity and evolution of these viruses. However, the baculoviruses are both double-stranded DNA viruses and pathogens of invertebrates, and thus are a 'non-model' organism, and although they may provide insights into some aspects of virology, there has been comparatively little investment to date to analyse strain variation within isolates, and the technology available may be either impractical, have high error rates or are not yet available This review describes the current literature on baculovirus, next generation sequencing (NGS) and bioinformatics approaches to genome assembly and 'metabarcode' analysis of communities, as well as the limitations of current approaches that this thesis addresses.

2.2 AN INTRODUCTION TO THE BACULOVIRUSES

2.2.1 Baculovirus Overview

Baculoviruses (family: *Baculoviridae*) are a family of invertebrate-specific viruses consisting of four genera and 66 species (Table 2 - 1) commonly used as biopesticides (G. F. Rohrmann, 2013c; "Virus Taxonomy: 2016 Release," 2016). Each baculovirus contains a circular dsDNA genome between 80 and 180 kbp in length, encoding between 90 and 180 genes of which 38 are core-genes common with all baculovirus species (Javed et al., 2017; McCarthy & Theilmann, 2008; Miele, Garavaglia, Belaich, & Ghiringhelli, 2011a).

The baculoviruses are categorised as either of two morphotypes, grouped into the four genera (Figure 2-1, Table 2-1); the granuloviruses (GVs) and the nucleopolyhedroviruses (NPVs). GVs contain a singly enveloped nucleocapsid occluded in the protein 'granulin, whereas NPVs contain singly- (SNPV) or multiply- (MNPV) enveloped nucleocapsids occluded in the

protein 'polyhedrin' (G. F. Rohrmann, 2013c, 2013d). The *Alphabaculovirus* NPVs are divided into an additional two types, group I and group II which are classed as either having the GP64 fusion protein and an additional 11 genes of various functions (group I) or the fusion (F) protein (group II) (Miele, Garavaglia, Belaich, & Ghiringhelli, 2011b; M. N. Pearson & Rohrmann, 2002; G. F. Rohrmann, 2013c, 2013d). Monsma *et al.* (1996) observed that deletion of the *gp64* gene renders the virus incapable of budding and infecting surrounding cells. F protein uses a different receptor to GP64 and does not aid entry into surrounding cells, however, has been shown to increase the infectivity factor of the budded virus (Westenberg, Uijtdewilligen, & Vlak, 2007).

Table 2-1: Genera of the baculoviruses.

Genus	Members	Total Species
Alphabaculovirus	Lepidopteran-specific NPVs	40
Betabaculovirus	Lepidopteran-specific GVs	23
Gammabaculovirus	Hymenopteran-specific NPVs	2
Deltabaculovirus	Dipteran-specific NPVs	1

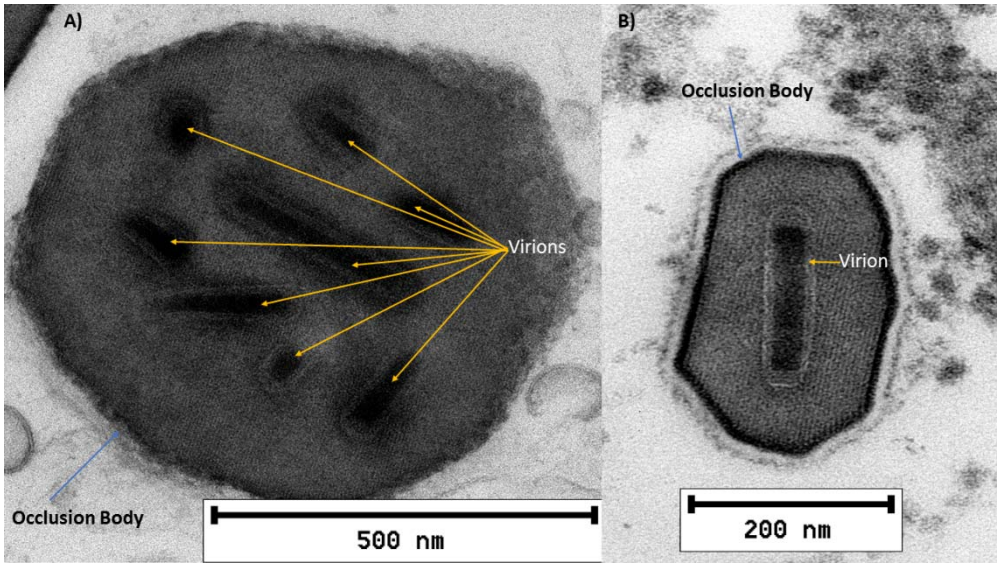


Figure 2-1: Transmission electron microscopy of the HaSNPV-AC53 isolate (Noune & Hauxwell, 2015) and the PxGV-C isolate (Spence, Noune, & Hauxwell, 2016) highlighting structural differences between NPVs and GVs.

Two distinct life-history stages have been exhibited by the baculoviruses during the infection cycle (Figure 2-2 and Figure 2-3). The first stage involves the occlusion derived virus (ODV), in which either a single (GV) or several thousand singly-enveloped or multiply-enveloped virions are embedded in protein occlusion bodies (OB), and responsible for horizontal (environmental) transmission and the primary infection (Blissard & Rohrmann, 1990; G. F.

Rohrmann, 2013c). The OB is broken down in the alkaline midgut lumen which releases the virions that fuse with the midgut epithelial cells and replication is initiated (Krell, 2008; G. F. Rohrmann, 2013d). A single OB may contain multiple genetically-related virus variants co-occluded within a single, protein body with differing phenotypic properties reducing the chance of insect resistance (Vicky Lynne Baillie & Bouwer, 2012b; Blissard & Rohrmann, 1990; Chateigner et al., 2015; Cory, Green, Paul, & Hunter-Fujita, 2005; Goulson & Hauxwell, 1995; Noune & Hauxwell, 2016a; Ogembo et al., 2007; Elizabeth M. Redman, Wilson, Grzywacz, & Cory, 2010; Reeson, Wilson, Gunn, Hails, & Goulson, 1998).

This is followed by the second stage involving the budded virus (BV). Single virions are produced in an infected cell and acquire a membrane from the basal side of the epithelial cell before exiting into the hemolymph for *in vivo* transmission (Krell, 2008). Subsequent infection in other cells requires fusion with the cell membrane mediated by one of the two previously mentioned 'fusion proteins' (Krell, 2008; G. F. Rohrmann, 2013c, 2013d).

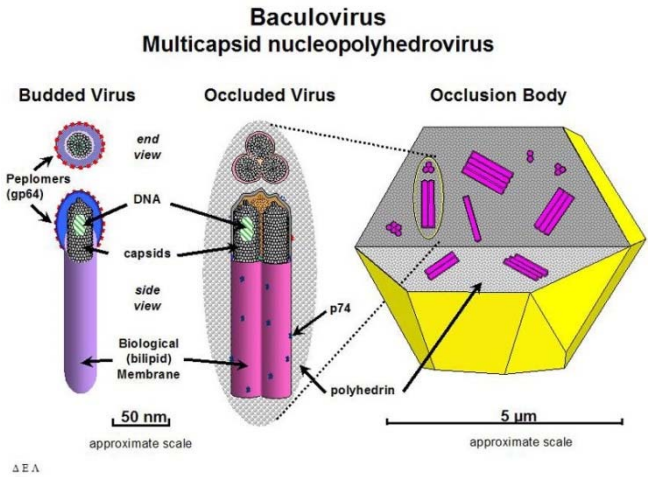


Figure 2-2: Structure and key features of an NPV showing the differences between the BV and occluded virus (Lynn, 2006).

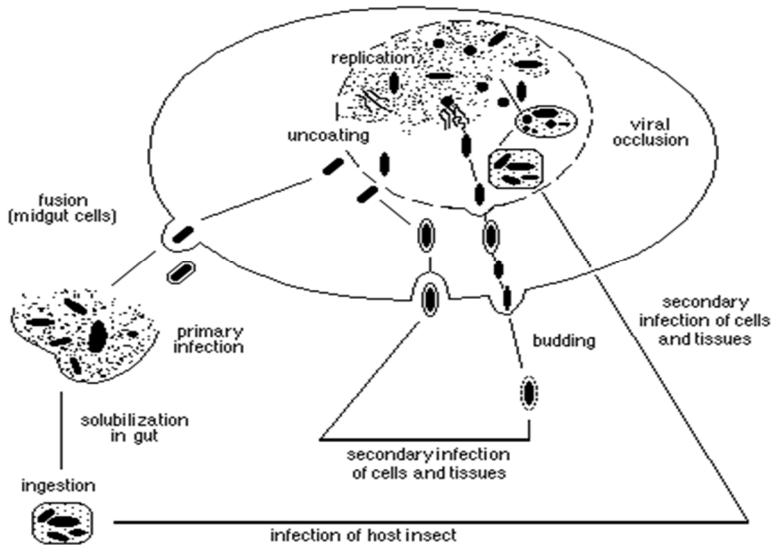


Figure 2-3: Overview of the baculovirus infection cycle (Murphy & Piwnica-Worms, 2001).

2.2.2 Biopesticide Use

Insecticide resistance and demands to reduce health and environmental impacts of insecticides has led to the advancement of biopesticide development (Buerger et al., 2007). The evolution of resistance to broad-range chemical insecticides and the negative effects of these chemicals to beneficial insects such as parasitoids (e.g. *Trichogramma* spp., *Microplitis* spp.) has given rise to the development and application of biopesticides such as the baculoviruses and the *Bacillus thuringiensis* toxin (G. Fitt, Cotter, & Sharma, 2005; Gary P. Fitt, 2000; G.P. Fitt, 2003; Wilson, Mensah, & Fitt, 2004).

In many cases biopesticides based on invertebrate pathogens are selected on the basis of possessing a narrow host range, therefore reducing or eliminating impact on non-target invertebrates and vertebrates (Bonning & Nusawardani, 2007; Copping & Menn, 2000; Gary P. Fitt, 2000; G.P. Fitt, 2003; Hunter-Fujita, Entwistle, Evans, & Crook, 1998; Wilson, Mensah, & Fitt, 2004). However, the high cost to manufacture and slower field efficacy compared to synthetic chemical insecticides has previously seen biopesticides viewed with some scepticism (Buerger et al., 2007; Burges, 1998; Huang, Hu, Pray, Qiao, & Rozelle, 2003; G. Zhang, 1994; GY Zhang & Bai, 1992).

An example of this scepticism had been observed in Australia during the 1970s as an example of growers unwilling to change perception due to the costs associated with the introduction of a baculovirus-based biopesticide derived from the *Helicoverpa zea* SNPV (Buerger et al., 2007). Essentially, the viewed scepticism led to the continued use of synthetic pyrethroids and ultimately country wide pyrethroid resistant insect populations (Daly & Murrary, 1988). In recent times the scepticism surrounding the use of biopesticides has lessened as manufacturing improvements allowed for more cost-effective production and greater control over quality, quantity and strain enhancements, establishing a credible commercial product (Moscardi, 1999). This led to decreased manufacturing costs, safer application and successful integration into pest management systems, and the rise in commercial use of insect pathogens as a means of insect control (Buerger et al., 2007; Erlandson, 2008; Gary P. Fitt, 2000; G.P. Fitt, 2003; Hauxwell, 2008a; Wilson et al., 2004).

In Australia, *Helicoverpa armigera* is a polyphagous pest, resistant to most chemical insecticides and costs an estimated AUD$562/ha to control (G. Fitt, Cotter, & Sharma, 2005). The Queensland Department of Agriculture and Fisheries (DAF) conducted the initial research and isolation of many native baculoviruses as an alternative to control *H. armigera* and *H. punctigera* (Buerger et al., 2007; Hauxwell, 2008a). This led to the eventual application of the most commonly used baculovirus-based biopesticide in Australia, the *Helicoverpa armigera* SNPV isolate AC53 (HaSNPV-AC53) and is commercially produced by AgBiTech Pty. Ltd and

marketed as 'Vivus Gold' and 'Vivus Max' (Buerger et al., 2007). As of 2013, the 'Vivus' brand has been further deployed in Brazil and the USA (AgBiTech, 2013, 2014) and in China, an indigenous isolate of HaSNPV is used to control both *H. armigera* and *H. assulta* (GY Zhang & Bai, 1992). Globally, the improvements have driven baculovirus-based biopesticide commercial applications to agro-economically important pests such as *Cydia pomonella* (codling moth), *Mamestra brassicae* (cabbage moth) and *Autographica californica* (alfalfa looper) (Burand, Nakai, & Smith, 2009).

2.3 DYNAMICS, DIVERSITY AND EVOLUTION

2.3.1 Dynamics and Diversity

The baculoviruses, as previously mentioned, are a highly diverse viral family consisting of 66 known species, countless isolates and contain a set of 38 core genes regardless of genus or type (Garavaglia, Miele, Iserte, Belaich, & Ghiringhelli, 2012; E.A. Herniou & Jehle, 2007; Elisabeth A Herniou, Olszewski, Cory, & O'Reilly, 2003; Javed et al., 2017; "Virus Taxonomy: 2016 Release," 2016). Furthermore, some of these genes are shared with the related viral families *Nudivirdae, Polydnaviridae*, *Ascoviridae*, *Iridoviridae* and *Nimaviridae* (Bideshi, Renault, Stasiak, Federici, & Bigot, 2003; Thézé, Bézier, Periquet, Drezen, & Herniou, 2011; Y.-j. Wang, Burand, & Jehle, 2007; Y. Wang & Jehle, 2009).

NPV isolates from *Helicoverpa* spp., for example, show high levels of genotypic (between 94% and 99% nucleotide homology) and phenotypic diversity between isolates at a local and global level (Figure 2-4), however were previously classed as separate species until the recent reclassification (Noune & Hauxwell, 2016a; Rowley et al., 2011; "Virus Taxonomy: 2016 Release," 2016), and contain multiple strains with differing genotypic and phenotypic diversity within a single host or OB (Chateigner et al., 2015; Cory et al., 2005; Noune & Hauxwell, 2016a; G. F. Rohrmann, 2013c). It's been suggested that the high levels of genotypic diversity within a single baculovirus isolate may allow them to overcome the evolution of immunity by the insect host (G. Clavijo, Williams, Muñoz, Caballero, & López-Ferber, 2010; Robert L. Harrison, 2009b; Noune & Hauxwell, 2016a; Rowley et al., 2011).

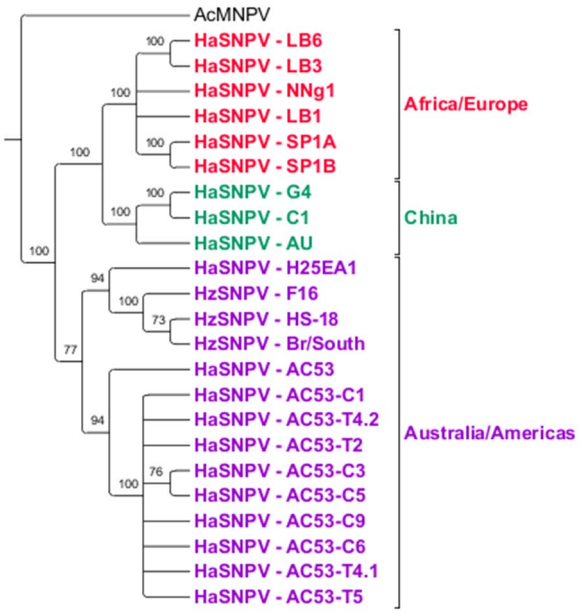


Figure 2-4: Phylogenetic relationships and global distribution of *Helicoverpa* spp. SNPV isolates (with bootstrap support as a percentage) and rooted to *Autographica californica* MNPV (AcMNPV) (Noune & Hauxwell, 2016a).

Many phenotypic variations have been observed across and within species such as field efficacy, loss of core-genes for improved environmental adaptation, speed of kill and OB production (Cory et al., 2005; Fleming-Davies et al., 2015; Lua, Pedrini, Reid, Robertson, & Tribe, 2002; Elizabeth M Redman et al., 2016; Spence et al., 2016). The impact of these phenotypic variations on subsequent vertical and horizontal transmission (within hosts and between hosts) are described within the virulence-transmission trade-off theory (Bull & Lauring, 2014; Fleming-Davies et al., 2015; van Baalen & Sabelis, 1995; White et al., 2012).

Virulence-transmission trade-off theory is a concept that suggests that as virulence (overall pathogenicity or speed of kill) increases, then transmission of the virus decreases, in the case of baculoviruses through reduced horizontal transmission (Bull & Lauring, 2014; Fleming-Davies et al., 2015; van Baalen & Sabelis, 1995; White et al., 2012). This is evident in baculoviruses where previous studies have shown that selection for variants with a fast speed of kill have reduced OB production which limits horizontal transmission and vice-versa (Elizabeth M Redman et al., 2016; White et al., 2012). Fast speed of kill variants suffer from a loss of fitness as distribution is limited and are unable to continuously infect new hosts, thus trade-offs are needed to achieve balance between virulence and transmission (Bull & Lauring, 2014; Fleming-Davies et al., 2015; van Baalen & Sabelis, 1995; White et al., 2012).

Artificial selection of viral variants using transmission bottlenecks by plaque-purification (*in vitro*) or through application of environmental pressures (*in vivo*) have been used to identify variants within NPV and GV strains and exploit virulence-transmission trade-offs for both commercial and research purposes (Arrizubieta, Simón, Williams, & Caballero, 2015b; Corsaro & Fraser, 1987; Cory et al., 2005; David J. Hodgson, Vanbergen, Hartley, Hails, & Cory, 2002;

Nguyen et al., 2011; I. R. Smith & Crook, 1988). An early example of this can be seen with the application of a low mortality dose infection for the *in vivo* isolation of *Pieris rapae* GV and *Lymantria dispar* MNPV strains, in which the first direct evidence of the independent action hypothesis for microbial pathogenicity was demonstrated (Meynell & Stocker, 1957; I. R. Smith & Crook, 1988).

Applying artificial selection relaxes the fitness cost of transmission which can be used to select for variants with high virulence, or select for slow variants with high OB production which can be co-occluded (combining variant formulations) to produce a viral formulation with improved insecticidal characteristics (Arrizubieta et al., 2015b; Hamblin, Van Beek, Hughes, & Wood, 1990).

This has been previously observed with a group of Iberian HaSNPV isolates in which two genotypic clusters were isolated using either plaque-purification or end-point dilution, with differing virulence-transmission trade-offs observed (Arrizubieta et al., 2015b). Individual genotypes with desirable traits can be co-occluded to produce binary mixtures with improved insecticidal properties, however, co-occluded mixtures require high lethal dosages to be maintained otherwise variants will begin to outcompete each other with a single variant eventually becoming the most prevalent (Arrizubieta et al., 2015b; Hamblin et al., 1990). Furthermore, low lethal dosages can lead to the establishment of a latent or covert, sub-lethal infections in an insect population in which virus is transmitted from parent to offspring (J. Fuxa, Weidner, & Richter, 1992; J. R. Fuxa & Richter, 1992).

Covert, sub-lethal infections affect the insect present's maladaptive traits such as a shorter adult life span, reduced number of eggs produced and reduced egg viability (J. Fuxa & Richter, 1989). It has been hypothesised that the apparently resistant insects generated in these experiments may carry covert, sub-lethal infections that interfere with subsequent infections, simulating 'resistance' (Goulson & Hauxwell, 1995). Analysis of *Lymantria dispar, Mamestra brassicae* and *Spodoptera litura* have shown that 'baculovirus resistant' populations are infected with a covert, sub-lethal virus and suffer from the effects as previously described (Burden et al., 2003; Goulson & Cory, 1995; D. S. Hughes, Possee, & King, 1993, 1997; John Kuzio et al., 1999; Vasconcelos, Cory, Speight, & Williams, 2002). It has been suggested that this mechanism of infection is vital in long-term persistence of baculoviruses (Goulson & Cory, 1995; Monobrullah & Shankar, 2008; Myers, Malakar, & Cory, 2000).

Furthermore, covert infections are reported to cause low levels of mortality in offspring of the infected parents, with the highest frequency of mortality observed at the second instar (Goulson & Cory, 1995). An example of this was observed with *Plodia interpunctella* subjected to a sub-lethal infection with the pest exhibiting reduced reproduction of eggs and a stage-dependent infection at the 4^{th} and 5^{th} instars in which development time is significantly longer (Sait, Begon, & Thompson, 1994). It's been suggested that covert infections may be activated by

stress-related factors such as temperature, poor diet and overcrowding, or by subsequent infection of the insect with an alternative baculovirus (D. S. Hughes et al., 1993; Longworth & Cunningham, 1968; K. M. Smith, 2012).

In some instances, the insect does not present any symptoms during covert infection and can only be detected by molecular techniques. An example of this was observed with *Mamestra brassicae*, which was infected by an MNPV resulting in the vertical transmission of the virus from generation to generation, presenting no symptoms, and identified throughout the entire insect life cycle (D. S. Hughes et al., 1993). The viral DNA observed in this example was of similar nature to the initial MNPV stock but not identical and infection was only activated when *M. brassicae* larvae were fed *Panolis flammea* NPV or *Autographica californica* NPV (D. S. Hughes et al., 1993).

Wild insect populations frequently contain baculovirus covert infections, and has been suggested that all baculoviruses adopt a strategy to covertly infect insects (Burden et al., 2003).

2.3.2 Evolution

The exact origins of the baculoviruses have not been determined, but evidence suggests that they co-evolved with their host insects (Elisabeth A Herniou, Olszewski, O'reilly, & Cory, 2004; Ikeda, Hamajima, & Kobayashi, 2015; GF Rohrmann, Pearson, Bailey, Becker, & Beaudreau, 1981; G. F. Rohrmann, 1992). It's hypothesised that baculoviruses co-evolved from a common ancestor shared with the nudiviruses ~310 million years ago (with the first insects), and recent studies have suggested diversification of these viruses occurred during the diversification of the Class Insecta (Bézier et al., 2009; Thézé et al., 2011). Extrapolated phylogenetic studies have suggested that the *Culex nigripalus* NPV from the genus *Deltabaculovirus* is the most ancient lineage of baculoviruses (Garcia-Maruniak et al., 2004; E.A. Herniou & Jehle, 2007; Elisabeth A Herniou et al., 2004).

Selection pressure, genetic bottlenecks, and highly mutagenic environments all drive the evolution of baculovirus (Marie Berling et al., 2009; Bull & Lauring, 2014; Burke, 2012; Gabriel Clavijo, Williams, Muñoz, López-Ferber, & Caballero, 2009; E.A. Herniou & Jehle, 2007; Elisabeth A Herniou et al., 2003; Ikeda et al., 2015; G. F. Rohrmann, 2013c; Zhou et al., 2011). An example of this has been shown with the SNPVs from *Helicoverpa* spp. where ORFs that were originally thought to only occur in *H. zea* SNPV variants were identified intact or as fragments in all *Helicoverpa* spp. SNPVs (Noune & Hauxwell, 2016a). Furthermore, phylogenetic analysis of whole-genomes and conserved regions, suggest that *Helicoverpa* spp. SNPVs potentially originated in Australia prior to global distribution via global wind-patterns and insect migrations (Noune & Hauxwell, 2016a). However, mutation rates are low in DNA virus replication (1.8×10^{-8} mutations per nucleotide per genomic replication) and there are no

current estimates for the rate of genetic mutations or recombination events for baculoviruses (Duffy, Shackelton, & Holmes, 2008; E.A. Herniou & Jehle, 2007).

Mutations may arise from several different mechanisms. Previous studies have observed that naturally occurring mutations in the form of substitutions are a potentially high source of variation in non-static environments but can have harmful impacts on the virus (Duffy et al., 2008). In addition, genetically modifying virus with deletion mutations have been found to improve pathogenicity against genetically variable insect populations and research into their biology has led to the discovery of new infectivity factors (Oihane Simón, Palma, Williams, López-Ferber, & Caballero, 2012; O. Simón, Williams, Caballero, & López-Ferber, 2006; O. Simón, Williams, López-Ferber, & Caballero, 2005).

Recombination through enzyme-mediated breakage-reunion increases baculoviruses genetic diversity without the loss of viability that occurs with high mutation rates (Duffy et al., 2008; Robert L. Harrison, 2009b; Posada, Crandall, & Holmes, 2002). This allows the virus to capture host genes which may also permit the blocking or mimic of host proteins, therefore assisting in the development of persistent infections (Chaston & Lidbury, 2001). Furthermore, related baculovirus isolates and within-isolate variants have been known to co-infect a single insect cell at the same time and produce recombinant variants (Cory et al., 2005; Elisabeth A Herniou et al., 2003).

In some instances, genetic material has been incorporated into the baculovirus genomes through horizontal transfer of genetic material from insects to virus via transposon-mediated mutations (Blissard & Rohrmann, 1990; Gilbert et al., 2014). Essentially, transposons are inserted near early baculovirus gene promoters leading to high genetic variability with promoters resembling the host insect (Blissard & Rohrmann, 1990; Gilbert et al., 2014; G. F. Rohrmann, 2013b). For example, the *piggyBac* transposon, commercially used as a 'cut and paste' system for genetic engineering, was identified within an AcMNPV 'few polyhedra' mutant during viral adaptation to *Trichoplusia ni* and *Spodoptera frugiperda* cell lines (M. Fraser, Smith, & Summers, 1983; M. J. Fraser, Cary, Boonvisudhi, & Wang, 1995; X. Li et al., 2013). An estimated frequency of a single transposable element integration using an AcMNPV isolate occurred once in every ~8500 genomes (Gilbert et al., 2014).

In previous studies in which baculoviruses have been used to infect insect cell lines, it was observed that some late-expressed proteins are not essential for baculovirus replication in tissue culture (Arif, 2005; Robert L. Harrison, 2009a; O'Reilly & Miller, 1991; van Oers, Pijlman, & Vlak, 2015). Furthermore, genetic manipulation of baculoviruses has led to the improvement in host pathogenicity through the insertion of insect-specific toxins (Bruce C Black et al., 1997; Gershburg et al., 1998). An example of this genetic manipulation was with the construction of a HaSNPV isolate expressing a cathepsin L-like cysteine protease from *Sarcophaga peregrina* (Xiulian. Sun et al., 2009) and HaSNPV expressing an insect-selective

neurotoxin from *Androctonus australis* (Xiulian Sun et al., 2004). These studies noted that the use of genetically engineered forms of HaSNPV controlled insect populations with enhanced mortality rates compared to the non-recombinant wild type isolate (Xiulian Sun et al., 2004; Xiulian. Sun et al., 2009). However, genetic modification of pathogens is controversial and rigorously regulated. As an alternative strategy, improving baculovirus formulation and application technology has led to improved efficacy through increased virus ingestion by the insect (Hauxwell, 2008b; M.-L. Johnson et al., 2000).

Resistance to NPVs has not been shown to occur in the field, however resistance to GVs has been observed, with a commercial GV isolate infecting *Cydia pomonella* following regular repeated application over several years (Sabine Asser-Kaiser et al., 2007; Asser-Kaiser, Heckel, & Jehle, 2010; Marie Berling et al., 2009; M Berling et al., 2009; Eberle & Jehle, 2006). However, repeated exposure of the GV towards resistant populations (applying selection pressures) produced a GV variant which overcame resistance (Graillot et al., 2014). Evolution of resistance to NPVs may be less probable due to the presence of numerous variants within NPV isolates.

2.3.3 Introduction to Quasispecies

As discussed above, baculoviruses typically contain a population of genotypes within a single isolate and, in the cases of NPV's, may contain a population of genotypes within a single occlusion body. Each of these genotypes may have different phenotypes with different biological activity that may collectively improve the success of the infection by infecting different host tissues, and maintenance of diversity (Blissard & Rohrmann, 1990; Hails et al., 2002; Elizabeth M Redman et al., 2016; G. F. Rohrmann, 2013a, 2013c, 2013d; White et al., 2012; Zwart et al., 2009). Observations of variants within a single isolate has been extensively studied in RNA viruses and some DNA viruses, where the genotypic variants have been referred to as 'viral quasispecies' or the 'quasispecies model' (Andino & Domingo, 2015; Domingo et al., 1998; Domingo, Sheldon, & Perales, 2012; Lauring & Andino, 2010; Wilke, 2005).

A viral quasispecies is a population of viruses (or genotypes) that behave as a single species, are related by similar mutations and in which selection acts upon genotype 'clouds' (Domingo et al., 2012; Eigen, 1978; Holland, De La Torre, & Steinhauer, 1992; Lauring & Andino, 2010; Solé, Ferrer, González-García, Quer, & Domingo, 1999; Wilke, 2005). In model RNA viruses, such as human immunodeficiency virus, poliovirus and rabies, the quasispecies models of ecology and evolution have been well characterised (Arbiza, Mirazo, & Fort, 2010; Ball, Gilchrist, & Coombs, 2007; Domingo et al., 2012; Solé et al., 1999).

Ecologically, two models have been known to occur within a quasispecies: niche differentiation and competitive exclusion principal, and are both essential in modelling the

interactions of genotypes within the quasispecies (Ball et al., 2007; Hardin, 1960; Pocheville, 2015; Solé et al., 1999; Vignuzzi, Stone, Arnold, Cameron, & Andino, 2006).

Niche differentiation can be summarised as viral variants within the population partitioning host resources, with a single dominant genotype occupying the most resources (Arbiza et al., 2010; Domingo et al., 1998; Eigen & Biebricher, 1988). This leads to cooperation between genotypes and has been observed in a poliovirus model in which genotypes of differing phenotypes break down host immune responses, allowing other genotypes to infect host tissues to which they would otherwise not have had access (Vignuzzi et al., 2006).

However, when two quasispecies of equal fitness coinfect a host, the ecological model, competitive exclusion principal is observed (D. K. Clarke et al., 1994; Solé et al., 1999). An arms race begins between the two quasispecies and eventually one of the quasispecies will become extinct, or in some cases host immunity can affect a single quasispecies population through selection pressures and lead to a loss in lower fitness genotypes (Arbiza et al., 2010; D. K. Clarke et al., 1994; Solé et al., 1999; Wilke, 2005).

Quasispecies exhibit mutational robustness or 'survival of the flattest' that improves long-term viability through the maintenance of a high diversity of genotypes that are equally fit on the fitness landscape (Wilke, Wang, Ofria, Lenski, & Adami, 2001). The quasispecies model suggests that if mutation rates are high, selection will act on a group of mutants or genotypes rather than individual genotypes within a population, and this is important for long-term survivability (Crotty, Cameron, & Andino, 2001; Domingo et al., 2012; Van Nimwegen, Crutchfield, & Huynen, 1999; Wilke, 2005).

It has been hypothesised that baculoviruses exhibit characteristics of a viral quasispecies through maintaining a genotypically and phenotypically diverse population of genotypes which may cooperate during the infection cycle (Chateigner et al., 2015; Cory et al., 2005). However, additional evidence such as observing the previously described ecological and evolutionary models during the infection cycle would be needed to validate the quasispecies hypothesis in baculoviruses.

2.4 ANALYSIS TECHNIQUES, APPLICATION OF BIOINFORMATICS AND LIMITATIONS

2.4.1 Conventional Techniques and the Introduction of Next Generation Sequencing

Sequencing of the human genome through the human genome project took approximately 13 years and cost roughly between $USD 125 million to $USD 250 million per year to complete (Collins et al., 1998). The high cost and time required highlighted two predominant issues, the high cost of low-throughput Sanger sequencing and hands-on labour requirements to clone DNA into plasmids. It was because of these two reasons, the project began

the development of high-throughput sequencing (HTS) technologies to reduce the cost and time associated with sequencing and to accelerate genomic research (Collins et al., 1998). This lead to the invention of second-generation sequencing or its more commonly known name, 'Next Generation Sequencing' (NGS), producing a technology which has advanced the genomic characterisation of all organisms including baculoviruses (X. Chen et al., 2001; Noune & Hauxwell, 2016a; Oulas et al., 2015; Xia et al., 2012). NGS is a non-Sanger based, HTS method that generates millions of short-fragment sequences between 36 bp to 1 Kbp (reads) that can be used to assemble full genomes, and, particularly in microbiology, to identify and describe community composition or 'metagenomes' (Capobianchi, Giombini, & Rozera, 2013; Liu et al., 2012; Oulas et al., 2015). The fast run times, low error rates and relatively cheap cost associated has seen NGS become the dominant technology for genomic analysis (Goodwin, McPherson, & McCombie, 2016; Liu et al., 2012).

The issues highlighted with the human genome project were also evident when sequencing baculoviruses (Afonso et al., 2001). Previous molecular characterisation of diversity within baculovirus populations was completed using either denaturing gradient gel electrophoresis (DGGE) or restriction fragment length polymorphism (RFLP) to identify genotypes (V.L. Baillie & Bouwer, 2011; Nealis, Turnquist, Morin, Graham, & Lucarotti, 2015). Both techniques have significant drawbacks including poor gel resolution, loss of genetic material during the excision and purification of PCR products after gel electrophoreses, primer bias and the lack of genotype abundance quantification (Brooks et al., 2015; Neilson et al., 2013). Since the introduction of NGS, baculovirus research has benefited from both genome assembly and analysis of strain variation with improved accuracy (Arrizubieta, Simón, Williams, & Caballero, 2015a; Chateigner et al., 2015; X. Chen et al., 2001; Gomi, Majima, & Maeda, 1999; Hayakawa et al., 1999; Noune & Hauxwell, 2015, 2016a, 2016b; Spence et al., 2016).

All NGS platforms (Table 2-2) are based on modified implementations of sequencing by synthesis (SBS), with three SBS implementations currently dominating the market: reversible dye-terminator sequencing (MiSeq, NextSeq, HiSeq - Illumina), phosphate release - pyrosequencing (GS FLX – Roche 454) and hydrogen ion semiconductor sequencing (Ion Torrent/Ion S5 - ThermoFisher) (Liu et al., 2012; Quail et al., 2012). Baculovirus research has predominantly used Illumina and 454 platforms however the Ion Torrent personal genome machine (PGM) has been applied to baculovirus research as a cost-effective alternative (Maghodia, Jarvis, & Geisler, 2014; Noune & Hauxwell, 2015).

More recently, single-molecule sequencing (third-generation sequencing) such as the PacBio RSII and Sequal, and Oxford Nanopore have been developed, which have increased read lengths (> ~10 kbp reads), but are quite costly, low throughput and have significant errors (> ~87% error per base) (Chin et al., 2013; Jain, Olsen, Paten, & Akeson, 2016; Quail et al., 2012). The short-read length generated by most second-generation platforms has been partially

overcome with third-generation techniques (Bleidorn, 2016) but until accuracy and throughput increases and costs decrease, second-generation techniques will continue to be widely adopted (Bleidorn, 2016; Chin et al., 2013; Jain et al., 2016; Quail et al., 2012). This thesis and literature review will focus on the second-generation sequencing platforms.

Table 2-2: A comparison of some of the most commonly used NGS platforms (Goodwin et al., 2016; Quail et al., 2012; van Dijk, Auger, Jaszczyszyn, & Thermes, 2014).

Platform	Read Length	Accuracy	Output	Run Time	Sequencing Technique	Advantages	Disadvantages	Manufacturer
MiSeq	Single Read: 50 bp – 300 bp Paired Read: 2x50 bp – 2x300 bp	~99.9%	540 Mb – 15 Gb	~4 hrs – 56 hrs	4-colour Reversible Dye Terminator	High accuracy, fast library preparation, paired reads	Lower throughput than a HiSeq, read quality drops near end of the read	Illumina
NextSeq	Single Read: 75 bp – 150 bp Paired Read: 2x75 bp – 2x150 bp		16.25 Gb – 120 Gb	~11 hrs – 29 hrs	2-colour Reversible Dye Terminator		As above, in addition to lowest accuracy out of all Illumina platforms due to 2-colour chemistry	
HiSeq	Single Read: 36 bp Paired Read: 2x50 bp – 2x 250 bp		9 Gb – 1 Tb	~7 hrs – 11 days	4-colour Reversible Dye Terminator		Read quality drops near end of the read	
Ion Torrent PGM	Single Read: 200 bp – 500 bp	~99%	30 Mb – 1 Gb	2.3 hrs – 7.3 hrs	Detection of Hydrogen ion release using a semiconductor	Fast sequencing runs, longer reads than Illumina platforms	Homopolymer stretches, lower accuracy than Illumina	ThermoFisher
Ion S5 & S5 XL	Single Read: 200 bp – 600 bp		600 Mb – 15 Gb	2.5 hrs – 4 hrs				
GS FLX	Single Read: Up to 1 Kbp		700 Mb	23 hrs	Detection of Phosphate release using chemiluminescence	Long reads, fast		Roche 454

2.4.2 Techniques, Algorithms and Applications of NGS

NGS has been used for two main analyses of baculoviruses: shotgun sequencing of the whole genome of isolates and 'ultra-deep' sequencing of amplicons (Chateigner et al., 2015; X. Chen et al., 2001; Gilbert et al., 2014; Noune & Hauxwell, 2017a, 2017b). Both techniques have advantages and disadvantages, but are primarily limited by the technologies utilised by current second-generation sequencing errors (Gomi et al., 1999; Goodwin et al., 2016; Hayakawa et al., 1999; Kõljalg et al., 2013; Quail et al., 2012; Tedersoo et al., 2015).

Shotgun sequencing is a limited bias approach in which whole genomes or metagenomic samples are sheared (mechanically or using enzymes) into fragments that are the approximate length of the intended sequencing platform read length, usually between 36 bp to 1 Kbp (Table 2-2) (Anderson, 1981). Shotgun sequencing is intended to provide a relatively high-resolution map of a whole genome or metagenome (the genomes of a mixed population), however several limitations have been identified (Gardner et al., 1981; Nayfach et al., 2015; F. Sun & Xia, 2015). This includes a loss in DNA products during purification in which DNA may be discarded during purification either through limitations with the kit used or human error, resulting in a misrepresentation of some species in a metapopulation. Furthermore, PCR bias through ligation of adapters onto sheared DNA can be introduced during library preparation which may over-represent species within a population (Jones et al., 2015).

Amplicon sequencing is a PCR-based approach that can be used to provide a high-resolution 'snapshot' of a population by amplifying a short 'barcode' region of taxonomic significance such as the 16S ribosomal RNA (16S rRNA) sub-unit in bacteria, the internal transcribed spacer (ITS) in fungi, or cytochrome oxidase (COI) in animals (Brittnacher et al., 2016; Janssen, 2006; Kõljalg et al., 2013; Oulas et al., 2015; Sanschagrin & Yergeau, 2014; D. W. Yu et al., 2012). In designing a 'barcode', it recommended that the target region should be encompassed by the span of a single NGS 'read' to avoid the need for shearing of the amplicon DNA or use of assembly algorithms on the data (Noune & Hauxwell, 2017b). In amplicon sequencing bias, may still be introduced during amplification by primer bias, which can reduce the number constituents of the population that are amplified and sequenced (Brooks et al., 2015; L. J. Clarke, Soubrier, Weyrich, & Cooper, 2014; P. L. Johnson & Slatkin, 2008; Salipante et al., 2014; Sipos et al., 2010; Tedersoo et al., 2015).

Advances in DNA sequencing as well as the need to analyse and assemble the read data into contigs and genomes (Figure 2-5) has given rise to interdisciplinary research in bioinformatics, which merges computer, statistical and biological sciences (Moody, 2004). Algorithms and analysis pipelines have been applied to assemble genomes from whole-genome shotgun sequencing, identify polymorphisms, model gene expression and predict protein structures and roles (McElroy et al., 2014; Moody, 2004; Oulas et al., 2015).

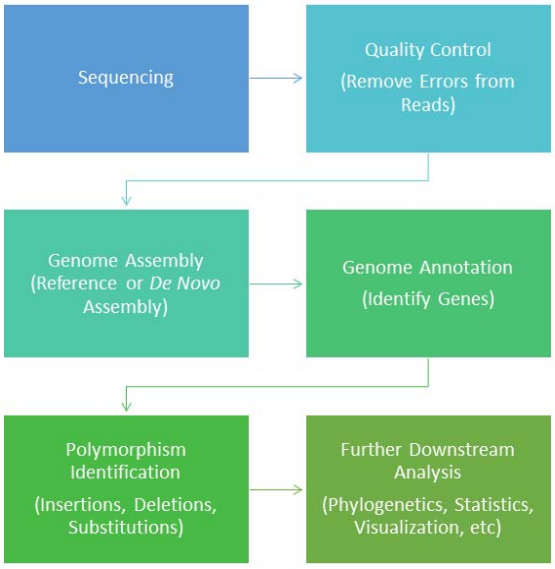


Figure 2-5: Typical steps involved with analysis of NGS datasets.

Quality Control

As discussed (Table 2-2), each sequencing platform has some form of error introduced to the read which can alter results, such as incorrect identification of polymorphisms or homopolymers included in a finished genome (Endrullat, Glökler, Franke, & Frohme, 2016) and low quality bases near the end of a read in Illumina datasets (Quail et al., 2012; Schirmer, D'Amore, Ijaz, Hall, & Quince, 2016; Schirmer et al., 2015).

Quality control of datasets is routinely performed using software packages such as FastQC, a visualisation tool that provides a detail report of data quality prior to algorithmic analysis of datasets (Andrews, 2010). Tools such as Fastx-Toolkit are applied to filter low quality reads and trim reads to expected sizes and remove low quality bases, barcodes and primer sequences from either end of the read (Gordon & Hannon, 2010).

Reference Assembly verses De Novo Assembly

Whole or partial genome assembly from short-reads is computationally demanding and requires efficient algorithms which can accurately identify patterns within reads and form a single or multiple genome sequence (Ekblom & Wolf, 2014). Currently, there are two genome assembly approaches (Figure 2-6) reference assembly and *de novo* assembly (Ekblom & Wolf, 2014; Marchant et al., 2016). Both approaches can be applied to either a single organism or a metapopulation (Sharpton, 2014). Reference assembly requires a previously sequenced genome with high nucleotide similarity for reads to be mapped to, while *de novo* assembly is required when no reference sequence is available and identifies patterns within reads which are joined together to form a contig.

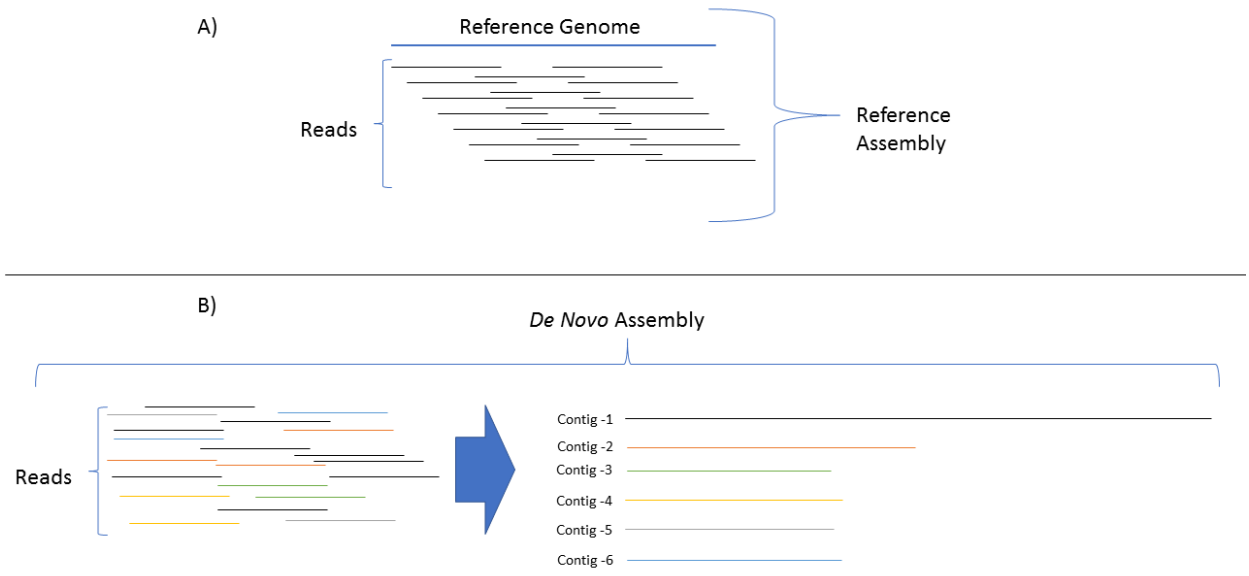


Figure 2-6: Comparison of reference assembly and *de novo* assembly. A) Reference assembly maps reads to a reference genome by identifying reads with similar nucleotides to the reference. Essentially a jigsaw puzzle. B) *De novo* assembly attempts to join reads together like a jigsaw puzzle but without a reference to compare reads to. This produces either one or more contigs (colour-coded) which are sections of one or multiple genomes and require further algorithmic techniques to form a whole genome.

Software packages commonly used for reference assembly are based upon the Burrows-Wheelers transform algorithm, which rearranges and aligns reads based on reference genome nucleotide similarity (Burrows & Wheeler, 1994; H. Li & Durbin, 2009). This is a highly efficient algorithm that can complete assembly relatively quickly, however, in some cases areas which are not covered by any reads can introduce gaps and require filling either using *de novo* assembly, resequencing the uncovered region or filling the gaps from the reference genome (Nadalin, Vezzi, & Policriti, 2012). This thesis uses a pipelines to overcome these limitations (Noune & Hauxwell, 2016a, 2016b), using a three-step assembly based on a Burrows-Wheelers transformation and *k*-mer based *de novo* assembly followed by an iterative reference assembly to produce complete genomes without gaps.

Most *de novo* assembly algorithms are either based on or both a De Bruijn graph, a directed graph representing overlaps between sequences, and *k*-mers which are short sequence fragments that can be overlapped to assemble a longer sequence (Chikhi & Medvedev, 2014; Marchant et al., 2016). However, *de novo* assembly in most cases is unable to produce a single contig representing a whole genome, especially with metapopulations, and therefore requires scaffolding (the linking of non-contiguous sequences corresponding to read overlaps (Hunt, Newbold, Berriman, & Otto, 2014)) to merge contigs together and gap filling algorithms (as above) to produce a complete genome (Fullwood, Wei, Liu, & Ruan, 2009; Nayfach et al., 2015; Sharpton, 2014; F. Sun & Xia, 2015).

Genome sequences, whether using a reference or *de novo* assembly technique, are only as good as the technology used for sequencing and can therefore still contain errors in the finished product (Beerenwinkel, Gunthard, Roth, & Metzner, 2012; Beerenwinkel, Günthard, Roth, & Metzner, 2012; Fox, Reid-Bayliss, Emond, & Loeb, 2014; Wall et al., 2014; Zook et al., 2014). It is because of these errors, that sequencing should be repeated multiple times to improve genome assembly accuracy and identification of polymorphisms within a population.

Polymorphism Identification

Identification of variants within a population or a single organism is a widely-discussed topic with multiple algorithms available that use either simple counting of alleles, probabilistic Bayesian approaches or apply a mixture of De Bruijn graphs, the Smith-Waterman algorithm for nucleotide alignment and Burrows-Wheelers transformation (McKenna et al., 2010; Nielsen, Paul, Albrechtsen, & Song, 2011; Van der Auwera et al., 2013; X. Yu & Sun, 2013).

Polymorphism identification has evolved from simple calling methods for more sophisticated techniques (Nielsen et al., 2011). Probabilistic approaches can assign quality and confidence scores for calling a polymorphism and account for errors in sequencing data, and this is particularly important with low coverage datasets as confidence can be low (Nielsen et al., 2011; X. Yu & Sun, 2013). High-coverage datasets may identify a larger number of polymorphisms with high confidence, however, depending on the NGS platform, low-frequency errors can be extrapolated leading to false positive detection (Wall et al., 2014).

Traditionally, polymorphism identification involves mapping reads to a reference sequence followed by the application of a polymorphism calling algorithm on the assembly (DePristo et al., 2011; McKenna et al., 2010; X. Yu & Sun, 2013). Recently, genotype likelihood methods have been developed which build upon probabilistic approaches (Martin et al., 2010; McKenna et al., 2010; Nielsen et al., 2011; Van der Auwera et al., 2013). This approach disregards the assembly for which it would normally call polymorphisms from and applies a mixture of different techniques and estimates error rates directly from the read data for each individual base.

An example of this is with the 'Genome Analysis Toolkit' (GATK) HaplotypeCaller algorithm (McKenna et al., 2010; Van der Auwera et al., 2013). The HaplotypeCaller essentially takes the initial assembly and re-assembles the reads using a De Bruijn graph, Burrows-Wheelers transformation and the Smith-Waterman algorithm and applies a hidden Markov model and Bayesian statistics to determine likelihoods. This allows for the accurate identification of polymorphisms by assigning a Phred-scaled likelihood to indicate the confidence and quality of each identified polymorphism. The HaplotypeCaller is considered to have the highest accuracy out of the most popular calling algorithms (Highnam et al., 2015; Pirooznia et al., 2014).

Applying NGS to Metapopulations and the Current Limitations

As previously mentioned, both amplicon and shotgun sequencing can be applied to analyse metapopulations, with amplicon sequencing being the primary technique used but both techniques have significant limitations (Brooks et al., 2015; P. L. Johnson & Slatkin, 2008; Oulas et al., 2015; Salipante et al., 2014; Sanschagrin & Yergeau, 2014; Tedersoo et al., 2015; Werner et al., 2012; D. W. Yu et al., 2012).

Shotgun sequencing contains the least amount of bias when analysing metapopulations (Tedersoo et al., 2015). However, the lower coverage obtained by sequencing the entire genome of every organism in an environmental sample can impact *de novo* assembly of whole or partial-genomes and population abundance cannot be accurately determined (Chateigner et al., 2015; Kõljalg et al., 2013; Tedersoo et al., 2015). Recently, techniques have been developed for shotgun sequencing datasets to estimate population abundance but rely on gene identification within a database for model organisms (E. Z. Chen, Bushman, & Li, 2016; Nayfach et al., 2015; Sharpton, 2014; F. Sun & Xia, 2015). In addition, these techniques focus on taxonomic classification and therefore can underestimate a population if unidentified organisms are present.

Non-model organisms do not have the luxury of reference databases for comparison to and require custom pipeline solutions that target specific organism types or sub-types such is the case with viruses and in some cases, not available for public use (McElroy et al., 2014; Zagordi, Bhattacharya, Eriksson, & Beerenwinkel, 2011). A previous study which focused on a *Autographica californica* MNPV isolate attempted to provide a solution for this and applied *k*-means clustering on the abundance of each identified polymorphism to create abundance clusters (Chateigner et al., 2015). However, this technique is only as good as the polymorphism identification algorithm applied and cannot determine the relative abundance of individual strains or abundance of strains that may contain multiple polymorphisms distributed across fragmented reads.

Amplicon sequencing of meta-barcode regions such as with 16S rRNA can provide a high-resolution snapshot of a metapopulation but introduces primer bias and is limited by the read length of the utilised NGS platform (Brooks et al., 2015; Salipante et al., 2014; Sanschagrin & Yergeau, 2014; Schloss et al., 2011; Werner et al., 2012). This is an issue with 16S rRNA datasets as specific regions are targeted instead of the entire gene as current second-generation techniques are unable to span the entire length and this leads to an underestimation of taxa (Lennon & Locey, 2016; Yarza et al., 2014). Furthermore, meta-barcoding is primarily aimed at taxonomic classification of model systems which have marker regions that can be distinguished from other taxa using databases of known taxa (Cole et al., 2013; DeSantis et al., 2006; Quast et al., 2013; Yarza et al., 2014).

Bioinformatic approaches such as the 'Quantitative Insights Into Microbial Ecology' (QIIME) pipeline has been applied to analyse 16S and ITS datasets (Caporaso et al., 2010). This

approach focuses on clustering reads at 97% sequence similarity to produce an operational taxonomic unit (OTU) and taxonomic identification and relative abundance calculation by comparison to a reference database. However, this approaches is limited by the clustering approached used, comparison to databases that contain errors, misannotated sequences, identification based on short or partial sequences and limited sequence availability for non-model organisms thus contributing to the underestimation of taxa (Ashelford, Chuzhanova, Fry, Jones, & Weightman, 2005; Clarridge, 2004; Lennon & Locey, 2016; Mignard & Flandrois, 2006; Poretsky, Rodriguez-R, Luo, Tsementzi, & Konstantinidis, 2014; Werner et al., 2012; Yarza et al., 2014).

Again, non-model organisms such as viruses, except for a few significant small RNA viruses, are unable to use these meta-barcoding approaches due to the lack of these reference databases (Chang et al., 2007; McElroy et al., 2014; Prosperi et al., 2011; Shafer, Stevenson, & Chan, 1999; Sharpton, 2014; Zagordi et al., 2011). This thesis describes the development of bioinformatic pipeline 'Meta-barcoding Genotyping and Abundance Pipeline' (MetaGaAP) that uses a meta-barcoding approach with a baculovirus as a model system (Noune & Hauxwell, 2017b). This approach requires amplicon sequencing of a region with a high density of polymorphisms that spans a single read and constructs a database containing sequences with different combinations of polymorphisms for which reads are assigned to. However, the approach is limited to regions that contain no more than 30 polymorphisms as it is computationally intensive but can be applied to both model and non-model organisms. This approach will be discussed in detail in chapter 6 of this thesis.

2.5 CONCLUSIONS

Baculovirus studies have only recently had NGS applied with techniques currently lacking the ability to appropriately handle the genetically mixed nature of baculoviruses, especially considering most NGS and bioinformatic techniques are designed for model organisms. The lack of knowledge regarding the genotypic effects when baculoviruses have had a selection pressure applied and change in baculovirus genotype abundance during the infection cycle will be addressed. Furthermore, identification of trait-specific mutations, although previously analysed with *in vitro* adapted baculoviruses, will be determined for both *in vitro* and *in vivo* selected viral strains. The hypothesis for this study predicts that trait-specific genetic mutations will be easily identifiable using an NGS approach, and a change in relative abundance of genotypes can be inferred from read copy number during the infection cycle.

Implications in the use of baculoviruses as biopesticides, particularly in registration of products and production quality control measures will be discussed. Furthermore, the NGS techniques and bioinformatic techniques applied and developed in this study may be extended to

other metapopulation and meta-barcoding studies, and used as a baseline to develop new techniques with the introduction of third-generation sequencing technologies.

Chapter 3: Complete Genome Sequences of *Helicoverpa armigera* Single Nucleopolyhedrovirus Strains AC53 and H25EA1 from Australia

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified* that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

1. **Noune, C., & Hauxwell, C. (2015). Complete Genome Sequences of Helicoverpa armigera Single Nucleopolyhedrovirus Strains AC53 and H25EA1 from Australia. Genome announcements, 3(5). doi:10.1128/genomeA.01083-15**

Contributor	Statement of contribution*
Christopher Noune Signature: Date: 27/10/17	Performed experimental design, conducted laboratory analysis, data analysis and wrote the manuscript.
Caroline Hauxwell	Contributed to experimental design, data analysis, edited and reviewed manuscript.

Principal Supervisor Confirmation		
I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
Caroline Hauxwell Name	CHauxwell Signature	27/10/17 Date

3.1 COMPLETE GENOME SEQUENCES OF *HELICOVERPA ARMIGERA* SINGLE NUCLEOPOLYHEDROVIRUS STRAINS AC53 AND H25EA1 FROM AUSTRALIA

genomeA Journals.ASM.org

Complete Genome Sequences of *Helicoverpa armigera* Single Nucleopolyhedrovirus Strains AC53 and H25EA1 from Australia

Christopher Noune, Caroline Hauxwell

School of Earth, Environmental and Biological Sciences, Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia

We report here the genome sequences of two alphabaculoviruses of *Helicoverpa* spp. from Australia: AC53, used in the biopesticides ViVUS and ViVUS Max, and H25EA1, used in *in vitro* production studies.

Received 6 August 2015 **Accepted** 11 August 2015 **Published** 24 September 2015

Citation Noune C, Hauxwell C. 2015. Complete genome sequences of *Helicoverpa armigera* single nucleopolyhedrovirus strains AC53 and H25EA1 from Australia. Genome Announc 3(5):e01083-15. doi:10.1128/genomeA.01083-15.

Address correspondence to Caroline Hauxwell, caroline.hauxwell@qut.edu.au.

Helicoverpa spp. (Lepidoptera, Noctuidae) are polyphagous pests of international significance (1). Widespread resistance to chemical insecticides has prompted the registration of biopesticides based on baculoviruses (*Baculoviridae*) (2).

Two species of group II nucleopolyhedroviruses (genus *Alphabaculovirus*) from *Helicoverpa* species have been designated *Helicoverpa armigera* single nucleopolyhedrovirus (HaSNPV) and *Helicoverpa zea* single nucleopolyhedrovirus (HzSNPV) (3–11). Strain AC53 (also known as A44WT [11–13]) is used in the biopesticides ViVUS and ViVUS Max (AgBiTech Pty. Ltd.) (2). It was originally isolated from an unspecified *Helicoverpa* species from a cadaver from Brookstead, Southeast Queensland, Australia, in 1974 (2, 11–13) and isolate P9/H25WT from an unspecified *Helicoverpa* species from a cadaver from Central Queensland in 1973 (14–19). Both isolates were passaged initially through *Helicoverpa punctigera* Wallengren and then repeatedly through *H. armigera* (Hübner) by the Queensland Department of Primary Industries (DAFF Qld); strain H25EA1, used *in vitro* baculovirus production, was selected *in vitro* by CSIRO from P9/H25WT (14–19).

AC53 (AgBiTech Pty. Ltd.) and H25EA1 (from S. Reid, University of Queensland) were passaged once through *H. armigera* larvae. Viral DNA was extracted from occlusion bodies, as previously described (7, 20), and sequenced using the Ion Torrent PGM (316 Chip, 200-bp chemistry). Read quality was determined using FastQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the qProfiler tool from the AdamaJava project (Queensland Centre for Medical Genomics) and trimmed using CLC Genomics Workbench 7.04 (CLC 7.04), with a final Phred score of 28.

AC53 contigs were assembled *de novo* using CLC 7.04 and compared with BLAST against all available *Helicoverpa* species SNPV genomes (GenBank accession numbers JN584482, NC011354, NC003349, NC003094, and NC002645). The HzSNPV (accession no. NC003349) genome was selected as a mapping reference and a consensus sequence for AC53 produced using Burrows-Wheeler aligner (BWA)-mem 0.7.5a, SAMtools 0.19, and the genome analysis toolkit GATK 3.1-1. *De novo* contigs were assembled to fill gaps (21–25). Assembly of the H25EA1 genome was conducted according to the same process for mapping to the AC53 sequence.

The AC53 and H25EA1 genomes were, 130,442 bp and 130,440 bp, with G+C contents of 39.2% and 39.1%, respectively. The homology between the strains was 99.60%. The homology to HzSNPV (accession no. NC003349) was 99.56% but ranged between 98.43% (accession no. NC003094) and 98.99% (accession no. NC011354) in comparison to HaSNPV genomes.

Both strains contain 138 open reading frames (ORFs), 5 homologous repeat (Hr) regions, and all 62 of the conserved genes were found in all lepidopteran baculoviruses (26). Of the 138 ORFs, 52 had 100% sequence homology between the two strains. The greatest differences between AC53 and H25EA1 were found in the baculovirus repeated open reading frames BRO-A (89.78% homology), and BRO-B (96.41% homology) and the 5 Hr regions (94.61% to 99.27%). However, they contained 100% homology in the BRO region located at ORF107. This is consistent with many baculoviruses (27–29).

Both isolates contained the HaSNPV ORF42 (typically located at ORF43 in HzSNPVs) (30, 31), but unlike published HaSNPV genomes, both contained the HzSNPV ORF79 (7, 28), located at ORF78.

We conclude that AC53 and H25EA1 are type II *Heliothine* SNPVs intermediate between HzSNPV and HaSNPV and support the argument that all *Heliothine* SNPVs are variants of a single species of HaSNPV (3–5, 7).

Nucleotide sequence accession numbers. The complete sequences of HaSNPV AC53 and HaSNPV H25EA1 were deposited to GenBank under the accession numbers KJ909666 and KJ922128, respectively.

ACKNOWLEDGMENTS

This work was supported by the Cotton Research Development Corporation (CRDC), the Australian Postgraduate Award (APA), AgBiTech Pty. Ltd., Steve Reid of the University of Queensland, Stephen Rudd of QFAB, Vincent Chand from the Molecular Genetics Research Facility (QUT), Peter Prentis from the School of Earth, Environmental, and Biological Sciences (QUT), and Peter Christian (CSIRO). The project work was carried out at Queensland University of Technology, Australia.

REFERENCES

1. Tay WT, Soria MF, Walsh T, Thomazoni D, Silvie P, Behere GT, Anderson C, Downes S. 2013. A brave new world for an old world pest: *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Brazil. PLoS One 8:e80134. http://dx.doi.org/10.1371/journal.pone.0080134.
2. Buerger P, Hauxwell C, Murray D. 2007. Nucleopolyhedrovirus introduction in Australia. Virol Sin 22:173–179. http://dx.doi.org/10.1007/s12250-007-0019-y.
3. Jehle JA, Lange M, Wang H, Hu Z, Wang Y, Hauschild R. 2006. Molecular identification and phylogenetic analysis of baculoviruses from *Lepidoptera*. Virol 346:180–193. http://dx.doi.org/10.1016/j.virol.2005.10.032.
4. Jehle JA, Blissard GW, Bonning BC, Cory JS, Herniou EA, Rohrmann GF, Theilmann DA, Thiem SM, Vlak JM. 2006. On the classification and nomenclature of baculoviruses: a proposal for revision. Arch Virol 151:1257–1266. http://dx.doi.org/10.1007/s00705-006-0763-6.
5. Herniou EA, Jehle JA. 2007. Baculovirus phylogeny and evolution. Curr Drug Targets 8:1043–1050. http://dx.doi.org/10.2174/138945007782151306.
6. Chen X, Zhang WJ, Wong J, Chun G, Lu A, McCutchen B, Presnail J, Herrmann R, Dolan M, Tingey S. 2002. Comparative analysis of the complete genome sequences of *Helicoverpa zeae* and *Helicoverpa armigera* single-nucleocapsid nucleopolyhedroviruses. J Gen Virol 83:673–684.
7. Rowley DL, Popham HJ, Harrison RL. 2011. Genetic variation and virulence of nucleopolyhedroviruses isolated worldwide from the *Heliothine* pests *Helicoverpa armigera*, *Helicoverpa zeae*, and *Heliothis virescens*. J Invertebr Pathol 107:112–126. http://dx.doi.org/10.1016/j.jip.2011.03.007.
8. Wardhaugh KG, Room PM, Greenup LR. 1980. The incidence of *Heliothis armigera* (Hübner) and *H. punctigera* Wallengren (Lepidoptera: Noctuidae) on cotton and other host-plants in the Namoi valley of New South Wales. Bull Entomol Res 70:113–131. http://dx.doi.org/10.1017/S0007485300009822.
9. Daly JC, Gregg P. 1985. Genetic variation in *Heliothis* in Australia: species identification and gene flow in the two pest species *H. armigera* (Hübner) and *H. punctigera* Wallengren (Lepidoptera: Noctuidae). Bull Entomol Res 75:169–184. http://dx.doi.org/10.1017/S0007485300014243.
10. Zhang G. 1989. Commercial viral insecticide *Heliothis armigera* viral insecticide in China. The IPM Practitioner 11:13.
11. Richards AR, Christian PD. 1999. A rapid bioassay screen for quantifying nucleopolyhedroviruses (*Baculoviridae*) in the environment. J Virol Methods 82:63–75. http://dx.doi.org/10.1016/S0166-0934(99)00080-4.
12. Richards A, Cory J, Speight M, Williams T. 1999. Foraging in a pathogen reservoir can lead to local host population extinction: a case study of a *Lepidoptera*-virus interaction. Oecologia 118:29–38. http://dx.doi.org/10.1007/s004420050700.
13. Christian PD, Gibb N, Kasprzak AB, Richards A. 2001. A rapid method for the identification and differentiation of *Helicoverpa* nucleopolyhedroviruses (NPV *Baculoviridae*) isolated from the environment. J Virol Methods 96:51–65. http://dx.doi.org/10.1016/S0166-0934(01)00318-4.
14. Lua LH, Reid S. 2000. Virus morphogenesis of *Helicoverpa armigera* nucleopolyhedrovirus in *Helicoverpa zeae* serum-free suspension culture. J Gen Virol 81:2531–2543.
15. Lua LH, Pedrini MR, Reid S, Robertson A, Tribe DE. 2002. Phenotypic and genotypic analysis of *Helicoverpa armigera* nucleopolyhedrovirus serially passaged in cell culture. J Gen Virol 83:945–955.
16. Pedrini MR, Christian P, Nielsen LK, Reid S, Chan LC. 2006. Importance of virus–medium interactions on the biological activity of wild-type *Heliothine* nucleopolyhedroviruses propagated via suspension insect cell cultures. J Virol Methods 136:267–272.
17. Nguyen Q, Qi YM, Wu Y, Chan LC, Nielsen LK, Reid S. 2011. *In vitro* production of *Helicoverpa baculovirus* biopesticides—automated selection of insect cell clones for manufacturing and systems biology studies. J Virol Methods 175:197–205. http://dx.doi.org/10.1016/j.jviromet.2011.05.011.
18. Nguyen Q, Nielsen LK, Reid S. 2013. Genome scale transcriptomics of baculovirus-insect interactions. Viruses 5:2721–2747. http://dx.doi.org/10.3390/v5112721.
19. Matindoost L, Nielsen LK, Reid S. 2015. Intracellular trafficking of baculovirus particles: a quantitative study of the HearNPV/HzAM1 cell and AcMNPV/Sf9 cell systems. Viruses 7:2288–2307. http://dx.doi.org/10.3390/v7052288.
20. Baillie VL, Bouwer G. 2011. Development of highly sensitive assays for detection of genetic variation in key *Helicoverpa armigera* nucleopolyhedrovirus genes. J Virol Methods 178:179–185. http://dx.doi.org/10.1016/j.jviromet.2011.09.009.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. http://dx.doi.org/10.1101/gr.107524.110.
23. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. http://dx.doi.org/10.1038/ng.806.
24. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altschuler D, Gabriel S, DePristo MA. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 11:11.10.1–11.10.33. http://dx.doi.org/10.1002/0471250953.bi1110s43.
25. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324.
26. Herniou EA, Olszewski JA, Cory JS, O'Reilly DR. 2003. The genome sequence and evolution of baculoviruses. Annu Rev Entomol 48:211–234. http://dx.doi.org/10.1146/annurev.ento.48.091801.112756.
27. Bideshi DK, Renault S, Stasiak K, Federici BA, Bigot Y. 2003. Phylogenetic analysis and possible function of bro-like genes, a multigene family widespread among large double-stranded DNA viruses of invertebrates and bacteria. J Gen Virol 84:2531–2544. http://dx.doi.org/10.1099/vir.0.19256-0.
28. Le TH, Wu T, Robertson A, Bulach D, Cowan P, Goodge K, Tribe D. 1997. Genetically variable triplet repeats in a RING-finger ORF of *Helicoverpa* species baculoviruses. Virus Res 49:67–77. http://dx.doi.org/10.1016/S0168-1702(97)01454-8.
29. Erlandson MA. 2009. Genetic variation in field populations of baculoviruses: mechanisms for generating variation and its potential role in baculovirus epizootiology. Virol Sin 24:458–469. http://dx.doi.org/10.1007/s12250-009-3052-1.
30. Chen X, IJkel WF, Tarchini R, Sun X, Sandbrink H, Wang H, Peters S, Zuidema D, Lankhorst RK, Vlak JM. 2001. The sequence of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus genome. J Gen Virol 82:241–257.
31. Chen X, Zhang W-J, Wong J, Chun G, Lu A, McCutchen B, Presnail J, Herrmann R, Dolan M, Tingey S. 2002. Comparative analysis of the complete genome sequences of *Helicoverpa zeae* and *Helicoverpa armigera* single-nucleocapsid nucleopolyhedroviruses. J Gen Virol 83:673–684.

Chapter 4: Complete Genome Sequences of Seven *Helicoverpa armigera* SNPV-AC53-Derived Strains

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified* that:

6. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
7. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
8. there are no other authors of the publication according to these criteria;
9. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
10. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

2. **Noune, C., & Hauxwell, C. (2016). Complete Genome Sequences of Seven Helicoverpa armigera SNPV-AC53-Derived Strains. Genome announcements, 4(3). doi:10.1128/genomeA.00260-16**

Contributor	Statement of contribution*
Christopher Noune Signature: Date: 27/10/17	Performed experimental design, conducted laboratory analysis, data analysis and wrote the manuscript.
Caroline Hauxwell	Contributed to experimental design, data analysis, edited and reviewed manuscript.

Principal Supervisor Confirmation		
I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
Caroline Hauxwell Name	Signature	27/10/17 Date

4.1 COMPLETE GENOME SEQUENCES OF SEVEN *HELICOVERPA ARMIGERA* SNPV-AC53 DERIVED STRAINS

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Complete Genome Sequences of Seven *Helicoverpa armigera* SNPV-AC53-Derived Strains

Christopher Noune, Caroline Hauxwell
Queensland University of Technology, Brisbane, Australia

Wild-type baculovirus isolates typically consist of multiple strains. We report the full genome sequences of seven alphabaculovirus strains derived by passage through tissue culture from *Helicoverpa armigera* SNPV-AC53 (KJ909666).

Received 25 February 2016 **Accepted** 11 March 2016 **Published** 5 May 2016
Citation Noune C, Hauxwell C. 2016. Complete genome sequences of seven *Helicoverpa armigera* SNPV-AC53-derived strains. Genome Announc 4(3):e00260-16. doi:10.1128/genomeA.00260-16.

Address correspondence to Caroline Hauxwell, caroline.hauxwell@qut.edu.au.

Wild-type baculovirus isolates typically consist of multiple strains (1). Seven strains were isolated from a single nucleopolyhedrovirus (SNPV), HaSNPV-AC53 (KJ909666) (2), using a modified tissue culture plaque assay (3, 4). Larvae of *Helicoverpa armigera* were infected with HaSNPV-AC53. Strains were isolated from hemolymph of infected larvae either by plaque purification in $HzAM_1$ cells, one passage through tissue culture, and one passage through larvae ("C-strains"), or by passage through tissue culture cells and then one passage of occlusion bodies produced through larvae, followed by plaque purification, passage in cell culture, and one passage through larvae as above ("T-strains"). Viral DNA was extracted from occlusion bodies using a Bioline Isolate II Genomic DNA kit (Bioline, USA) following published methods (2, 5, 6).

Isolated strains and HaSNPV-AC53 were prepared using a NexTera kit (Illumina, USA) and sequenced using the Illumina NextSeq 500 with 150-bp paired-end reads. Trimming was completed using the FASTX-Toolkit version 0.0.13 (7). An eight-step technique to assemble the genomes without gaps was established using a combination of open-source and commercial software. The strains were initially mapped to the HaSNPV-AC53 reference using the Burrows-Wheeler aligner "mem" algorithm (BWA-mem) version 0.7.12 (8) and converted and sorted into the BAM format using SAMtools version 1.2 (9). A gapped-consensus sequence was produced using SAMtools version 1.2, BEDtools2 (10), BCFtools (as part of SAMtools), Picard Tools version 1.140 (http://broadinstitute.github.io/picard) and the Genome Analysis Toolkit version 3.4-46 (11–13). The mapped reads were filtered using bam2fastx as part of TopHat version 2.1.0 (14) and loaded into KmerGenie version 1.6982 (15) to determine the *k*-mer size of the mapped data and then assembled *de novo* using Tadpole (BBMap 35.59 package) (16). The mapped reads, *de novo*-assembled contigs, and the consensus sequence (with gaps) were merged into a single fasta file and mapped against the HaSNPV-AC53 reference using the Geneious R9 mapper with medium-low sensitivity and 5× iterations (17). The final consensus sequence and annotations were completed using Geneious R9.

The HaSNPV-AC53 sequence produced had 100% sequence identity to the published HaSNPV-AC53 reference sequenced on the Ion Torrent PGM (2). One strain was identical in length to the parent HaSNPV-AC53 sequence (130,442 bp): HaSNPV-AC53-C5 (130,442 bp). Four strains were between 5 bp and 7 bp shorter; HaSNPV-AC53-C6 (130,435 bp), HaSNPV-AC53-C9 (130,437 bp), HaSNPV-AC53-T2 (130,437 bp), and HaSNPV-AC53-T5 (130,439 bp). Two strains, HaSNPV-AC53-C3 (130,443 bp) and HaSNPV-AC53-C1 (130,460 bp) were, respectively, 1 bp and 18 bp longer. All the strains contain the 138 open reading frames (ORFs) and 5 homologous repeat regions found within HaSNPV-AC53 (2). Comparison of strain and parent HaSNPV-AC53 sequences shows differences within HOAR, ORF5, ORF7, ORF61, BRO-A, DNA-polymerase, ORF78, 38.7-K protein, ORF128, and PKIP-1, and in all 5 homologous repeat regions. Nonsynonymous mutations were identified in ORF5 (HaSNPV-AC53-C3), BRO-A (HaSNPV-AC53-T2), and DNA-polymerase (HaSNPV-AC53-T2 and HaSNPV-AC53-C5).

Nucleotide sequence accession numbers. The complete sequences of HaSNPV-AC53C1, HaSNPV-AC53C3, HaSNPV-AC53C5, HaSNPV-AC53C6, HaSNPV-AC53C9, HaSNPV-AC53T2, and HaSNPV-AC53T5 were deposited to GenBank under the accession numbers KU738896, KU738897, KU738898, KU738899, KU738900, KU738901, and KU738904, respectively.

ACKNOWLEDGMENTS

This work was funded in part by the Cotton Research Development Corporation, the Australian Postgraduate Award, and AgBiTech Pty., Ltd. We thank the staff of the Molecular Genetics Research Facility (QUT) for their assistance with sequencing. The sequencing was carried out in the Central Analytical Research Facility (CARF) at Queensland University of Technology, Australia.

FUNDING INFORMATION

This work, including the efforts of Christopher Noune and Caroline Hauxwell, was funded by Cotton Research and Development Corporation (CRDC) (QUT1402).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. Blissard GW, Rohrmann GF. 1990. Baculovirus diversity and molecular biology. Annu Rev Entomol 35:127–155. http://dx.doi.org/10.1146/annurev.en.35.010190.001015.
2. Noune C, Hauxwell C. 2015. Complete genome sequences of *Helicoverpa armigera* single nucleopolyhedrovirus strains AC53 and H25EA1 from Australia. Genome Announc 3(5):e01083-15. http://dx.doi.org/10.1128/genomeA.01083-15.
3. Brown M, Faulkner P. 1978. Plaque assay of nuclear polyhedrosis viruses in cell culture. Appl Environ Microbiol 36:31–35.
4. BD Biosciences. Plaque assay. http://www.bdbiosciences.com/br/resources/baculovirus/protocols/plaque_assay.jsp. Accessed 16 March 2016.
5. Rowley DL, Popham HJ, Harrison RL. 2011. Genetic variation and virulence of nucleopolyhedroviruses isolated worldwide from the heliothine pests *Helicoverpa armigera*, *Helicoverpa zea*, and *Heliothis virescens*. J Invertebr Pathol 107:112–126. http://dx.doi.org/10.1016/j.jip.2011.03.007.
6. Baillie VL, Bouwer G. 2011. Development of highly sensitive assays for detection of genetic variation in key *Helicoverpa armigera* nucleopolyhedrovirus genes. J Virol Methods 178:179–185. http://dx.doi.org/10.1016/j.jviromet.2011.09.009.
7. Gordon A, Hannon G. 2010. FASTX-Toolkit: FASTQ/A short-reads preprocessing tools. http://hannonlab.cshl.edu/fastx_toolkit. Accessed 16 March 2016.
8. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997. http://arxiv.org/abs/1303.3997.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
10. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. BioInformatics 26:841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.
11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. http://dx.doi.org/10.1038/ng.806.
12. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. http://dx.doi.org/10.1101/gr.107524.110.
13. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 11:. http://dx.doi.org/10.1002/0471250953.bi1110s43.
14. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. http://dx.doi.org/10.1186/gb-2013-14-4-r36.
15. Chikhi R, Medvedev P. 2014. Informed and automated *k*-mer size selection for genome assembly. BioInformatics 30:31–37. http://dx.doi.org/10.1093/bioinformatics/btt310.
16. Bushnell B. BBMap short read aligner. https://sourceforge.net/projects/bbmap. Accessed 16 March 2016.
17. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. BioInformatics 28:1647–1649. http://dx.doi.org/10.1093/bioinformatics/bts199.

Chapter 5: Comparative Analysis of HaSNPV-AC53 and Derived Strains

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified* that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

3. **Noune, C., & Hauxwell, C. (2016). Comparative Analysis of HaSNPV-AC53 and Derived Strains. Viruses, 8(11), 280.**

Contributor	Statement of contribution*
Christopher Noune Signature:  Date: 27/10/17	Performed experimental design, conducted laboratory analysis, data analysis and wrote the manuscript.
Caroline Hauxwell	Contributed to experimental design, data analysis, edited and reviewed manuscript.

Principal Supervisor Confirmation		
I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
Caroline Hauxwell	CHauxwell	27/10/17
Name	Signature	Date

5.1 COMPARATIVE ANALYSIS OF HASNPV-AC53 AND DERIVED STRAINS

*For supplementary material refer to section 12.1

viruses

MDPI

Article

Comparative Analysis of HaSNPV-AC53 and Derived Strains

Christopher Noune and Caroline Hauxwell *

Queensland University of Technology, Brisbane 4000, Australia; chris.noune@connect.qut.edu.au
* Correspondence: caroline.hauxwell@qut.edu.au; Tel.: +61-07-3138-8062

Academic Editor: Karyn Johnson
Received: 12 September 2016; Accepted: 21 October 2016; Published: 31 October 2016

Abstract: Complete genome sequences of two Australian isolates of *H. armigera* single nucleopolyhedrovirus (HaSNPV) and nine strains isolated by plaque selection in tissue culture identified multiple polymorphisms in tissue culture-derived strains compared to the consensus sequence of the parent isolate. Nine open reading frames (ORFs) in all tissue culture-derived strains contained changes in nucleotide sequences that resulted in changes in predicted amino acid sequence compared to the parent isolate. Of these, changes in predicted amino acid sequence of six ORFs were identical in all nine derived strains. Comparison of sequences and maximum likelihood estimation (MLE) of specific ORFs and whole genome sequences were used to compare the isolates and derived strains to published sequence data from other HaSNPV isolates. The Australian isolates and derived strains had greater sequence similarity to New World SNPV isolates from *H. zea* than to Old World isolates from *H. armigera*, but with characteristics associated with both. Three distinct geographic clusters within HaSNPV genome sequences were identified: Australia/Americas, Europe/Africa/India, and China. Comparison of sequences and fragmentation of ORFs suggest that geographic movement and passage in vitro result in distinct patterns of baculovirus strain selection and evolution.

Keywords: baculovirus; SNPV; *Helicoverpa*; Next Generation Sequencing; strain selection; virus evolution

1. Introduction

Baculoviruses (family *Baculoviridae*) are double-stranded DNA (dsDNA) viruses with a genome of between 80,000 and 180,000 base pairs [1,2]. The genetic diversity of the Group II nucleopolyhedroviruses (genus *Alphabaculoviruses*) from Lepidoptera of the genus *Helicoverpa* are of importance due to their worldwide distribution and widespread use as biopesticides against these significant polyphagous pests [3]. Group II singly-enveloped nucleopolyhedroviruses from species of the genus *Helicoverpa* (Lepidoptera: Noctuidae) were originally classified into two species; Old World *H. armigera* single nucleopolyhedrovirus (HaSNPV), isolated from *H. armigera* (Hübner) and New World *H. zea* single nucleopolyhedrovirus (HzSNPV) isolated from *H. zea* (Boddie) [3–13]. This has been recently revised to classify both types as a single species, HaSNPV, with similarities in DNA sequence and biological activity [12].

Old world isolates of HaSNPV, and New World isolates from *H. zea* are widely used in Australia as biopesticides against both *H. armigera* and *H. punctigera* (Wallengren) in a range of crops including sorghum, chickpea and cotton [14] and are also registered in South Africa and the USA. Two Australian HaSNPV isolates, H25EA1 and AC53, are of international interest as biopesticides. HaSNPV isolate HaSNPV-AC53 (AC53) is manufactured in Australia and included in the commercial biopesticides "Vivus" and "Vivus Max" (AgBiTech Pty Ltd., Brisbane, Queensland, Australia). H25EA1 was selected

by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) from a wild type isolate, and was used by the University of Queensland for in vitro baculovirus production [10,15–17].

Significant genotypic and phenotypic diversity exists within nucleopolyhedroviruses (NPV) isolates, which can be identified by cloning in vivo or in vitro [11,18–22]. For example, 25 of the 162 tissue culture clones isolated from field populations in Kenya, South Africa, Zimbabwe and Thailand were unique variants of HaSNPV [23,24]. Classification and origin of baculovirus species and strains remain important due to restrictions on import of non-native species and concerns over variation between strains during registration of biopesticides, particularly in Australia [25].

Baculovirus species have been described using restriction endonuclease digestion profile and Sanger sequencing, and more recently by Next Generation Sequencing (NGS) [10,11,16,17,23,26–29]. Previous research has shown that HaSNPV and HzSNPV share sequence similarity of up to 99.9%, but could be distinguished by a small number of nucleotide substitutions and by open reading frame (ORF) insertions and deletions in the published consensus genome [17,30,31]. However, we know little about the strain diversity within these isolates and their taxonomic relationship to the Old and New World wild type strains.

This paper examines the sequence similarity and relationships of two Australian HaSNPV isolates from larvae of unidentified *Helicoverpa* sp. and of nine strains derived by passage in tissue culture and insects. We compare whole genome sequences and sequences of selected hypothetical and functional ORFs to determine patterns of strain selection and evolution [12,17] in comparison to sequences from both Old and New World isolates. Throughout, we use HaSNPV to refer to the *Helicoverpa* SNPV virus species but identify isolates from the insect *H. zea* as HzSNPV to differentiate isolates from that of the host and where sequences use the old nomenclature.

2. Materials and Methods

2.1. Virus Source and Passage

HaSNPV isolate AC53, also known as A44WT [10,16], was obtained from AgBiTech and isolate H25EA1 was selected in vitro by CSIRO from P9/H25WT [15,32–35], and obtained from the University of Queensland [17]. Both were originally isolated from cadavers of an unspecified *Helicoverpa* species in Queensland, Australia in 1973 and 1974, respectively, and passaged once through *H. punctigera* before repeated passage through *H. armigera*. This isolation predates the introduction of New World isolates from *H. zea* and use of commercial biopesticides in Australia [10,16]. Both isolates were passaged once by infection of third instar *H. armigera* larvae using a modified droplet method [36]. Insects were fed a suspension of virus with the addition of 10% blue food dye (Queen Fine Foods©, Brisbane, Queensland, Australia) to visualise ingestion and then maintained in individual cups with fresh modified tobacco hornworm diet at constant 26 °C ± 1 °C with 16 h light/8 h dark periods and 70% ± 5% humidity until death.

Occlusion bodies were extracted from cadavers by maceration in 0.1% sodium dodecyl sulphate (SDS), filtration through muslin and centrifugation at 500 rpm and 4 °C for 5 min to remove insect debris, followed by centrifugation at 4000 rpm and 4 °C for 20 min in a swing-out rotor (Sorvall Legend RT®, Sorval Heraeus Rotor). The supernatant was discarded and the pellet resuspended in MilliQ water (Merck Millipore, Boston, MA, USA).

2.2. Test for Latent Virus

The possible presence of latent or sub-lethal (covert) HaSNPV infection in the *H. armigera* insects was investigated. A total of 20 instar larvae were collected for examination by PCR [37,38]. A single AC53 infected larvae was used as a positive control. Each larva was homogenized in a 1.5 mL microcentrifuge tube with 1 mL cold buffer (Tris 10 mM, magnesium chloride 1.5 mM, sodium chloride 140 Mm and 80 µL of 5% Tergitol) and centrifuged at 3800 rpm (Eppendorf Minispin, Hamburg, Germany) for 10 min. The supernatant was collected and spun at 4000 rpm for a further 10 min to pellet cellular material. The supernatant was then discarded, the pellet resuspended and cells lysed

by addition of 200 µL of 0.05 M sodium chloride, 200 µL of 2× Tris/EDTA (TE) buffer and 40 µL of 0.1% SDS, vortexed until clear and incubated at 50 °C for 10 min, then centrifuged at 13,000 rpm for 10 min in order to pellet cellular debris.

Proteins were precipitated by adding 400 µL of supernatant from each sample to 200 µL of ice cold 2.5 M potassium acetate and vortexed thoroughly, then left on ice for 5 min. The samples were centrifuged at 13,000 rpm for 10 min to precipitate the protein. DNA was precipitated from the supernatant by addition of isopropanol and centrifugation at 13,000 rpm for 10 min. The DNA pellet was washed twice with 1 mL of 70% analytical grade ethanol/MilliQ water, centrifuged at 13,000 rpm for 10 min and then air-dried for 10 min. The DNA pellet was resuspended with 60 µL of 1× Tris/EDTA. Degenerate rPol and A44-RIX PCR primers and reaction conditions were as previously described [16]. The PCR amplification was carried out using a Mango Taq kit (Meridian Bioscience Inc., Cincinnati, Ohio, USA) and an Eppendorf Pro S thermocycler, and underwent electrophoresis using a 1% *w*/*v* agarose gel with 0.0001% GelRed (Biotium Inc., Fremont, California, USA) in 1× Tris-acetate EDTA buffer for 1 h at 100 volts.

2.3. Strain Isolation by Passage and Selection in Tissue Culture

Strains were isolated from AC53 using a modified tissue culture plaque assay [39–41]. HzAM1 tissue culture cells were obtained from Dr. Steven Reid (University of Queensland, Australia) and cultured in Ex-Cell 420 Serum Free Insect Medium (Sigma-Aldrich, St. Louis, MO, USA) supplemented with 10% Fetal Bovine Serum (Invitrogen, Thermo-Fisher, Waltham, MA, USA). Second instar *H. armigera* larvae were infected with 1.11×10^5 OB/mL (LC_{90}) of AC53 and hemolymph was harvested at 48 h and 72 h post-infection (pi) by nicking the cuticle of the rear dorsal surface with a scalpel blade. Between 2 and 10 µL was harvested from lots of 10 neonates into glass vials containing 200 µL of Ex-Cell 420 Serum Free Insect Medium (Sigma-Aldrich) with 0.005% phenylthiourea in ethanol (to prevent melanization).

Two methods of tissue culture selection were used (Figure 1). The first method used conventional plaque selection with an agar overlay from infected insect haemolymph. The second used an initial passage of virus from infected haemolymph in tissue culture without agar overlay in order to generate occlusion bodies of strains more adapted to tissue culture. The occlusion bodies generated were used to infect insects from which plaques were selected by the conventional method.

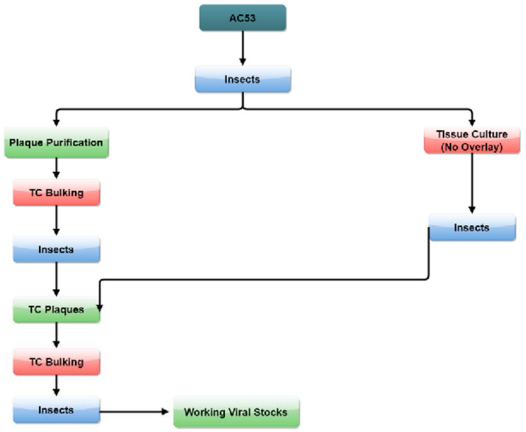


Figure 1. Isolation of *H. armigera* single nucleopolyhedrovirus (HaSNPV) strains in tissue culture (TC).

In Method 1, a total of 100 µL diluted haemolymph was used to infect tissue culture cells at 1.5×10^5 cells per plate (30 mm tissue culture treated petri dishes (Corning Inc., Corning, NY, USA) with an overlay of 1% low temperature gelling agar (SeaPrep LE; Lonza Group, Basel, Switzerland)

in Ex-Cell 420 Serum Free Insect Medium and the addition of 10% Fetal Bovine Serum (Invitrogen) and incubated for 7 days at 28 °C [39,42]. Plaques were visualized by incubation overnight with 25% neutral red in water and picked using a pipette into 200 µL of Ex-Cell 420 Serum Free Insect Medium. Plaque suspensions were then used to infect cells at 2×10^6 cells per plate without overlay and incubated for 7 days as above to produce occlusion bodies. Cells and occlusion bodies were scraped into 2 mL Eppendorf tubes and pelleted at 13,000 rpm for 10 min. Supernatant was poured off and resuspended with 0.05% Tween 80 in MilliQ water and pelleted again. Final pellets were resuspended in 50% glycerol:50% MilliQ and used to infect second instar *H. armigera* larvae as above. The larvae were bled as above, along with a second round of plaque purification and occlusion body production. This was used to infect second instar larvae, from which occlusion bodies were harvested and five strains were produced ("C" strains).

For Method 2, haemolymph, harvested between 48 and 72 h (Table 1) as described above, was first passaged through tissue culture without an agar overlay and occlusion bodies harvested at 7 days. Occlusion bodies were then used to infect second instar larvae that were bled at intervals from 48 to 120 h pi (Table 1). The haemolymph was used for plaque selection and occlusion body production in tissue culture and second instar insects as above (Figure 1), and occlusion bodies were extracted from cadavers to produce four strains ("T" strains).

In the final passage in insects, two strains, T4.1 and T4.2, were selected from two distinct peaks in larval mortality at 168 and 288 h pi in strain T4 to give a total of nine strains (Table 1).

Table 1. HaSNPV-AC53 (AC53) Isolated Strains.

Strain	Time (h) Post-Infection (pi)	First Round Isolation Method
AC53-C1	48 and 48	Agar Overlay
AC53-C5	48 and 48	Agar Overlay
AC53-C6	48 and 48	Agar Overlay
AC53-C9	48 and 48	Agar Overlay
AC53-C3	72 and 72	Agar Overlay
AC53-T2	48 and 48	Tissue Culture—No Overlay
AC53-T4.1 *	72 and 96	Tissue Culture—No Overlay
AC53-T4.2 *	72 and 96	Tissue Culture—No Overlay
AC53-T5	72 and 120	Tissue Culture—No Overlay

Eight isolates were harvested from haemolymph at different times post-infection (passages 1 and 2). Strains C1, C5, C6, C9 and C3 were selected by Method 1, and strains T2, T4 and T5 by Method 2; * Selected from strain T4 at two time points (168 h and 288 h pi) during final passage in neonate larvae.

2.4. DNA Extraction, Next Generation Sequencing Library Preparation, Sequencing and Genome Assembly

DNA was extracted from occlusion bodies using a modification of the method of Doyle et al. [43]. Analytical-grade 0.05 M sodium carbonate was added to the virus pellet to release virions from occlusion bodies. Then, 0.1% SDS in TE buffer was added to disrupt virion membranes. Isolate II Genomic DNA kits (Bioline) were used from Step 4 of the manufacturer's instructions. DNA concentration was determined, (Qubit assay; Invitrogen) and then diluted to 1 ng/µL in MilliQ water. Library preparation was completed using a Nextera XT kit (Illumina, San Diego, CA, USA) and sequences using 150 base pair (bp) paired-end chemistry on an Illumina NextSeq 500 [11]. The assembly method was completed as previously described [11]. This assembly method was developed into a Bash software pipeline, Invertebrates & Microbiology Group-Assembly Pipeline (IMG-AP).

2.5. Sequence Analysis and Maximum-Likelihood Estimation (MLE)

Twenty-one published full genome sequences of HaSNPV isolates [17,23,24,31,44–48] and three genome sequences from HzSNPV isolates [13,49] were aligned and rooted to the *Autographica californica* MNPV (AcMNPV) [50] using MAFFT (Version 7.222) with the FFT-NS-2 algorithm, default settings [51], and any polymorphism including gaps caused by insertions and deletions were classed as mismatches. Alignments were visualized using Geneious R9.1.5. Maximum-likelihood tree construction was

completed with RAxML (Version 7.2.8) with the GTR GAMMA model, rapid bootstrapping and searching for the best-scoring maximum-likelihood tree with 1000 bootstrap replicates [52]. Tree visualisation and editing was completed using TreeGraph 2.9.2-622 beta [53]. This was repeated for the available baculovirus repeated open reading frame (BRO)-A, BRO-B, ORF42 (ORF43 homolog), ORF61 (HzSNPV ORF62 homolog), ORF78 including ORF78a and ORF78b (HzSNPV ORF79 homolog) *lef*-8, *lef-9* and *polh* sequences on Genbank including *Busseola fusca* SNPV isolate A2-4 [54], *Helicoverpa gelotopoeon* SNPV [55], *Helicoverpa assulta* SNPV [56], *Mamestra configurata* NPV-A (MacoNPV-A) [57,58], *Heliothis virescens* Ascovirus 3e [59] and *Plasmodium falciparium* 3D7. Accession numbers and country of origins of each analysed isolate and ORF are shown in Tables S3–S7.

Comparisons of derived strains ORF and homologous repeat mutations were analysed with MAFFT, with the FFT-NS-2 algorithm, and a local copy of BLAST+ (Version 2.5.0) using the Megablast algorithm, to identify nucleotide mutations and the blast algorithm for amino acid mutations [60–62].

3. Results

3.1. Test for Latent Virus

No latent virus was detected within the colony using both rPol (Figure S1) and A44-RIX (Figure S2) primers. The infected positive control tested positive.

3.2. Strain Isolation

Eight strains were selected, 5 "C" strains from Method 1 and 3 "T" strains from Method 2 (Table 1). Strain T4 was split into two strains, T4.1 and T4.2, from cadavers in two distinct peaks of larval mortality: 40% mortality at 168 h and 60% mortality at 288 h.

3.3. Sequence Analysis

All of the derived strains exhibited differing whole genome sequence lengths of between 130,435 bp and 130,460 bp (Table 2) compared to 130,442 bp for the parent strain AC53 and 130,440 bp for H25EA1. Most of the variation in length was found in non-coding regions, but some variation was found within ORFs (Tables 3 and 4 and Table S1).

Table 2. Sequenced *H. armigera* single nucleopolyhedrovirus (HaSNPV) and *H. zea* single nucleopolyhedrovirus (HzSNPV) nucleotide identity compared to AC53 with sequence identity ranging between 81.692% (L1 strain) and 99.604% (AC53-T5).

Common Name	Genbank Accession	Sequence Length (bp)	Nucleotide Identity to AC53 (%)	Country/Region of Origin
HaSNPV-AC53-C1	KU738896	130,460	99.624	Australia
HaSNPV-AC53-C5	KU738898	130,439	99.600	Australia
HaSNPV-AC53-C6	KU738899	130,435	99.601	Australia
HaSNPV-AC53-T4.1	KU738902	130,440	99.602	Australia
HaSNPV-AC53-T5	KU738904	130,442	99.603	Australia
HaSNPV-AC53-C9	KU738897	130,437	99.599	Australia
HaSNPV-AC53-T2	KU738901	130,440	99.596	Australia
HaSNPV-AC53-T4.2	KU738896	130,443	99.530	Australia
HaSNPV-AC53-C3	KU738897	130,437	99.595	Australia
HaSNPV-H25EA1	KJ922128	130,436	99.423	Australia
HzSNPV-HS18	KJ004000	130,890	99.220	Unknown—Sequenced in Russia
HzSNPV-F16 (Elcar-derived)	AF334030	130,869	99.208	USA—Sequenced in China
HzSNPV-Br/South	KM596835	129,694	98.277	Brazil
HaSNPV-NNg1	AP010907	132,425	96.203	Kenya
HaSNPV-LB1	KJ701029	131,966	96.012	Iberia
HaSNPV-SP1A	KJ701032	132,481	95.961	Iberia
HaSNPV-SP1B	KJ701033	132,265	95.810	Iberia
HaSNPV-LB3	KJ701030	130,949	95.799	Iberia
HaSNPV-LB6	KJ701031	130,992	95.798	Iberia
HaSNPV-C1	AF303045	130,759	95.353	China
HaSNPV-AU	JN584482	130,992	94.860	Australia—Sequenced in China
HaSNPV-G4	AF271059	131,405	94.442	China

AC53-derived, H25EA1, and HzSNPV strains are all within 2% nucleotide identity, whereas the remaining HaSNPV strains are within 5.5% nucleotide identity—excluding the L1 strain, which seems to be an outlier.

Viruses **2016**, *8*, 280

Table 3. Amino acid (AA) and nucleotide (N) identity (%) of the regions that are not identical to AC53. The main difference between AC53 and the derived strains occur with both baculovirus repeated open reading frame (BRO)-A and BRO-B, Hr1–Hr5 and HOAR, which is to be expected due to the known hypervariability of the regions. However, open reading frame (ORF) 7 and the hypothetical ORF contains an early stop resulting in a smaller sequence, whereas ORF61 is longer due to the derived strains containing an early stop.

ORF	Protein	AC53-C1		AC53-C3		AC53-C5		AC53-C6		AC53-C9		AC53-T2		AC53-T4.1		AC53-T4.2		AC53-T5		Notes
		AA	N	AA	N	AA	N	AA	N	AA	N	AA	N	AA	N	AA	N	AA	N	
4	HOAR	95.6	95.7	95.8	96.1	96.9	97.1	96.8	96.9	96.9	96.9	96.7	96.9	96.7	96.9	96.9	96.9	96.7	96.9	
5		34.9	97.2	98.3	98.9	34.9	97.2	34.9	97.2	34.9	97.2	34.9	97.2	34.9	97.2	34.9	97.2	34.9	97.2	AC53 and AC53-C3 have identical length
6		99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	
7		94.1	99.3	94.1	99.3	94.1	99.3	94.1	99.3	94.1	99.3	94.1	99.3	94.1	99.3	94.1	99.3	94.1	99.3	AC53 is 85 bp shorter
	Hr1	N.A	99.7	N.A	99.7	N.A	99.7	N.A	99.8	N.A	99.8	N.A	99.7	N.A	99.7	N.A	99.8	N.A	99.7	
	Hr2	N.A	95.5	N.A	95.2	N.A	95.2	N.A	95.2	N.A	95.3	N.A	95.2	N.A	95.2	N.A	95.2	N.A	95.2	
Hypothetical ORF	Hypothetical Protein	70.7	98.0	70.7	98.0	70.7	98.0	70.7	98.0	70.7	98.0	70.7	98.0	70.7	98.0	70.7	98.0	70.7	98.0	AC53 is 24 bp shorter
59	BRO-A	90.9	92.1	90.9	92.1	90.9	92.1	90.9	92.1	90.9	92.1	91.3	92.1	90.9	92.1	90.9	92.1	90.9	92.1	
60	BRO-B	90.4	94.0	90.4	94.0	90.4	94.0	90.4	94.0	90.4	94.0	90.4	94.0	90.4	94.0	90.4	94.0	90.4	94.0	
	Hr3	N.A	99.6	N.A	99.6	N.A	99.6	N.A	99.6	N.A	99.6	N.A	99.69	N.A	99.6	N.A	99.8	N.A	99.6	
61		80.0	86.9	86.8	86.9	80.0	86.9	80.0	86.9	80.0	86.9	80.0	86.9	80.0	86.9	80.0	86.9	80.0	86.9	AC53 is 41 bp longer
68	DNA polymerase	100	100	99.9	99.9	99.9	99.9	100	100	100	100	99.9	99.9	100	100	100	100	100	100	
78a/78b (ORF78 in all other strains)		100	100	100	100	100	100	76.3	99.4	100	100	100	100	100	100	100	100	100	100	Split in two with AC53-C6
	Hr4	N.A	99.5	N.A	99.0	N.A	99.0	N.A	99.0	N.A	99.0	N.A	99.0	N.A	99.0	N.A	99.0	N.A	99.0	
	Hr5	N.A	99.3	N.A	99.1	N.A	99.1	N.A	99.5	N.A	99.2	N.A	99.1	N.A	99.4	N.A	99.1	N.A	99.4	
126	38.7K	100	100	100	100	100	100	100	100	100	100	99.7	99.4	100	100	100	99.8	100	100	
128a/128b (ORF128 in all other strains)		100	100	100	100	100	100	100	100	100	100	100	100	100	100	23.2	87.6	100	100	Split in two with AC53-T4.2
133	PKIP-1	100	100	100	99.8	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
137		97.2	99.3	97.2	99.3	97.2	99.3	97.2	99.3	97.2	99.3	97.2	99.3	97.2	99.3	97.2	99.3	97.2	99.3	
Total regions with sequence polymorphisms		9	14	10	16	10	15	10	15	8	14	11	16	9	14	10	16	9	14	

N.A. = not applicable.

Table 4. Comparison of the nucleotide and amino acid sequence similarity of AC53 derived strains to each other. The greatest diversity was within the five homologous repeat regions, DNA polymerase and HOAR. Only AC53-T2 contained an amino acid difference in BRO-A. ORF5 is shorter in length in all strains except AC53-C3 than in AC53. AC53-C6 and AC53-T4.2 contained unique mutations within ORF78 and ORF128, respectively, due to an inserted stop.

ORF/Region	Nucleotide Similarity and Clusters within Derived Strains	Amino Acid Similarity and Clusters within Derived Strains
HOAR	- AC53-T4.1, AC53-T5 = 100% - AC53-C6, AC53-C9, AC53-T4.2 = 100% - Remaining 4 strains all different at 96.7% to 99.9%	- AC53-T4.1 and AC53-T5 = 100% - AC53-C6, AC53-C9 ,AC53-T4.2 = 100% - Remaining 4 strains all different at 95.8% to 99.8%
ORF5 *	- AC53-C3 = 96.1% - Remaining strains all identical	- AC53-C3 = 33.3% - Remaining strains all identical
BRO-A	- AC53-T2 = 99.9% - Remaining strains all identical	- AC53-T2 = 99.5% - Remaining strains all identical
DNA-Polymerase	- AC53-T5, AC53-T4.2, AC53-T4.1, AC53-C9, AC53-C6, AC53-C1 = 100% - AC53-C5, AC53-C3 = 100% - AC53-T2 = 99.9%	- AC53-C3, AC53-C5, AC53-T2 = 100% - AC53-T5, AC53-T4.2, AC53-T4.1, AC53-C9, AC53-C6 - AC53-C1 = 100%
ORF78/ORF78a and 78b in AC53-C6	- AC53-C6 = 99.4% - Remaining strains all identical	- AC53-C6 = 77.9% - Remaining strains all identical
38.7K Protein	- AC53-T2 = 99.4% to other 7 strains and 99.6% to AC53-T4.2 - AC53-T4.2 = 99.8% to other 7 strains - Remaining 7 strains all identical	- AC53-T2 = 99.7% - Remaining 8 strains all identical
ORF128/ORF128a and 128b in AC53-T4.2	- AC53-T4.2 = 87.6% - Remaining strains all identical	- AC53-T4.2 = 23.2% - Remaining strains all identical
PKIP-1	- AC53-C3 = 99.8% - Remaining strains all identical	- All strains = 100%
Hr1	- AC53-T4.2, AC53-C9, AC53-C6 = 100% - AC53-T5, AC53-T4.1, AC53-T2, AC53-C5, AC53-C3, AC53-C1 = 100% - 99.9% when both groups compared	- Not Applicable
Hr2	- AC53-C6, AC53-T4.1 = 100% - Remaining strains all identical	- Not Applicable
Hr3	- AC53-T4.2 = 99.8% - Remaining strains all identical	- Not Applicable
Hr4	- AC53-C1 = 99.2% - Remaining strains all identical	- Not Applicable
Hr5	- AC53-T2, AC53-C5, AC53-C3 = 100% - Remaining strains all identical	- Not Applicable

* ORF5 is 87 bp longer in AC53-C3.

Both AC53 and H25EA1, and the derived strains shared sequence similarities with both Old World isolates from *H. armigera* and New World isolates from *H. zea*. Overall sequence similarity between AC53 and other isolates was greatest with *H. zea* isolates (98% to 99%) and 94%–98% with *H. armigera* isolates (Table 2). Sequence similarity to the L1 isolate from India was 81%, which contained a significant rearrangement in the genome and was excluded from whole genome comparisons. AC53 and H25EA1 had 99.4% overall sequence similarity (Table 2). The greatest nucleotide differences within reading frames were in BRO-A (10%) and BRO-B (4%). ORF136 of H25EA1 was 432 bp shorter than in AC53 but had 99% nucleotide sequence similarity to that of AC53.

Comparison of the AC53 genome with the derived strains identified nucleotide base changes in up to 16 different regions, resulting in predicted amino acid changes in 11 ORFs (Table 3, Table S2). Nine ORFs contained predicted amino acid changes in every tissue culture-derived strain (i.e., HOAR, ORF5, ORF6, ORF7, Hypothetical ORF, BRO-A, BRO-B, ORF61, and ORF137). Of these, 6 ORFs (ORF6, ORF7, Hypothetical ORF, BRO-B, ORF61, ORF137) had identical predicted amino acid changes. ORF7 was 85 bp longer in all of the derived strains than in the AC53 parent isolate (Figure S3).

The remaining 3 ORFs (HOAR, ORF5 and BRO-A) had differences between strains as well as from the parent AC53 isolate (Table 4). In the AC53-T2 strain, BRO-A had a single nucleotide change that resulted in a different predicted amino acid sequence from that of the other eight derived strains. Strain AC53-C3 had the full-length ORF5 found in AC53 isolate, but ORF5 was 87 bp shorter in the other eight derived strains resulting in changes in predicted amino acid sequence. HOAR differed from the AC53 parent strain in every derived strain but the strains contained multiple and different polymorphisms resulting in six different genotypes with differences in nucleotide and predicted amino acid sequence (Table 4, Figure S5).

Nucleotide base changes were identified in seven ORFs when comparing the genomes of the derived strains (Table 4), but mutations resulting in amino acid changes were only found in the sequences of six ORFs (HOAR, ORF5, DNA polymerase and BRO-A, ORF61, ORF78 and ORF128). All polymorphisms are listed in Tables 3 and 4 (no polymorphisms were observed in other ORFs). The complete genome sequences of HaSNPV-H25EA1, AC53 and seven derived isolates, contained 139 ORFs and five homologous repeat (hr) regions. Of these, 138 ORFs were shared with the HzSNPV-F16 isolate [13].

Two derived strains (AC53-C6 and AC53-T4.2) contained 140 ORFs as a result of insertions of early stop codons. A single base pair deletion at position 69,728 in ORF78 of AC53-C6 split the ORF into two smaller hypothetical ORFs of 73 bp (ORF78a) and 81 bp (ORF78b). Similarly, in AC53-T4.2, ORF128 splits into two hypothetical ORFs of 267 bp (ORF128a) and 141 bp (ORF128b), and a 26 bp non-coding region resulting from 50 substitutions to five deletions (Figure S6). This split did not occur in AC53-T4.1 or in the AC53 wild type. HaSNPV-H25EA1, AC53 and the nine derived strains all contained an additional hypothetical ORF identified in the antisense direction between ORF54 and *lef-9* (Figure 2). This ORF was 99 bp in the AC53 parent strain and H25EA1, and 120 bp in the derived strains, which contained an additional 21 bp CA-repeat region.

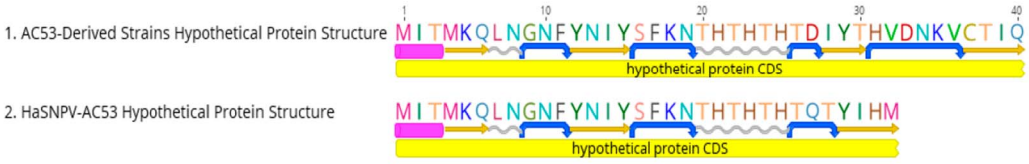


Figure 2. Amino acid sequence and predicted protein structure of hypothetical proteins identified using the EMBOSS garnier [63]. Alpha helix (purple rectangle), beta strands (yellow arrows), coils (gray wavy lines), and the turns (blue curved arrows) are depicted here. AC53-derived strains contain an additional turn and an additional beta strand; CDS, coding DNA sequence.

The Australian isolates contained ORFs similar SNPV isolates from both *H. armigera* and *H. zea*. Both AC53 and H25EA1 and all nine derived strains contained ORF42, which is reported to be found

in some HaSNPV isolates, and as a homolog at ORF43 in some *H. zea* SNPVs [13,48]. Likewise, in both isolates and eight of the derived strains, ORF78 had 99.4% sequence similarity to an ORF79 that is reported to be specific to isolates from *H. zea*, and no similarity to the ORF78 annotated in published sequences of isolates from *H. armigera* [3]. In derived strain AC53-C6, ORF78 is split into ORF78a (73 bp) and ORF78b (81 bp; as above).

There was a pattern of increasing fragmentation of the ORF78 homologs in comparison to published HaSNPV genomes (Figure S7). The full-length *H. zea* ORF 79 homolog in all published *H. zea* isolates, except for HzSNPV-Br/South, is split into two 84 bp and 81 bp hypothetical ORFs (not annotated on Genbank) that are homologs of AC53-C6 ORF78a and ORF78b, respectively. The Iberian (SP1A, SP1B, LB1, LB3 and LB6) and Kenyan (NNg1) HaSNPV isolates contained a similar, unannotated 81 bp homolog of the ORF78b found in AC53-C6 that resulted from either a 16 bp or 14 bp deletion in a 28 bp AT-repeat, as had been previously observed in other strains [3]. The Chinese C1, G4 and AU HaSNPV isolates did not contain a homolog to ORF78b but contained two overlapping, unannotated hypothetical ORFs that are 45 bp and 48 bp in length within the region of the genome analogous to ORF78b of AC53-C6.

Two of the Iberian isolates (LB6 and LB3) also contained an unannotated 73 bp hypothetical ORF78a homolog while the remaining Iberian, NNg1, Chinese C1, G4 and AU isolates contained a much smaller 57 bp hypothetical ORF homolog of ORF78a. The pattern of increased fragmentation of isolates can also be observed in ORF61 (Figure S8). The ORF61 of AC53 and H25EA1 shared 99.4% sequence similarity, with a single substitution in H25EA1 from T to C, at position 52,591. There were again greater similarities with *H. zea* isolates: ORF61 in AC53 had 100% identity to ORF62 reported in *H. zea* isolate 35036 derived from the commercial biopesticide Gemstar, and H25EA1 had 100% sequence similarity to ORF62 in three HzSNPV isolates (isolate F16 from the commercial product Elcar, isolate 35022 from Gemstar, and *H. zea* isolate HS-18 sequenced in Russia), and 100% similarity to ORF66 in the Brazilian HzSNPV isolate Br/South [13,49]. ORF61 of AC53 had 99.4% sequence similarity to ORF62 of the Kenyan isolate NNg1, which has a substitution from T to A at position 52,516 (of AC53). ORF61 of H25EA1 has 98.8% similarity to ORF62 of NNg1, but with a second substitution (T to C) at position 52,591. The ORF61 from AC53 and H25EA1 has 98.9% (two substitutions) and 99.4% (one substitution) sequence similarity to a hypothetical ORF of the samelength (180 bp) in four Iberian *H. armigera* isolates that is not annotated as an ORF in the sequences on Genbank (KJ701030 to KJ701033).

ORF61 in eight of the derived strains has a 27 bp insertion, and one (AC53-C1) has a 20 bp insertion, that truncates ORF61 by 41 bp when compared to AC53. The remaining 72 bp of the ORF have 100% sequence similarity to the same region of ORF61 in AC53. The 72 bp truncated ORF61 also shares 100% sequence similarity to a 72 bp hypothetical ORF in the Iberian HaSNPV isolate LB1, but the truncation was caused by a 1 bp insertion and two substitutions and not the 27 bp insertion. This hypothetical ORF has not been annotated in the LB1 sequence on Genbank (KJ701029). The truncated, 72 bp, hypothetical ORF61 has 98.7% sequence similarity (1 bp substitution) to an unannotated hypothetical ORF in the Chinese *H. armigera* isolate C1, which also contains the 27 bp insertion. Homologs of ORF61 are not found in the Chinese G4 or AU isolates but a 107 bp fragment with 99.1% sequence similarity is found the Hr3 region of those two isolates.

3.4. Maximum Likelihood Estimation

Comparison of the AC53, H25EA1 and derived strain whole genome sequences to all *Helicoverpa* SNPV genomes available on Genbank identified three distinct clusters of isolates based on geographic origin but not on host species (Figure 3). AC53, H25EA1 and the derived strains all clustered with isolates from *H. zea*. This cluster could be further divided (with 77% support) into a cluster containing AC53 and its derived strains, and a second cluster containing H25EA1 and all the HzSNPV isolates. The Old World HaSNPV isolates form two distinct clusters with 100% support: a group of three isolates from China (G4, C1 and the Chinese 'AU' isolate), and a second cluster containing the Iberian and Kenyan isolates.

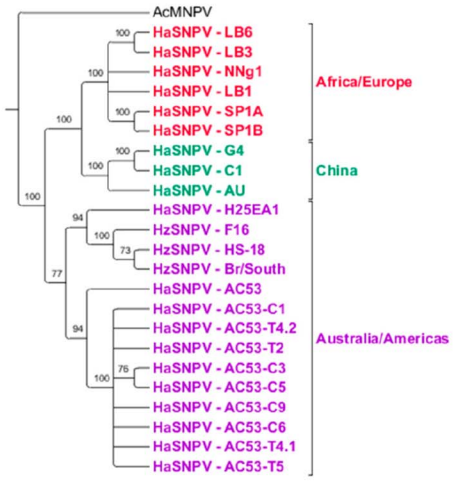


Figure 3. Phylogenetic Relationship of all of the HaSNPV and *H. zea* single nucleopolyhedrovirus (HzSNPV) strains (with bootstrap support as a percentage) and rooted to *Autographica californica* multiple nucleopolyhedrovirus (AcMNPV). Tree has been collapsed based on a minimum of 70% support. Geographically, three distinct groups can be observed; an Australian/American group consisting of AC53, H25EA1, AC53-derived strains and the HzSNPV isolates (purple), a Chinese group containing the HaSNPV C1, G4 and AU strains (green), and a third African/European group containing the Iberian (SP, LB) and Kenyan (NNg1) isolates (red).

BRO-A homologs are only found in AC53, H25EA1, strains derived from AC53, and HzSNPV isolates. Maximum likelihood estimation (MLE) of BRO-A identified four clusters with 100% support: the *H. zea* isolates and H25EA1, isolate AC53, derived strain AC53-T2 and the remaining strains derived from AC53 (Figure 4). BRO-B homologs are found in the majority of other sequenced HaSNPVs except the Chinese AU, Iberian SP1A and Iberian SP1B isolates. Four clusters were identified: *H. zea* isolates, *H. armigera* isolates (including H25EA1), AC53, and the strains derived from AC53, which again clustered separately from all other isolates including AC53.

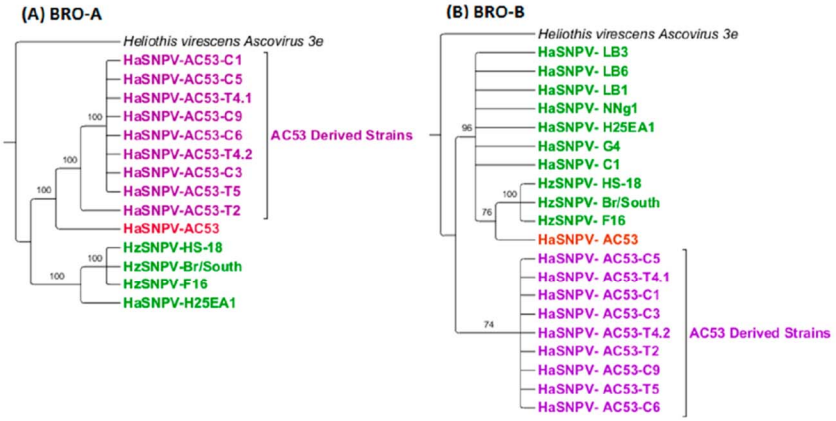


Figure 4. The root for both trees is the *Heliothis virescens* Ascovirus 3e isolate and have been collapsed based on 70% bootstrap support. (**A**) baculovirus repeated open reading frame (BRO)-A maximum likelihood estimation (MLE) identified two distinct clusters, AC53 derived and non-AC53 derived; (**B**) BRO-B also identified two distinct clusters, however the parent AC53 strain was grouped with non-AC53 derived. A third group in BRO-B consisting of HzSNPV isolates could also be identified.

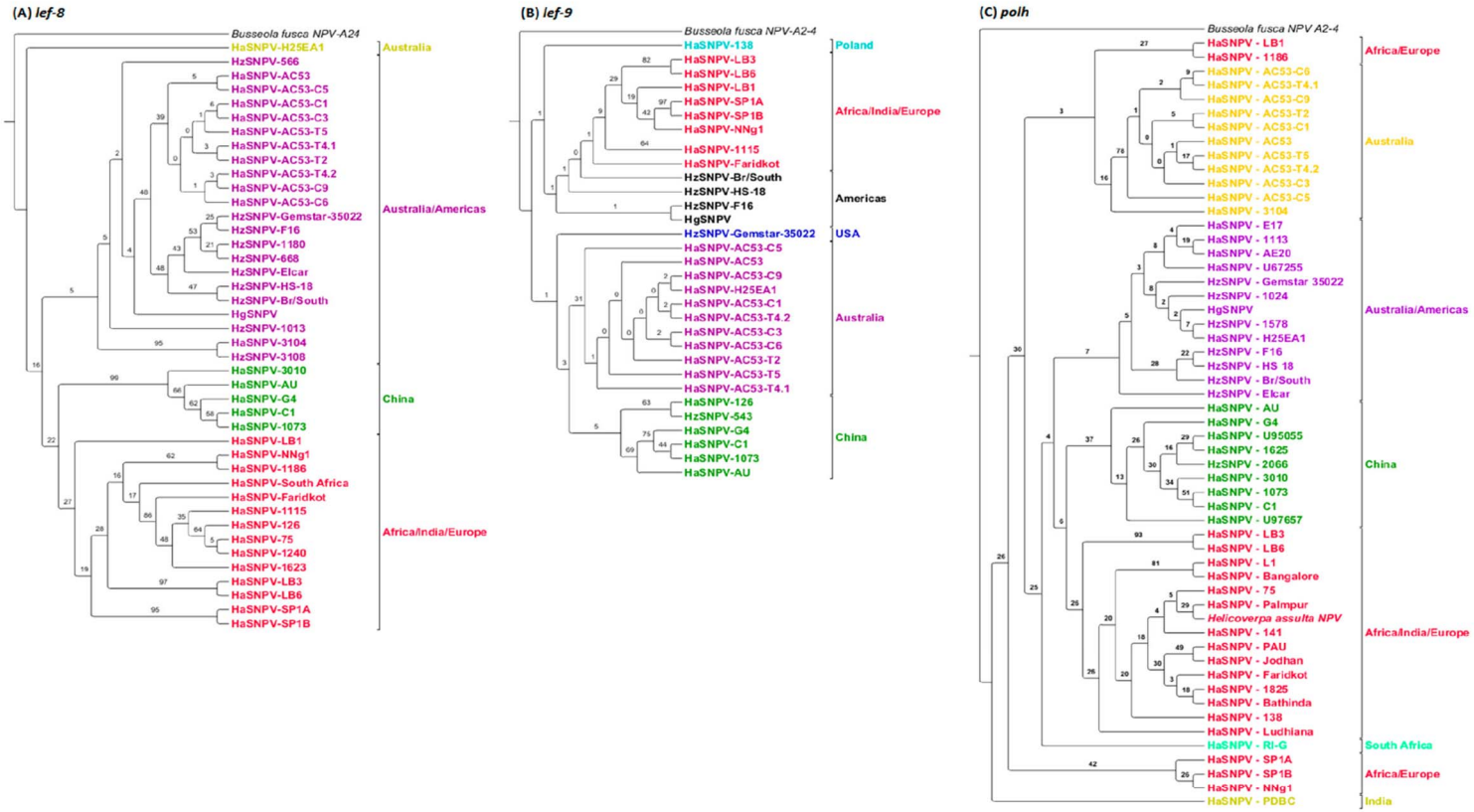


Figure 5. All trees have not been collapsed, due to poor bootstrapping support, and would not be differentiated otherwise, and have all been rooted to *Busseola fusca* NPV A2-4. (**A**) The geographic relationships of *lef-8* have at least three clusters, Africa/India/Europe (red), China (green), Australia/Americas (purple) and the Australian H25EA1 isolate (yellow-green); (**B**) The geographic relationships of *lef-9* have at least four clusters, Africa/India/Europe (red), Poland (aqua), the United States of America (blue) and North and South America (black), Australia (purple) and China (green); (**C**) The geographic relationships of *polh* have at least four clusters, India (yellow-green), Africa/Indian/Europe (red), China (green), South Africa (light green), Australia (yellow) and Australia/Americas (purple).

Analysis of the baculovirus ORFs *lef-8*, *lef-9* and *polh*, commonly used as markers, identified the similar geographic clusters as identified using whole-genome phylogenetic analysis. However, the high sequence similarity in these ORF led to very low levels of support (Figure 5). Isolates from India consistently clustered with the African/European isolates using all three ORFs.

The same three clusters of isolates by geographic origin (Australia/Americas, Europe and Africa, and China) were identified by MLE of *lef-8*, but with very low levels of support. Analysis of *polh* and *lef-9* was less powerful in resolving clusters. North and South American isolates were grouped separately from most of the Australian isolates except H25EA1, but again with very low levels of support. Clusters based on *polh* had particularly low levels of support.

Maximum likelihood analysis of the ORF42 and ORF78 identified the same three geographic clusters: Australia and Americas, Europe/Africa and China (Figure 6). The Indian isolate 'Faridkot' was separate from the clusters using ORF42. Indian isolate L1 was included with the European and African isolates using ORF78. ORF61 did not support resolution based on geographic origin or insect species (Figure 7), although two Indian isolates could be separated with 100% support, and the Iberian isolate LB1 was grouped with the Sudanese (African) isolate.

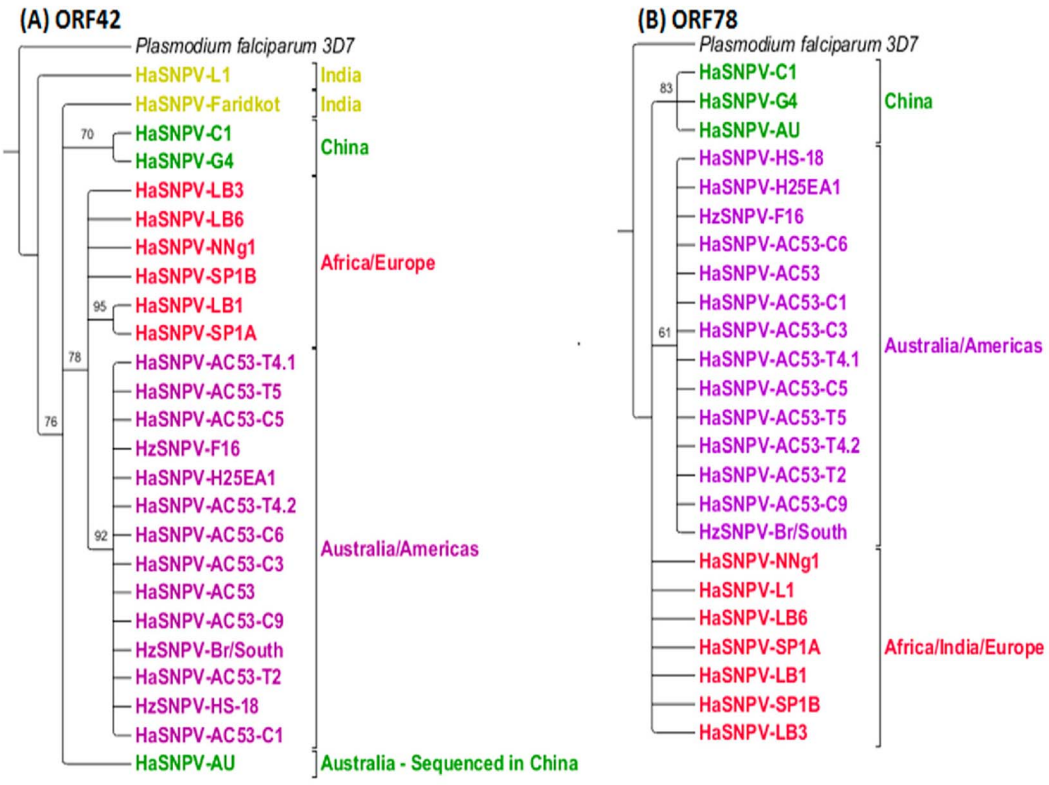


Figure 6. MLEs using (**A**) open reading frame (ORF) 42 and its homologs; and (**B**) ORF78 and its homologs identify the same three geographic clusters as the whole-genome sequences. Both trees have been rooted to *Plasmodium falciparum* 3D7 with a similar nucleotide sequence to ORF42 and ORF78, and collapsed to 60% minimum support values. Coloring code is identical to Figure 5.

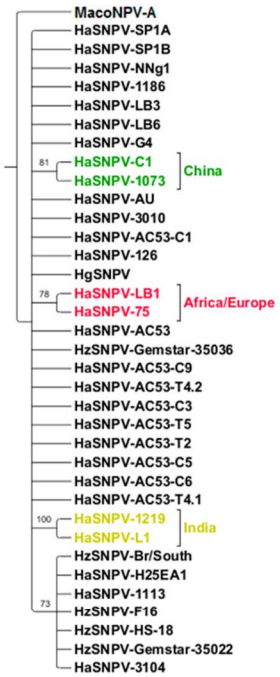


Figure 7. ORF61 (ORF62 homolog), rooted to MacoNPV-A, showing only six strains clustering based on geography.

4. Discussion

The strains derived from AC53 contained a number of polymorphisms but shared high sequence similarity with the parent (over 99.5%) and with each other (over 99.9%; Table S2). These could not be separated into clusters by maximum likelihood analysis.

Previous studies have shown significant deletions to occur in the *ecdysteroid UDP-glucosyltransferase* (*egt*) gene during passage in tissue culture [64,65]. However, no differences were found within the *egt* gene in the AC53-derived strains, but nine ORFs contained nucleotide sequences that differed from the parent AC53 isolate that resulted in changes in predicted amino acid sequence, of which changes in six ORFs were identical in all nine strains. We speculate that this is a result of selection of strains or of mutations arising as a result of passage of the virus through tissue culture.

The two strains, AC53-T4.1 and AC53T4.2, were isolated from two distinct peaks in time to death during passage in vivo of strain AC53-T4. In AC53-T4.2, ORF128 was split into two hypothetical ORFs of 267 bp (ORF128a) and 141 bp (ORF128b), and a 26 bp non-coding region. This split was not observed in AC53-T4.1 or the AC53 parent isolate. We speculate that changes in ORF128 result in differences in pathogenicity and speed of kill.

Sequencing of a number of ORFs has been used to attempt to differentiate isolates of HaSNPV by geographic or insect of origin [3]. However, maximum likelihood analysis using *polh*, *lef-9* and ORF61/62 homologs did not resolve either insect species or geographic origin of isolation with any significant degree of support, although not all the ORF61/62 isolates described in the published comparisons are available through Genbank [3]. On the other hand, maximum likelihood comparison of whole-genome sequences and sequences of *lef-8*, ORF42 and ORF78 from the Australian isolates and derived strains with sequences of homologs of other isolates available through Genbank identified a consistent pattern of clusters based on geographic origin: Australia and the Americas, Europe, Africa and India, and China. The Chinese 'AU' isolate was produced and sequenced in the same facility as

the other Chinese isolates (G4 and C1) and shares a high degree of sequence similarity with these isolates. We have classified it with these isolates.

There was no support for classification of the viruses based on species of isolation. The Australian isolates and derived strains showed high levels of sequence similarity with isolates from *H. zea*. Several hypothetical ORFs have been used to differentiate between isolates from *H. zea* and *H. armigera*, on the basis of presence or absence, but close inspection of the whole genome sequences on Genbank found that these hypothetical ORFs (ORF42, 62 and 78) were present in all sequences either as whole or truncated hypothetical ORFs that are not annotated [3,13]. A previously unannotated hypothetical ORF located between ORF54 and *lef-9* was identified but not annotated in all published HaSNPV genomes. Clustering of *H. zea* and *H. armigera* isolates based on MLEs from whole-genome and ORF sequences supported the classification of the viruses as a single species [12].

There was a consistent pattern of fragmentation of ORFs 78 and 61 of isolates AC53 and H25EA1 in comparison to homologs in other isolates. All isolates from the Americas had intact homologs with 100% sequence similarity to ORF61 and to ORF78 except for the Brazilian isolate Br/South which had a fragmentation of ORF78 similar to that of the AC53-derived strain AC53-C6.

The fragmentation of these ORFs became more pronounced in Iberian and African isolates. ORF61 was present as a full-length homolog in the African NNg1 and four Iberian strains, but Iberian isolate LB1 was truncated to a 72 bp unannotated hypothetical ORF homolog with 100% sequence similarity to a 72 bp truncated hypothetical ORF found all the strains derived from AC53. All homologs of ORF78 in the Iberian and African isolates were split into two, again similar to that found in AC53-C6. Fragmentation was most strongly found in the Chinese isolates. A truncated homolog of ORF61 with a 27 bp insertion identical to that identified in the AC53 derived strains was identified in isolate C1. In the other Chinese isolates, a 107 bp homolog of ORF61 was found in Hr3. ORF78 was further fragmented into three short, unannotated hypothetical ORFs.

This pattern of fragmentation suggests that strains have been selected following evolutionary bottlenecks. We speculate that the clustering of isolates based on MLE, in combination with the observed pattern of fragmentation suggests an origin of the HaSNPV species in Australia with subsequent movement to the Americas, then to Africa, Europe and India, and more recently to China, a pattern which follows global wind patterns [66–68]. The strains derived from AC53 also contained evidence of selection. Distinct clustering by MLE, using the sequences of ORFs BRO-A and BRO-B, and truncation and fragmentation of ORFs 61 and 78, suggest that plaque selection and passage through tissue culture led to the selection of new dominant genotypes from the AC53 isolate.

In general, the lack of support for clusters identified by MLE, based on ORFs 61, *polh*, *lef-9* and *lef-8* suggest that these regions cannot be used for taxonomic comparison within a virus species. Similarly, ORFs 42 and 78 were able to differentiate HaSNVP isolates based on geographic origin, but only a moderate level of support. Our analysis supports the conclusion that HaSNPV and HzSNPV are variants of the same species, and that host insect cannot be used to predict isolate identity [3,5,13,20], and suggests a possible origin for HaSNPVs isolates in Australia.

Supplementary Materials: The following tables and figures are available online at www.mdpi.com/1999-4915/8/11/280/s1, Figure S1: PCR detection for all NPV using rPol primer set; Figure S2: PCR detection of HaSNPV using the A44-RIX primer set; Figure S3: Nucleotide comparison of ORF7 within AC53 and its derived strains; Figure S4: Nucleotide comparison of ORF5 within AC53 and its derived strains; Figure S5: Nucleotide comparison of the four regions (A, B, C and D) containing mutations within the HOAR nucleotide sequence of AC53 and its derived strains; Figure S6: Nucleotide comparison of ORF128 with AC53 and its derivatives to the AC53-T4; Figure S7: Nucleotide comparison of fragmentation occurring within ORF78/79 with 10 distinct genotypes observed across all HaSNPV and HzSNPV strains; Figure S8: Nucleotide comparison of fragmentation occurring within ORF61/62 with 16 distinct genotypes observed across all HaSNPV and HzSNPV strains; Table S1: Nucleotide and Amino Acid comparison of the AC53 and H25EA1 strains; Table S2: Nucleotide distance matrix of AC53 and its derived strains; Table S3: *Lef-8* analysed strains; Table S4: *Lef-9* analysed strains; Table S5: *polh* analysed strains; Table S6: BRO-A and BRO-B analysed strains; Table S7: ORF42, ORF61 and ORF78 analysed strains.

Acknowledgments: This work was funded in part by the Cotton Research Development Corporation, by funding from the Queensland University of Technology (QUT), Australia; and materials were supplied by AgBiTech. Some of the data reported in this paper was generated in the Central Analytical Research Facility (CARF) operated by

the Institute for Future Environments at QUT. Access to CARF is supported by generous funding from the Science and Engineering Faculty at QUT. We would like to thank staff of the Molecular Genetics Research Facility and the Invertebrate Microbiology Group at QUT for their assistance with sequencing and technical support.

Author Contributions: Christopher Noune and Caroline Hauxwell contributed equally to this paper.

Conflicts of Interest: The authors declare a conflict of interest. The Cotton Research Development Corporation has funded the work by Christopher Noune through a post-graduate student scholarship. AgBiTech provided the sample of HaSNPV-AC53 and previously funded consultancy and research work with Caroline Hauxwell, but did not contribute financially to this study.

Script Availability: The IMG-AP is available for download at https://github.com/CNoune/IMG_pipelines.

References

1. Rohrmann, G. Introduction to the baculoviruses and their taxonomy. In *Baculovirus Molecular Biology*, 2nd ed.; National Center for Biotechnology Information: Bethesda, MD, USA, 2011.
2. Blissard, G.W.; Rohrmann, G.F. Baculovirus diversity and molecular biology. *Annu. Rev. Entomol.* **1990**, *35*, 127–155. [CrossRef] [PubMed]
3. Rowley, D.L.; Popham, H.J.R.; Harrison, R.L. Genetic variation and virulence of nucleopolyhedroviruses isolated worldwide from the heliothine pests *Helicoverpa armigera*, *Helicoverpa zea*, and *Heliothis virescens*. *J. Invertebr. Pathol.* **2011**, *107*, 112–126. [CrossRef] [PubMed]
4. Jehle, J.A.; Lange, M.; Wang, H.; Hu, Z.; Wang, Y.; Hauschild, R. Molecular identification and phylogenetic analysis of baculoviruses from Lepidoptera. *Virology* **2006**, *346*, 180–193. [CrossRef] [PubMed]
5. Jehle, J.A.; Blissard, G.W.; Bonning, B.C.; Cory, J.S.; Herniou, E.A.; Rohrmann, G.F.; Theilmann, D.A.; Thiem, S.M.; Vlak, J.M. On the classification and nomenclature of baculoviruses: A proposal for revision. *Arch. Virol.* **2006**, *151*, 1257–1266. [CrossRef] [PubMed]
6. Herniou, E.A.; Jehle, J.A. Baculovirus phylogeny and evolution. *Curr. Drug Targets* **2007**, *8*, 1043–1050. [CrossRef] [PubMed]
7. Wardhaugh, K.G.; Room, P.M.; Greenup, L.R. The incidence of *Heliothis armigera* (Hübner) and *H. punctigera* Wallengren (Lepidoptera: Noctuidae) on cotton and other host-plants in the Namoi Valley of New South Wales. *Bull. Entomol. Res.* **1980**, *70*, 113–131. [CrossRef]
8. Daly, J.C.; Gregg, P. Genetic variation in Heliothis in Australia: Species identification and gene flow in the two pest species *H. armigera* (Hübner) and *H. punctigera* Wallengren (Lepidoptera: Noctuidae). *Bull. Entomol. Res.* **1985**, *75*, 169–184. [CrossRef]
9. Zhang, G. Commercial viral insecticide *Heliothis armigera* viral insecticide in China. *IPM Pract.* **1989**, *11*, 13.
10. Richards, A.R.; Christian, P.D. A rapid bioassay screen for quantifying nucleopolyhedroviruses (Baculoviridae) in the environment. *J. Virol. Methods* **1999**, *82*, 63–75. [CrossRef]
11. Noune, C.; Hauxwell, C. Complete genome sequences of seven *Helicoverpa armigera* SNPV-AC53-derived strains. *Genome Announc.* **2016**, *4*, e00260-16. [CrossRef] [PubMed]
12. Harrison, R.L.; Herniou, E.A.; Theilmann, D.A.; Becnel, J.J.; Arif, B.; Jehle, J.A.; Burand, J.P.; Oers, M.V. *Removal of Species Helicoverpa zea Single Nucleopolyhedrovirus from the Genus Alphabaculovirus*; International Committee on Taxonomy of Viruses: Edinburgh, UK, 2013.
13. Chen, X.; Zhang, W.-J.; Wong, J.; Chun, G.; Lu, A.; McCutchen, B.; Presnail, J.; Herrmann, R.; Dolan, M.; Tingey, S.; et al. Comparative analysis of the complete genome sequences of *Helicoverpa zea* and *Helicoverpa armigera* single-nucleocapsid nucleopolyhedroviruses. *J. Gen. Virol.* **2002**, *83*, 673–684. [CrossRef] [PubMed]
14. Buerger, P.; Hauxwell, C.; Murray, D. Nucleopolyhedrovirus introduction in Australia. *Virol. Sin.* **2007**, *22*, 173–179. [CrossRef]
15. Nguyen, Q.; Qi, Y.M.; Wu, Y.; Chan, L.C.L.; Nielsen, L.K.; Reid, S. In vitro production of *Helicoverpa baculovirus* biopesticides—Automated selection of insect cell clones for manufacturing and systems biology studies. *J. Virol. Methods* **2011**, *175*, 197–205. [CrossRef] [PubMed]
16. Christian, P.D.; Gibb, N.; Kasprzak, A.B.; Richards, A. A rapid method for the identification and differentiation of *Helicoverpa* nucleopolyhedroviruses (NPV *Baculoviridae*) isolated from the environment. *J. Virol. Methods* **2001**, *96*, 51–65. [CrossRef]

17. Noune, C.; Hauxwell, C. Complete genome sequences of *Helicoverpa armigera* single nucleopolyhedrovirus strains AC53 and H25EA1 from Australia. *Genome Announc.* **2015**, *3*, e01083-15. [CrossRef] [PubMed]
18. Baillie, V.L.; Bouwer, G. High levels of genetic variation within *Helicoverpa armigera* nucleopolyhedrovirus populations in individual host insects. *Arch. Virol.* **2012**, *157*, 2281–2289. [CrossRef] [PubMed]
19. Erlandson, M.A. Genetic variation in field populations of baculoviruses: Mechanisms for generating variation and its potential role in baculovirus epizootiology. *Virol. Sin.* **2009**, *24*, 458–469. [CrossRef]
20. Gettig, R.R.; McCarthy, W.J. Genotypic variation among wild isolates of *Heliothis* spp. nuclear polyhedrosis viruses from different geographical regions. *Virology* **1982**, *117*, 245–252. [CrossRef]
21. Crawford, A.M.; Zelazny, B.; Alfiler, A.R. Genotypic variation in geographical isolates of oryctes baculovirus. *J. Gen. Virol.* **1986**, *67*, 949–952. [CrossRef]
22. Corsaro, B.G.; Fraser, M.J. Characterization of genotypic and phenotypic variation in plaque-purified strains of HzSNPV elkar isolate. *Intervirology* **1987**, *28*, 185–198. [PubMed]
23. Ogembo, J.G.; Caoili, B.L.; Shikata, M.; Chaeychomsri, S.; Kobayashi, M.; Ikeda, M. Comparative genomic sequence analysis of novel *Helicoverpa armigera* nucleopolyhedrovirus (NPV) isolated from Kenya and three other previously sequenced *Helicoverpa* spp. NPVs. *Virus Genes* **2009**, *39*, 261–272. [CrossRef] [PubMed]
24. Ogembo, J.G.; Chaeychomsri, S.; Kamiya, K.; Ishikawa, H.; Katou, Y.; Ikeda, M.; Kobayashi, M. Cloning and comparative characterization of nucleopolyhedroviruses isolated from African bollworm, *Helicoverpa armigera*, (Lepidoptera: Noctudiae) in different geographic regions. *J. Insect Biotechnol. Sericol.* **2007**, *76*, 39–49.
25. Kabaluk, J.T.; Svircev, A.M.; Goettel, M.S.; Woo, S.G. (Eds.) *The Use and Regulation of Microbial Pesticides in Representative Jurisdictions Worldwide*; IOBC Global, 2010; p. 99. Available online: http://www.iobc-global.org/download/Microbial_Regulation_Book_Kabaluk_et_al_2010.pdf (accessed on 17 October 2016).
26. Baillie, V.L.; Bouwer, G. Development of highly sensitive assays for detection of genetic variation in key *Helicoverpa armigera* nucleopolyhedrovirus genes. *J. Virol. Methods* **2011**, *178*, 179–185. [CrossRef] [PubMed]
27. Craveiro, S.R.; Inglis, P.W.; Togawa, R.C.; Grynberg, P.; Melo, F.L.; Ribeiro, Z.M.; Ribeiro, B.M.; Báo, S.N.; Castro, M.E. The genome sequence of Pseudoplusia includens single nucleopolyhedrovirus and an analysis of p26 gene evolution in the baculoviruses. *BMC Genom.* **2015**, *16*, 127. [CrossRef] [PubMed]
28. Chateigner, A.; Bézier, A.; Labrousse, C.; Jiolle, D.; Barbe, V.; Herniou, E.A. Ultra deep sequencing of a baculovirus population reveals widespread genomic variations. *Viruses* **2015**, *7*, 3625–3646. [CrossRef] [PubMed]
29. Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three Next Generation Sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom.* **2012**, *13*, 341. [CrossRef] [PubMed]
30. Harrison, R.L. Structural divergence among genomes of closely related baculoviruses and its implications for baculovirus evolution. *J. Invertebr. Pathol.* **2009**, *101*, 181–186. [CrossRef] [PubMed]
31. Zhang, C.-X.; Ma, X.-C.; Guo, Z.-J. Comparison of the complete genome sequence between C1 and G4 isolates of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus. *Virology* **2005**, *333*, 190–199. [CrossRef] [PubMed]
32. Pedrini, M.R.; Christian, P.; Nielsen, L.K.; Reid, S.; Chan, L.C. Importance of virus—Medium interactions on the biological activity of wild-type *Heliothine* nucleopolyhedroviruses propagated via suspension insect cell cultures. *J. Virol. Methods* **2006**, *136*, 267–272. [CrossRef] [PubMed]
33. Nguyen, Q.; Nielsen, L.K.; Reid, S. Genome scale transcriptomics of baculovirus-insect interactions. *Viruses* **2013**, *5*, 2721–2747. [CrossRef] [PubMed]
34. Lua, L.H.; Reid, S. Virus morphogenesis of *Helicoverpa armigera* nucleopolyhedrovirus in *Helicoverpa zea* serum-free suspension culture. *J. Gen. Virol.* **2000**, *81*, 2531–2543. [CrossRef] [PubMed]
35. Lua, L.H.; Pedrini, M.R.; Reid, S.; Robertson, A.; Tribe, D.E. Phenotypic and genotypic analysis of *Helicoverpa armigera* nucleopolyhedrovirus serially passaged in cell culture. *J. Gen. Virol.* **2002**, *83*, 945–955. [CrossRef] [PubMed]
36. Hughes, P.R.; Wood, H.A. A synchronous peroral technique for the bioassay of insect viruses. *J. Invertebr. Pathol.* **1981**, *37*, 154–159. [CrossRef]
37. Hughes, D.S.; Possee, R.D.; King, L.A. Evidence for the presence of a low-level, persistent baculovirus infection of Mamestra brassicae insects. *J. Gen. Virol.* **1997**, *78*, 1801–1805. [CrossRef] [PubMed]

38. Hughes, D.S.; Possee, R.D.; King, L.A. Activation and detection of a latent baculovirus resembling *Mamestra brassicae* nuclear polyhedrosis virus in *M. brassicae* insects. *Virology* **1993**, *194*, 608–615. [CrossRef] [PubMed]
39. Brown, M.; Faulkner, P. Plaque assay of nuclear polyhedrosis viruses in cell culture. *Appl. Environ. Microbiol.* **1978**, *36*, 31–35. [PubMed]
40. BDBiosciences. Plaque Assay. Available online: http://www.bdbiosciences.com/br/resources/baculovirus/protocols/plaque_assay.jsp (accessed on 17 September 2012).
41. Hauxwell, I.C. *Evaluation of Potential Baculovirus Insecticides: Studies of the Infection Process and Host Susceptibility*; Imperial College London (University of London): London, UK, 1999.
42. Matsuura, Y.; Possee, R.D.; Overton, H.A.; Bishop, D.H. Baculovirus expression vectors: The requirements for high level expression of proteins, including glycoproteins. *J. Gen. Virol.* **1987**, *68*, 1233–1250. [CrossRef] [PubMed]
43. Doyle, C.J.; Hirst, M.L.; Cory, J.S.; Entwistle, P.F. Risk assessment studies: Detailed host range testing of wild-type cabbage moth, *Mamestra brassicae* (Lepidoptera: Noctuidae), nuclear polyhedrosis virus. *Appl. Environ. Microbiol.* **1990**, *56*, 2704–2710. [PubMed]
44. Zhang, H.; Yang, Q.; Qin, Q.L.; Zhu, W.; Zhang, Z.F.; Li, Y.N.; Zhang, N.; Zhang, J.H. Genomic sequence analysis of *Helicoverpa armigera* nucleopolyhedrovirus isolated from Australia. *Arch. Virol.* **2014**, *159*, 595–601. [CrossRef] [PubMed]
45. Arrizubieta, M.; Williams, T.; Caballero, P.; Simón, O. Selection of a nucleopolyhedrovirus isolate from *Helicoverpa armigera* as the basis for a biological insecticide. *Pest Manag. Sci.* **2014**, *70*, 967–976. [CrossRef] [PubMed]
46. Arrizubieta, M.; Simón, O.; Williams, T.; Caballero, P. A novel binary mixture of *Helicoverpa armigera* single nucleopolyhedrovirus genotypic variants has improved insecticidal characteristics for control of cotton bollworms. *Appl. Environ. Microbiol.* **2015**, *81*, 3984–3993. [CrossRef] [PubMed]
47. Arrizubieta, M.; Simón, O.; Williams, T.; Caballero, P. Genomic sequences of five *Helicoverpa armigera* nucleopolyhedrovirus genotypes from Spain that differ in their insecticidal properties. *Genome Announc.* **2015**, *3*, e00548-15. [CrossRef] [PubMed]
48. Chen, X.; IJkel, W.F.; Tarchini, R.; Sun, X.; Sandbrink, H.; Wang, H.; Peters, S.; Zuidema, D.; Lankhorst, R.K.; Vlak, J.M. The sequence of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus genome. *J. Gen. Virol.* **2001**, *82*, 241–257. [CrossRef] [PubMed]
49. Ardisson-Araújo, D.M.; Sosa-Gomez, D.R.; Melo, F.L.; Báo, S.N.; Ribeiro, B.M. Characterization of *Helicoverpa zea* single nucleopolyhedrovirus isolated in Brazil during the first old world bollworm (Noctuidae: *Helicoverpa armigera*) nationwide outbreak. *Virus Rev. Res.* **2015**, *20*, 2. [CrossRef]
50. Ayres, M.D.; Howard, S.C.; Kuzio, J.; Lopez-Ferber, M.; Possee, R.D. The complete DNA sequence of Autographa californica nuclear polyhedrosis virus. *Virology* **1994**, *202*, 586–605. [CrossRef] [PubMed]
51. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
52. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
53. Stöver, B.C.; Müller, K.F. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinf.* **2010**, *11*. [CrossRef] [PubMed]
54. Lange, M.; Wang, H.; Zhihong, H.; Jehle, J.A. Towards a molecular identification and classification system of lepidopteran-specific baculoviruses. *Virology* **2004**, *325*, 36–47. [CrossRef] [PubMed]
55. Ferrelli, M.; Taibo, C.; Fichetti, P.; Sciocco-Cap, A.; Arneodo, J. Characterization of a new *Helicoverpa armigera* nucleopolyhedrovirus variant causing epizootic on a previously unreported host, *Helicoverpa gelotopoeon* (Lepidoptera: Noctuidae). *J. Invertebr. Pathol.* **2015**, *138*, 89–93. [CrossRef] [PubMed]
56. Woo, S.-D.; Choi, J.Y.; Je, Y.H.; Jin, B.R. Characterization of the *Helicoverpa assulta* nucleopolyhedrovirus genome and sequence analysis of the polyhedrin gene region. *J. Biosci.* **2006**, *31*, 329–338. [CrossRef] [PubMed]
57. Li, Q.; Donly, C.; Li, L.; Willis, L.G.; Theilmann, D.A.; Erlandson, M. Sequence and organization of the Mamestra configurata nucleopolyhedrovirus genome. *Virology* **2002**, *294*, 106–121. [CrossRef] [PubMed]

58. Li, S.; Erlandson, M.; Moody, D.; Gillott, C. A physical map of the *Mamestra configurata* nucleopolyhedrovirus genome and sequence analysis of the polyhedrin gene. *J. Gen. Virol.* **1997**, *78 Pt 1*, 265–271. [CrossRef] [PubMed]
59. Asgari, S.; Davis, J.; Wood, D.; Wilson, P.; McGrath, A. Sequence and organization of the *Heliothis virescens* ascovirus genome. *J. Gen. Virol.* **2007**, *88*, 1120–1132. [CrossRef] [PubMed]
60. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
61. Morgulis, A.; Coulouris, G.; Raytselis, Y.; Madden, T.L.; Agarwala, R.; Schaffer, A.A. Database indexing for production MegaBLAST searches. *Bioinformatics* **2008**, *24*, 1757–1764. [CrossRef] [PubMed]
62. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinf.* **2009**, *10*, 421. [CrossRef] [PubMed]
63. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
64. Harrison, R.L. Genomic sequence analysis of the Illinois strain of the *Agrotis ipsilon* multiple nucleopolyhedrovirus. *Virus Genes* **2009**, *38*, 155–170. [CrossRef] [PubMed]
65. Simon, O.; Palma, L.; Beperet, I.; Munoz, D.; Lopez-Ferber, M.; Caballero, P.; Williams, T. Sequence comparison between three geographically distinct *Spodoptera frugiperda* multiple nucleopolyhedrovirus isolates: Detecting positively selected genes. *J. Invertebr. Pathol.* **2011**, *107*, 33–42. [CrossRef] [PubMed]
66. Parrish, J.T.; Peterson, F. Wind directions predicted from global circulation models and wind directions determined from eolian sandstones of the western United States—A comparison. *Sediment. Geol.* **1988**, *56*, 261–282. [CrossRef]
67. Trenberth, K.E.; Large, W.G.; Olson, J.G. The mean annual cycle in global ocean wind stress. *J. Phys. Oceanogr.* **1990**, *20*, 1742–1760. [CrossRef]
68. Parrish, J.T.; Curtis, R.L. Atmospheric circulation, upwelling, and organic-rich rocks in the Mesozoic and Cenozoic eras. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **1982**, *40*, 31–66. [CrossRef]

Chapter 6: MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified* that:

6. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
7. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
8. there are no other authors of the publication according to these criteria;
9. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
10. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

4. **Noune, C., & Hauxwell, C. (2017). MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data. Biology, 6(1), 14.**

Contributor	Statement of contribution*
Christopher Noune Signature: Date: 27/10/17	Performed experimental design, conducted laboratory analysis, data analysis and wrote the manuscript.
Caroline Hauxwell	Contributed to experimental design, data analysis, edited and reviewed manuscript.

Principal Supervisor Confirmation		
I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
Caroline Hauxwell	CHauxwell	27/10/17
Name	Signature	Date

6.1 METAGAAP: A NOVEL PIPELINE TO ESTIMATE COMMUNITY COMPOSITION AND ABUNDANCE FROM NON-MODEL SEQUENCE DATA

*For supplementary material refer to section 12.2. For source code refer to: https://github.com/CNoune/IMG_pipelines/tree/master/Legacy

biology

MDPI

Article

MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data

Christopher Noune and Caroline Hauxwell *

School of Earth, Environmental and Biological Sciences, Queensland University of Technology, Brisbane City QLD 4000, Australia; chris.noune@connect.qut.edu.au
* Correspondence: caroline.hauxwell@qut.edu.au; Tel.: +61-3-138-8062

Academic Editor: Jukka Finne
Received: 1 December 2016; Accepted: 7 February 2017; Published: 17 February 2017

Abstract: Next generation sequencing and bioinformatic approaches are increasingly used to quantify microorganisms within populations by analysis of 'meta-barcode' data. This approach relies on comparison of amplicon sequences of 'barcode' regions from a population with public-domain databases of reference sequences. However, for many organisms relevant 'barcode' regions may not have been identified and large databases of reference sequences may not be available. A workflow and software pipeline, 'MetaGaAP,' was developed to identify and quantify genotypes through four steps: shotgun sequencing and identification of polymorphisms in a metapopulation to identify custom 'barcode' regions of less than 30 polymorphisms within the span of a single 'read', amplification and sequencing of the 'barcode', generation of a custom database of polymorphisms, and quantitation of the relative abundance of genotypes. The pipeline and workflow were validated in a 'wild type' *Alphabaculovirus* isolate, *Helicoverpa armigera* single nucleopolyhedrovirus (HaSNPV-AC53) and a tissue-culture derived strain (HaSNPV-AC53-T2). The approach was validated by comparison of polymorphisms in amplicons and shotgun data, and by comparison of predicted dominant and co-dominant genotypes with Sanger sequences. The computational power required to generate and search the database effectively limits the number of polymorphisms that can be included in a barcode to 30 or less. The approach can be used in quantitative analysis of the ecology and pathology of non-model organisms.

Keywords: bioinformatics; baculoviruses; metapopulation; meta-barcoding; MetaGaAP; HaSNPV-AC53; community analysis

1. Introduction

Culture-independent molecular techniques to identify and quantify components of microbial communities have been facilitated by the use of next generation sequencing (NGS) [1,2].

Shotgun sequencing and whole or partial genome assembly uses algorithms comparing sequence data to public sequence databases (such as Genbank) [2–6]. 'Barcode' analysis uses PCR amplification of well-characterized regions (e.g., the 16S rRNA sub-unit in bacteria, internal transcribed space (ITS) of fungi or cytochrome oxidase) and comparison to sequence databases specific to those regions to determine taxonomic assignment and relative abundance of taxa in the community [2,7–11].

Both approaches are limited by available sequencing technology that relies on partial genome 'reads', and by the scope and accuracy of sequences in the reference databases. Shotgun sequencing and partial genome assembly is biased towards identification of dominant genotypes or taxa as a result of the limited read depth across multiple whole genomes [10,12,13]. Amplicon sequencing introduces bias resulting from gene copy number, selection of primers, and classification based on limited span of

the genome [2,7,12,14]. Both depend on reference databases which contain sequences from the small proportion of organisms that have been sequenced and variable standards of validation. Furthermore, non-model organisms, for which sequence databases are not available or for which marker regions have not been identified, require custom solutions. This is a particular issue in analysis of viral metapopulations and quasispecies [15].

Baculoviruses (*Baculoviridae*) are invertebrate-specific double-stranded DNA viruses with a genome of between 80 kb to 180kb [16]. The nucleopolyhedroviruses (*Alphabaculoviruses*) are known to contain high levels of genotypic and phenotypic diversity within a single isolate [17–21].

Previous studies on within-isolate diversity used techniques such as in vitro and in vivo isolation of sub-populations to identify strains [19,22–25], but such culture-dependent approaches themselves select a sub-set of strains that are adapted to the selection method, such as growth in tissue culture [19,26,27]. Molecular approaches include restriction fragment length polymorphism (RFLP) in combination with quantitative polymerase chain reaction (qPCR) [28–31], and denaturing gradient gel electrophoresis (DGGE) [32–34]. DGGE cannot be used reliably to quantify relative abundance and both qPCR and DGGE rely on primers that may not detect all variants [14,33–37].

Shotgun sequencing can be used to assemble a consensus sequence for an isolate containing multiple strains, and the same data can then be used to identify polymorphisms across the genome to determine the relative abundance of a single polymorphism [38–40]. Shotgun data can also be used to infer an approximate total number of strains within an isolate and the relative abundance of taxonomic clusters of strains within this population [13,19], but cannot determine the relative abundance of individual strains or abundance of strains that may contain multiple polymorphisms distributed across fragmented reads.

In this paper, we describe the application and validation of stepwise sequencing and a metabarcoding software pipeline to identify and quantify within-isolate strain variants within a baculovirus model.

2. Materials and Methods

2.1. Viruses

The baculovirus isolate HaSNPV-AC53 was obtained from AgBiTech Pty Ltd., passaged once in *H. armigera* larvae and DNA extracted as previously described [17,39].

The strain variant HaSNPV-AC53-T2 was derived from the AC53 wild type by passage in tissue culture and DNA extracted as previously described [19,41].

2.2. Identification of High Density Polymorphic Regions in Shotgun Data

DNA extraction from the HaSNPV-AC53 wild-type isolate, shotgun sequence generation using the Ion Torrent PGM, and assembly of a consensus sequence (Genbank accession: KJ909666) were completed as previously described [42]. The Genome Analysis Toolkit v3.5 (GATK) (Broad Institute, Cambridge, MA, USA) 'best practices' pipeline was used to identify substitutions, insertions and deletions (polymorphisms) in the shotgun data which were filtered to exclude those with a minimum genotype quality of below 60 (0.0001% error) and minimum read depth of 20x coverage [38]. Polymorphisms were annotated using Geneious R9.1.5 (Biomatters, Auckland, New Zealand) and snpEff 4.2 [43,44].

2.3. Amplicon Sequencing and Validation of Sequence Polymorphisms

Primers were designed to amplify custom 'barcode' regions of 325 and 365 bp (i.e., less than the span of a single Ion Torrent PGM read) within each of two ORFs with different polymorphism density (Table 1): Baculovirus Repeated ORF-A (BRO-A) and DNA polymerase.

Table 1. Primers used for amplification of selected regions within the ORFs BRO-A and DNA polymerase.

Target Gene	Primer	Fragment Size
BRO-A	* 5′-CATTTGCAAGGATATTGGAGT-3′ # 5′-AAGCTCGTTGGTTATCACAT-3′	365 bp
DNA Polymerase	* 5′-GTATGACTTATCACGACAATTGC-3′ # 5′-CGGTTTGCATATGTACTCTG-3′	325 bp

* An adapter, BarcodeX barcode adaptor and random hexamer is attached to the forward primer in the 5′ direction;
# trP1 adapter is attached to the reverse primer in the 5′ direction.

Both BRO-A and DNA Polymerase 'barcode' regions of the AC53 wild-type isolate and the BRO-A region of the HaSNPV-AC53-T2 strain were amplified from DNA using the Platinum Taq High Fidelity Super Mix kit (Life Technologies, Thermo-Fisher, Waltham, MA, USA) and an Eppendorf Pro S thermocycler (Eppendorf, Hamburg, Germany) as per the Platinum Taq standard method (Life Technologies, Thermo-Fisher, Waltham, MA, USA). NGS amplicon preparation and clean-up was completed as per the Life Technologies (Thermo-Fisher, Waltham, MA, USA) Ion Torrent PGM fusion primer manual. Shotgun sequencing was completed using an Ion Torrent PGM with a 318v2 chip and 400 bp chemistry.

Read quality was determined using FastQC 0.11.4 (Babraham Institute, Cambridge, UK) and any reads containing artefacts and/or quality less than Q20 were removed. Reads were trimmed to the expected amplicon size (Table 1) to remove primer regions using Fastx-toolkit 0.0.14 (Hannon Laboratory, Cold Spring Harbor, New York, NY, USA) [45,46]. Polymorphisms within the amplicon reads data were identified as described for the shotgun data and validated by comparison using vcf-compare within the VCFtools package (version 0.1.14) [47].

2.4. Sanger Sequencing

Both 'barcode' regions of the AC53 isolate and the BRO-A region of the HaSNPV-AC53-T2 strain were amplified using the forward primer in Table 1, the Mango Taq kit (Bioline, Meridian Bioscience, Cincinnati, OH, USA) and an Eppendorf Pro S thermocycler (Eppendorf, Hamburg, Germany). PCR products were then cleaned using an Isolate II PCR clean-up kit (Bioline, Alexandria, Australia) and labelled using a Big Dye Terminator (BDT) v3.1 kit (Applied Biosystems, Thermo-Fisher, Waltham, MA, USA). Labelled products were then precipitated using EDTA/ethanol as per the BDT v3.1 kit insert. Products were then sequenced using an ABI 3500 Genetic Analyzer (Applied Biosystems, Thermo-Fisher, Waltham, MA, USA).

2.5. Genotyping and Abundance Pipeline

Amplicon reads were mapped to the relevant consensus sequence for the 2 ORFs in the HaSNPV-AC53 genome and the BRO-A sequence for the HaSNPV-AC53-T2 strain using the BWA mem 0.7.12 algorithm with default settings to produce unsorted SAM files [48]. These unsorted SAM files were converted to sorted BAM files using SAMtools 1.3 [49]. The BAM headers were then corrected and the reference sequences and BAM files were indexed and a sequence dictionary was produced using samtools 1.3 and picard-tools 2.5.0 (Broad Institute, Cambridge, Massachusetts, USA) [50].

Polymorphisms were identified within the genome using the GATK HaplotypeCaller to produce a 'genomic variant call format' (gVCF) file with the following parameters: maximum read depth per site of 300,000 reads, 100 maximum alternate alleles per site, genotyping mode set to 'discovery',down-sampling set to 'none' and emit reference confidence set to 'GVCF'. These files were then sorted to genotype and converted to a standard 'variant call format' (VCF) file using the GATK GenotypeGVCFs tool and hard-filtered using the GATK VariantFiltration tool to include only polymorphisms with a genotype quality (GQ) score greater than 60 (minimum 0.0001% error on a Phred scale). The final VCF file containing all the filtered polymorphisms within each ORF and the consensus

sequence of each ORF were imported respectively into the Biostars 175929 tool as part of the JVarkit package [51] to produce a compressed fasta file database for each amplicon containing generated references sequences with every possible combination of identified polymorphisms (Figure 1). All generated reference sequences were renamed using the BBmap Renamer tool [52] to include the ORF from which they were derived and a numerical identification number.

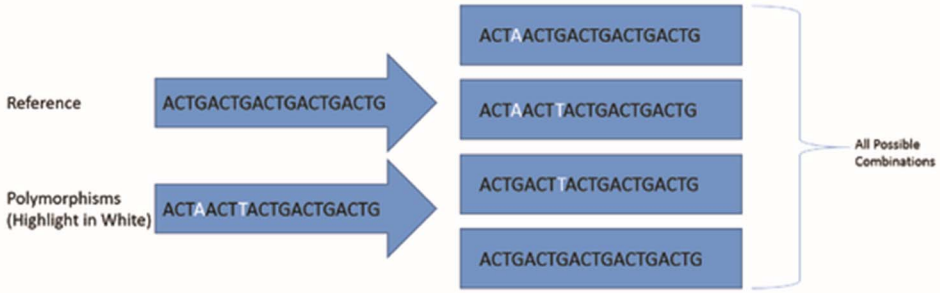


Figure 1. A visual representation of how the Biostars 175929 tool produces sequences containing all polymorphism combinations.

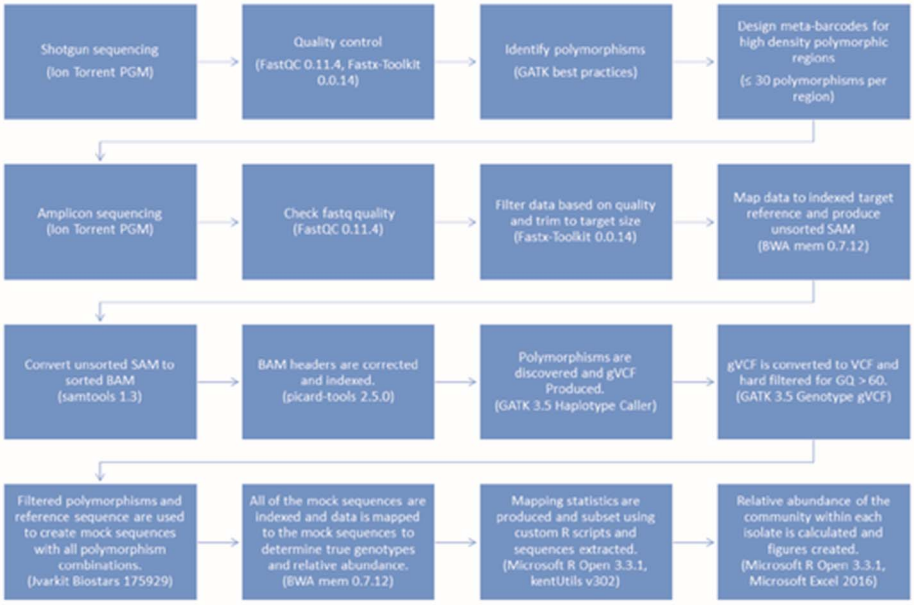


Figure 2. The MetaGaAP workflow to identify genotypes and the relative abundance of the community composition within a single isolate.

The generated reference sequences were then indexed and used as the references to map amplicon reads using the BWA mem 0.7.12 algorithm with default settings to produce a SAM file [48]. The SAM files were then imported into Tablet 1.15.09.01 for visual and statistical comparison of the mapped amplicon sequence reads and the generated reference sequences [53,54]. The mapping statistics were produced using samtools 1.3 and sequences that contained less than 20x coverage (equivalent to a 1% error on a phred scale) or sequences with imperfect mapping (containing gaps) were excluded using kentUtils (version 302) and custom R scripts built with Microsoft R Open 3.3.1 [55–57]. Relative abundances of each identified genotype were calculated using Microsoft R Open 3.3.1 (Microsoft, Redmond, WA, USA) [55,57].

The pipeline was coded using Bash and Microsoft R Open 3.3.1 with a text-based interface to improve versatility and ease of use and named the Meta-barcoding Genotyping and Abundance Pipeline (MetaGaAP) [58]. A schematic of the pipeline is presented in Figure 2.

2.6. Comparison of amplicon and Sanger sequences

Genotype sequences identified using MetaGaAP and chromatograms from Sanger sequencing were visualized using Geneious R9.1.5 and aligned using MAFFT v7.222 (Kyoto University, Kyoto, Japan) with the FFT-NS-2 algorithm and default settings [59]. The dominant genotype and abundant minor genotypes predicted from the mapped NGS amplicon sequences were compared visually at each predicted SNP locus with the Sanger chromatographs.

3. Results

3.1. Identification of Polymorphisms in Shotgun Sequence Data

A total of 438 polymorphisms were identified within the Ion Torrent PGM shotgun dataset of the AC53 isolate, equivalent to 1 nucleotide change every 297 bases. Within the 139 ORFs in the AC53 consensus genome sequence, 37 ORFs contained no polymorphisms and 102 ORFs contained polymorphisms. Polymorphisms were identified within exons, intergenic regions and all five homologous repeat (Hr) regions (Table S1): 53 were insertions, 339 were deletions and 46 were substitutions. Most ORFs contained 9 or fewer polymorphisms. The ORF with the highest number of polymorphisms was BRO-A (30) and had a mix of substitutions, insertions and deletions and was selected for amplicon sequencing. DNA polymerase contained 5 polymorphisms across the entire 3 kb ORF: a 325 bp region within the ORF that contained no polymorphisms was selected as the negative control.

3.2. Validation by Comparison of Amplicon Sequence Variants to Shotgun Sequence Data

AC53 shotgun sequencing predicted 25 polymorphism in the targeted 'barcode' region within BRO-A. All 25 polymorphisms were identified in the amplicon sequences (Figure 3). No polymorphisms were detected in the amplicon data of the DNA polymerase region, as predicted from the shotgun data (Figure S1). A single polymorphism (an 'A' substitution at position 293) was detected in the BRO-A amplicon sequences of the derived strain AC53-T2. This polymorphism was confirmed as one of the 25 polymorphisms in AC53 wild-type isolate shotgun and amplicon data.

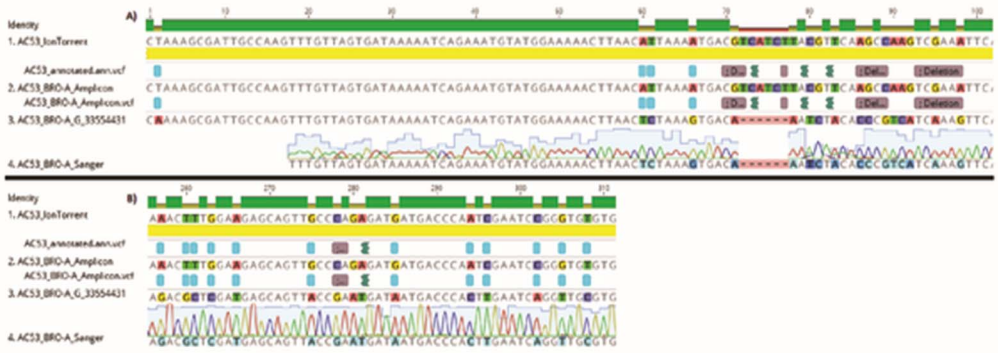


Figure 3. Fragments A and B of the HaSNPV-AC53 BRO-A target region containing the identified polymorphisms, showing identical alignment of polymorphisms in the amplicon, shotgun and Sanger sequences of the dominant BRO-A variant identified by MetaGaAP within the wild type baculovirus isolate HaSNPV-AC53.

3.3. Genotype Sequence Construction, Abundance Mapping and Validation by Sanger Sequencing

The 25 polymorphisms identified in the BRO-A amplicon 'barcode' region of isolate AC53 generated 3.4×10^7 possible combinations of polymorphisms in reference sequences in the custom database. Mapping of the amplicon sequences data to this database identified 329 of these possible sequences were present in the amplicon sequencing, with a minimum of 1 read mapping to them. Of these, 28 amplicon sequences with between 21× and 258,084× coverage were identified (Table 2). Genotype abundance was estimated from the number of reads mapping to each of these 28 hypothetical variants.

Table 2. Relative abundance of the identified AC53 BRO-A community composition that were above the 20x coverage threshold with G_33554431 identified as the dominant strain in the population.

Genotype	Reads	Relative Abundance %
G_33554431 #	258084	97.03
G_33554303	1643	0.62
G_33552383	787	0.30
G_16777215	666	0.25
G_33554423	533	0.20
G_25165823	437	0.16
G_33554430	437	0.16
G_33292287	400	0.15
G_31457279	393	0.15
G_33554429	261	0.10
G_33554399	228	0.09
G_33554427	213	0.08
G_33553919 *	138	0.05
G_33554175	129	0.05
G_33546239	123	0.05
G_33554367	105	0.04
G_29360127	103	0.04
G_33030143	103	0.04
G_33550335	92	0.03
G_33552255	68	0.03
G_33521663	62	0.02
G_33554415	56	0.02
G_33554428	55	0.02
G_20971519	52	0.02
G_33553407	48	0.02
G_23068671	35	0.01
G_33554239	28	0.01
G_33538047	21	0.01

Equivalent to the AC53-T2 BRO-A G_1; * Equivalent to the AC53-T2 BRO-A G_0.

Genotype G_33554431 accounted for 97% of the reads and was thus predicted to be the dominant genotype, while the second most abundant genotype (G_33554303) accounted for 0.62% of the reads (Table 2). The dominant genotype G_33554431 was confirmed by Sanger sequencing, with 100% sequence similarity (Figure 3).

The single polymorphism detected in the BRO-A amplicon sequences of the tissue-culture derived strain AC53-T2 resulted in generation of two reference sequences: with the A substitution or without the substitution (i.e., with a T). Mapping of amplicon sequence data to the two reference sequences showed that both were present in similar abundance: the T genotype accounted for 54% of reads and the A genotype for 46% of reads (Table 3). The dominant T genotype of the derived strain AC53-T2 had 100% sequence similarity with the dominant G_33554431 genotype of the AC53 wild type isolate. The minor A genotype of AC53-T2 had 100% homology to genotype G_33553919 of the AC53 wild type isolate, which accounted for only 0.05% of the reads in the wild type isolate (Table 3).

The Sanger chromatogram detected both genotypes in strain AC53-T2, with both A and T detected in approximately equal intensity in the chromatogram at position 293 (Figure 4).

Table 3. Relative abundance of the two BRO-A genotypes within AC53-T2 BRO-A.

Genotype	Reads	Relative Abundance %
AC53-T2 BRO-A G_1	104,065	54.27
AC53-T2 BRO-A G_0	87,689	45.73

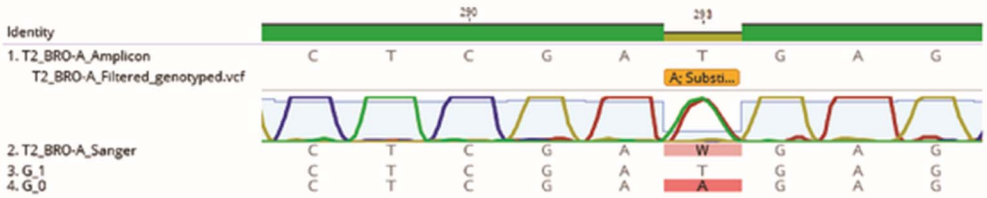


Figure 4. Comparison of the AC53-T2 reference sequence to the Sanger sequence and the two identified genotypes within HaSNPV-AC53-T2. The Sanger chromatogram at position 293 shows the two competing genotypes which were identified with MetaGaAP and validates relative abundance result.

The single genotype predicted by the amplicon sequencing of DNA polymerase was confirmed to have 100% homology with the Sanger sequencing (Figure S1).

4. Discussion

Shotgun sequencing and identification of polymorphisms was used to identify of custom 'barcode' regions in the viral metapopulation of the wild type baculovirus isolate HaSNPV-AC53. Hr and non-coding regions were excluded to reduce possible primer bias and sequencing errors. The highest number of polymorphisms was identified in an AT-rich region of the BRO-A ORF. The high abundance of mutations in BRO-A has been previously described in published whole-genome sequences of SNPV isolates from *Helicoverpa spp.* [19,60,61] and a polymorphism rich 'custom barcode' region was selected within the ORF [62–64].

In contrast, only five polymorphisms were identified across the entire 3 kbp DNA polymerase ORF. A previous study using a different HaSNPV isolate identified 60 polymorphisms in the DNA polymerase ORF using 454 pyrosequencing with only 30× coverage, but the authors had expected DNA polymerase to be much more highly conserved [33]. Our results support this expectation and we suggest that the low coverage of the 454 pyrosequencing may have led to overestimation of polymorphisms in that study [65–67].

Comparison of the amplicon sequence data identified the same 25 polymorphisms in BRO-A and absence of polymorphisms in DNA polymerase, as predicted from the shotgun sequence data of HaSNPV-AC53. In contrast, a single polymorphism was predicted in the amplicon data of the tissue culture derived strain AC53-T2, which was also confirmed as one of the 25 polymorphisms predicted from the shotgun data of the parent isolate. This validates the use of amplicon and shotgun sequence to compare polymorphisms using the GATK best practices guidelines [38–40,68].

Comparison of amplicon data with the database of all possible combinations of polymorphisms using MetaGaAP identified 28 variants within the HaSNPV-AC53 wild type viral metapopulation at the level of 20× read coverage. A dominant variant within the wild type HaSNPV-AC53 accounted for 97% of the population. In contrast, two variants of approximately equal abundance were identified in the derived strain AC53-T2. The slightly more abundant variant in AC53-T2 had 100% sequence similarity to the dominant variant in the parent isolate, and the other variant had 100% sequence similarity to a minor variant accounting for 0.05% of abundance in the parent isolate. This supports the sensitivity of MetaGaAP to detect and identify minor variants as low as 129× coverage. We suggest including

strains with a minimum 20× coverage threshold (to eliminate 'false positives' due to sequencing error). However, the presence of minor genotypes with coverage below 129× would require confirmation by, for example, detection in multiple deep sequencing of the isolate during different stages of infection, or large scale sequence or virus cloning and characterisation.

Sanger sequencing is the 'gold-standard' for validation of NGS datasets and has the lowest error rates [69]. Sanger sequencing confirmed the identification of the predicted dominant variant in both the BRO-A and DNA Polymerase amplicons of the HaSNPV wild type metapopulation. Furthermore, Sanger sequencing detected both the predicted variants within the derived strain AC53-T2 in the approximately equal proportions calculated by MetaGaAP. This confirmed the validity both of the identification of variants and the calculation of their relative abundance by MetaGaAP.

Current tools for 16S based taxonomic classification of clinical isolates use either pairwise or non-pairwise alignments to a very limited set of sequences from culture collections. Most meta-barcode analyses of microbial communities use partial regions of 16S and 18S ribosomal RNA and, to a lesser degree, the ITS region of fungi, while 'barcodes' for viruses are limited to a few significant virus types such as small RNA viruses [1–3,6,7,15,70,71]. These approaches are primarily used for taxonomic classification and rely on either phylogenetic clustering or alignment scores in comparison to sequences in reference databases such as Greengenes for 16S [9,72–76] However, these approaches are limited by errors such as submission of misannotated sequences or identification based on short or partial sequences, in addition to the limited sequence availability for non-model organisms [77–79].

5. Conclusions

MetaGaAP accurately identified and estimated abundance of variants in a virus metapopulation by generating a custom database from sequence data and comparison with ultra-deep sequencing of amplicons of novel, polymorphism-rich 'barcode' regions in the viral metagenome. However, the computer data handling and processing time increases as the number of polymorphisms increases and the number of possible combinations generated in the database increases by 2^y, where y = number of polymorphisms. The application is thus practically limited regions with 30 or fewer polymorphisms.

Despite this limitation, MetaGaAP has potential application in analysis of community composition where suitable reference sequence databases are not available, complete or accurately assigned, and can be used to identify and quantify strain variants in pathology, ecology and evolutionary studies without the need for viral cloning. MetaGaAP is publicly available for download on GitHub [58].

Supplementary Materials: The following are available online at www.mdpi.com/2079-7737/6/1/14/s1, Figure S1: Comparison of the AC53 DNA polymerase Sanger sequence and the AC53 DNA polymerase reference sequence showing 100% nucleotide identity and no polymorphisms identified., Table S1: Polymorphisms detected within ORFs. BRO-A has the highest number of polymorphisms (30) and HOAR and P74 have the second highest (13).

Acknowledgments: This work was funded by Queensland University of Technology, the Cotton Research Development Corporation and an Australian Government Research Training Program Scholarship. We would like to acknowledge the support of AgBiTech Pty.Ltd in supplying insects and the virus isolate and to thank staff of the Molecular Genetics Research Facility and the Invertebrate & Microbiology Group at QUT for their assistance with sequencing and technical support. Some of the data reported in this paper was obtained at the Central Analytical Research Facility (CARF) operated by the Institute for Future Environments (QUT). Access to CARF is supported by generous funding from the Science and Engineering Faculty (QUT).

Conflicts of Interest: The authors declare a conflict of interest. The Cotton Research Development Corporation has funded the work by C. Noune through a post-graduate student scholarship. AgBiTech Pty Ltd. Provided the sample of HaSNPV-AC53 and previously funded consultancy and research work with C. Hauxwell but did not contribute financially to this study.

Software and Dataset Availability: MetaGaAP is available for download at https://github.com/CNoune/IMG_pipelines. Genotypes described in this paper are available for download at https://researchdatafinder.qut.edu.au/display/n14806.

References

1. Gilbert, J.A.; Dupont, C.L. Microbial metagenomics: Beyond the genome. *Annu. Rev. Mar. Sci.* **2011**, *3*, 347–371. [CrossRef]
2. Oulas, A.; Pavloudi, C.; Polymenakou, P.; Pavlopoulos, G.A.; Papanikolaou, N.; Kotoulas, G.; Arvanitidis, C.; Iliopoulos, I. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights* **2015**, *9*, 75–88.
3. Sharpton, T.J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **2014**, *5*. [CrossRef] [PubMed]
4. Xia, L.C.; Cram, J.A.; Chen, T.; Fuhrman, J.A.; Sun, F. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS ONE* **2011**, *6*, e27992. [CrossRef] [PubMed]
5. Chen, E.Z.; Bushman, F.D.; Li, H. A model-based approach for species abundance quantification based on shotgun metagenomic data. *Stat. Biosci.* **2016**. [CrossRef]
6. Kunin, V.; He, S.; Warnecke, F.; Peterson, S.B.; Martin, H.G.; Haynes, M.; Ivanova, N.; Blackall, L.L.; Breitbart, M.; Rohwer, F. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.* **2008**, *18*, 293–297. [CrossRef] [PubMed]
7. Sanschagrin, S.; Yergeau, E. Next-generation sequencing of 16S ribosomal RNA gene amplicons. *J. Vis. Exp.* **2014**, *29*, e51709. [CrossRef] [PubMed]
8. Brittnacher, M.J.; Heltshe, S.L.; Hayden, H.S.; Radey, M.C.; Weiss, E.J.; Damman, C.J.; Zisman, T.L.; Suskind, D.L.; Miller, S.I. Gutss: An alignment-free sequence comparison method for use in human intestinal microbiome and fecal microbiota transplantation analysis. *PLoS ONE* **2016**, *11*, e0158897. [CrossRef] [PubMed]
9. Yu, D.W.; Ji, Y.; Emerson, B.C.; Wang, X.; Ye, C.; Yang, C.; Ding, Z. Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* **2012**, *3*, 613–623. [CrossRef]
10. Kõljalg, U.; Nilsson, R.H.; Abarenkov, K.; Tedersoo, L.; Taylor, A.F.; Bahram, M.; Bates, S.T.; Bruns, T.D.; Bengtsson-Palme, J.; Callaghan, T.M. Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **2013**, *22*, 5271–5277. [CrossRef] [PubMed]
11. Janssen, P.H. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl. Environ. Microbiol.* **2006**, *72*, 1719–1728. [CrossRef] [PubMed]
12. Tedersoo, L.; Anslan, S.; Bahram, M.; Põlme, S.; Riit, T.; Liiv, I.; Kõljalg, U.; Kisand, V.; Nilsson, H.; Hildebrand, F. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* **2015**, *10*, 1–43. [CrossRef]
13. Chateigner, A.; Bézier, A.; Labrousse, C.; Jiolle, D.; Barbe, V.; Herniou, E.A. Ultra deep sequencing of a baculovirus population reveals widespread genomic variations. *Viruses* **2015**, *7*, 3625–3646. [PubMed]
14. Sipos, R.; Székely, A.; Révész, S.; Márialigeti, K. Addressing PCR biases in environmental microbiology studies. *Bioremediat. Methods Protoc.* **2010**, *599*, 37–58.
15. McElroy, K.; Thomas, T.; Luciani, F. Deep sequencing of evolving pathogen populations: Applications, errors, and bioinformatic solutions. *Microb. Inform. Exp.* **2014**, *4*, 1–14. [CrossRef] [PubMed]
16. Rohrmann, G. Introduction to the Baculoviruses and Their Taxonomy. In *Baculovirus Molecular Biology*; National Center for Biotechnology Information: Bethesda, MD, USA, 2011.
17. Rowley, D.L.; Popham, H.J.R.; Harrison, R.L. Genetic variation and virulence of nucleopolyhedroviruses isolated worldwide from the heliothine pests Helicoverpa armigera, Helicoverpa zea, and Heliothis virescens. *J. Invertebr. Pathol.* **2011**, *107*, 112–126. [CrossRef] [PubMed]
18. Van Oers, M.M.; Vlak, J.M. Baculovirus Genomics. *Curr. Drug Targets* **2007**, *8*, 1051–1068. [CrossRef]
19. Noune, C.; Hauxwell, C. Comparative analysis of HaSNPV-AC53 and derived strains. *Viruses* **2016**, *8*, 280–297. [CrossRef] [PubMed]
20. Vignuzzi, M.; Stone, J.K.; Arnold, J.J.; Cameron, C.E.; Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **2006**, *439*, 344–348. [CrossRef] [PubMed]
21. Cory, J.S.; Green, B.M.; Paul, R.K.; Hunter-Fujita, F. Genotypic and phenotypic diversity of a baculovirus population within an individual insect host. *J. Invertebr. Pathol.* **2005**, *89*, 101–111. [CrossRef] [PubMed]

22. Brown, M.; Faulkner, P. A plaque assay for nuclear polyhedrosis viruses using a solid overlay. *J. Gen. Virol.* **1977**, *36*, 361–364. [CrossRef]
23. Graillot, B.; Berling, M.; Blachere-López, C.; Siegwart, M.; Besse, S.; López-Ferber, M. Progressive adaptation of a CpGV isolate to codling moth populations resistant to CpGV-M. *Viruses* **2014**, *6*, 5135–5144. [CrossRef] [PubMed]
24. Vanarsdall, A.L.; Okano, K.; Rohrmann, G.F. Characterization of the replication of a baculovirus mutant lacking the DNA polymerase gene. *Virology* **2005**, *331*, 175–180. [CrossRef] [PubMed]
25. Redman, E.M.; Wilson, K.; Cory, J.S. Trade-offs and mixed infections in an obligate-killing insect pathogen. *J. Anim. Ecol.* **2016**, *85*, 1200–1209. [CrossRef] [PubMed]
26. Simon, O.; Palma, L.; Beperet, I.; Munoz, D.; Lopez-Ferber, M.; Caballero, P.; Williams, T. Sequence comparison between three geographically distinct Spodoptera frugiperda multiple nucleopolyhedrovirus isolates: Detecting positively selected genes. *J. Invertebr. Pathol.* **2011**, *107*, 33–42. [CrossRef] [PubMed]
27. Harrison, R.L. Genomic sequence analysis of the Illinois strain of the Agrotis ipsilon multiple nucleopolyhedrovirus. *Virus Genes* **2009**, *38*, 155–170. [CrossRef] [PubMed]
28. Christian, P.D.; Gibb, N.; Kasprzak, A.B.; Richards, A. A rapid method for the identification and differentiation of *Helicoverpa* nucleopolyhedroviruses (NPV *Baculoviridae*) isolated from the environment. *J. Virol. Methods* **2001**, *96*, 51–65. [CrossRef]
29. Lightner, D.V.; Redman, R.M.; Bell, T.A. Observations on the geographic distribution, pathogenesis and morphology of the baculovirus from Penaeus monodon Fabricius. *Aquaculture* **1983**, *32*, 209–233. [CrossRef]
30. Crawford, A.M.; Zelazny, B.; Alfiler, A.R. Genotypic variation in geographical isolates of oryctes baculovirus. *J. Gen. Virol.* **1986**, *67*, 949–952. [CrossRef]
31. Gettig, R.R.; McCarthy, W.J. Genotypic variation among wild isolates of Heliothis spp nuclear polyhedrosis viruses from different geographical regions. *Virology* **1982**, *117*, 245–252. [CrossRef]
32. Baillie, V.L.; Bouwer, G. High levels of genetic variation within Helicoverpa armigera nucleopolyhedrovirus populations in individual host insects. *Arch. Virol.* **2012**, *157*, 2281–2289. [CrossRef] [PubMed]
33. Baillie, V.L.; Bouwer, G. High levels of genetic variation within core *Helicoverpa armigera* nucleopolyhedrovirus genes. *Virus Genes* **2012**, *44*, 149–162. [CrossRef] [PubMed]
34. Baillie, V.L.; Bouwer, G. Development of highly sensitive assays for detection of genetic variation in key Helicoverpa armigera nucleopolyhedrovirus genes. *J. Virol. Methods* **2011**, *178*, 179–185. [CrossRef] [PubMed]
35. Neilson, J.W.; Jordan, F.L.; Maier, R.M. Analysis of artifacts suggests DGGE should not be used for quantitative diversity analysis. *J. Microbiol. Methods* **2013**, *92*, 256–263. [CrossRef] [PubMed]
36. Lueders, T.; Friedrich, M.W. Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. *Appl. Environ. Microbiol.* **2003**, *69*, 320–326. [CrossRef]
37. Schloss, P.D.; Gevers, D.; Westcott, S.L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **2011**, *6*, e27310. [CrossRef] [PubMed]
38. Van Der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**. [CrossRef]
39. Yu, X.; Sun, S. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinform.* **2013**, *14*, 274. [CrossRef] [PubMed]
40. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]
41. Noune, C.; Hauxwell, C. Complete genome sequences of seven helicoverpa armigera SNPV-AC53-Derived strains. *Genome Announc.* **2016**, *4*. [CrossRef] [PubMed]
42. Noune, C.; Hauxwell, C. Complete genome sequences of helicoverpa armigera single nucleopolyhedrovirus strains AC53 and H25EA1 from Australia. *Genome Announc.* **2015**, *3*. [CrossRef] [PubMed]
43. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef] [PubMed]

44. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [CrossRef] [PubMed]
45. Andrews, S. FASTQC: A Quality Control Tool for High Throughput Sequence Data. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 29 September 2014).
46. Gordon, A.; Hannon, G.J. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. 2010, unpublished work.
47. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef] [PubMed]
48. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. Available online: https://arxiv.org/abs/1303.3997 (accessed on 26 May 2013).
49. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
50. Institute, B. Picard. Available online: http://broadinstitute.github.io/picard/ (accessed on 9 September 2016).
51. Pierre, L. JVarkit: Java Utilities for Bioinformatics. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.822.1547&rep=rep1&type=pdf (accessed on 26 May 2015).
52. Bushnell, B. BBMap Short Read Aligner. Available online: http://sourceforge.net/projects/bbmap (accessed on 18 September 2016).
53. Milne, I.; Stephen, G.; Bayer, M.; Cock, P.J.A.; Pritchard, L.; Cardle, L.; Shaw, P.D.; Marshall, D. Using tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **2013**, *14*, 193–202. [CrossRef] [PubMed]
54. Milne, I.; Bayer, M.; Cardle, L.; Shaw, P.; Stephen, G.; Wright, F.; Marshall, D. Tablet-next generation sequence assembly visualization. *Bioinformatics* **2010**, *26*, 401–402. [CrossRef] [PubMed]
55. Microsoft R Open. Available online: https://mran.revolutionanalytics.com/rro/ (accessed on 6 May 2016).
56. Kent, J. kentUtils. Available online: https://github.com/ENCODE-DCC/kentUtils (accessed on 12 September 2014).
57. Team, R.C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
58. Noune, C. The Invertebrates & Microbiology Group Pipelines, GitHub, Queensland University of Technology. Available online: https://github.com/CNoune/IMG_pipelines (accessed on 5 September 2016).
59. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
60. Chen, X.; Zhang, W.-J.; Wong, J.; Chun, G.; Lu, A.; McCutchen, B.; Presnail, J.; Herrmann, R.; Dolan, M.; Tingey, S.; et al. Comparative analysis of the complete genome sequences of Helicoverpa zea and Helicoverpa armigera single-nucleocapsid nucleopolyhedroviruses. *J. Gen. Virol.* **2002**, *83*, 673–684. [CrossRef] [PubMed]
61. Chen, X.; IJkel, W.F.; Tarchini, R.; Sun, X.; Sandbrink, H.; Wang, H.; Peters, S.; Zuidema, D.; Lankhorst, R.K.; Vlak, J.M. The sequence of the Helicoverpa armigera single nucleocapsid nucleopolyhedrovirus genome. *J. Gen. Virol.* **2001**, *82*, 241–257. [CrossRef] [PubMed]
62. Nelson, M.R.; Marnellos, G.; Kammerer, S.; Hoyal, C.R.; Shi, M.M.; Cantor, C.R.; Braun, A. Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Res.* **2004**, *14*, 1664–1668. [CrossRef] [PubMed]
63. Piepho, H.-P. Optimal marker density for interval mapping in a backcross population. *Heredity* **2000**, *84*, 437–440. [CrossRef] [PubMed]
64. Beissinger, T.M.; Hirsch, C.N.; Sekhon, R.S.; Foerster, J.M.; Johnson, J.M.; Muttoni, G.; Vaillancourt, B.; Buell, C.R.; Kaeppler, S.M.; De Leon, N. Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* **2013**, *193*, 1073–1081. [CrossRef] [PubMed]
65. Gilles, A.; Meglécz, E.; Pech, N.; Ferreira, S.; Malausa, T.; Martin, J.-F. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genom.* **2011**, *12*, 245–255. [CrossRef] [PubMed]
66. Luo, C.; Tsementzi, D.; Kyrpides, N.; Read, T.; Konstantinidis, K.T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* **2012**, *7*, e30087. [CrossRef]
67. Van Dijk, E.L.; Auger, H.; Jaszczyszyn, Y.; Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **2014**, *30*, 418–426. [CrossRef] [PubMed]

68. Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom.* **2012**, *13*, 341–353. [CrossRef] [PubMed]
69. Hoff, K.J. The effect of sequencing errors on metagenomic gene prediction. *BMC Genom.* **2009**, *10*, 520–528. [CrossRef] [PubMed]
70. Schoch, C.L.; Seifert, K.A.; Huhndorf, S.; Robert, V.; Spouge, J.L.; Levesque, C.A.; Chen, W.; Bolchacova, E.; Voigt, K.; Crous, P.W. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl. Acad. Sci. USA* **2012**, *109*, 6241–6246. [CrossRef] [PubMed]
71. Prosperi, M.C.; Prosperi, L.; Bruselles, A.; Abbate, I.; Rozera, G.; Vincenti, D.; Solmone, M.C.; Capobianchi, M.R.; Ulivi, G. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinform.* **2011**, *12*, 5–17. [CrossRef] [PubMed]
72. Puente-Sánchez, F.; Aguirre, J.; Parro, V. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic Acids Res.* **2016**, *44*, e40. [CrossRef] [PubMed]
73. Caporaso, J.G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.D.; Costello, E.K.; Fierer, N.; Pena, A.G.; Goodrich, J.K.; Gordon, J.I.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7*, 335–336. [CrossRef] [PubMed]
74. DeSantis, T.Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E.L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G.L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **2006**, *72*, 5069–5072. [CrossRef] [PubMed]
75. Cole, J.R.; Wang, Q.; Fish, J.A.; Chai, B.; McGarrell, D.M.; Sun, Y.; Brown, C.T.; Porras-Alfaro, A.; Kuske, C.R.; Tiedje, J.M. Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **2013**, *42*, D633–D642. [CrossRef] [PubMed]
76. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef] [PubMed]
77. Clarridge, J.E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* **2004**, *17*, 840–862. [CrossRef] [PubMed]
78. Mignard, S.; Flandrois, J. 16S rRNA sequencing in routine bacterial identification: A 30-month experiment. *J. Microbiol. Methods* **2006**, *67*, 574–581. [CrossRef] [PubMed]
79. Werner, J.J.; Koren, O.; Hugenholtz, P.; DeSantis, T.Z.; Walters, W.A.; Caporaso, J.G.; Angenent, L.T.; Knight, R.; Ley, R.E. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J.* **2012**, *6*, 94–103. [CrossRef] [PubMed]

6.2 THE PORTING OF METAGAAP TO PYTHON (METAGAAP-PY)

*For source code refer to: https://github.com/CNoune/IMG_pipelines/tree/master/MetaGaAP-Py

#Published as a white-paper in *bioRxiv* – see appendix B.

6.2.1 Introduction

In the previous section, the publishing and development of the MetaGaAP was described as a new method for the analysis of non-model systems and publicly released. Since the publishing of the paper and releasement of the pipeline fundamental flaws in the usability, lack of cross-platform operating system compatibility and the limited capabilities of the Bash and R programming languages have been identified. In addition, issues such as the large dependency requirements (as highlighted by a reviewer during the peer-review process) to utilise MetaGaAP have hampered the adoption of the pipeline.

During the peer-review process and since publication, an undertaking to port MetaGaAP to the Python programming language began. This aimed to improve usability, reduce the number of dependencies required to execute the pipeline and improve computational efficiency wherever possible.

6.2.2 Method and Implementations

The Python version 3.6 programming language was selected as the language of choice as it is a non-compiled language (interpreted language) and is backwards compatible with all versions of Python 3. Implementing Python required MetaGaAP to be re-coded from scratch to ensure no redundant Bash or R code would be carried over. The coding was completed using the Anaconda 3.6.1 environment and the Spyder 3.1.4 integrated developer environment ("Anaconda Software Distribution," 2017; R. Pierre, 2009).

A necessary core set of packages and dependencies were kept: BWA, Samtools, GATK, fastx-toolkit, Biostars175929, Oracle Java 1.8, awk, sed, cat and picard-tools. The BBmap renamer and duplicate sequence removal tools, kentUtils, zenity and the R coded back-end scripts Subset_Stats.R and Seq_List.R were discarded in favour of pure Python implementations coded directly into the MetaGaAP source code (Table 6-1). Furthermore, the implemented Python packages except for BioPython are all natively installed at the same time Python 3.6 is installed and removes the need to write an installation script.

In addition to the replaced packages and scripts, new features were implemented such as multi-reference, multi-sample analysis and single-reference, multi-sample analysis, automatic creation of directories, multi-threaded processing of sequence duplicate removal (Peter JA. Cock, 2010), sequence combinations database compression by converting multi-line fasta sequences to a single-line fasta sequences (L. Pierre, 2015), automatic sequence length and read number counting and garbage collection to optimise 'Random Access Memory' (RAM) usage.

Table 6-1: Python implemented packages in MetaGaAP.

Package/Library	Function
TKinter	Implements a simple graphical user interface for file selection.
BioPython (Peter JA Cock et al., 2009)	A Python library to manipulate fasta sequences.
The Python Data Analysis Library	Removes the need to use R code.
Multiprocessing	A standard Python library to implement multi-threading.
Sys	Captures operating system type i.e. Windows, Linux or Mac. This package was implemented to fix an issue in multi-threading on a Linux system.
Garbage Collection	Optimises RAM utilisation.
Getpass	Captures user information.
OS	Basic Python functions which allow Python to communicate some functions with the operating system.
Subprocess	Python functions which allow for the execution of non-Python packages.

6.2.3 Discussion

The various improvements implemented and porting to Python has produced a highly-refined iteration of MetaGaAP that has an optimised workflow. Less user-interactions have been implemented as MetaGaAP now assumes or pre-calculates a lot of parameters such as target depth and sequence length which improves the overall experience of running MetaGaAP as it can be 'set and forget'. In addition, RAM optimisations and the implementation of a Python native, multi-threaded duplicate sequence removal tool has reduced the required amount of RAM needed and overall processing time but is dependent on how many cores the central processing unit has, whether the storage unit is a solid-state drive or a mechanical hard-disk drive and how large the dataset is.

Furthermore, the reduction in required packages and dependencies and coding in a cross-platform compatible language has enabled MetaGaAP to now be executed in Mac OSX, the Microsoft Windows 10 Linux Subsystem and all Linux distributions. However, some caveats still exist within this new version of MetaGaAP such as the computational bottleneck caused by the Biostars175929 package and an issue when completing a single-reference, multi-sample analysis in which fastq files and sample names need to be re-specified when completing the final mapping stage and calculating abundance results.

Overall the optimisations, reduced dependencies and package requirements and the new and replaced features of MetaGaAP has extended its capabilities and allows for an easier experience when executing MetaGaAP. In addition, the enabling of cross-platform compatibility and improvements in usability can potentially increase its adoption rate as an essential tool to analyse non-model systems and one-day become the standard-method for this type of analysis.

Chapter 7: Time-Course Analysis of Strains with Polymorphisms in the *BRO-A* ORF During *In Vivo* Infection by the HaSNPV-AC53 isolate.

7.1 ABSTRACT

A novel bioinformatics pipeline that generates a custom synthetic sequence database using custom sequence 'barcodes' and ultra-deep sequence data was used to identify and quantify the diversity and relative abundance of Baculovirus strain variants at time points during infection by Helicoverpa armigera single nucleopolyhedrovirus (HaSNPV). A total of 289 shared nucleotide genotypes encoding 107 amino acid genotypes were identified, using a 365-base pair 'barcode' region within the Baculovirus repeated open reading frame (BRO-A) within the 'wild type' inoculum, in larvae during the infection, and in the post-infection occlusion body. Two evolutionary effects typical of viral quasispecies populations were identified; weak-negative selection with mutation bias, causing twelve positions within the barcode region to evolve faster than neutral, and a 'drift barrier' that caused the remaining positions to evolve more slowly than neutral, limiting the effects of genetic drift. A significant change in relative abundance of the nucleotide and amino acid genotypes was observed both during the infection cycle and between the initial inoculum and the post-infection occlusion bodies, indicating inadvertent selection has been applied during the infection cycle. The results support the application of 'quasispecies' model in baculoviruses and demonstrate the application of a new bioinformatics approach to analyse quasispecies abundance and diversity in a non-model viral system. They highlight the potential for inadvertent selection pressure to alter the composition of strains within an isolate during commercial production of baculoviruses biopesticides.

7.2 INTRODUCTION

A viral quasispecies is a population of viruses (or genotypes) that behave as a single species, are related by similar mutations and in which selection acts upon genotype 'clouds' (Domingo et al., 2012; Eigen, 1978; Holland et al., 1992; Lauring & Andino, 2010; Solé et al., 1999; Wilke, 2005). In model RNA viruses, such as human immunodeficiency virus, poliovirus and rabies, the quasispecies models of ecology and evolution have been well characterised (Arbiza et al., 2010; Ball et al., 2007; Domingo et al., 2012; Solé et al., 1999).

Ecologically, two models have been known to occur within a quasispecies: niche differentiation and competitive exclusion principal, and are both essential in modelling the

interactions of genotypes within the quasispecies (Ball et al., 2007; Hardin, 1960; Pocheville, 2015; Solé et al., 1999; Vignuzzi et al., 2006). Niche differentiation can be summarised as viral variants within the population partitioning host resources, with a single dominant genotype occupying the most resources (Arbiza et al., 2010; Domingo et al., 1998; Eigen & Biebricher, 1988). This leads to cooperation between genotypes and has been observed in a poliovirus model in which genotypes of differing phenotypes break down host immune responses, allowing other genotypes to infect host tissues to which they would otherwise not have had access (Vignuzzi et al., 2006). However, when two quasispecies of equal fitness coinfect a host, the ecological model, competitive exclusion principal is observed (D. K. Clarke et al., 1994; Solé et al., 1999). An arms race begins between the two quasispecies and eventually one of the quasispecies will become extinct, or in some cases host immunity can affect a single quasispecies population through selection pressures and lead to a loss in lower fitness genotypes (Arbiza et al., 2010; D. K. Clarke et al., 1994; Solé et al., 1999; Wilke, 2005).

Quasispecies exhibit mutational robustness or 'survival of the flattest' that improves long-term viability through the maintenance of a high diversity of genotypes that are equally fit on the fitness landscape (Wilke et al., 2001). The quasispecies model suggests that if mutation rates are high, selection will act on a group of mutants or genotypes rather than individual genotypes within a population, and this is important for long-term survivability (Crotty et al., 2001; Domingo et al., 2012; Van Nimwegen et al., 1999; Wilke, 2005).

The Nucleopolyhedroviruses (family; *Baculoviridae*, genus; *Alphabaculovirus*) are a group of Lepidopteran infecting, double-stranded DNA viruses commonly used as commercially produced biological controls (B. C. Black, L. A. Brennan, P. M. Dierks, & I. Gard, 1997; Buerger et al., 2007; Copping & Menn, 2000; G. Zhang, 1989). They exhibit two distinct life-history stages; budded virus (BV) which are single virions that are produced during *in vivo* transmission between cells during infection, and occlusion bodies (OB), in which several thousand singly- (SNPV) or multiply-enveloped (MNPV) virions are embedded in protein occlusion bodies, which are responsible for horizontal transmission of the virus (Blissard & Rohrmann, 1990; George Rohrmann, 2011a, 2011b). Each OB may contain multiple genetically-related virus variants co-occluded within a single, protein body with differing phenotypic properties reducing the chance of insect resistance (Vicky Lynne Baillie & Bouwer, 2012b; Blissard & Rohrmann, 1990; Chateigner et al., 2015; Cory et al., 2005; Goulson & Hauxwell, 1995; Noune & Hauxwell, 2016a; Ogembo et al., 2007; Elizabeth M. Redman et al., 2010; Reeson et al., 1998). Identification of baculovirus strains within an isolate or OB typically uses selection of 'cloned' strains *in vitro* by selection in tissue culture of plaques from budded virus or OB derived virions (Brown & Faulkner, 1977, 1978; Corsaro & Fraser, 1987; Noune & Hauxwell, 2016a; Ogembo et al., 2007; Elizabeth M. Redman et al., 2010; Simon et al., 2011). Alternatively, strains may be selected by repeated infection *in vivo* with end-point dilution of the

inoculum (theoretically resulting in infection by a single occlusion body) (Brown & Faulkner, 1977; I. R. Smith & Crook, 1988; Vlak, 1979)

A typical infection cycle involves two stages. Stage 1 is the ingestion of the OB by the larva, allowing the OB to be broken down in the alkaline midgut lumen, releasing the virions which fuse with the midgut epithelial cells and initiate replication (Krell, 2008; G. F. Rohrmann, 2013d). This is followed by the second stage involving a second phenotype, the budded virus (BV). Single virions (BV) are produced in an infected cell and acquire a membrane from the basal side of the epithelial cell before exiting into the hemolymph for *in vivo* transmission (Krell, 2008). Each of these genotypes may have different phenotypes with different biological activity that may collectively improve the success of the infection by infecting different host tissues, and maintenance of diversity (Blissard & Rohrmann, 1990; Hails et al., 2002; Elizabeth M Redman et al., 2016; G. F. Rohrmann, 2013a, 2013c, 2013d; White et al., 2012; Zwart et al., 2009).

The identification of genotypes during a baculovirus infection cycle has relied on measuring mRNA expression or cDNA (differential expression) of genes to determine when transcription is initiated and the relative abundance of genes (Kang, Suzuki, Zemskov, Okano, & Maeda, 1999; J Kuzio, Jaques, & Faulkner, 1989; Rohel & Faulkner, 1984; Soneson & Delorenzi, 2013). However, once transcription has terminated, expression of the gene ends and therefore monitoring of the infection cycle ends. Furthermore, baculoviruses and most viruses lack standard methods or sequence databases for which 'meta-barcode' regions have been identified to analyse a full infection cycle (Capobianchi et al., 2013; McElroy et al., 2014; Noune & Hauxwell, 2017b).

The 'Meta-barcoding, Genotyping and Abundance Pipeline' (MetaGaAP) is a novel bioinformatics pipeline that generates a custom synthetic sequence database from ultra-deep sequence data using custom sequence 'barcodes' (Noune & Hauxwell, 2017b). It can be used to quantify virus variants from genomic DNA, facilitating analysis of genotype abundance throughout a complete infection-cycle.

It has been hypothesised that baculoviruses exhibit characteristics of a viral quasispecies through maintaining a genotypically and phenotypically diverse population of genotypes which may cooperate during the infection cycle (Chateigner et al., 2015; Cory et al., 2005). However, additional evidence such as observing the previously described ecological and evolutionary models during the infection cycle would be needed to validate the quasispecies hypothesis in baculoviruses. This study applies the latest iteration of MetaGaAP, MetaGaAP-Python (MetaGaAP-Py) (Noune & Hauxwell, 2017a, 2017b) to model the viral variation within the *Helicoverpa armigera* SNPV – AC53 isolate (HaSNPV-AC53) by targeting the custom 'meta-barcode' previously identified region within the BRO-A gene (Noune & Hauxwell, 2016a, 2017b). We quantify viral genotypes in the inoculum, in occluded (OB) and non-occluded (BV)

virus during infection in larvae of its Lepidopteran host *Helicoverpa armigera*, and use statistical analysis to describe the ecology and evolution of the identified viral genotypes.

7.3 MATERIALS AND METHODS

7.3.1 Virus Source

The AC53 isolate was obtained and OBs purified as previously described (Noune & Hauxwell, 2016a).

7.3.2 Time-Course Sampling and DNA Extraction

A total of forty-two 2nd instar *H. armigera* larvae were infected with $1.11x10^5$ OB/mL (LC_{90}) of AC53 using a droplet assay and reared in individual containers as previously described (P. R. Hughes & Wood, 1981; Noune & Hauxwell, 2016a). Six insects were collected every 24hrs post-infection (P.I.) until 144hrs P.I., and stored at -20°C in individual 1.5 mL microcentrifuge tubes for 24hrs. In addition, six insects from the infected forty-two were left to succumb to the infection and collected at death. Of these six insects, two died at 96hrs P.I., two died at 120hrs P.I., and two died at 144hrs P.I.

Budded virus (BV) DNA was extracted from individual insects by pouring liquid nitrogen onto the insect and grinding with a sterile mortar and pestle. The grindate was transferred to clean, individual 1.5 mL microcentrifuge tubes containing a 1 mL DNA extraction buffer consisting of tris/hydrochloric acid (50mM), sarkosyl (0.5%), ethylenediaminetetraacetic acid (EDTA) (25mM) and proteinase K (100µg/mL) and vortexed briefly. Samples were incubated at 50°C for 2hrs. Post incubation, samples were centrifuged in an Eppendorf 5424R microcentrifuge at 3,800 rpm for 10 min to pellet the insect debris. The supernatant was collected and processed using an Isolate II Genomic DNA kit (Bioline) continuing from step 4 of the manufacturer's instructions.

The six insects that succumbed to the infection were processed using the modified OB extraction protocol described by Noune and Hauxwell (Noune & Hauxwell, 2016a). OBs were extracted from cadavers by maceration in 0.1% sodium dodecyl sulphate (SDS), filtration through muslin and centrifugation at 500 rpm at 4 °C for 5 min to remove insect debris, followed by centrifugation at 4000 rpm at 4 °C for 20 min in a swing-out rotor (Sorvall Legend RT®, Sorval Heraeus Rotor). The supernatant was discarded and 200µL of analytical-grade 0.05 M sodium carbonate and 200µL of 1x tris-EDTA (TE) buffer was added to the virus pellet to disrupt virion membranes and release virions from OBs. Samples were vortexed briefly and incubated at 50°C for 30 min. Sample processing was then continued using an Isolate II Genomic DNA kit (Bioline) from Step 4 of the manufacturer's instructions.

7.3.3 Amplicon Sequencing and Ion Torrent PGM Library Preparation

The BRO-A region was targeted for sequencing as we have previously reported it to have a high degree of variability and a high density of polymorphisms within the AC53 isolate (Noune & Hauxwell, 2016a, 2017b). The BRO-A primers used, PCR amplification and reaction conditions were as previously described (Noune & Hauxwell, 2017b). Prior to sequencing, PCR products were visualized to confirm successful amplification using a 1.5% agarose gel made with 1x tris-acetate-EDTA (TAE) buffer, 1x GelRed (Biotium) and electrophoresed in 1x TAE buffer for 30 min at 100 volts. This confirmed 39 of the 42 infected insects successfully amplified with five out of the six insects collected at 72hrs and four OB samples (two per time point) isolated at 96hrs and 120hrs producing a product. No 144hr OB samples were sequenced.

Ion Torrent PGM library preparation and amplicon clean-up was completed as per the Life Technologies Ion Torrent PGM fusion primer manual (ThermoFisher). Sequencing was completed using 400 bp chemistry and two 318v2 chips with the OB samples on one chip and the BV samples on another.

The previously described AC53 BRO-A OB dataset (Noune & Hauxwell, 2017b) was used as the time-course inoculum. This dataset was sequenced on the same 318v2 chip as the OB samples but underwent bulking and extraction as previously described (Noune & Hauxwell, 2016a).

7.3.4 Data Analysis

MetaGaAP

MetaGaAP-Py build 3.3.1 was used to analyse and apply QC to each individual BRO-A dataset with reads trimmed to lengths between 300 bp and 365 bp and a Q20 threshold (~99% read accuracy) applied. A single database containing 2^{25} unique, synthetic BRO-A sequences was produced and MetaGaAP-Py mapped each individual BRO-A dataset to the database to identify the common nucleotide genotypes with greater than 1x coverage. These sequences were extracted for downstream analysis. In addition, the identified nucleotide genotype sequences were substituted into the full AC53 BRO-A consensus sequence to observe, and produce amino acid genotypes.

Nucleotide Similarity and Evolutionary Analysis

A distance matrix was produced using the 289 nucleotide genotypes by aligning using MAFFT v7.308 with the FFT-NS-2 algorithm, a 200 point accepted mutation (200PAM) scoring matrix, a 1.53 gap open penalty and a 0.123 offset value (Dayhoff, Schwartz, & Orcutt, 1978; Katoh & Standley, 2013). A second matrix was produced with the 108 amino acid genotypes using the previously described parameters but with the blocks substitution matrix 62 (BLOSUM62) as the scoring matrix (Henikoff & Henikoff, 1992). Protein structures were

predicted using the European Molecular Biology Open Software Suite (EMBOSS) version 6.5.7 garnier tool (Rice, Longden, & Bleasby, 2000).

Tajima D and Fay and Wu H evolutionary statistics were calculated using the previously completed nucleotide alignment with VariScan 2.0.3 and the AC53 BRO-A consensus sequence used as the outgroup (Fay & Wu, 2000; Hutter, Vilella, & Rozas, 2006; Koboldt et al., 2012; Tajima, 1989; Vilella, Blanco-Garcia, Hutter, & Rozas, 2005). VariScan was run using the total number of mutations in sliding-window mode with Tajima's D and Fay and Wu's H calculated every 20 bp with a 10 bp jump and sites with gaps included in the analysis. The statistics were plotted using Microsoft R Open version 3.4.0, RStudio version 1.0.136 and the R packages ggplot2 version 2.1.0, reshape2 version 1.4.1 and plyr version 1.8.4 (Microsoft, 2016; R. Team, 2014; Wickham, 2009, 2014, 2016).

Mean (relative) evolutionary rates were calculated per individual nucleotide with MEGA7 (Kumar, Stecher, & Tamura, 2016; Nei & Kumar, 2000). Parameters for this analysis used an automatically produced neighbor-joining tree with maximum-likelihood estimation, a gamma distributed general time reversible substitution model using all sites (including gaps) with 5 distinct gamma categories and no branch swap filter.

Statistical Analysis

To simplify the analysis and enhance clarity of the visualization, hierarchical clustering and heat-mapping was limited to both DNA and amino acid genotypes with at least 0.1% relative abundance. Hierarchical clustering and heat-mapping of the mean relative abundance per time point was completed using an unweighted pair group method with arithmetic mean (bottom-up clustering) and Bray-Curtis dissimilarity (Bray & Curtis, 1957; Kolde, 2012; Murrell, 2002; Noune, 2016; Oksanen et al., 2007; Sokal & Rohlf, 1962).

Statistical analysis of the BV and OB genotypes were completed separately to limit any bias caused by the alternative extraction methods applied (previously described), and sequencing on two different 318v2 chips. Scatterplots of the infection cycle were produced, and the analysis was undertaken within a generalized linear modelling framework (GLM) and a quasi-Poisson likelihood distribution to account for over-dispersion (McCullagh, 1984; McCullagh & Nelder, 1989; K. Pearson, 1900; Warton & Guttorp, 2011; Wedderburn, 1974). The analysis was applied to the two genotypes with the highest relative abundance and three minor genotypes. Two different formulas were applied and have been defined below; equation 7-1 for BV genotype abundance, presence-absence and viral copy number proxy, and equation 7-2 for OB genotype abundance. Furthermore, the analysis was limited to the five DNA and amino-acid genotypes with the highest relative abundance per dataset. This was completed using Microsoft R Open version 3.4.0, RStudio version 1.0.136 and the R packages, 'Modern applied-statistics with S-PLUS' (MASS) version 7.3-45 and ggplot2 version 2.1.0 (Microsoft, 2016; R. Team, 2014; R. C.

Team, 2013; Venables & Ripley, 2013). In addition, raw read counts during the infection cycle were analysed as an estimated proxy for virus copy number using the previously described parameters.

$$\log y_i = \beta_0 + \beta_1 t_i + \beta_2 z_i + \beta_3 w_i$$

Equation 7-1: Parameters are defined as follows: y = reads (or present genotypes), β_0 = intercept, t = time such that β_1 represents the linear trend in abundance, read counts or present genotypes over time for genotype 1, observed reads or observed genotypes, z is an indicator of genotype (0 = genotype 1, 1= genotype 2) such that β_2 represents a change in abundance, read counts or present genotypes over time and w = is time for genotype 2 such that β_3 represents the change in linear trend for genotype 2 in comparison to genotype 1 over time.

$$\log y_i = \delta_0 + \delta_1 t_i + \delta_2 t_i^2 + \delta_3 z_i + \delta_4 w_i + \delta_5 w_i^2$$

Equation 7-2: This equation was applied to analyse the OB genotype abundance to accommodate for the low number of samples (five samples in total: the AC53 inoculum, and the four OB samples). Parameters are defined as above with the addition of: t^2 = time^2 such that δ_2 represents the curvature in abundance for genotype 1 over time, and w^2 is defined such that δ_5 represents the change in curvature in abundance in genotype 2 relative to genotype 1 over time.

Presence-absence of genotypes within each dataset was analysed by converting reads associated to genotypes into a binary matrix using Microsoft R Open 3.4.0 (Microsoft, 2016; R. Team, 2014; R. C. Team, 2013). Genotypes containing more than a single read were classed as a 1 (genotype present), and genotypes with no reads classed as a 0 (genotype absent). A scatterplot and model predictions of genotype presence-absence during the infection cycle was produced using each individual BV dataset as previously described. Presence-absence was tallied per dataset and the R package Pheatmap version 1.0.8 applied to produce a presence-absence heat-map with Sørensen–Dice coefficient clustering using Microsoft R Open 3.4.0 (Dice, 1945; Kolde, 2012; Microsoft, 2016; Sørensen, 1948; R. Team, 2014; R. C. Team, 2013).

7.4 RESULTS

7.4.1 MetaGaAP

Sequencing of the BRO-A barcode region and comparison with the synthetic sequence library identified 289 common nucleotide genotypes with at least 1x coverage in a single dataset and between 0.000327% and 97.0624% relative abundance (for the complete spreadsheet see: https://researchdatafinder.qut.edu.au/display/n13986, for the genotype sequences see: https://researchdatafinder.qut.edu.au/display/n14806). The 289 nucleotide genotypes encoded 107 amino-acid genotypes with relative abundances between 0.0011% and 98.7457%. (for the complete spreadsheet and nucleotide groups encoding an amino-acid genotype see: https://researchdatafinder.qut.edu.au/display/n13986, for the genotype sequences see: https://researchdatafinder.qut.edu.au/display/n14806). Of the 107 amino-acid genotypes, 29 encoded predicted functional BRO-A proteins and 78 encoded non-functional proteins (Table 12-1, Figure 12-1).

A total of 17 nucleotide genotypes were identified above the 0.1% relative abundance threshold, with a single dominant genotype observed (G_33554431). This single genotype was found to have a mean relative abundance between 94.9020% to 95.3692% during the infection cycle, 97.0624% abundance within the AC53 inoculum and a mean relative abundance of 81.9973% in the P.I. OBs (Table 7-1). In addition, 29 amino-acid genotypes were identified above this threshold with 13 encoding a predicted functional protein, with the dominant amino-acid genotype encoding a functional protein (A.A_1). The result is similar to the nucleotide genotypes as the mean abundance of the dominant genotype was between 97.5543% and 97.8753% during the infection cycle, 98.7457% within the AC53 inoculum, and a mean relative abundance of 86.4124% within the P.I. OBs (Table 7-2).

Table 7-1: Mean relative abundance of nucleotide genotypes across each sampled point and with at least 0.1% abundance in a single dataset. A single variant genotype G_33554431 was identified to be the dominant genotype in the populations prior, during and post infection cycle. Abundance results reported in this table do not total 100% as the result has been subset.

Genotype	Inoculum (%)	24hrs P.I. (%)	48hrs P.I. (%)	72hrs P.I. (%)	96hrs P.I. (%)	120hrs P.I. (%)	144hrs P.I. (%)	Final P.I. OBs (%) (96hrs and 120hrs P.I. OBs)
G_33554431	97.0624	94.9020	94.9968	95.2925	95.3692	95.4204	94.9486	81.9973
G_33554303	0.6179	0.7588	0.7528	0.7592	0.7523	0.7336	0.8521	1.3761
G_33552383	0.2960	0.5583	0.7000	0.7434	0.6808	0.6604	0.7504	0.9959
G_33554423	0.2005	0.3778	0.5362	0.4903	0.4404	0.4160	0.5266	0.9288
G_16777215	0.2505	0.1945	0.1371	0.1880	0.1933	0.1790	0.1797	0.2735
G_33292287	0.1504	0.5689	0.6139	0.6393	0.6174	0.6704	0.6954	0.1830
G_33554399	0.0857	0.1566	0.2456	0.1404	0.1315	0.1162	0.1323	0.1799
G_33554429	0.0982	0.1345	0.1099	0.1157	0.1219	0.1217	0.1121	0.1722
G_33554427	0.0801	0.1381	0.1622	0.1225	0.1060	0.1277	0.1090	0.1712
G_33552255	0.0256	0.0155	0.0324	0.0508	0.0386	0.0484	0.0475	0.1636
G_33554430	0.1644	0.0896	0.1360	0.0760	0.0689	0.0726	0.0881	0.1570
G_25165823	0.1644	0.2864	0.1474	0.1817	0.1752	0.1740	0.1655	0.1551
G_33553919	0.0519	0.8353	0.2896	0.2605	0.2479	0.2398	0.2419	0.1467
G_33554175	0.0485	0.1908	0.1874	0.0791	0.0753	0.0736	0.0780	0.1451
G_31457279	0.1478	0.2000	0.2361	0.1692	0.2499	0.2318	0.3278	0.1228
G_33554415	0.0211	0.0320	0.0894	0.0446	0.0434	0.0420	0.0463	0.1144
G_33554367	0.0395	0.0428	0.0644	0.0733	0.0839	0.0799	0.0855	0.1036

Table 7-2: Mean relative abundance of amino-acid genotypes across each sampled point and with at least 0.1% abundance in a single dataset. A single genotype A.A_1 was identified to be the dominant genotype in the populations prior, during and post infection cycle. This result is similar to what has been observed with the nucleotide genotypes. Abundance results reported in this table do not total 100% as the result has been subset.

Genotype	Inoculum (%)	24hrs P.I. (%)	48hrs P.I. (%)	72hrs P.I. (%)	96hrs P.I. (%)	120hrs P.I. (%)	144hrs P.I. (%)	Final P.I. OBs (%) (96hrs and 120hrs P.I. OBs)
A.A_1	98.7457	97.8753	97.5543	97.6488	97.6504	97.6728	97.4140	86.4124
A.A_8	0.3071	0.5805	0.7270	0.7778	0.7141	0.6856	0.7781	1.4620
A.A_2	0.0511	0.0338	0.0946	0.0475	0.0514	0.0479	0.0523	1.3112
A.A_4	0.0951	0.1607	0.2538	0.1482	0.1380	0.1206	0.1381	0.7343
A.A_3	0.0098	0.0000	0.0021	0.0038	0.0078	0.0030	0.0025	0.6967
A.A_5	0.0053	0.0097	0.0000	0.0027	0.0014	0.0014	0.0009	0.5801
A.A_6	0.0045	0.0007	0.0007	0.0005	0.0004	0.0021	0.0024	0.5750
A.A_7	0.0159	0.0264	0.0424	0.0427	0.0509	0.0483	0.0453	0.5499
A.A_9	0.0038	0.0008	0.0011	0.0021	0.0003	0.0011	0.0005	0.4331
A.A_11	0.0102	0.0004	0.0025	0.0007	0.0036	0.0022	0.0018	0.3472
A.A_10	0.0027	0.0000	0.0012	0.0005	0.0007	0.0003	0.0025	0.3375
A.A_12	0.0223	0.0324	0.0905	0.0457	0.0451	0.0431	0.0471	0.2628
A.A_24	0.0261	0.0160	0.0327	0.0519	0.0395	0.0491	0.0490	0.2178
A.A_13	0.0015	0.0000	0.0005	0.0018	0.0005	0.0007	0.0019	0.1984
A.A_14	0.0015	0.0000	0.0000	0.0007	0.0003	0.0006	0.0000	0.1934
A.A_15	0.0015	0.0000	0.0000	0.0021	0.0019	0.0000	0.0000	0.1930
A.A_19	0.1519	0.2057	0.2402	0.1748	0.2557	0.2365	0.3342	0.2933
A.A_16	0.0015	0.0000	0.0007	0.0005	0.0010	0.0003	0.0013	0.1881
A.A_17	0.0019	0.0000	0.0000	0.0011	0.0005	0.0000	0.0006	0.1874
A.A_38	0.1515	0.5794	0.6206	0.6462	0.6239	0.6775	0.7033	0.1855
A.A_18	0.0019	0.0000	0.0000	0.0010	0.0010	0.0004	0.0004	0.1837
A.A_39	0.1655	0.2913	0.1489	0.1837	0.1771	0.1758	0.1674	0.1572
A.A_21	0.0019	0.0000	0.0000	0.0000	0.0005	0.0003	0.0004	0.1457

A.A_20	0.0027	0.0000	0.0003	0.0007	0.0008	0.0000	0.0000	0.1457
A.A_22	0.0019	0.0000	0.0000	0.0012	0.0000	0.0005	0.0005	0.1433
A.A_25	0.0011	0.0004	0.0004	0.0011	0.0003	0.0014	0.0007	0.1059
A.A_26	0.0015	0.0008	0.0034	0.0022	0.0018	0.0013	0.0025	0.1045
A.A_28	0.0015	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.1002
A.A_27	0.0019	0.0000	0.0020	0.0006	0.0012	0.0023	0.0009	0.1001

7.4.2 Nucleotide Similarity and Evolution

Distance Matrix

Alignment of all 289-nucleotide genotype identified nucleotide similarity between 94.985% and 99.702% (for the complete distance matrix see: https://researchdatafinder.qut.edu.au/display/n13986). Nucleotide similarity to the AC53 BRO-A consensus sequence was between 88.253% and 91.867%. Similarity of the amino-acid genotypes was between 22.13% and 99.52%, with similarity to the AC53 BRO-A consensus sequence between 25.93% and 95.73% (for the complete distance matrix see: https://researchdatafinder.qut.edu.au/display/n13986).

Tajima's D, Fay and Wu's H and Mean Evolutionary Rate

Tajima's D was calculated to be 0.46, suggesting balancing selection with sudden population contraction has occurred with rare-alleles present in low frequencies (Figure 7-1). However, Fay and Wu's H was calculated to be -3 suggesting an excess of high-frequency derived SNPs are in the population (Figure 1).

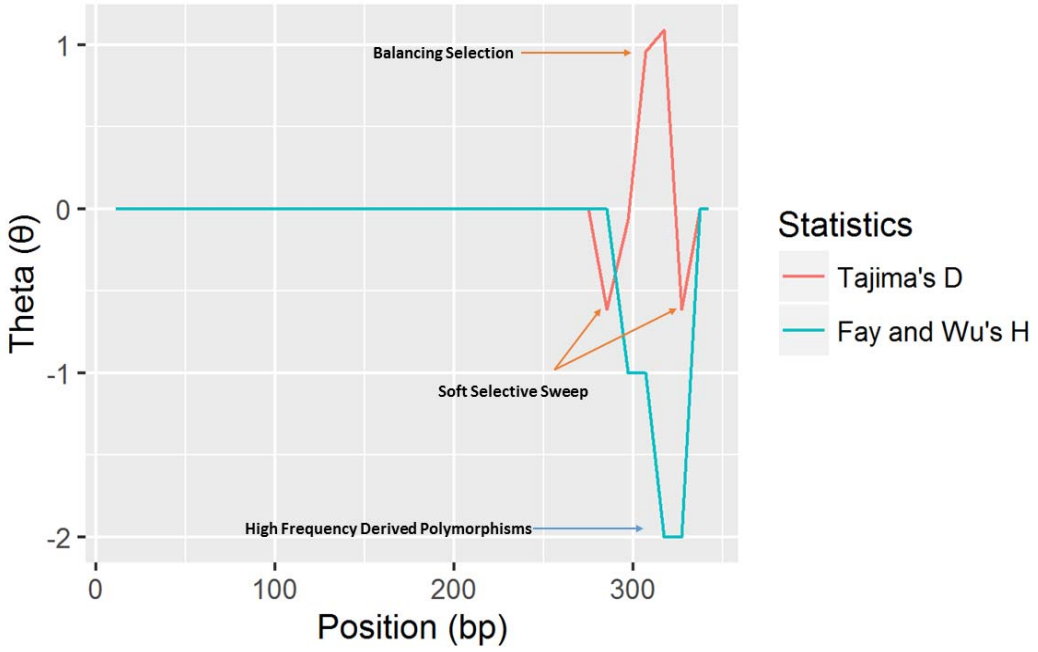


Figure 7-1: Comparison of Tajima's D and Fay and Wu's H indicating where selective sweeps are occurring and location of high-frequency derived SNPs. This result is indicative of a false-positive bottleneck and cannot be used to infer evolutionary affects occurring within the population.

Mean relative evolutionary rate analysis suggests that the population is evolving slower than neutral evolution. However, twelve positions were identified to be evolving faster than neutral evolution (Table 7-3).

Table 7-3: The twelve positions identified to be evolving faster than neutral evolution. Mean rates of evolution >1 are indicative of faster than neutral evolution.

Nucleotide Position Relative to Alignment (bp)	Mean Relative Rate of Evolution	Polymorphism
71	1.28	G deletion
79	1.12	T deletion
98	1.43	T deletion
99	32.57	C deletion
265	32.77	T to C substitution
279	32.77	G to A substitution
291	32.77	G to A substitution
300	32.77	A to C substitution
302	32.77	C to T substitution
308	32.77	C to A substitution
311	32.77	G to T substitution
314	32.77	T to C substitution

7.4.3 Statistical Analysis

Hierarchical Clustering and Heat-mapping

Hierarchical clustering of the mean relative abundance per time-point identified two distinct branches, one containing the final OB products and the second containing the inoculum and each analysed time-point (Figure 7-2 and Figure 7-3). The results indicate that the 120hrs, 96hrs and 72hrs P.I. samples are the most similar regarding relative abundance, whereas the 48hrs and 144hrs P.I. are most similar. The difference in relative abundance of the dominant genotype with the final OB products and the inoculum indicates that the OB produced after an infection cycle differs from the initial stock.

Phylogenetic analysis was excluded due to the high nucleotide similarity observed, which resulted in no separation of genotypes but hierarchical clustering of the relative abundance of individual genotypes could separate the genotypes with high-resolution. At least three distinct branches were identified, but the high abundance of the dominant genotype (G_33554431 and A.A_1) masked the resolution of minor genotypes when included in the heat-map (Figure 7-2 and Figure 7-3). Furthermore, the dominant genotype was separated from all the minor genotypes.

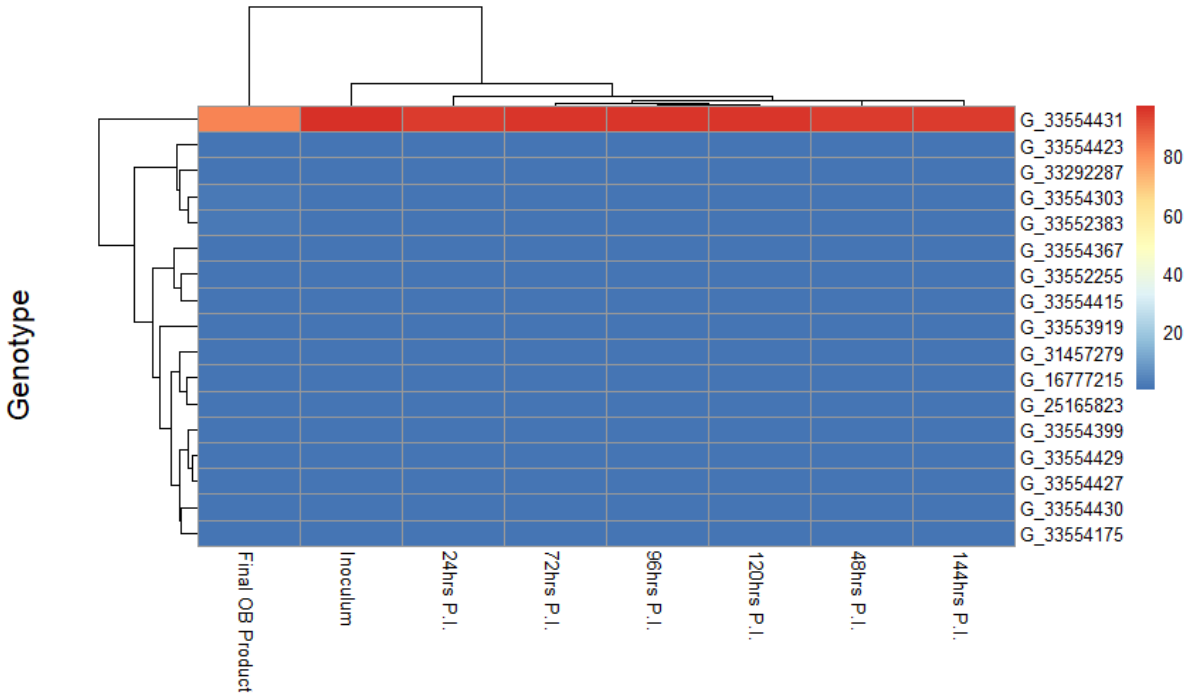


Figure 7-2: Hierarchical clustering and heat-mapping of the mean relative abundance of nucleotide genotypes above the 0.1% threshold. The dominant genotype abundance masks the minor genotypes and therefore cannot be visualised accurately, however, at least three distinct genotype clusters can be observed. In addition, clustering of the samples has highlighted the P.I. OB samples to be clustering separately whereas the inoculum and time-points are clustered together. Abundance scale is a percentage.

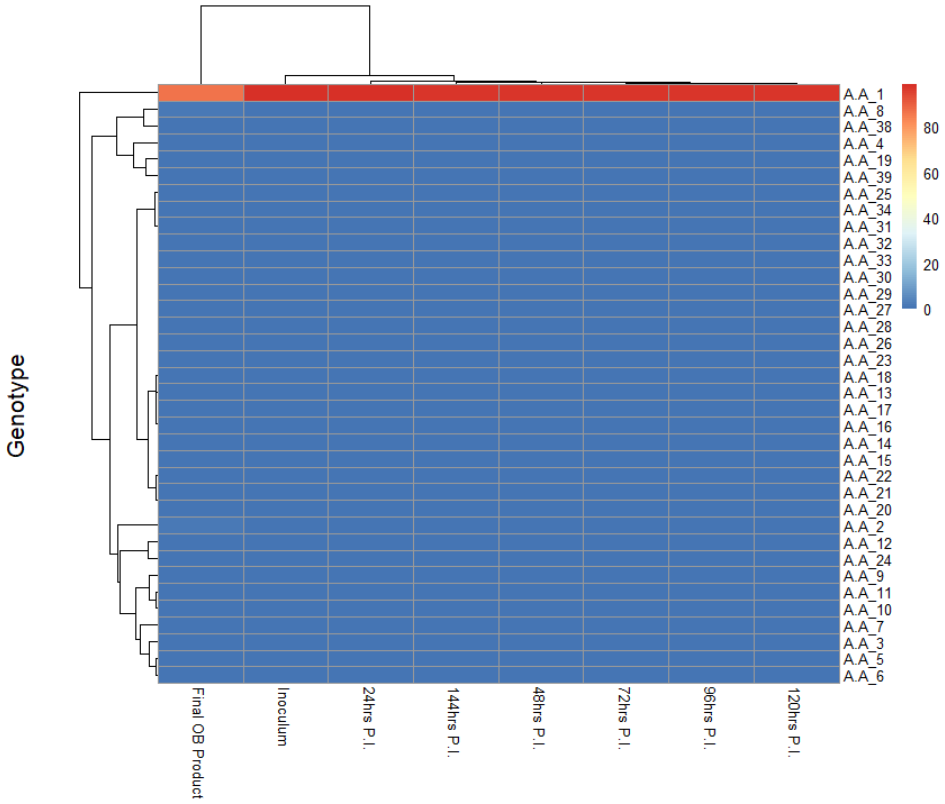


Figure 7-3: Hierarchical clustering and heat-mapping of the mean relative abundance of amino-acid genotypes above the 0.1% threshold. The result mirrors the nucleotide genotype clustering in addition to the dominant genotype masking the minor genotype abundance. Abundance scale is a percentage.

Removing the dominant strain from the hierarchical clustering and heat-mapping confirmed the differences in composition of genotypes between the final OB produced and the initial inoculum (Figure 7-4 and 7-5). Clustering and heat-mapping of minor genotypes were more easily distinguishable and produced a result which mirrors both figures 7-2 and 7-3, however, nucleotide clustering identified the inoculum and the final OB product branches had switched (Figures 7-4). This was not observed with the amino-acid clustering of minor genotypes (Figure 7-5).

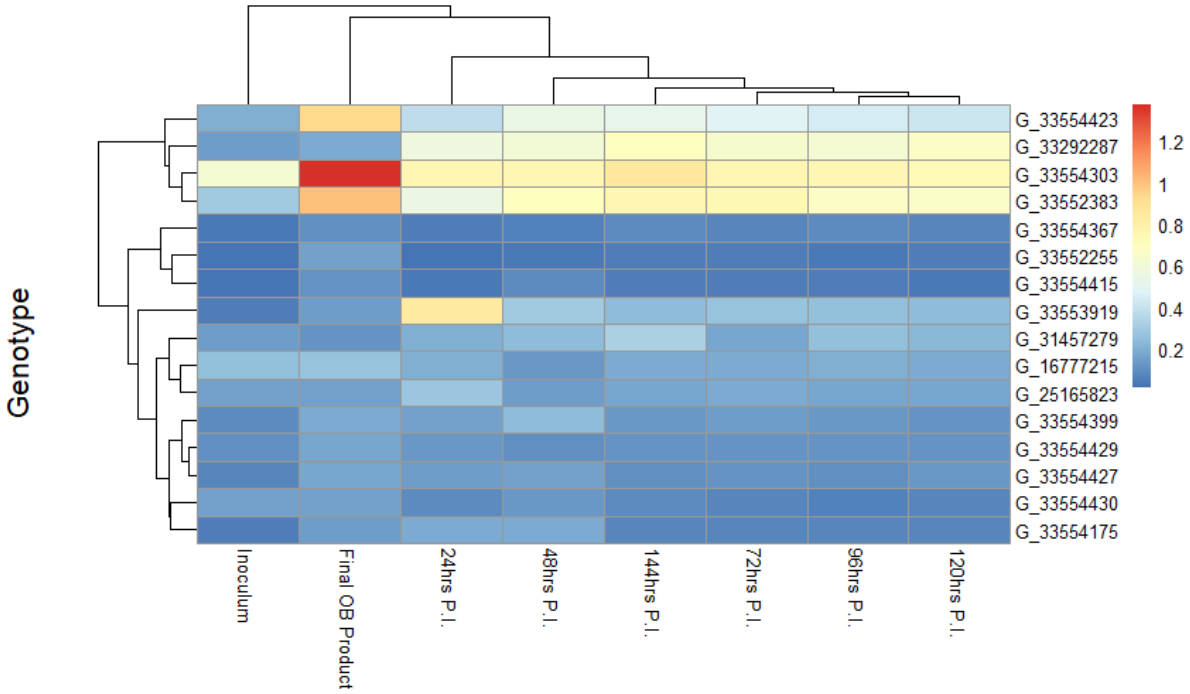


Figure 7-4: Mean relative abundance per time-point of the hierarchically clustered heat-map excluding the dominant genotype. Removing the dominant genotype from the cluster analysis could highlight the change in relative abundance per time-point for the minor genotypes. Clustering of the genotypes indicated at least three distinct branches which mirrors figure 3. However, excluding the dominant genotype has altered the clustering of the samples with the inoculum and the P.I. OB samples switching branches. Abundance scale is a percentage.

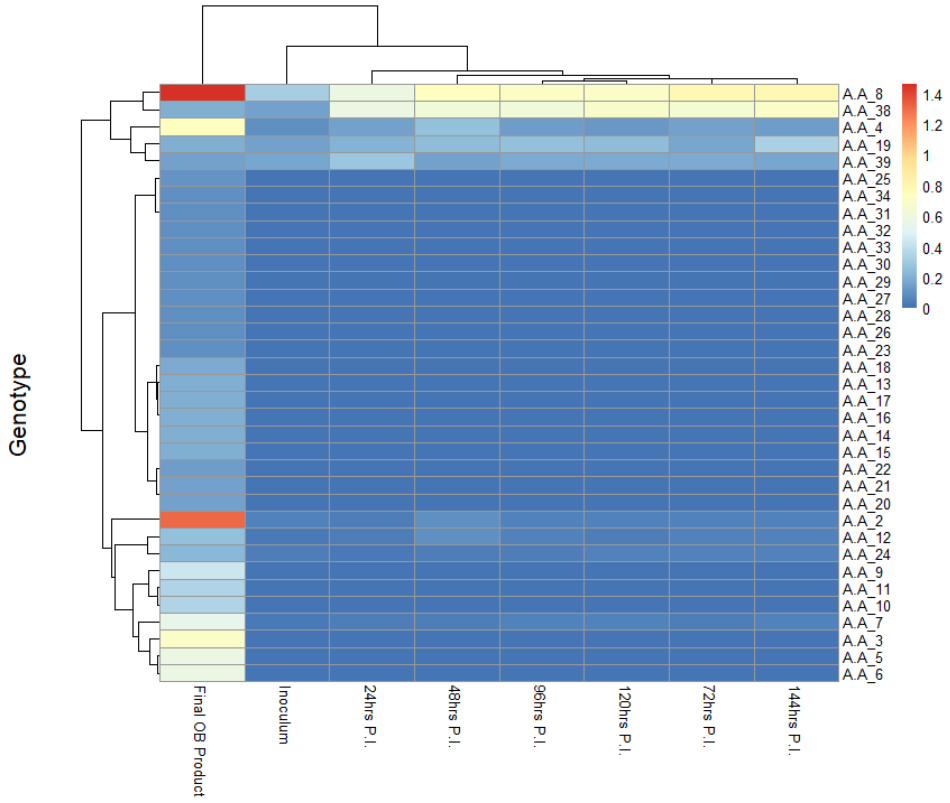


Figure 7-5: Hierarchical clustering and heat-mapping of the mean relative abundance of amino-acid excluding the dominant A.A_1. The result mirrors the nucleotide genotype clustering except the inoculum and the final OB product did not switch branches. Abundance scale is a percentage.

Scatterplots and GLM

*For GLM outputs and diagnostic plots please refer to this repository: https://researchdatafinder.qut.edu.au/display/n13986

Analysis of read counts during the BV infection was consistent with a significant linear increase in virus titre over time during early infection (Table 12-2), before plateauing at 120 hrs and 144 hrs i.e. as the virus reaches saturation (Figure 7-6).

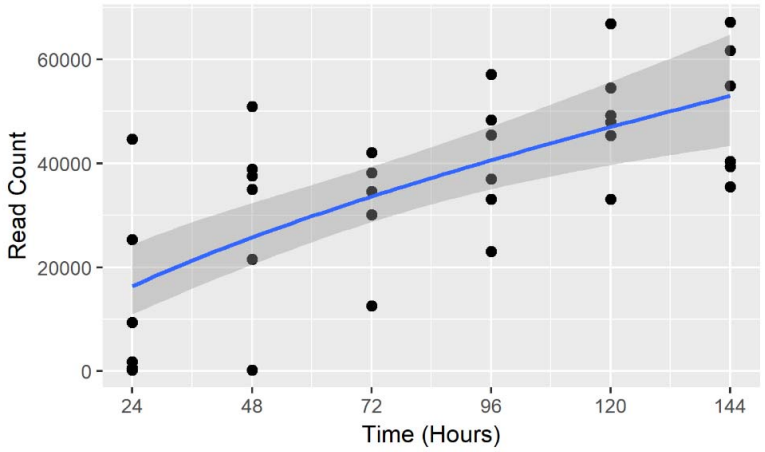


Figure 7-6: Scatterplot of read count during the infection cycle indicating an almost linear increase in virus production prior to reaching saturation at 120 hrs and 144 hrs P.I. A GLM with a quasi-Poisson distribution line of fit has been applied to visualise the trend in relative abundance. Shading indicates 95% confidence intervals.

The dominant genotype (G_33554431) was found to have a statistically significant (see β_1 and β_2 parameters – Table 12-3) decline in relative abundance during the infection cycle, but was found to be a non-linear (Figure 7-7 A, Table 12-3). Furthermore, this result is mirrored with the dominant A.A_1 genotype (Figure 7-7 B, Table 12-3). This is in contrast with the total read count during infection cycle result, which is consistent with a significant increase in virus production.

A significant, non-linear increase in the abundance of reads of the minor genotypes G_33554303, G_33552383 and G_33554423 was observed (Table 12-3), with a peak observed at 144 hrs P.I. in the scatterplot (Figure 7-8 A), with the exception of G_16777215 for which no significant change in abundance was observed. However, the amino-acid genotypes A.A_2, A.A_3 and A.A_4 were found to have non-significant results, with the exception of A.A_8 for which a significant, non-linear increase in abundance was observed (Figure 7-8 B, Table 12-3).

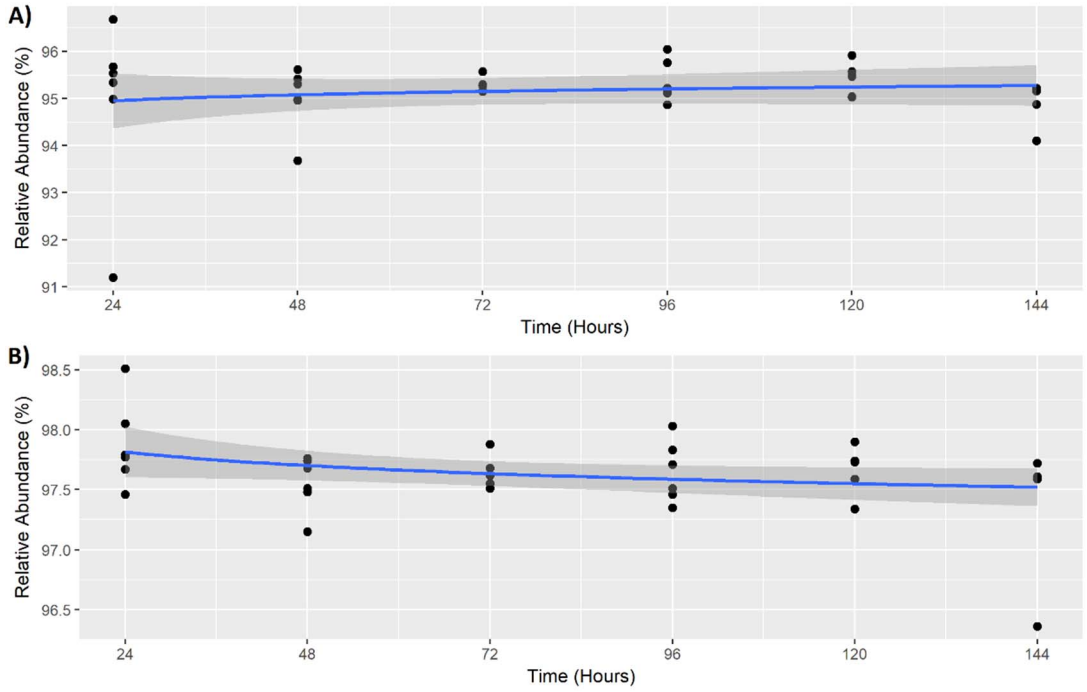


Figure 7-7: Scatterplot of the dominant genotypes A) G_33554431 and B) A.A_1. A statistically significant, albeit, minor reduction in relative abundance is observed during the infection cycle and contrasts from the read count during infection cycle result which identified a significant increase in virus production. A GLM with a quasi-Poisson likelihood distribution line of fit has been applied to visualise the trend in relative abundance. Shading indicates 95% confidence intervals.

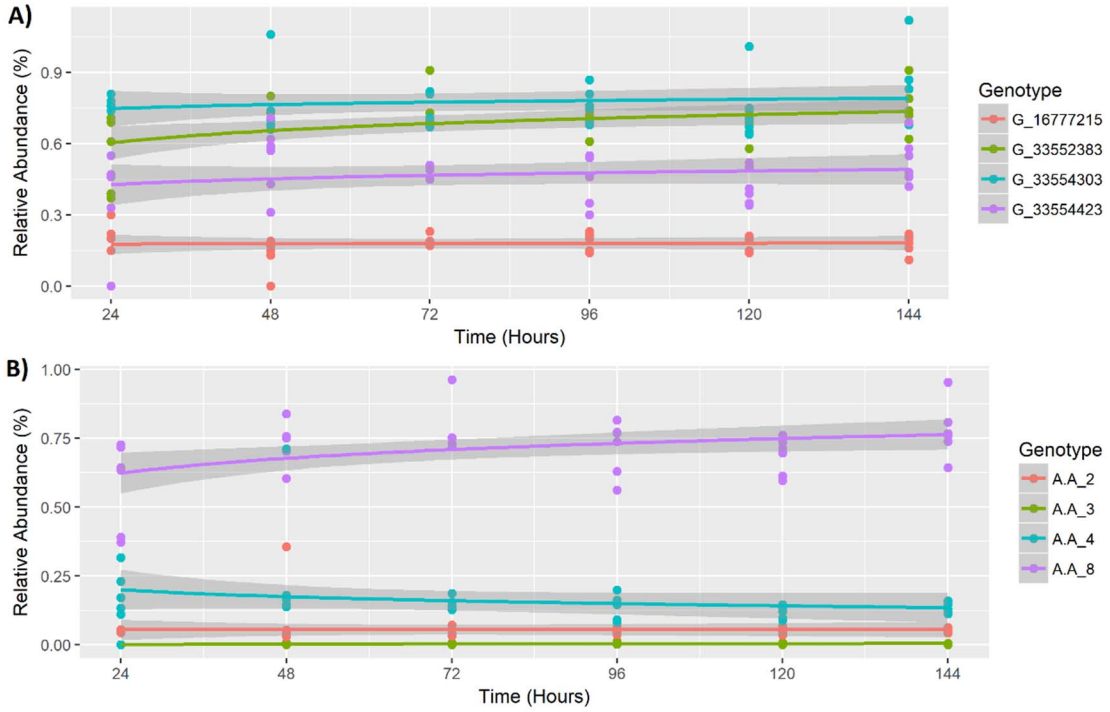


Figure 7-8: Scatterplot of a subset of A) minor nucleotide genotypes and B) minor amino-acid genotypes. A) Significant increase in abundance is observed between the initial infection stock for genotypes G_33554303, G_33552383 and G_33554423. G_16777215 was found to have non-significant changes in abundance. B) Non-significant results were observed for A.A_2, A.A_3 and A.A_4, however, A.A_8 was shown to have significant non-linear increase in abundance. The trend for all the minor genotypes was found to be non-linear. A GLM with a quasi-Poisson likelihood distribution line of fit has been applied to visualise the trend in relative abundance. Shading indicates 95% confidence intervals.

Analysis of reads indicated a significant reduction in both the nucleotide and amino-acid dominant genotype abundances in P.I. OB samples in comparison to read abundance in the inoculum (Figure 7-9; Table 12-4). In addition, reads of all the minor nucleotide and amino-acid genotypes had significant increases in relative abundance between the inoculum and cadavers, but were found to be non-linear, with no significant change to the curvature over time, with the exception of the nucleotide genotype G_16777215 which was found to have no significant changes in abundance (Figure 7-10, Table 12-4). The result for the amino-acid genotypes contrasts from the BV results as non-significant changes were previously identified.

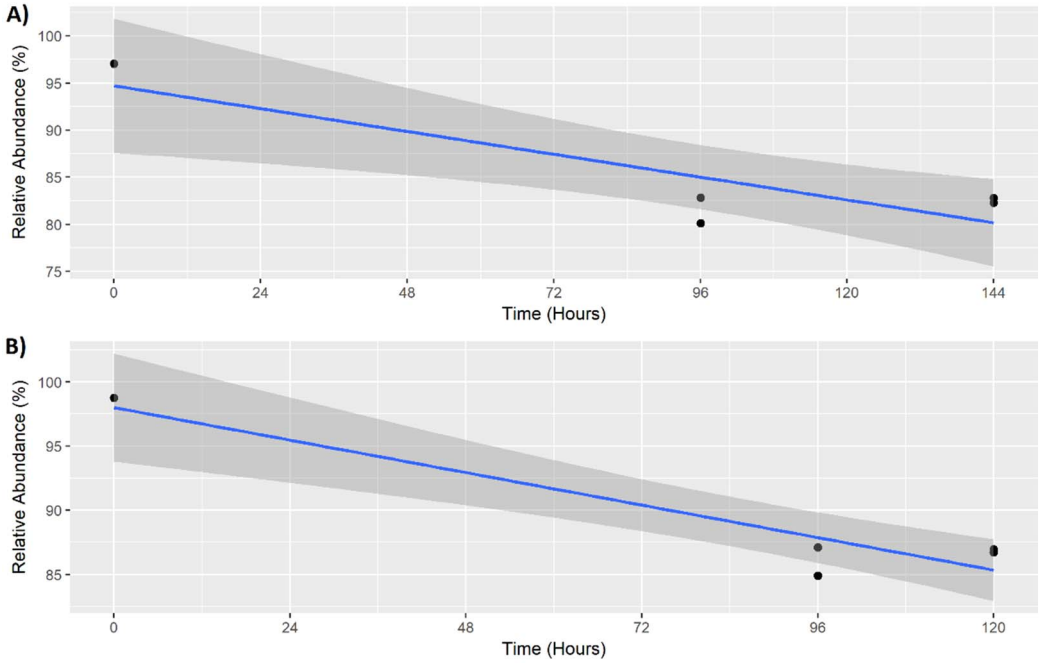


Figure 7-9: Scatterplot of the dominant A) nucleotide genotype G_33554431 and B) amino-acid genotype A.A_1when the inoculum (time point 0) is compared to the final OB products produced. A statistically significant decrease in abundance of both dominant genotypes was observed. A GLM with a quasi-Poisson likelihood distribution line of fit has been applied to visualise the trend in relative abundance. Shading indicates 95% confidence intervals.

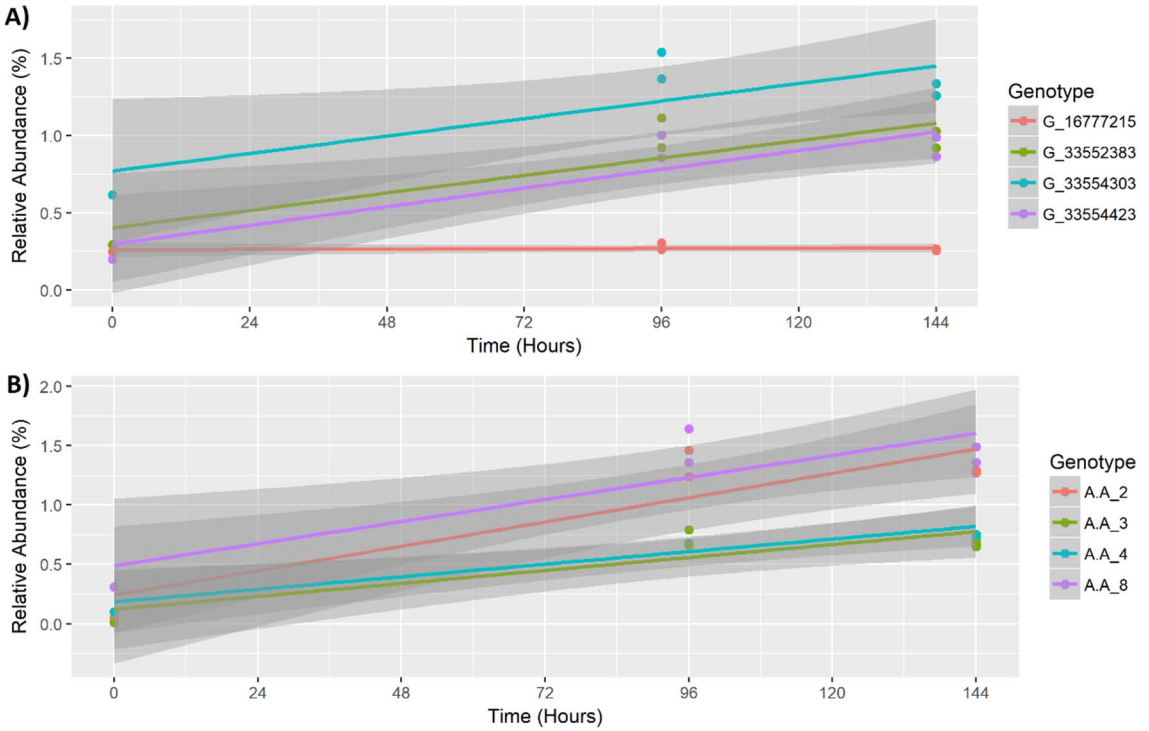


Figure 7-10: Scatterplot of minor A) nucleotide genotypes and B) amino-acid genotypes when the inoculum (time point 0) is compared to the final OB products produced. All minor genotypes had significant increases in relative abundance except for the nucleotide genotype, G_16777215 which was found to have non-significant changes in abundance, and mirrored the BV results. A GLM with a quasi-Poisson likelihood distribution line of fit has been applied to visualise the trend in relative abundance. Shading indicates 95% confidence intervals.

Presence-Absence Analysis

*For GLM outputs and diagnostic plots please refer to this repository: https://researchdatafinder.qut.edu.au/display/n13986

A significant linear increase in present genotypes was observed during the infection cycle (Table 12-5) with mean nucleotide genotype presence increasing from 13.09% at 24 hrs P.I. to 26.07% at 144 hrs P.I (Figure 7-11, Table 12-6, full binary matrix see: https://researchdatafinder.qut.edu.au/display/n13986), and mean amino-acid genotype presence increasing from 19.29% at 24hrs P.I. to 37.65% at 144hrs P.I. (Figure 7-11, Table 12-7, full binary matrix see: https://researchdatafinder.qut.edu.au/display/n13986). This was highlighted further with figure 7-12 and 7-13 as three distinct clusters which encompassed genotypes present in most datasets (green), genotypes present in randomly throughout the infection cycle (yellow), and genotypes present in the inoculum and OB datasets exclusively (orange). All OB datasets contained 100% of both the nucleotide and amino-acid genotypes.

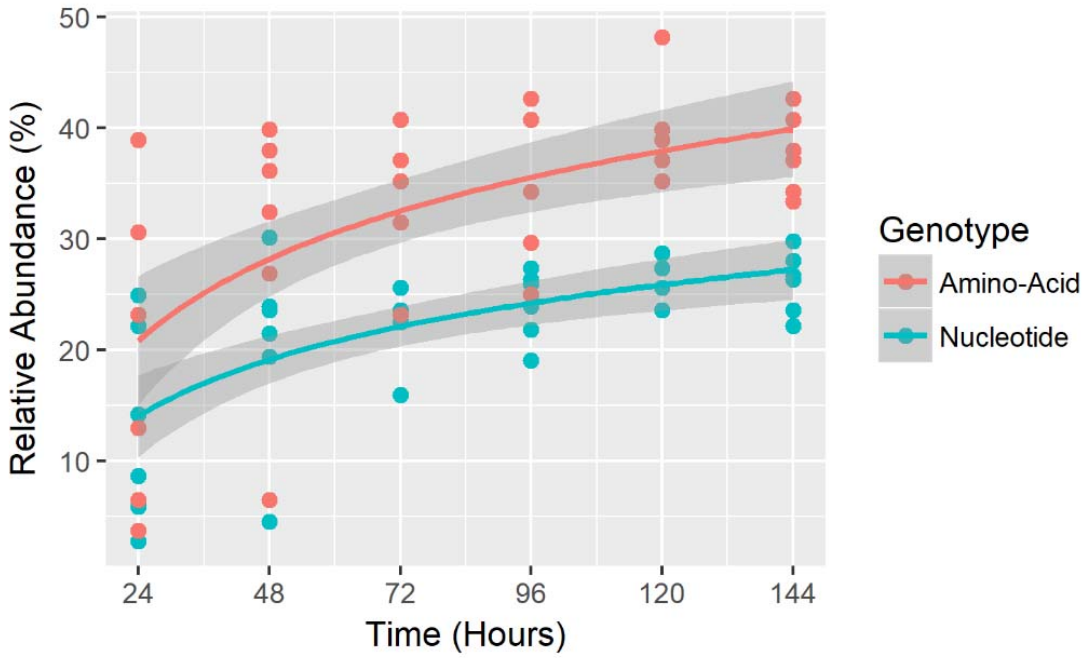


Figure 7-11: Scatterplot of present amino-acid (red) and nucleotide (aqua) genotypes during the infection cycle showing an increase in present genotypes. More amino-acid genotypes were identified (mean presence between 19.29% at 24hrs P.I. and 37.65% at 144hrs P.I.) over the course of the infection than nucleotide genotypes (mean presence between 13.09% 24 hrs P.I. to 26.07% at 144 hrs P.I), with a peak in present genotypes at 144hrs P.I. The inoculum and OB products (not shown) contained the 100% of the population. A GLM with a quasi-Poisson likelihood line of best fit has been applied to visualise the trend in relative abundance. Shading indicates 95% confidence intervals.

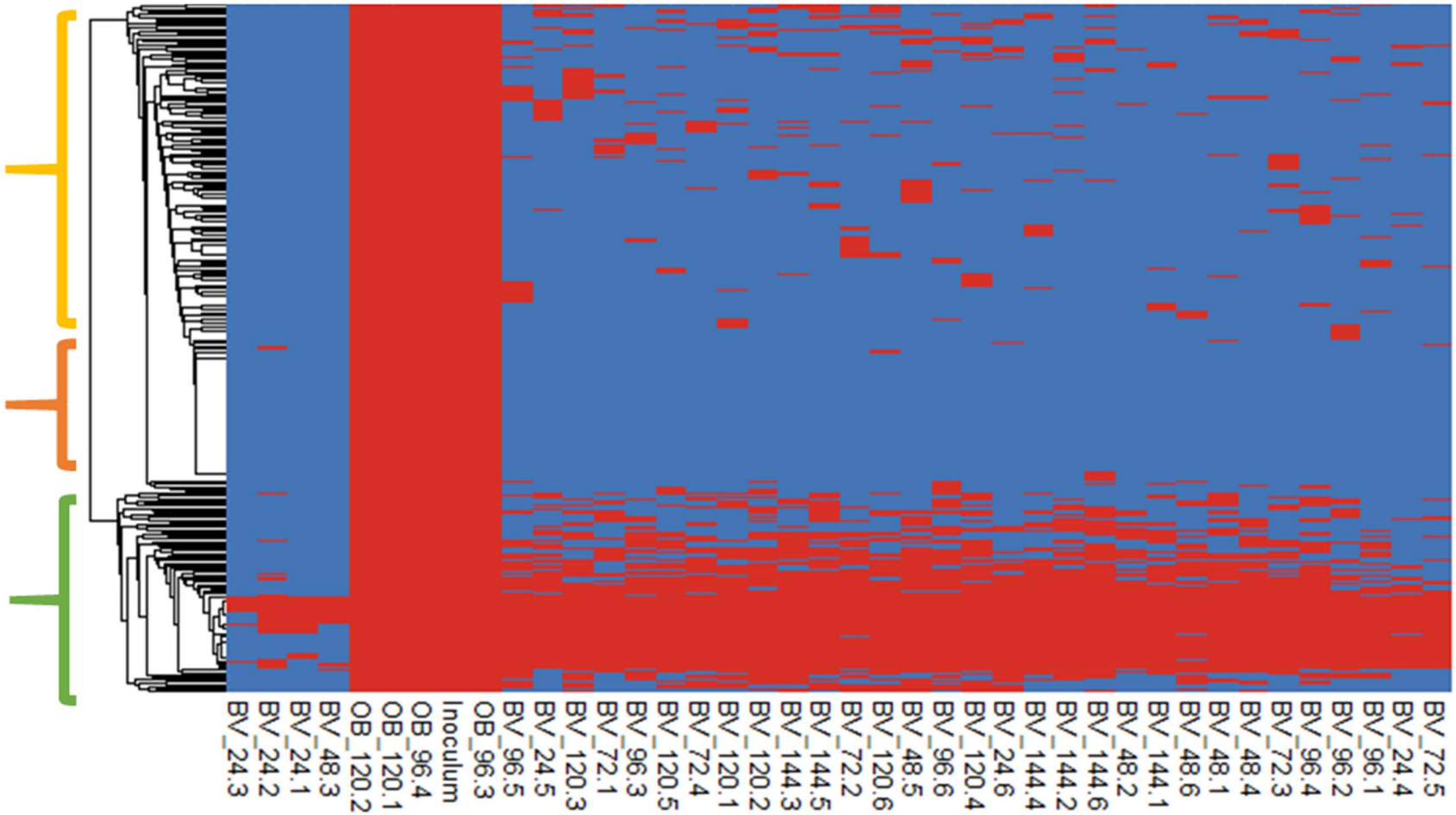


Figure 7-12: Heat-mapping of present (red) and absent (blue) nucleotide genotypes within every analysed dataset. Sørensen–Dice coefficient clustering identified three distinct groups: a group present in most datasets (green), a group present in the inoculum and OB products exclusively (orange) and a third group which consisted of genotypes randomly appearing and disappearing during the infection cycle (yellow).

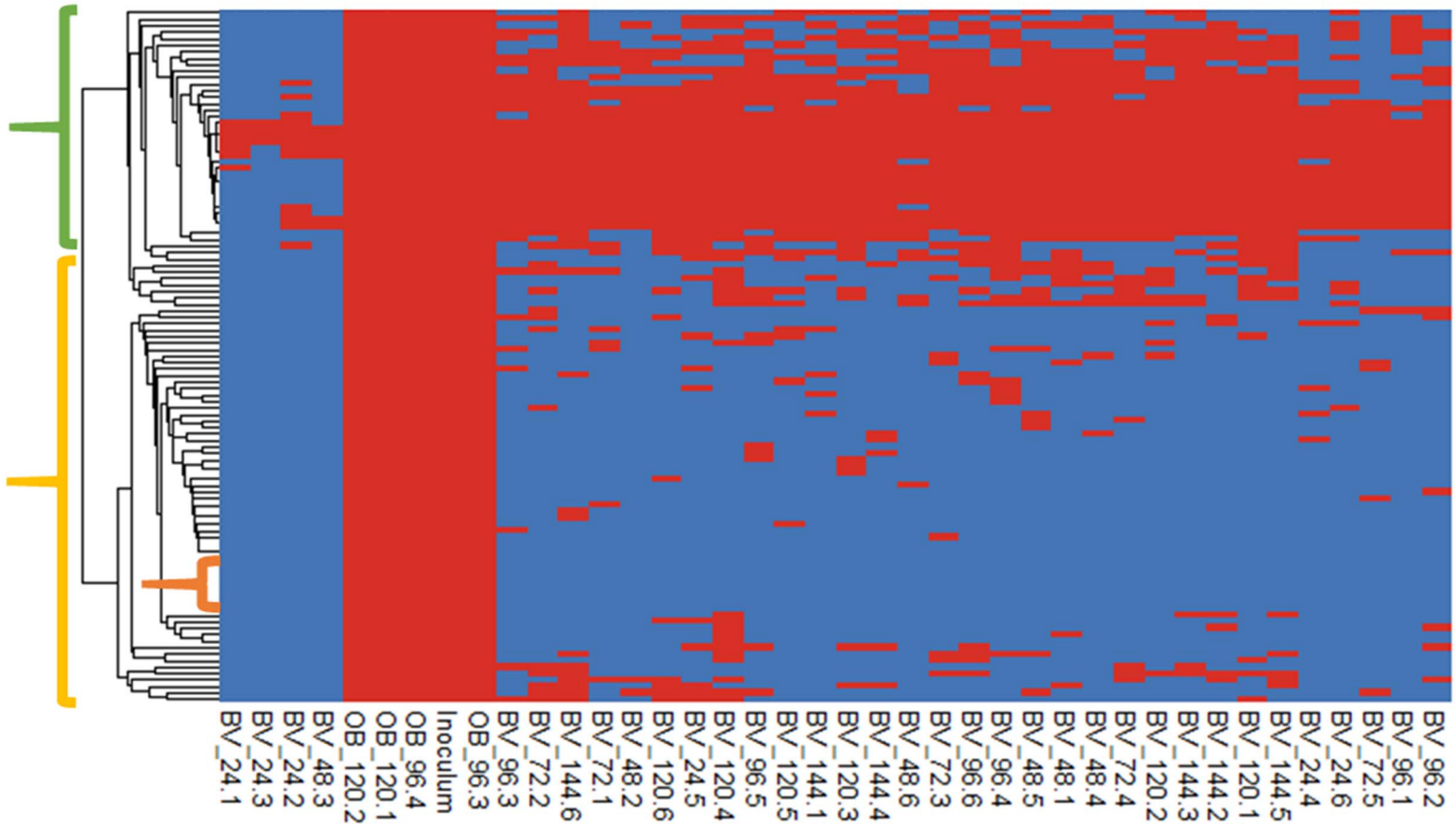


Figure 7-13: Heat-mapping of present (red) and absent (blue) amino-acid genotypes within every analysed dataset. The amino-acid Sørensen–Dice coefficient clustering mirrored the nucleotide result in which three distinct groups were identified: a group present in most datasets (green), a group present in the inoculum and OB products exclusively (orange) and a third group which consisted of genotypes randomly appearing and disappearing during the infection cycle (yellow).

7.5 DISCUSSION

The AC53 OB stock used for infection had previously been identified to contain an estimated 329 genotypes within BRO-A (Noune & Hauxwell, 2017b), and previously been suggested that the validation of these genotypes would require multiple ultra-deep sequencing runs throughout the infection cycle. Following the previous study, this number has been revised as 289 common genotypes with a minimum of 1x coverage were identified during the infection cycle and the final OB product. The 41 unique genotypes identified within the AC53 stock were all below 10x coverage, and were either outcompeted during the infection cycle or the result of sequencing error. Furthermore, a revision of the minimum coverage threshold required to call a genotype could be lowered to 1x coverage instead of the previously suggested 20x coverage (Noune & Hauxwell, 2017b) as some of the common genotypes identified had at least 1x coverage. Furthermore, 107 amino-acid genotypes were encoded from the 289 common nucleotide genotypes, of which 29 of the amino-acid genotypes produced a predicted functional protein. Future enhancements to MetaGaAP-Py may focus on identifying amino-acid genotypes to identify functional mutations occurring within a population and, to improve computational performance and the remove redundancy caused by multiple nucleotide sequences encoding a single amino-acid genotype.

The high nucleotide similarity observed with the common genotypes was expected as they had been identified within a single isolate. Furthermore, the Tajima's D and Fay and Wu's H results were are at odds, indicating a false-positive bottleneck and therefore could not be used to determine if the identified genotypes have undergone a selective sweep or balancing selection (Jensen, Kim, DuMont, Aquadro, & Bustamante, 2005).

Mean relative evolution rate analysis identified twelve positions within the BRO-A targeted region that were evolving faster than neutral evolution. A previous study has suggested profiles similar to what has been observed in this study is caused by weak negative selection and mutation bias (Lawrie, Petrov, & Messer, 2011). Furthermore, the 'drift barrier' theory could explain why most of the sequence appeared to be evolving slower than neutral (Lynch & Walsh, 2007; Sung, Ackerman, Miller, Doak, & Lynch, 2012).

The analysis of BV during the infection cycle identified significant changes in relative abundance and a significant increase in read counts, but were found to be non-linear. However, this may not necessarily reflect the actual virus copy number due to sequencing errors and primer bias, but such counts could provide a reasonable approximation. In addition, relative abundance of both nucleotide and amino-acid genotypes changes very little regardless of the increase in virus production. To put into perspective, the rapid increase in read counts within 24 hrs of infection can be attributed to the early infection cycle in which the virus has infected gut-epithelial cells and begins producing BV (Nguyen, Chan, Nielsen, & Reid, 2013; G. F.

Rohrmann, 2013a). The production in BV increases during the cycle as the BV enters the haemolymph and spreads throughout the insect into fat bodies and organs, until peak viral load is reached at 120 hrs and 144 hrs P.I (Hauxwell, 1999). It would be at these points that BV production ceases and OB production begins for environmental transmission.

Hierarchical clustering and scatterplots of the relative abundance suggested that time-points during the infection can be clustered into two categories early/very-late infection (24hrs to 48hrs and 144hrs), and mid-late infection (72hrs - 120hrs). This inexplicable link maybe the due to a potential cycle in OB vs BV genotypes during the infection, with a peak in BV genotypes during the early stages of infection and a peak in OB genotypes during the mid-late infection. Furthermore, even though read counts increased during the infection cycle, the relative abundance of genotypes during the infection was relatively flat and may be explained by the competitive exclusion principle or niche differentiation (Hardin, 1960; Pocheville, 2015).

Essentially, the dominant genotype has a long-term competitive advantage over the minor genotypes caused either by some previous ecological or evolutionary shift, with the minor genotypes occupying an ecological niche. An alternative theory maybe that the minor genotypes are hitchhikers and the dominant genotype is a brute-force, go-to genotype that can overcome host immunity but once resistance is developed, a minor genotype which can overcome this resistance could begin to dominate. Alternatively, the 'viral quasispecies' model may explain these results through the concept known as mutational robustness or 'survival of the flattest' in which each genotype within the quasispecies has equal fitness and can exploit selection to preserve population diversity (Van Nimwegen et al., 1999; Wilke et al., 2001).

Analysis of the infection cycle demonstrated a clear, statistically significant change in abundance between the inoculum and four OB samples isolated at the completion of the infection cycle (96hr and 120hrs). The significant difference in relative abundance between the inoculum and the four P.I. OB samples suggests that potentially, the initial abundance of genotypes will not be at the same proportions as the final OB produced. This was demonstrated with a mean reduction of 15% in abundance of the dominant nucleotide genotype, and mean reduction of 12% of the dominant amino-acid genotype in all four final OB samples and a significant increase in relative abundance of minor genotypes. However, this may be attributed to the production method used to produce the inoculum (Noune & Hauxwell, 2016a) and the final OB samples during this infection cycle. Potentially, this may implicate commercial production of baculoviruses as the change in production method has resulted in a different final product.

Analysis of genotype presence-absence demonstrated a distinct difference between the nucleotide and amino-acid BV infection cycle, as more amino-acid genotypes were detected. However this is indicative of the redundancy introduced by the large number of nucleotide genotypes and could be explained as a drift barrier (Sung et al., 2012). The increase in

identifiable genotypes observed during the infection cycle may correlate with the increase in read counts as low abundance genotypes could potentially be detected.

Furthermore, the BV infection cycle differed significantly from the inoculum and final OB products as they all contained the entire 289 nucleotide genotypes and 107 amino-acid genotypes. Heat-mapping and Sørensen–Dice coefficient clustering identified three distinct groups in both the amino-acid and nucleotide genotypes: a group present in most datasets, a group present in the inoculum and OB products exclusively and a third group which consisted of genotypes randomly appearing and disappearing during the infection cycle. This significant discrepancy between OB and BV samples could potentially be explained by three theories: 1) The techniques used to extract the BV DNA has resulted in a significant loss of genotypes whereas sequencing the OB has provided a snapshot of the entire genotype population, but this is pure speculation and would need to be validated. 2) OB genotypes occupy an undetectable resource niche during the BV infection cycle but "switch-on" and rapidly replicate to play an unknown role in OB production or mediation, and may be explained by the 'viral quasispecies model' and niche differentiation. 3) BV samples were densely packed onto a single 318v2 chip resulting in lower coverage per dataset than the OB samples which were sequenced together on a separate chip that was not densely packed and therefore sequencing sensitivity was reduced.

In conclusion, the large population of genotypes observed within BRO-A and variations in evolutionary rates is indicative of a population that may be enacting a 'drift barrier' to reduce the effects of genetic drift. Furthermore, these results potentially imply that NPVs act as a 'viral quasispecies' as selection could be acting on mutant clouds rather than individual genotypes, as indicated by the large proportions of nucleotide genotypes encoding a single amino-acid genotype. This has previously been suggested with NPVs but with little empirical evidence (Chateigner et al., 2015; Cory et al., 2005; Domingo et al., 2012; Vignuzzi et al., 2006; Wilke, 2005). In addition, implications in the production of NPV-based biopesticides have been observed in this study as the abundance of genotypes in the final OB product was significantly different to the initial starting OB product. Additional rounds of the infection cycle using the final OB product produced from each round may need to be completed to accurately determine if the relative abundance returns to the same levels as the inoculum used for the round 1 infection.

Chapter 8: Strain Selection & Trade Offs Between Virulence and Transmission Under Selection Pressure during *In Vivo* passage of the HaSNPV-AC53 isolate.

8.1 ABSTRACT

The Nucleopolyhedroviruses (family; Baculoviridae, genus; Alphabaculovirus) are invertebrate-specific obligate pathogens in which a host is infected by a community of phenotypically and genetically diverse virus strains that are co-occluded within a protein body during horizontal transmission. Baculoviruses are widely used in biological control of Lepidopteran pests, and understanding isolate dynamics, diversity and evolution is important in resistance management strategies and developing next generation biopesticides with desired phenotypic traits. In this study, we apply three selection pressures over five generations; fast speed-of-kill, slow speed-of-kill and maximum virus production to the commercial wild-type *Helicoverpa armigera* single nucleopolyhedrovirus isolate AC53 to determine virulence-transmission trade-offs and relative performance of each trait. The wild-type isolate was identified to be phenotypically superior compared to the trait-specific derived strains as the wild-type can balance virulence-transmission effectively. However, each trait-specific derived strain was identified to have improved trait-specific pathogenic characteristics but contain significant virulence-transmission trade-offs. Genotypic analysis of these derived strains is yet to be investigated.

8.2 INTRODUCTION

Nucleopolyhedroviruses (genus *Alphabaculovirus*: Family. *Baculoviridae*) or 'NPVs' are a genus of invertebrate-specific obligate pathogens (George Rohrmann, 2011a; White et al., 2012). They are commonly used as biopesticides, especially in Australia within integrated pest management systems to reduce insecticide resistance (Buerger et al., 2007; G. Fitt et al., 2005; Gary P. Fitt, 2000; Hauxwell, 2008a).

NPVs have two distinct life-history stages; budded virus (BV) which are single virions that are produced during *in vivo* transmission between cells during infection, and occlusion bodies (OB), in which several thousand singly- (SNPV) or multiply-enveloped (MNPV) virions are embedded in protein occlusion bodies, which are responsible for horizontal transmission of

the virus (Blissard & Rohrmann, 1990; G. F. Rohrmann, 2013c, 2013d). Each OB may contain multiple genetically-related virus variants co-occluded within a single, protein body with differing phenotypic properties reducing the chance of insect resistance (Vicky Lynne Baillie & Bouwer, 2012b; Blissard & Rohrmann, 1990; Chateigner et al., 2015; Cory et al., 2005; Goulson & Hauxwell, 1995; Noune & Hauxwell, 2016a; Ogembo et al., 2007; Elizabeth M. Redman et al., 2010; Reeson et al., 1998). Identification of baculovirus strains within an isolate typically uses selection of 'cloned' strains *in vitro* by selection in tissue culture of plaques from budded virus or occlusion-body derived virions (Brown & Faulkner, 1977, 1978; Corsaro & Fraser, 1987; Noune & Hauxwell, 2016a; Ogembo et al., 2007; Elizabeth M. Redman et al., 2010; Simon et al., 2011). Alternatively, strains may be selected by repeated infection *in vivo* with end-point dilution of the inoculum (theoretically resulting in infection by a single occlusion body) (Brown & Faulkner, 1977; I. R. Smith & Crook, 1988; Vlak, 1979).

An early example of this can be seen with the application of a low mortality dose infection for the *in vivo* isolation of *Pieris rapae* GV and *Lymantria dispar* MNPV strains, in which the first direct evidence of the independent action hypothesis for microbial pathogenicity was demonstrated (Meynell & Stocker, 1957; I. R. Smith & Crook, 1988).

Strain selection is of commercial interest: increased speed of kill might reduce crop damage, while increased yield might maximise yields during production. (Marie Berling et al., 2009; Evans & O'Reilly, 1998; D.J Hodgson et al., 2004; Noune & Hauxwell, 2016a). However, baculoviruses with low diversity can cause resistance after repeated insect exposure such as with the granulovirus (genus *Betabaculoviruses*) infecting *Cydia pomonella* (codling moth) (Arrizubieta et al., 2015b; S. Asser-Kaiser et al., 2007; Bernal, Simón, Williams, Muñoz, & Caballero, 2013; Cary et al., 1989; Gebhardt, Eberle, Radtke, & Jehle, 2014; Gilbert et al., 2014; Hamblin et al., 1990; Elisabeth A Herniou et al., 2004; J. Jehle, Eberle, Asser-Kaiser, Schulze-Bopp, & Schmitt, 2010). In this example, constant exposure of a CpGV isolate to resistant codling moth populations over several generations produced a new strain that overcame the codling moth resistance (Marie Berling et al., 2009; Graillot et al., 2014). The changes in NPV diversity, strain dynamics and phenotypic characteristics in response to selection provides a useful model in which to study viral evolution and ecology, but may also have benefits in the commercial use of NPVs as biopesticides through selection of strains with enhanced pathogenic traits (Arrizubieta et al., 2015b; Hamblin et al., 1990; Noune & Hauxwell, 2016a; Elizabeth M Redman et al., 2016; White et al., 2012).

Both *in vivo* and *in vitro* strain selection introduces transmission bottlenecks, reducing the genetic diversity and phenotypic composition within an isolate, (Elizabeth M Redman et al., 2016; Adam L. Vanarsdall, Okano, & Rohrmann, 2005). Reducing isolate composition through strain selection causes trade-offs between speed of kill ('virulence'), percentage kill ('pathogenicity') and transmission which leads to reduced efficacy (Elizabeth M Redman et al.,

2016; White et al., 2012). This is known as the virulence-transmission trade-off hypothesis (Bull & Lauring, 2014; Elizabeth M Redman et al., 2016; van Baalen & Sabelis, 1995; White et al., 2012). The hypothesis is evident with fast-killing strains as they usually undergo fewer replication rounds, reducing OB production or kill so rapidly that OBs are not produced thus reducing the chance of secondary infections (horizontal transmission) (Fleming-Davies et al., 2015; Elizabeth M Redman et al., 2016; White et al., 2012). In a natural environment, these strains would have reduced fitness caused by the lack of transmission (van Baalen & Sabelis, 1995). However, previous studies have demonstrated serial passaging relaxes selection on pathogen transmission and virulence tends to increase because no cost is associated with killing rapidly (van Baalen & Sabelis, 1995; White et al., 2012). Slow strains tend to have increased rounds of replication as they drag the infection-cycle out, increasing OB production and increasing viral distribution for horizontal transmission, again demonstrating that increased transmission, reduces virulence (Fleming-Davies et al., 2015; Elizabeth M Redman et al., 2016; White et al., 2012). On the other hand, tissue culture plaque selection (*in vitro*) may bias towards phenotypic traits more suited to within-host interactions (vertical transmission) (Elisabeth A Herniou et al., 2003; Lua et al., 2002; Lua & Reid, 2000; Noune & Hauxwell, 2016a, 2016b), which include the loss of the moulting inhibitor, *ecdysteroid UDP-glucosyltransferase* (*egt*) gene (Evans & O'Reilly, 1998; Robert L. Harrison, 2009a; Lua et al., 2002; Lua & Reid, 2000; O'Reilly & Miller, 1991; Simon et al., 2011).

Commercially, all three examples (virulence verse OB production, virulence verse transmission and virulence verse pathogenicity) of trade-offs must be balanced to create a viable product as strains which are too slow allow for increased crop-damage whereas fast and plaque purified strains have reduced between-host transmission (Hails et al., 2002; Elizabeth M Redman et al., 2016; van Baalen & Sabelis, 1995; White et al., 2012). Co-occlusion of strains offers a solution to balancing differing phenotypic properties and have been shown to have improved insecticidal properties but require high lethal doses otherwise the most competitive strain will become the most prevalent (Arrizubieta et al., 2015b; Hamblin et al., 1990).

In this study, three different selection pressures (fast speed of kill, slow speed of kill and maximum OB production) were applied over 5 rounds of infection ('passages') of a baculovirus of commercial importance, the *Helicoverpa* spp. specific baculovirus isolate HaSNPV-AC53 (Noune & Hauxwell, 2015, 2016a). Changes in speed of kill, viral yield in the progeny virus and evolutionary trade-off caused by the shift in community composition through the application of three different pressures is discussed.

8.3 MATERIALS AND METHODS

8.3.1 Virus and Insect Source

The wild-type isolate HaSNPV-AC53 (AC53) was obtained and purified as previously described (Noune & Hauxwell, 2015). *Helicoverpa armigera* neonates were obtained from AgBiTech Pty Ltd and reared on a wheat germ and soy flour based diet in a temperature controlled growth room at 25°C ± 1°C with 16hrs light and 8hrs dark periods.

8.3.2 Infection and Selection of Viral strains

In the first round of infection, 252 neonate *H. armigera* larvae were inoculated using a droplet assay with a suspension of $8x10^5$ (LC_{100}) occlusion bodies (OB) per mL (OB/mL) with 10% sucrose and 10% blue food-dye (P. R. Hughes & Wood, 1981). Insects were placed into individual 30 mL plastic cups containing approximately 1 cm^3 of diet and put into a temperature controlled growth room with the above conditions. Insects that had died up to and including 24hrs post-infection (P.I.) were classed as manual handling deaths and removed. All deaths after 24hrs were recorded and cadavers collected.

'Fast' strains of virus were selected by collection of all larvae that died after 24hrs and up to and including 72hrs P.I. These were harvested individually and OB count per cadaver was determined before OBs were pooled and used for the next passage, as below.

'Slow' strains of virus were selected by isolation from the last insect to die. The OB count per dry weight of the cadaver was determined and the OBs extracted were used for the next passage as below.

Strains with the maximum virus yield (maxOB) were selected from the insect with the highest OB count per mass of dry weight of the cadaver. Cadavers from all time points were freeze dried for 24 – 48hrs in individual 1.5mL microcentrifuge tubes, then weighed. A 1mL volume of 0.1% sodium dodecyl sulphate (SDS) was added to each cadaver and the sample was homogenised using a sterile plastic mortar and pestle. OB count per cadaver was determined using a standard haemocytometer in triplicate (Arrizubieta et al., 2015b) and OB/µg was calculated from the total count/dry mass of cadaver. The single cadaver with the highest OB/µg was used for the next passage.

The 'fast', slow' and 'maxOB' cadavers in SDS were centrifuged for 10 min in a microcentrifuge at 10,000 rcf to form a pellet and supernatant was carefully discarded. The pellet was resuspended in 500µL of 50% analytical grade glycerol mixed with MilliQ water (Merck Millipore, Massachusetts, United States of America), counted again using a haemocytometer and used as the viral stocks for the next passage.

Subsequent passages infected 54 neonates per treatment. The generation selection and concentration of inoculum used varied at each generation depending upon the yield of virus produced in the previous passage are summarised in figure 8-1.

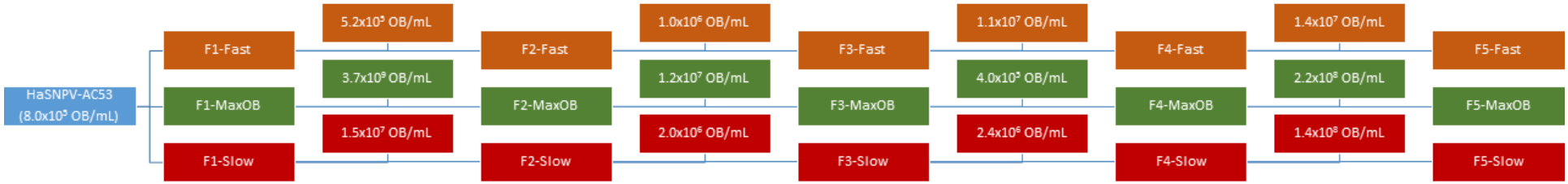


Figure 8-1: Summary and doses used for each generation of selection. Fast strains (orange) were selected using all cadavers between 24-72hrs P.I. MaxOB strains (green) were selected using the cadaver with the highest OB/µg. Slow strains (red) were selected using the cadaver that had died last.

8.3.3 Biological Performance Characterisation

The biological performance was assessed using the F5 generation of each selection using a dose-range finding droplet assay with *H. armigera* neonates (P. R. Hughes & Wood, 1981). The negative control for both LC_{50} and ST_{50} consisted of 42 neonates dosed with a 20% sucrose and 10% blue food dye solution.

Median lethal concentration (LC_{50}) was analysed using 42 neonates per dose and a set of eight standard doses (Table 8-1) for each selection pressure with the parent AC53 strain used as the positive control.

ST_{50} was determined using the two highest doses (dose 1 and 2) and 42 neonates per dose for all three strains and at $1.80x10^6 \pm 1\%$ OB/mL for the F5-Fast strain with AC53 used as the positive control. Manual handling deaths were removed 24hrs P.I. and mortality was calculated from 36hrs P.I. at intervals every 8hrs. OB counts, and dry weights were measured using insect deaths from dose 1 and the method previously listed.

Table 8-1: Concentration of viral doses for LC50 performance measurement.

Dose Number	Concentration (OB/mL)	Percentage kill of dose using AC53
1	$1.52x10^5 \pm 1\%$	LC_{95}
2	$1.44x10^5 \pm 1\%$	LC_{90}
3	$1.12x10^5 \pm 1\%$	LC_{70}
4	$8.0x10^4 \pm 1\%$	LC_{50}
5	$4.8x10^4 \pm 1\%$	LC_{30}
6	$1.6x10^4 \pm 1\%$	LC_{10}
7	$8.0x10^3 \pm 1\%$	LC_5
8	$1.6x10^3 \pm 1\%$	LC_1

8.3.4 Statistical Analysis

Passage variations were analysed using a Pearson product-moment correlation coefficient and covariance's were calculated using the Stats 3.4.0 package and principal component analysis (PCA) completed using the devtools 1.12.0, data.table 1.10.4 and ggbiplot 0.55 packages within Microsoft Open R 3.4.0 (Dowle, Short, & Lianoglou, 2013; Microsoft, 2016; Onwuegbuzie, Daniel, & Leech, 2007; R. Team, 2014; R. C. Team, 2013; Vu, 2011; Wickham & Chang, 2015).

Standardised LC_{50} (Equation 8-1) was determined from a bioassay of the final passage virus isolate using Microsoft Open R 3.4.0 and a generalised linear model (GLM) with a quasi-binomial likelihood and logit-link function within the 'Modern applied-statistics with S-PLUS' (MASS) package, and Abbott's mortality correction using the custom script RBassay version 2.0 as part of the 'Invertebrates & Microbiology Group' pipelines suite (Abbott, 1925; Microsoft, 2016; Noune, 2016; R. Team, 2014; R. C. Team, 2013; Venables & Ripley, 2013; Wedderburn, 1974). Estimates of LC_{50} at doses causing less than 60% mortality were extrapolated.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \log_{10}(dose_i)$$

Equation 8-1: Parameters are defined as follows: p_i = probability of i^{th} insect death given dose$_i$, β_0 = intercept, dose$_i$ = dose such that β_1 represents the linear trend in insect deaths for i = 1…,N.

During each passage, an estimate of 'median survival time' (ST_{50}) was calculated from the mortality in each of the dosages listed in figure 1. ST_{50} was calculated with a Kaplan-Meier estimator, using Microsoft Open R 3.4.0, and the survival version 2.41-3 package. Survival curves were produced using the survminer version 0.3.1 package.

Three measures of virus production over time were compared for each of the F5 selected strains and AC53 parent isolate (Equation 8-2): virus yield, virus density and dry weight of cadavers. Virus yield: cumulative total virus concentration (OB/mL) per total number of dead insects at each time point. Virus capacity: cumulative dry cadaver weight (µg) per total number of dead insects at each time point. Virus density: cumulative virus density (OB/µg) per total number of dead insects at each time point. Statistical significance was evaluated via a GLM with a quasi-Poisson likelihood to account for over-dispersion using Microsoft R Open 3.4.0 and the MASS package (Haight, 1967; Microsoft, 2016; Venables & Ripley, 2013; Wedderburn, 1974).

$$\log y_i = \delta_0 + \delta_1 t_i$$

Equation 8-2: Parameter i^{th} are defined as follows: y_i = virus yield, or virus density, or insect weight (viral capacity), δ_0 = intercept, t_i = time such that δ_1 represents the linear trend in virus yield, or virus density or insect weight (viral capacity) for i = 1…,N.

Relative potency and relative survival time was calculated using a GLM with a quasi-binomial likelihood and logit link function and a Bonferroni-adjustment of Fieller confidence intervals (C.I.) at a 95% confidence level within the sci.ratio.gen function, as part of the mratios R package (Djira, Hasler, Gerhard, Schaarschmidt, & Schaarschmidt, 2011; Dunn, 1961; Dunnett, 1964; Fieller, 1954; Wedderburn, 1974). LC_{50} and viral production scatterplots were produced using the R package ggplot2 version 2.1.0 and Microsoft R Open 3.4.0 with a line of best fit applied using equation 8-1 (LC_{50}) or equation 8-2 (viral production) and the formula: 1 – y ~ poly(x, 2) (Microsoft, 2016; Wickham, 2016).

8.4 RESULTS

8.4.1 Initial Observations of Generational Selection

A total of fifteen selected viral strains were isolated over five generations (five strains per pressure) from the initial AC53 stock.

Comparisons of the F1 selected strain ST_{50} estimations (Table 8-2 to Table 8-4) identified the F5 fast strain to be 36hrs faster, F5 slow strain to be 12hrs slower and the F5 maxOB strain to be 36hrs faster. Mean viral density decreased per generation regardless of applied pressure with an exception at F4 where a spike in OB production was observed but could be correlated to the decreased dosage. The maxOB strains isolation time point varied between 96hrs and 144hrs with the highest viral density and the last insect death observed for the slow strains varied between 120hrs to 192hrs.

Table 8-2: Fast selected strains isolation metrics showing a 49hr improvement in ST_{50} between the F1 and F5 generation. Total deaths at 72hrs has a linear increase at per generation but can be attributed to the dosage.

Generation	ST_{50} (Hrs)	Dosage (OB/mL)	Total Deaths (Up to 72hrs)	Mean Density (OB/µg)	Mean Weight (µg)	Mean Total Yield (OB/mL)
F1	99	$8.00x10^5$	7	$2.33x10^4$	456	$3.94x10^5$
F2	97	$5.20x10^5$	14	$1.39x10^3$	240	$2.48x10^5$
F3	60	$1.00x10^6$	30	$1.59x10^3$	195	$2.44x10^5$
F4	66	$1.10x10^7$	33	$2.05x10^3$	197	$3.19x10^5$
F5	50	$1.40x10^7$	44	$1.26x10^3$	270	$2.45x10^5$

Table 8-3: Slow selected strains isolation metrics showing a 10hr increase in ST_{50} between the F1 and F5 generation and a spike in OB/µg and total OB/mL at F4. Isolation time point is relatively stable at 144hrs.

Generation	ST_{50} (Hrs)	Dosage (OB/mL)	Isolation Time Point (Hrs)	Selected Density (OB/µg)	Selected Weight (µg)	Selected Total Yield (OB/mL)
F1	99	$8.0 x10^5$	192	$2.33x10^4$	1090	$5.00x10^7$
F2	95	$1.5 x10^7$	144	$1.20x10^4$	500	$6.00x10^6$
F3	91	$2.0 x10^6$	120	$4.80x10^3$	400	$1.92x10^6$
F4	91	$2.4 x10^6$	144	$8.00x10^5$	300	$2.40x10^8$
F5	109	$8.0 x10^5$	144	$2.60x10^4$	500	$1.30x10^7$

Table 8-4: MaxOB selected strains isolation metrics showing a 32hr improvement in ST_{50} between the F1 and F5 generation, and again a spike in OB/µg and total OB/mL at F4. The time point with the highest OB/µg varied due to the dosage.

Generation	ST_{50} (Hrs)	Dosage (OB/mL)	Mean Density (OB/µg)	Mean Weight (µg)	Mean Total Yield (OB/mL)	Time Point with Maximum OB/µg (Hrs)
F1	99	8.0×10^5	2.33×10^4	410	8.00×10^7	144
F2	83	3.7×10^9	2.86×10^3	200	6.87×10^5	120
F3	76	1.2×10^7	3.45×10^3	186	6.22×10^5	96
F4	94	4.0×10^5	1.19×10^5	2262	3.62×10^7	144
F5	67	2.2×10^8	2.40×10^3	195	9.21×10^5	96

PCA of key metrics (Tables 12-8 to 12-10) for all the selected strains could be explained with component 1 (PC1) and 2 (PC2), suggesting that dosage is the contributing factor for all the observed variation.

Fast strains correlation (Table 12-11) identified the ST_{50} moderately positively correlated to the average viral density, average weight and average total viral yield, whereas the dosage has a strong positive correlation to the total insect deaths up-to 72hrs. Covariance statistics (Table 12-12) elaborated on these results and suggested that as the dosage and total deaths increased, all other metrics decreased. Slow and maxOB strains correlations and covariance (Tables 12-13 – 12-16) mirrored the fast strains result with dosage causing the variation.

8.4.2 Standardised Performance Metrics of Selection

Standardised LC_{50} and ST_{50} Results

LC_{50} values estimated from bioassays using the same concentration range for all strains (Table 12-17, Table 8-5) identified the parent strain (AC53) with the lowest lethal dosage requirement to kill 50% of the effective number treated (ENT). The LC_{50} for the fast and maxOB strains were a forecasted estimate since mortality was less than 60% at the highest dose (Table 8-5, Figure 8-2). Relative potency to AC53 identified all three strains to be less potent (Table 8-5).

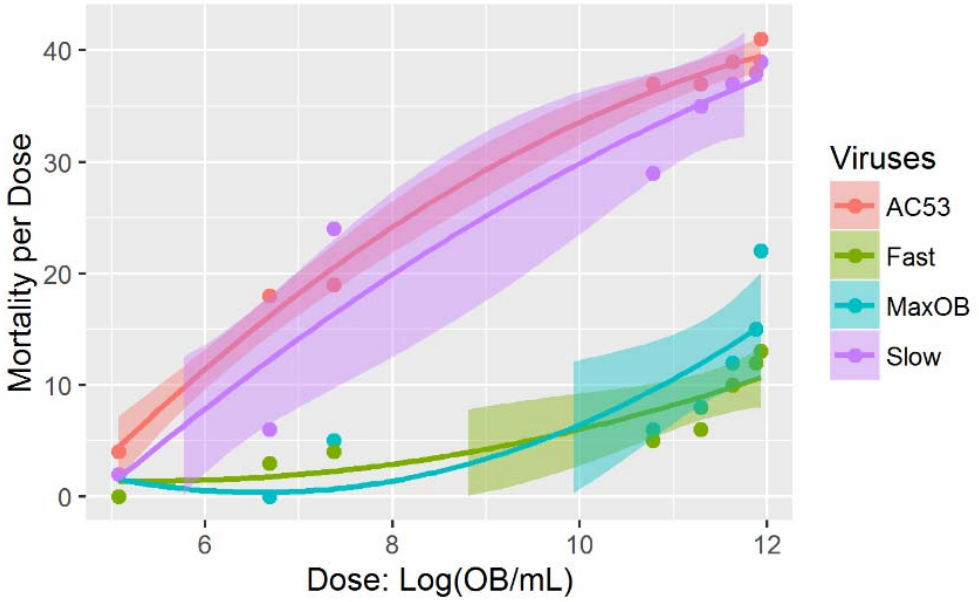


Figure 8-2: Standardized dose range bioassay (LC_{50}) for F5 selected strains. All selected strains performed worse than AC53, which suggests a reduction in community composition, causing a loss in pathogenic efficacy. A quasi-binomial GLM line of best-fit using equation 1 and the formula: $1 - y \sim \text{poly}(x, 2)$ has been applied to visualise the change in mortality, and shading indicates 95% confidence intervals.

Table 8-5: LC statistical summaries of the F5 strains to AC53 indicating significant results. The slow strain is the most comparable to the parent while both maxOB and fast are the least potent. The reduced potency can be attributed to the reduction of community composition and pathogenic diversity as strains have been pressured to specific pathogenic traits.

Strain	LC_{50} (OB/mL)	Standard Error (OB/mL)	β_1 Estimate (Standard Error)	β_1 Pr(>|t|)	Dispersion Parameter	Relative LC to AC53		
						Lower 95% C.I.	Potency (LC)	Upper 95% C.I.
AC53	$2.12x10^3$	1.28	1.5139 (0.1108)	$9.53x10^{-6}$	0.541	N.A.	**1.0**	N.A.
Slow	$4.49x10^3$	1.29	1.4506 (0.2631)	0.0015	3.436	**0.62**	**0.94**	**1.40**
MaxOB	$3.95x10^5$	1.51	1.3465 (0.4552)	0.0253	2.845	**-0.11**	**0.22**	**0.53**
Fast	$2.82x10^6$	2.92	0.8409 (0.2142)	0.0078	1.039	**-0.25**	**0.09**	**0.39**

Standardised ST_{50} metrics (Table 12-18 – 12-20 and Table 8-6) for the fast and maxOB strains were estimated as less than 60% mortality was observed. The slow, maxOB and fast strains (Table 8-6, Figure 8-3) were found to be slower than AC53 at doses 1 and 2 with a relative survival time of 0.92x to 0.88x (slow), 0.44x to 0.38x (maxOB) and 0.37x to 0.43x (fast) respectively, suggesting the strains have traded virulence for OB production and longer survival times. The fast strain comparison to AC53 at $1.80x10^6$ OB/mL identified a 1.17x speed improvement in survival performance but with 24% less mortality suggesting a virulence trade-off has occurred between kill speed, pathogenic efficacy and OB production (Figure 8-4).

Table 8-6: ST statistical summaries of F5 selected virus compared to AC53. Comparison of the fast strain to AC53 at the highest dose identified the fast strain to be 1.17x faster but with 24% less mortality. The slow, fast and maxOB strains were all found to be slower than AC53 at dose 1 and 2.

Dose (OB/mL)	Strain	Mortality of ENT* (%)	Lower 95% C.I. (Hrs)	ST_{50} (Hrs)	Upper 95% C.I. (Hrs)	Relative ST to AC53		
						Lower 95% C.I.	Potency (ST)	Upper 95% C.I.
$1.80x10^6$	AC53	100	80	80	88	N.A.	1.0	N.A.
	Fast	76	56	64	80	1.05	1.17	1.32
Dose 1 ($1.52x10^5$ ± 1%)	AC53	95	88	96	104	N.A.	1.0	N.A.
	Slow	95	104	104	112	0.83	0.92	1.01
	MaxOB	52	128	160	208	0.36	0.44	0.53
	Fast	31	152	200	Not Applicable	0.28	0.37	0.46
Dose 2 ($1.44x10^5$ ± 1%)	AC53	95	96	104	112	N.A.	1.0	N.A.
	Slow	93	104	112	112	0.77	0.88	1.00
	MaxOB	36	136	176	Not Applicable	0.30	0.38	0.46
	Fast	29	72	168	Not Applicable	0.32	0.43	0.54

* Table 12-18, Table 12-19, Table 12-20.

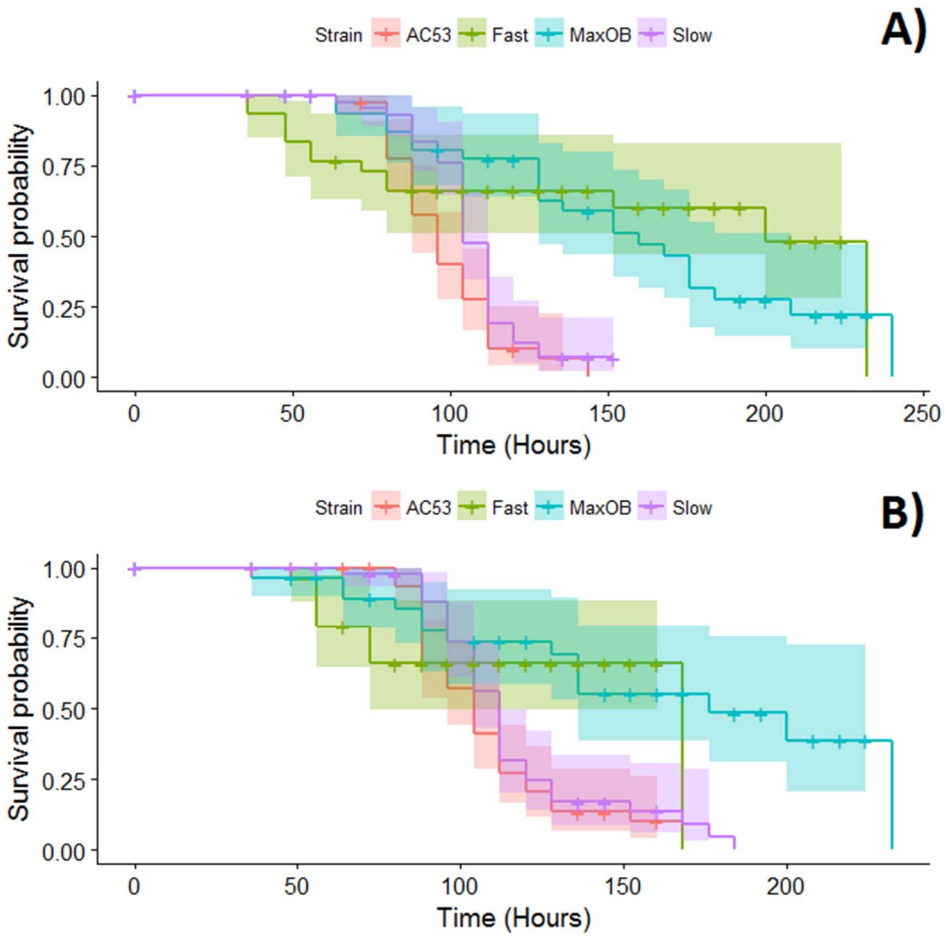


Figure 8-3: Kaplan-Meier ST metrics at A) Dose 1, and B) Dose 2. The slow strain was performing similarly to AC53, albeit with a slower lag phase between time of infection and the first insect death. The fast strain had the quickest lag phase but was the worst performing, caused by the poor efficacy of the strain, whilst the maxOB strain had the longest infection cycle suggesting that the strain is dominated with late-infection genotypes that may favour OB production. Shading indicates 95% confidence intervals.

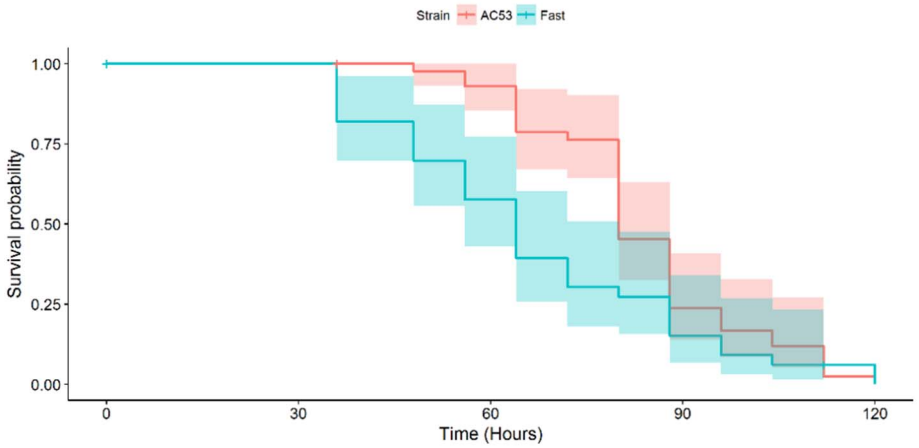


Figure 8-4: Kaplan-Meier ST metrics for AC53 and the fast strain at 1.80×10^6 OB/mL. The fast strain has a reduced lag-phase between infection and the first insect death before becoming inactive. Furthermore, the fast strain was observed to kill the effective number treated (ENT) faster than AC53. This suggests the fast strain is an early replicator with reduced efficacy caused by the loss of late replication genotypes. Shading indicates 95% confidence intervals.

Standardised OB Counts

Modelling of the standardized OB counts identified the maxOB strain to have the highest total virus yield (Figure 8-5, Table 12-21 – Table 12-22) and was followed by the slow strain, AC53 and the fast strain. This result mirrored the total viral capacity per insect (Figure 8-6, Table 12-21 – Table 12-22). Total viral density per insect (Figure 8-7, Table 12-21 – Table 12-22) identified the slow strain with the highest density, followed by AC53 and the fast and maxOB strains. A saturation curve was observed with the viral density and suggests that once peak density is reached, insect weight exponentially increases to increase viral capacity. These results suggest the maxOB strain has improved viral capacity and total OB production but poor viral density.

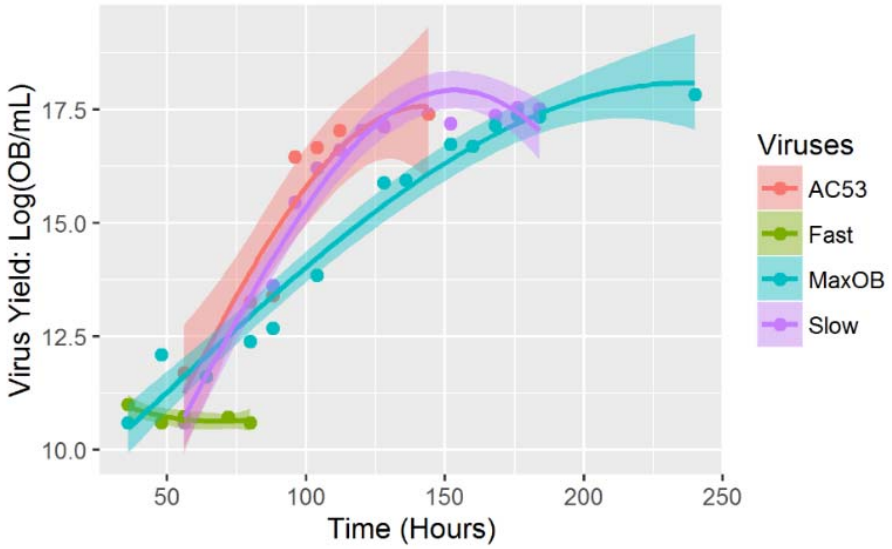


Figure 8-5: The longer infection time for the maxOB strain resulted in higher total viral yield, however, both the slow strain and wild-type isolate produced higher viral yield faster before insects succumbed to infection. The fast strain had low viral yield and suggests that virulence traits are delaying or preventing OB production. A quasi-Poisson GLM line of best-fit using equation 2 and the formula: $1 - y \sim \text{poly}(x, 2)$ has been applied to visualise the change in mortality, and shading indicates 95% confidence intervals.

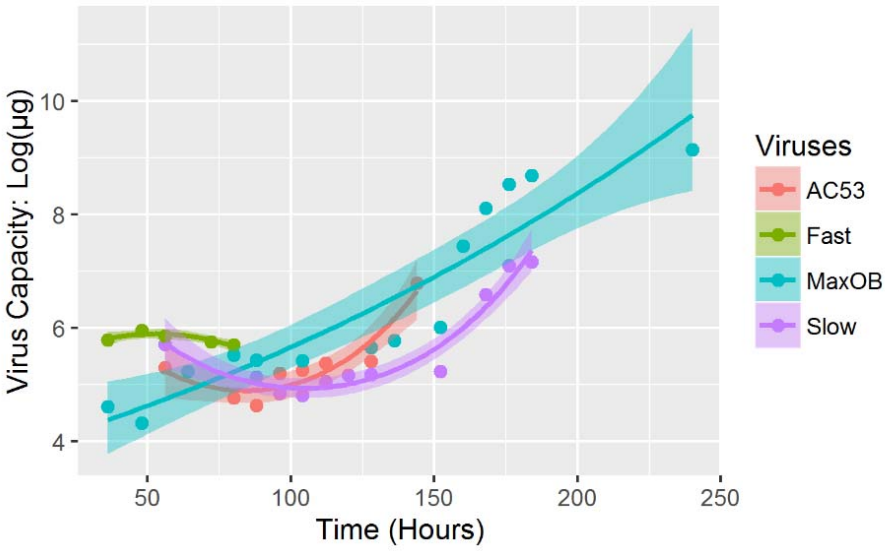


Figure 8-6: The maxOB strain had the highest viral capacity (insect weight) allowing for higher viral yield but at the cost of longer infection times. Both the slow strain and wild-type isolate viral capacity does not begin to increase until ~120hrs P.I. A quasi-Poisson GLM line of best-fit using equation 2 and the formula: $1 - y \sim$ poly(x, 2) has been applied to visualise the change in mortality, and shading indicates 95% confidence intervals.

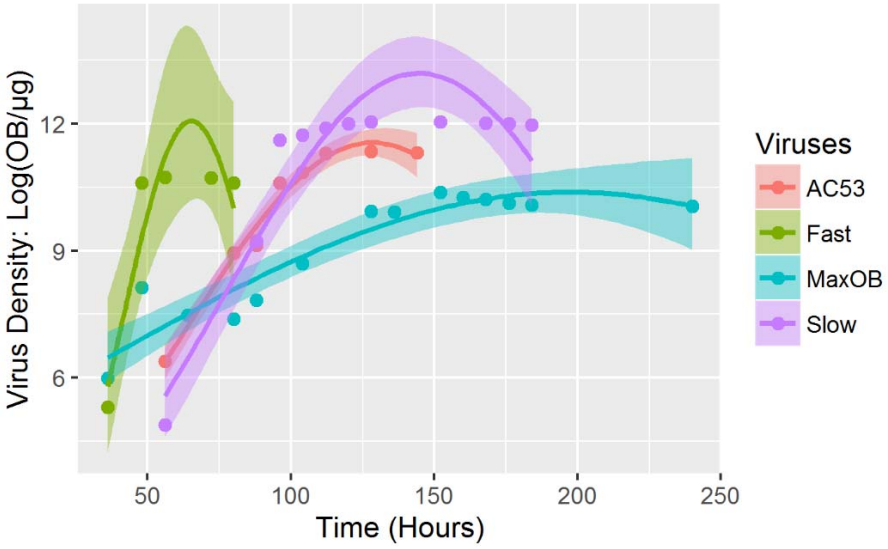


Figure 8-7: Viral density was highest in the slow strain and wild-type isolate (i.e. total virus/µg of insect). The density peaks at roughly the same time viral capacity begins to increase and is observed with all three pressured strains, however, the maxOB strain has the lowest viral density. A quasi-Poisson GLM line of best-fit using equation 2 and the formula: $1 - y \sim$ poly(x, 2) has been applied to visualise the change in mortality, and shading indicates 95% confidence intervals.

8.5 DISCUSSION

Selection of strains with specific pathogenic traits has demonstrated rapid adaptation towards one of the three pressures applied but as expected, a virulence-transmission trade-off occurred with selected strains (Bull & Lauring, 2014; Fleming-Davies et al., 2015; Elizabeth M Redman et al., 2016; van Baalen & Sabelis, 1995; White et al., 2012).

Estimation of performance metrics for each generation of selection verified the dosage as the main contributing factor to the variances observed and therefore the results were unreliable.

However, generational selection suggested virulence-transmission trade-offs were occurring between selected strains. Removing viral dose bias through standardised bioassays validated the trade-off occurrence: LC_{50} is significantly higher in all three trait-selections and relative potency is significantly reduced. This reduction in efficacy was also observed in the ST_{50} metrics for the maxOB and slow strains, which had a higher ST_{50} than the wild type.

The fast strain had a 1.17x increase in speed-of-kill when compared to the wild type, but at the cost of pathogenicity. The LC_{50} of the fast strain was 3 orders of magnitude greater than that of the wild type. This is a far greater reduction in pathogenicity than previously reported studies where a 'fast-killing' strain has been derived, e.g. a 3x increase in LC_{50} relative to the parental strain in a 'fast-killing' plaque-isolated variant of *Agrotis ipsilon* MNPV (Robert L Harrison, 2013). The fast strain had a maximum percentage kill of only 31% at $1.52x10^5$ OB/mL, and the LC_{50} was calculated by extrapolation, however, the variance around the estimated LC_{50} was not great (2.92 OB/mL), and mortality at the $1.8x10^6$ OB/mL used in the ST_{50} assays was still only 76%, suggesting that this remarkable reduction in efficacy is valid.

Essentially, three trade-offs were observed with each trait-specific strain. Fast strains had improved virulence but at the cost of efficacy and viral production, mirroring previous studies suggesting fast traits may prevent or limit viral production, thus reduced fitness (Bull & Lauring, 2014; Cory et al., 2005; Fleming-Davies et al., 2015; White et al., 2012). However, selecting for fast mortality traits removes the fitness cost associated with transmission (van Baalen & Sabelis, 1995; White et al., 2012). Slow strains and maxOB strains had similar observed trade-offs, increased transmission at the cost of efficacy and virulence, however, this was more predominant in the maxOB strains. We suspect the slow strains are phenotypically similar to the wild-type as they may contain a closer resemblance of the full community composition resulting from selection of the last insect to die.

Although the maxOB strain selection was based on peak viral density, this was not reflected in the final product: instead, improved viral capacity and longer infection times were observed. Interestingly, the fast and slow strains and wild-type isolate all had higher maximum viral density than the maxOB strain, reached maximum viral capacity (observed as a plateau in viral density per unit mass) just before time of death, suggesting that once viral density reaches peak saturation, increases in OB production can be explained purely by increases in capacity (i.e. larval mass). However, the maximum viral density in maxOB strains remains below that of the fast, slow and wild types, and peak viral capacity (the plateau in viral density per unit mass) is lower in the maxOB strains.

AC53 was found to be phenotypically superior when LC_{50} and ST_{50} results were compared to the selected strains as it balances virulence-transmission trade-offs and can be attributed to it containing the full population. This reflects previous studies which demonstrated

genetically mixed isolates have improved overall performance compared to single-genotype infections (D.J Hodgson et al., 2004; Elizabeth M Redman et al., 2016).

In conclusion, we have demonstrated that selecting for specific phenotypic traits produces strains with desirable traits, but the virulence-transmission trade-offs need to be taken into consideration for commercialisation. This may be resolved through strain mixing or if selection was to continue, negative traits may eventually be offset (Bull & Lauring, 2014). Furthermore, we speculate that selecting for specific traits has removed poor competitors from the population or are present in low levels. The genetic composition of these selected strains is still to be investigated, however we hypothesise that the reduction in community composition through transmission bottlenecks has reduced the population diversity and introduced a selective sweep and potential genetic-hitchhiking with new mutant genotypes reflecting the desired pathogenic trait (Burke, 2012; Cory et al., 2005; Gilbert et al., 2014; E.A. Herniou & Jehle, 2007; Lua & Reid, 2000; Messer & Petrov, 2013; J. M. Smith & Haigh, 1974; White et al., 2012).

Chapter 9: Genetic Analysis of Trait-Specific *In Vivo* Derived Strains from HaSNPV-AC53

9.1 ABSTRACT

Application of trait-specific *in vivo* selection can produce baculovirus strains with desirable traits, however, with virulence-transmission trade-offs. Current studies have focused primarily on the genetic analysis of *in vitro* derived strains which has left a gap in knowledge regarding genetic mutations potentially occurring as pressures are applied. In this study, NGS and several bioinformatic techniques were applied to fifteen previously derived selected strains obtained from the commercial isolate, HaSNPV-AC53 to understand the genetic relationships to the previously observed phenotypes. Nucleotide sequence comparisons of each strain's whole-genome could identify fast, slow and maximum OB production specific-mutations as well as phylogenetic clustering of each trait-specific strain. In addition, 8 diverged ORFs were identified to be common across all the analysed strains. Molecular clocking estimated divergence times which corresponded to what was expected of each trait-specific strain with fast strains diverging approximately 13 hrs post infection, maximum OB strains approximately 40 – 104 hrs and slow strains approximately 80 hrs. Estimation of polymorphism abundance within each strain could be completed however, detailed polymorphic comparisons and some evolutionary statistics could not be completed due to software and scale-up limitations.

9.2 INTRODUCTION

Phenotypic effects of *in vitro* and *in vivo* selection on baculoviruses is well studied with some genetic analysis showing the loss of *egt* during persistent *in vitro* infections (Robert L. Harrison, 2009a). However, no in depth genetic studies have been completed on *in vivo* derived selected strains.

This is especially important as chapter 7 shows the initial abundance of genotypes within AC53 changing throughout the infection cycle. Furthermore, previously described studies observed isolate mutations occurring within individual insects highlighting that baculovirus-host interactions are not static (Vicky Lynne Baillie & Bouwer, 2012b).

Essentially these studies suggest that baculoviruses fit the definition of a 'viral quasispecies' and by extension should be considered as one (Domingo et al., 2012; Vignuzzi et al., 2006). The quasispecies model suggests that if mutation rates are high, selection will act on a group of mutants or genotypes rather than individual genotypes within a population (Domingo et al., 2012; Wilke, 2005). This was evident in chapter 7 where mutation rates were identified to be

higher in specific regions of the targeted BRO-A region. It can be predicted from this that application of a selection pressure is acting on a subpopulation of genotypes within the AC53 isolate which have the desired trait-specific characteristics.

Therefore, the genetic characterisation of these derived strains is an important aspect in understanding the fundamental mutations associated with each applied selection pressure and which can lead to improved strain optimisations (Vicky Lynne Baillie & Bouwer, 2012a, 2012b; Gebhardt et al., 2014; Graillot et al., 2014; Noune & Hauxwell, 2016a; Elizabeth M Redman et al., 2016).

In the previous chapter, three selection pressures were applied over five generations – fast speed of kill, slow speed of kill and maximum OB production to the HaSNPV-AC53 isolate and derived fifteen strains (5 strains per pressure). Based on the lack of detailed genetic studies on *in vivo* derived strains, no hypothesis was previously given nor suggested on the ORFs which may potentially mutate, however, Hr regions should mutate as expected and as previously reported (Noune & Hauxwell, 2016a). This chapter aims to apply NGS to assemble each selected strain to determine the trait-specific mutations occurring, analyse and identify the phylogenetic relationships of these strains including core-SNPs, estimate viral divergence time and polymorphic abundance within each strain.

9.3 MATERIALS AND METHODS

9.3.1 DNA Purification, Sequencing and Assembly

The genomic DNA of the fifteen selected strains previously isolated in chapter 8 and of HaSNPV-AC53 (AC53) was purified as previously described (Noune & Hauxwell, 2016a). Sequencing library preparation was completed using a Nextera XT kit with an Illumina MiSeq and V3 chemistry (600-cycle). Genomes for all selected strains, including a new AC53 consensus genome (AC53 MiSeq) were assembled using the 'Invertebrates and Microbiology Groups Assembly Pipeline' version 1.5.4 with the AC53 (KJ909666) reference genome for reference mapping (Institute; Kearse et al., 2012; Heng Li, 2013; H. Li et al., 2009; Noune, 2016; Noune & Hauxwell, 2015, 2016a; Van der Auwera et al., 2013). Genomes were annotated as previously described (Kearse et al., 2012; Noune & Hauxwell, 2016a, 2016b).

9.3.2 Genome and Evolutionary Analysis

Comparison of the AC53 MiSeq and AC53 original genomes

The AC53 MiSeq genome and AC53 original genomes were compared to determine nucleotide and amino acid differences. Whole genomes and Sanger sequences of BRO-A and DNA polymerase from chapter 6 were aligned using MAFFT v7.222 with the FFT-NS-2 algorithm and default settings (Katoh & Standley, 2013; Noune & Hauxwell, 2017b).

Nucleotide and amino acid sequences of ORFs and Hr regions where compared using a local copy of BLAST+ version 2.5.0 with the Megablast algorithm to identify nucleotide mutations, and the blastp algorithm for amino acid mutations (Altschul, Gish, Miller, Myers, & Lipman, 1990; Camacho et al., 2009; Morgulis et al., 2008).

Distance, Recombination, Evolution and Phylogeny

All of the assembled genomes were aligned using MAFFT v7.222 with the FFT-NS-2 algorithm and default settings (Katoh & Standley, 2013). A nucleotide distance matrix of this alignment was produced using Geneious R9.1.7 (Kearse et al., 2012).

Comparisons of selected strains ORF and homologous repeat regions mutations were analysed using a local copy of BLAST+ version 2.5.0 with the Megablast algorithm to identify nucleotide mutations, and the blastp algorithm for amino acid mutations (Altschul et al., 1990; Camacho et al., 2009; Morgulis et al., 2008).

Recombination was analysed using the previously produced alignment with bratNextGen (Marttinen et al., 2012). A shared ancestry tree was produced, clustering set to a 0.01 threshold, and result significance determined by completing 100 replicate runs with a *p* value set at 0.05.

Tajima's D and Fay and Wu's H was calculated using the previously produced alignment and parameters as per chapter 7, however, with the sliding-window approach disabled. Mean (relative) evolutionary rates were calculated using MEGA7 with the previously described parameters as per chapter 7.

Maximum-likelihood estimation (MLE) tree construction was completed with RAxML version 7.2.8 using previously described parameters with the AC53 MiSeq genome used as the root (Noune & Hauxwell, 2016a; A. Stamatakis, 2006; Alexandros Stamatakis, 2014).

Using the previously constructed MLE tree as a guide, a time-tree was produced with MEGA-7.0.18 with the ReltimeML algorithm (Kumar et al., 2016; Tamura et al., 2012). Parameters were set to use a MLE statistical method, analytical variance estimation method, a gamma distributed GTR substitution model with 5 categories and data set to use all sites including gaps. Tree nodes were calibrated to estimate divergence times with time calibration constraints based on the described strain isolation times in chapter 8 (in hours).

A SNP MLE tree was produced using the previously described genome alignment. All common bases belonging to the 15 selected strains and AC53 genome were removed to identify the core SNPs belonging to each virus. These SNPs were re-aligned using MAFFT v7.222 with the FFT-NS-2 algorithm and default settings, and a MLE tree produced using the previously described parameters with the AC53 MiSeq SNPs used as the root.

All trees were visualised and edited using TreeGraph version 2.11.1-654 (Müller & Müller, 2004; Stöver & Müller, 2010).

Identification and Analysis of Polymorphisms

A partial implementation of MetaGaAP was applied to analyse polymorphic abundances from shot-gun sequencing data. Within-isolate (AC53) and within-strain (selected strains) polymorphisms were identified by anchoring all datasets to the new AC53 MiSeq genome using the GATK best-practices guidelines with GATK version 3.6 using the previously described parameters (Cingolani et al., 2012; McKenna et al., 2010; Noune & Hauxwell, 2017b; Van der Auwera et al., 2013). However, the HaplotypeCaller 'emit reference confidence' parameter was set to BP_RESOLUTION to determine the abundance of each individual polymorphism. Polymorphisms were visualised using Geneious R9.1.7 (Kearse et al., 2012). VCF files were then converted into tab-delimited files using the GATK version 3.6 VariantsToTable tool. Following conversions, polymorphic relative abundance was calculated by dividing the number of reads found to contain the polymorphism in a single position by the total reads assigned to that position using Microsoft Excel 2016.

SNP trees were produced from the polymorphic data and edited using the previously described tree construction parameters, however, data was pre-processed by manually extracting the polymorphisms from the previously converted tab-delimited files and manually creating fasta files.

K-means clustering was applied to the relative abundance of each polymorphism to estimate genotype population clusters within each virus. This was completed using a custom *k*-means clustering script developed and written in Microsoft Open R 3.3.1 using a Gaussian mixture model with an expectation-maximisation algorithm as part of the mclust function with the Mclust version 5.2.2. package to determine *k*, and the kmeans function as part of the stats 3.3.1 package (Dempster, Laird, & Rubin, 1977; Fraley & Raftery, 1999, 2006; Microsoft, 2016; Noune, 2016). The result was bootstrapped 100 times to validate clusters. In addition, trait-specific polymorphic abundances were monitored to observe abundance changes per generation of selection.

9.4 RESULTS

9.4.1 Genome Features and Nucleotide Distance

Comparison of the AC53 MiSeq Genome to the AC53 Original Genome

The newly assembled AC53 MiSeq genome (130,327 bp) contained the exact 139 ORFs and 5 Hr regions as the original AC53 genome (130,442 bp), however, was found to be 115 bp shorter with nucleotide similarity of 99.44%. The Sanger sequenced BRO-A fragment identified 100% nucleotide similarity to the AC53 MiSeq genome and 87.91% to the original AC53 sequence. DNA polymerase Sanger sequenced fragment comparison was 100% to both genomes.

In addition, 19 ORFs and all 5 Hr regions were identified to have between 61.789% and 99.9% nucleotide similarity and amino acid similarity between 92.891% and 100%, in addition to differing lengths (Table 9-1).

Table 9-1: Comparison of the AC53 MiSeq and AC53 original nucleotide and amino acid sequence similarity between the ORFs and Hr regions identified to be different.

ORF/Hr Region	AC53 MiSeq Length (bp)	AC53 Original Length (bp)	Nucleotide Distance (%)	Amino Acid Distance (%)
DNA Polymerase	3,063	3,063	99.935	99.902
Calyx/Pep	1,023	1,023	99.902	99.706
ORF 132	2,844	2,844	99.895	99.789
ODV-EC27	855	855	99.766	99.648
P49	1,407	1,407	99.716	99.573
GP19	282	282	99.645	98.936
ORF 136	2,034	2,034	99.558	100
BRO-B	1,092	1,092	99.542	99.432
ME53	1,080	1,080	99.537	100
ORF 137	546	546	99.451	97.238
ORF 17	168	168	99.405	98.182
ORF 6	873	879	99.317	99.298
ORF 131	801	801	99.752	97.744
38.7K	1,170	1,179	98.558	98.980
HOAR	2,296	2,334	97.188	96.846
BRO-A	708	714	93.697	92.891
ORF 78	165	177	93.220	93.103
ORF 7	246	156	61.789	94.118
ORF 5	138	180	76.667	97.778
Hr1	1,928	1,926	99.689	N.A.
Hr2	2,401	2,377	93.175	N.A.
Hr3	480	482	99.378	N.A.
Hr4	2,177	2,177	99.449	N.A.
Hr5	1,387	1,385	98.847	N.A.

*Not Applicable (N.A.)

Selected Strain Features and Distance to AC53 MiSeq

The selected strains were between 83 bp and 108 bp longer than the AC53 MiSeq genome with genome lengths between 130,410 bp and 130,435 bp (Table 9-2). Nucleotide distance of the selected strains (Table 9-2) when compared to the new AC53 MiSeq genome identified all the F1 strains to have the highest nucleotide similarity (99.667% to 99.669%) and the F4 fast strain to be the most divergent (99.518%).

All selected genomes contained the same 139 ORFs and 5 Hr regions as AC53. However, the F2-F5 fast strains and F4-F5 maxOB strains contained an additional ORF caused by ORF 128 splitting into two and ORF 130 splitting into two, respectively (Table 9-2).

Table 9-2: Nucleotide similarity of all selected strains to AC53 MiSeq. F1 strains have the highest nucleotide similarity to AC53, whereas the F3, F4 and F5 fast strains are the most divergent. Furthermore, all fast strains from the F2 generation and F4 and F5 maxOB strains contain 140 ORFs.

Strain	Nucleotide Distance (%)	Genome Length (bp)	Total ORFs	Total Hr regions
F1 Fast	99.669	130,434	139	5
F1 MaxOB	99.664	130,434	139	5
F1 Slow	99.667	130,434	139	5
F3 Slow	99.667	130,434	139	5
F5 Slow	99.665	130,434	139	5
F2 Slow	99.668	130,434	139	5
F2 MaxOB	99.665	130,434	139	5
F4 Slow	99.664	130,434	139	5
F3 MaxOB	99.663	130,434	139	5
F4 MaxOB	99.637	130,413	140	5
F5 MaxOB	99.620	130,410	140	5
F2 Fast	99.631	130,435	140	5
F3 Fast	99.531	130,433	140	5
F5 Fast	99.529	130,433	140	5
F4 Fast	99.518	130,433	140	5

ORF and Hr comparison to the AC53 MiSeq genome identified 18 ORFs that differed among the strains. All 5 Hr regions exhibited substitutions, with the highest number of mutant regions occurring within the F4, F5 maxOB and F5 slow strains (Table 9-3). Both BRO-A and BRO-B were identified to contain a significant number of mutations as previously reported in the tissue-culture derived strains (Noune & Hauxwell, 2016a). The F5 maxOB strain was the only strain identified to contain the same hypothetical ORF (located between ORF 54 and ORF 55) as AC53 and the F5 slow strain had a unique, non-synonymous mutation occurring within *egt*. Of the 18 ORFs with mutations, the following eight ORFs were mutated with non-synonymous mutations in all selected strains; 38.7K, BRO-A, BRO-B, DNA polymerase, ODV-EC27, ORF17, P49 and P74.

Comparison of the selected strains independent of AC53 indicated that, of the eight mutant ORFs that were common across all strains, six were identical; BRO-A, BRO-B, ODV-EC27, ORF17, P49 and P74, with DNA polymerase and 38.7K containing non-synonymous mutations in the F5 fast and F4 fast strains respectively (Table 9-4).

In addition, some mutations were identified to be selection pressure specific. ORF132, ORF131, ORF128/128a/128b were exclusive to the fast strains. Synonymous mutations in ME53 and non-synonymous mutations in GP19, Calyx/Pep and 38.7K were exclusive to the F4 fast strain and a synonymous mutation in DNA polymerase found to be exclusive to the F5 fast strain. The non-synonymous mutations causing ORF130 to split into ORF130a and ORF130b was unique to the F4 and F5 maxOB strains, in addition, the F5 maxOB strain contained a

hypothetical ORF that was identical to the AC53 MiSeq genome. Slow strains did not contain any selection specific mutations except for a single non-synonymous *egt* mutant identified in the F5 slow strain.

Table 9-3: Comparison of the selected strains nucleotide (N) and amino acid (A.A.) similarity (%) to the AC53 MiSeq genome. Regions with 100% similarity have been highlighted in red. The F4 and F5 maxOB and F5 slow strains have the highest number of regions containing nucleotide (21) and amino acid (14) mutations.

ORF/Hr Region	F1 Fast		F2 Fast		F3 Fast		F4 Fast		F5 Fast		F1 MaxOB		F2 MaxOB		F3 MaxOB		F4 MaxOB		F5 MaxOB		F1 Slow		F2 Slow		F3 Slow		F4 Slow		F5 Slow	
	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.	N	A.A.
38.7K	98.47	98.98	98.47	98.98	98.47	98.98	98.47	99.23	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98	98.47	98.98
BRO-A	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55	97.13	96.55
BRO-B	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73	89.81	87.73
Calyx/Pep	99.90	99.71	99.90	99.71	99.90	99.71	99.71	100.00	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71	99.90	99.71
DNA polymerase	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.97	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90	99.94	99.90
EGT	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.94	99.81
GP19	99.65	98.94	99.65	98.94	99.65	98.94	99.65	100.00	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94	99.65	98.94
ME53	99.54	100.00	99.54	100.00	99.54	100.00	99.63	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00	99.54	100.00
Hypothetical ORF	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	100.00	100.00	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50	97.98	97.50
ODV-EC27	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65	99.77	99.65
ORF 17	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18	99.40	98.18
ORF 128/128a/128b	100.00	100.00	39.96	16.06	36.38	16.06	36.38	16.06	36.38	16.06	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
ORF 130/130a/130b	100.00	100.00	99.84	99.51	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.26	48.84	91.71	48.84	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
ORF 131	98.75	97.74	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74	98.75	97.74
ORF 132	99.89	99.79	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79	99.89	99.79
ORF 136	99.56	100.00	99.56	100.00	99.56	100.00	99.66	100.00	99.61	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00	99.56	100.00
P49	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57	99.72	99.57
P74	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86	99.95	99.86
Hr1	99.95	N.A.	99.95	N.A.	99.95	N.A.	99.79	N.A.	99.95	N.A.	99.95	N.A.	99.95	N.A.	99.95	N.A.	99.95	N.A.	99.95	N.A.	99.95	N.A.	99.9	N.A.	99.95	N.A.	99.95	N.A.	99.95	N.A.
Hr2	91.58	N.A.	91.49	N.A.	91.24	N.A.	91.78	N.A.	91.37	N.A.	91.24	N.A.	91.33	N.A.	91.16	N.A.	91.41	N.A.	91.03	N.A.	91.45	N.A.	91.58	N.A.	91.45	N.A.	91.33	N.A.	91.45	N.A.
Hr3	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.	99.79	N.A.
Hr4	99.54	N.A.	99.54	N.A.	99.59	N.A.	99.4	N.A.	99.49	N.A.	99.59	N.A.	99.59	N.A.	99.63	N.A.	99.54	N.A.	99.59	N.A.	99.59	N.A.	99.59	N.A.	99.59	N.A.	99.59	N.A.	99.54	N.A.
Hr5	99.64	N.A.	99.64	N.A.	99.64	N.A.	98.78	N.A.	99.21	N.A.	99.64	N.A.	99.64	N.A.	99.64	N.A.	99.64	N.A.	99.57	N.A.	99.64	N.A.	99.64	N.A.	99.64	N.A.	99.64	N.A.	99.64	N.A.
Total Regions with Sequence Mutations	20	13	20	13	19	12	19	10	19	12	20	13	20	13	20	13	21	14	20	13	20	13	20	13	20	13	20	13	21	14

*Not Applicable (N.A.). Hr regions are non-coding.

Table 9-4: Nucleotide and amino acid comparison of the selected strains to themselves. BRO-A, BRO-B, ODV-EC27, ORF17, P49 and P74 are identical in each strain.

ORF/Region	Nucleotide Similarity and Clusters of Selected Strains	Amino Acid Similarity and Clusters of Selected Strains
38.7K	• F4 Fast – 99.41% • Remaining strains all identical	• F4 Fast – 99.74% • Remaining strains all identical
BRO-A	• All identical	• All identical
BRO-B	• All identical	• All identical
Calyx/Pep	• F4 Fast – 99.61% • Remaining strains all identical	• F4 Fast – 99.71% • Remaining strains all identical
DNA polymerase	• F5 Fast – 99.97% • Remaining strains all identical	• All identical
GP19	• F4 Fast – 99.30% • Remaining strains all identical	• F4 Fast – 98.94% • Remaining strains all identical
ME53	• F4 Fast – 99.91% • Remaining strains all identical	• All identical
Hypothetical ORF	• F5 MaxOB – 97.98% • Remaining strains all identical	• F5 MaxOB – 87.50% • Remaining strains all identical
ODV-EC27	• All identical	• All identical
ORF17	• All identical	• All identical
ORF128/128a/128b	• F3 to F5 Fast – 100% • F2 to F5 Fast – 87.56% • F2 Fast – 39.96% • F3-F5 Fast – 36.38% • Remaining strains all identical	• F3 to F5 Fast – 100% • F2 to F5 Fast – 87.56% • F2-F5 Fast – 16.06% • Remaining strains all identical
ORF130/130a/130b	• F4 to F5 MaxOB – 97.17% • F4 MaxOB – 94.26% • F5 MaxOB – 91.27% • F2 Fast – 98.84% • Remaining strains all identical	• F4 to F5 MaxOB – 96.00% • F4-F5 MaxOB – 48.84% • F2 Fast to F5 MaxOB – 45.49% • F2 Fast to F4 MaxOB – 48.37% • Remaining strains all identical
ORF131	• F2, F3, F5 Fast – 100% • F2, F3, F5 Fast to remaining strains – 98.75% • Remaining strains all identical	• F2, F3, F5 Fast – 100% • F2, F3, F5 Fast to remaining strains – 97.74% • Remaining strains all identical

ORF132	• F2, F3, F5 Fast – 100% • F2, F3, F5 Fast to remaining strains – 99.89% • Remaining strains all identical	• F2, F3, F5 Fast – 100% • F2, F3, F5 Fast to remaining strains – 99.79% • Remaining strains all identical
ORF136	• F5 Fast – 99.95% • F4 Fast – 99.90% • F4 to F5 Fast – 99.85% • Remaining strains all identical	• All identical
P49	• All identical	• All identical
P74	• All identical	• All identical
egt	• F5 Slow – 99.94% • Remaining strains all identical	• F5 Slow – 99.81% • Remaining strains all identical
Hr1	• F2 Fast – 99.95% • F2 Slow – 99.95% • F4 Fast – 99.84% • Remaining strains all identical	N.A.
Hr2	• F1, F3, F5 Slow – 100% • Remaining strains – 98.61% to 99.96%	N.A.
Hr3	• All identical	N.A.
Hr4	• F2 Fast, F4 MaxOB, F5 Slow – 100% • F1, F2, F5 MaxOB, F1-F4 Slow, F3 Fast – 100% • Remaining strains – 99.40% - 99.95%	N.A.
Hr5	• F5 Fast – 98.56% to 99.57% • F4 Fast – 98.56% to 99.57% • F5 MaxOB – 98.92% to 99.93% • Remaining strains all identical	N.A.

Recombination and Evolutionary Statistics

bratNextGen identified potential recombination events to be occurring within the respective genomes of AC53 MiSeq, all fast strains and the F5 maxOB strain (Table 9-5). Interestingly, the F1-F4 maxOB and all the slow strains did not contain any foreign genomic segments. Much of AC53 MiSeq contains segments from the F2-F5 fast strains, whereas a single AC53 segment was identified within the F1 fast strain. In addition, the F5 maxOB strain contained a small segment originating from the F2 fast strain.

The Tajima's D (0.33) and Fay and Wu's H (-8237) evolutionary statistics indicate a false-positive bottleneck as per chapter 7. In addition, evolutionary rate statistics failed to complete due to limitations associated with MEGA7.

Table 9-5: Recombination events occurring with the AC53 MiSeq genome and the selected strains. AC53 MiSeq has greater than half of its genome containing genetic segments originating from the fast strains.

Virus	Segment Start (bp) – Relative to Alignment	Segment End (bp) – Relative to Alignment	Segment Length (bp)	Origin Virus(s)	Genetic Region	Relative Proportion of Genome (%)
AC53 MiSeq	1	19,183	19,183	F3/F5 Fast	Polyhedron – P26	50.88%
	41,063	75,656	34,593		ORF 49 – ORF 85	
	114,840	116,782	1,942		ORF 125 – LEF-1	
	119,903	130,504	10,601		ORF 131 – ORF 138	
	91,512	91,622	110	F2 Fast	Hr4	
F1 Fast	91,512	100,087	8,575	AC53 MiSeq	Hr4 – SOD	6.57%
F2 Fast	116,783	117,444	661	F3/F5 Fast	LEF-1 – ORF 128	3.90%
	119,903	124,113	4,210		ORF 131 – ORF 132	
	117,445	117,467	22	F4 Fast	ORF 128	
	119,357	119,555	198		ORF 130	
F3 Fast	116,783	124,113	7,330	F5 Fast	LEF-1 – ORF 128	5.62%
F4 Fast	41,063	49,057	7,994	F3/F5 Fast	ORF 49 – Hr 2	21.05%
	100,088	119,555	19,467		SOD - *egt*	
F5 Fast	116,783	124,113	7,330	F3 Fast	LEF-1 – ORF 128	5.62%
F5 MaxOB	119,690	119,902	212	F2 Fast	ORF 130	0.16%

Phylogenetics

MLE tree construction of all the full genomes identified three distinct clusters corresponding to each applied pressure (Figure 9-1). Cross-over of the F1 fast and F4 slow strains into the maxOB cluster is indicative of the high nucleotide similarity previously indicated (Table 9-2), however, poor bootstrap support (less than 85%) has been observed for most of the branches of the tree. Again, this is indicative of the high nucleotide similarity observed between sequences. In addition, the MLE tree was unable to indicate accurately if the F1 strains gave rise to the F2 strains and so on and so forth.

Core SNP analysis identified between 571 (AC53) and 679 (F2 fast) unique SNPs in each consensus genome sequence (Table 9-5). The lower number of unique SNPs in the AC53 genome suggests that each derivative strain has produced new mutations caused by the applied selection pressure. MLE of these core SNPs clustered the strains in the same way as the whole-genome MLE (Figure 9-1), albeit with slightly improved bootstrap support.

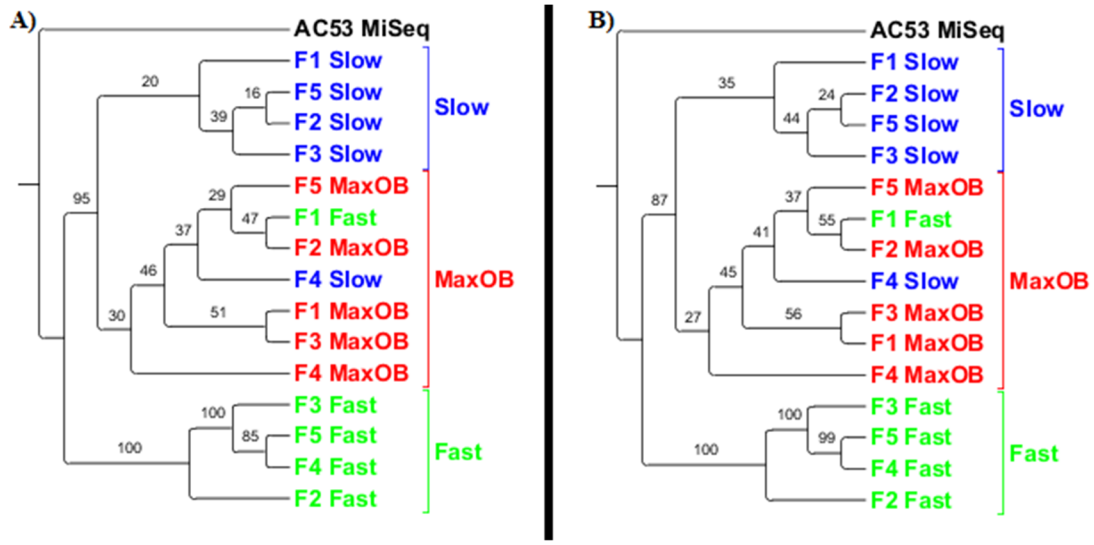


Figure 9-1: A) Whole-genome phylogenetic relationships of the selected strains rooted to the AC53 MiSeq genome. B) Core SNPs phylogenetic relationships of the selected strains rooted to the AC53 MiSeq genome. Slow strains (blue), MaxOB strains (red) and Fast strains (green) have been clustered together, however, cross-over of the F4 slow and F1 fast into the MaxOB cluster has been observed.

Table 9-6: Total core SNPs identified in each consensus genome. These results suggest that all selected strains have produced new mutations as the total number of identified SNPs increased.

Virus	Total Core SNPs
AC53 MiSeq	571
F1 Fast	678
F2 Fast	679
F3 Fast	677
F4 Fast	677
F5 Fast	677
F1 Slow	678
F2 Slow	678
F3 Slow	678
F4 Slow	678
F5 Slow	678
F1 MaxOB	678
F2 MaxOB	678
F3 MaxOB	678
F4 MaxOB	657
F5 MaxOB	654

Time tree analysis estimated the time needed for all strains to diverge from AC53 to be approximately 365 hrs P.I. (Figure 9-2). F2-F5 fast strains were estimated to diverge the quickest between 12 hrs and 13 hrs, whereas the F1 fast strain was estimated to have diverged at approximately 40 hrs. MaxOB strains were estimated to have diverged slower than the slow strains with complete divergence at 104 hrs, whereas slow strains completed divergence at approximately 80 hrs.

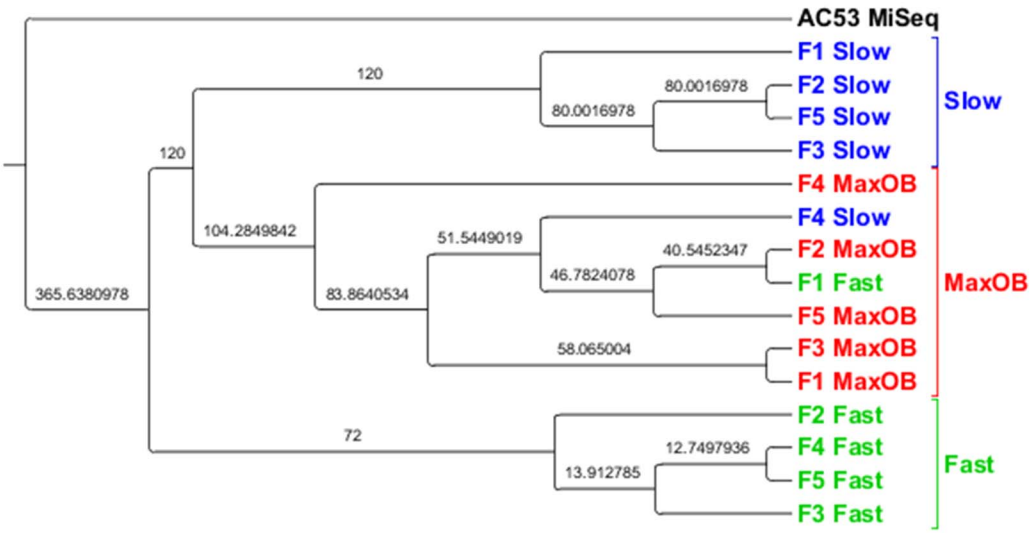


Figure 9-2: Time tree analysis using the ReltimeML algorithm with time points in hours and node recalibrated to isolation time points of each strain described in chapter 8. Fast strains (green) were estimated to have diverged between 12 hrs and 40 hrs P.I. Slow strains (blue) had diverged between 51 hrs and 80 hrs, while maxOB strains (red) diverged between 40 hrs and 104 hrs.

9.4.2 Within-Isolate and Within-Strain Polymorphic Diversity

Polymorphism Locations and Analysis

Of the total polymorphisms identified within each virus, the predominant source of variation was attributed to substitution based genotypes with a small number of insertion and deletion genotypes (Figure 9-3). Identification of within-strain and within-isolate polymorphisms identified between 78 (AC53) and 156 (F3 Fast) with the highest number of polymorphisms identified within P49, 38.7K, BRO-A, ORF131 and Hr2 except for Hr4 in the F3 maxOB strain (Figure 9-4, Table 9-7, Table 12-23).

High P49 polymorphic counts were predominantly identified within early generation fast strains as opposed to 38.7K, however, F4 and F5 fast strains have higher counts within 38.7K. Hr2 was identified as the most variable Hr region. Some mutations identified were within trait-specific populations: *lef-8* within F2-F5 slow strains and F2 fast, ORF130/130a/130b within F2-F5 maxOB strains, ORF12, ORF13 and IE-1 within F1-F4 fast strains, ODV-E56 within F1-F3 fast strains and ORF13 and IE-1 within F2 maxOB and F2-F3 maxOB strains respectively (Figure 9-4, Table 12-23).

Table 9-7: Total polymorphisms identified within the AC53 MiSeq genome and the selected strains highlighting ORFs and Hr regions with the highest polymorphic count.

Virus	Total Polymorphisms	Highest Polymorphic Count (ORF)	Highest Polymorphic Count (Hr)
AC53 MiSeq	78	P49 - 7	Hr2 - 17
F1 Fast	148	P49 - 13	Hr2 - 23
F2 Fast	150	P49 - 13	Hr2 - 25
F3 Fast	156	P49 - 13	Hr2 - 23
F4 Fast	149	BRO-A/38.7K/ORF131/P49 - 10	Hr2 - 31
F5 Fast	86	38.7K - 10	Hr2 - 22
F1 Slow	102	38.7K - 10	Hr2 - 24
F2 Slow	103	38.7K - 10	Hr2 - 29
F3 Slow	111	BRO-A - 13	Hr2 - 32
F4 Slow	125	38.7K/ORF131 - 10	Hr2 - 30
F5 Slow	122	38.7K/ORF131 - 10	Hr2 – 31
F1 MaxOB	106	38.7K - 10	Hr2 - 28
F2 MaxOB	122	38.7K - 10	Hr2 - 27
F3 MaxOB	131	BRO-A - 13	Hr4 - 34
F4 MaxOB	109	38.7K/ORF131 - 10	Hr2 - 23
F5 MaxOB	120	38.7K/ORF131 - 10	Hr2 - 29

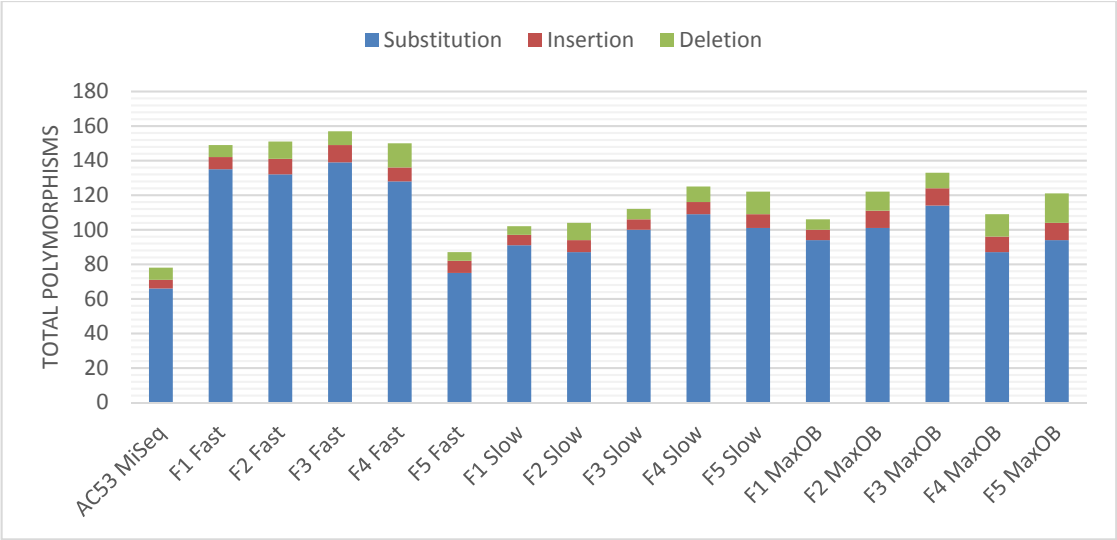


Figure 9-3: A summary of the total substitutions, insertions and deletions identified within each selected strain and the AC53 MiSeq genome.

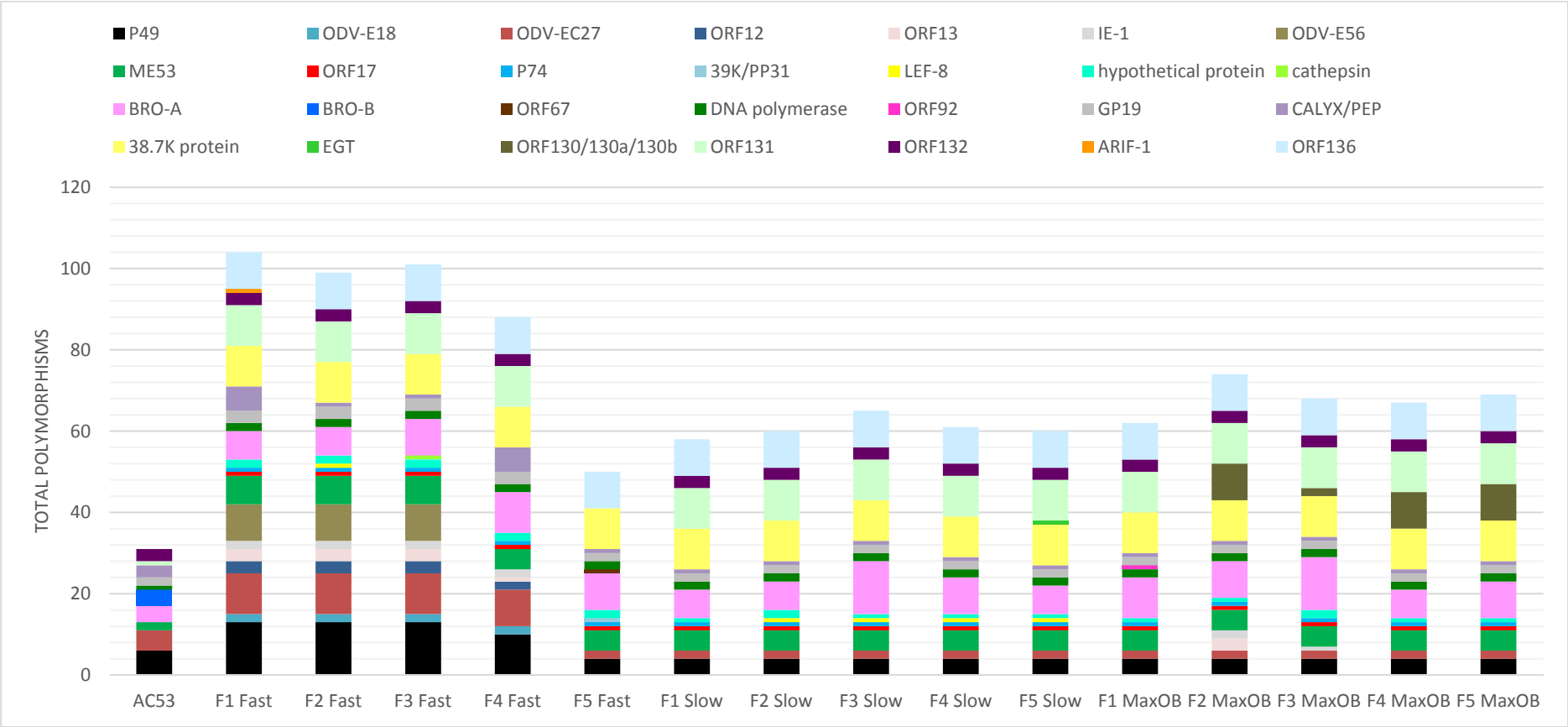


Figure 9-4: A summary of ORFs containing polymorphisms within each selected strain and the AC53 MiSeq genome.

MLE analysis of the identified polymorphisms (Figure 9-4) identified some clusters, albeit with poor bootstrap support, suggesting that fast and slow strain polymorphisms are more closely related to each than to those of the maxOB strains. This is further suggested by maxOB strains clustering together as opposed to the fast and slow strains which overlap. However, MLE analysis suggests F5 slow and maxOB strains are more closely related.

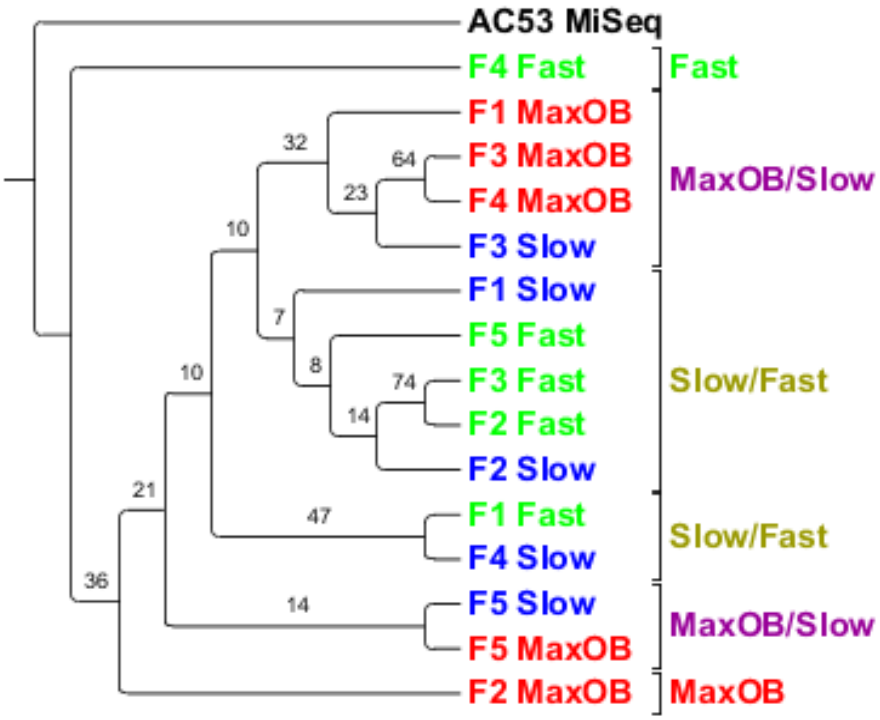


Figure 9-5: MLE analysis of polymorphisms within each selection strain (MaxOB – Red, Slow – Blue, Fast – Green) rooted to the AC53 MiSeq polymorphisms. Results indicate polymorphisms within slow and fast strains maybe more closely related than those within the maxOB strains, except for the F5 slow and F5 maxOB polymorphism.

Estimation of Genotype Abundance

K-means clustering of polymorphic abundance metrics produced between 2 and 9 clusters within each analysed genome, however, no distinct trend was observed within clusters (Table 9-8, Table 12-24). A trend was observed with all genomes when total mean abundance of the reference and alternative polymorphisms (allele) was analysed, with an increase in alternative allele abundance identified with the exception for the F1 slow, F2 fast, F2 and F3 maxOB strains (Table 9-8).

Each identified cluster encompassed multiple different positions throughout each genome. However, due to the large number of polymorphisms identified, lack of scalability with analysing several different datasets and limitations in visualisation and representation of genomic positions within each cluster, these results have been omitted and attached in separate spreadsheets (for the individual spreadsheets, see: https://researchdatafinder.qut.edu.au/display/n13986).

Table 9-8: Summary of *k*-means clustering and mean total abundance of the reference and alternative alleles. Clustering trends were not observed; however, alternative allele abundance was higher in most analysed genomes.

Virus	Total Clusters (*k*)	AC53 Allele Mean Total Abundance (%)	Alternative Allele Mean Total Abundance (%)
AC53 MiSeq	6	45.96	54.04
F1 Fast	3	38.35	61.05
F1 MaxOB	5	38.20	61.32
F1 Slow	6	58.13	41.87
F2 Fast	9	51.10	48.90
F2 MaxOB	4	50.33	49.67
F2 Slow	6	30.50	68.59
F3 Fast	2	49.09	50.38
F3 MaxOB	4	51.07	48.17
F3 Slow	4	29.22	70.36
F4 Fast	2	47.45	51.86
F4 MaxOB	5	48.30	51.70
F4 Slow	4	48.54	51.46
F5 Fast	2	29.28	68.78
F5 MaxOB	2	34.92	64.38
F5 Slow	5	31.71	68.26

Analysis of trait-specific polymorphism abundance changes per generation targeted the previously described locations; ORF130/130a/130b (9 polymorphisms) – maxOB strains, *lef-8* (1 polymorphism) – slow strains, ORF12 (3 polymorphisms), ORF13 (3 polymorphisms), IE-1 (2 polymorphisms) and ODV-E56 (9 polymorphisms) – fast strains.

Every trait-specific polymorphism analysed identified the dominant genotype outcompeting the minor genotype over the five generations of selection (Figures 9-6 to 9-8). In addition, abundance fluctuations were observed in all analysed polymorphisms prior to the dominant genotype outcompeting the minor genotype (Tables 12-25 to 12-30).

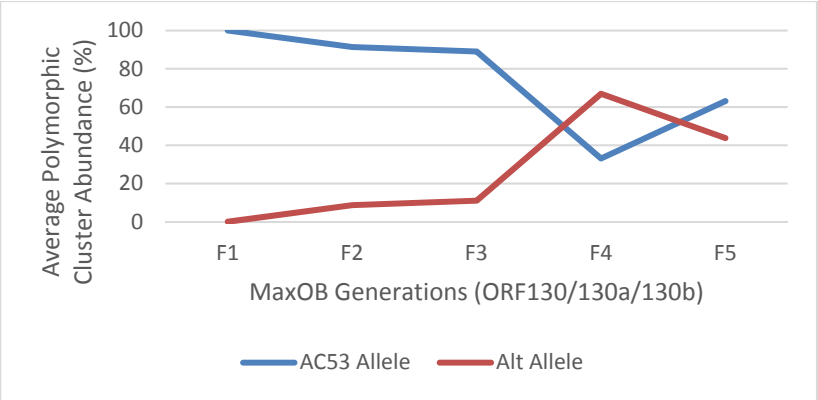


Figure 9-6: MaxOB strains ORF130/130a/130b changes in polymorphic abundances over each generation of selection. The AC53 allele is observed to have a downward trend but begins to outcompete the alternative allele after the slight fluctuation observed in F4.

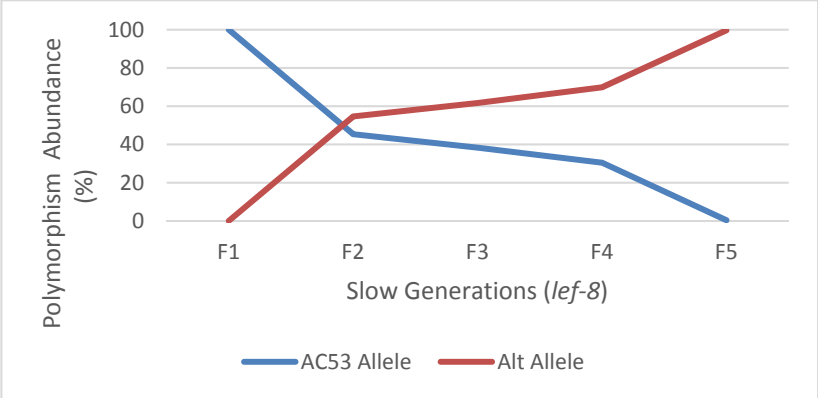


Figure 9-7: Slow strains *lef-8* observed abundance change. In this case, the alternative allele outcompetes and excludes the AC53 reference allele during each round of selection.

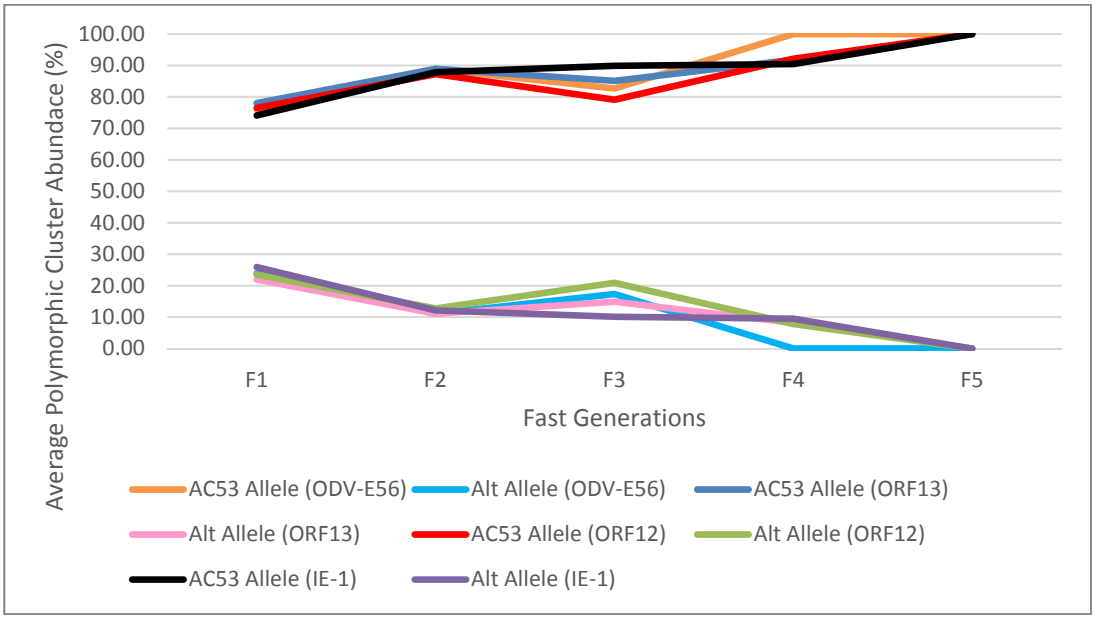


Figure 9-8: Fast strains ORF12, ORF13, IE-1 and ODV-E56 polymorphic clusters abundance changes per generation of selection. All four analysed ORFs are showing the AC53 reference allele outcompeting the alternative allele over the five generations of selection.

9.5 DISCUSSION

In the previous chapter, we demonstrated a technique to derive trait-specific strains from a parent isolate and analysed the biological activity of each strain, however, correlation of trait-specific mutations occurring within each genome was still to be completed. This chapter extends that study and identified the trait-specific mutations occurring within the selected strains, and MLE analysis could cluster strains based on trait-specificity (albeit, with poor bootstrap support).

Prior to the genetic analysis of the selected strains, a new AC53 genome (AC53 MiSeq) was sequenced and assembled for comparison with the whole genome sequences of the *in vivo* - selected strains generated using the Illumina MiSeq in order to eliminate possible errors and platform-specific bias that might be introduced by use of the original consensus genome from the Ion Torrent PGM sequencing previously described (Noune & Hauxwell, 2015). Comparison of the AC53 MiSeq genome to the original AC53 consensus sequence identified significant differences indicative of potential sequencing errors and bias, or resulting from assembly of a consensus genome using shotgun sequencing of a mixed population (Bragg, Stone, Butler, Hugenholtz, & Tyson, 2013; McElroy et al., 2014; Quail et al., 2012; Steven & Salzberg, 2005). In addition, baculovirus genomes are AT-rich and contain multiple repeat sequences which are known sources of limitations and errors with current sequencing platforms (Bragg et al., 2013; Hoff, 2009; P. L. Johnson & Slatkin, 2008; McElroy et al., 2014; Quail et al., 2012; Schirmer et al., 2016; Treangen & Salzberg, 2012; Wall et al., 2014). Until errors and limitations in current platforms are overcome, reproducibility of genome assemblies will continue to be a hot-topic.

Analysis and comparisons of the selected strains to the AC53 MiSeq genome identified high nucleotide similarity between all analysed sequences as expected and previously observed with tissue-culture derived strains (Noune & Hauxwell, 2016a, 2016b). All trait-specific strains contained the same 139 ORFs and 5 Hr regions as the parent isolate, however, the fast strains and maxOB strains contained 140 ORFs as ORF 128 and 130 split in two respectively.

Of the ORFs identified, a total of 18 ORFs had diverged from AC53, with 8 ORFs common across all strains; 38.7K, BRO-A, BRO-B, DNA polymerase, ODV-EC27, ORF17, P49 and P74. The functional roles of the ORFs 38.7K, BRO-A, BRO-B and DNA polymerase are DNA binding, enhancement and replication proteins, ODV-EC27 is a multifunctional cyclin, P74 is a *per os* infectivity factor which mediates binding of ODV to midgut cells, and P49 is associated with nucleocapsid formation (Belaich et al., 2006; Belyavskyi, Braunagel, & Summers, 1998; Bideshi et al., 2003; Kang et al., 1999; J Kuzio et al., 1989; Z. Li et al., 2006; Peng, van Oers, Hu, van Lent, & Vlak, 2010; Adam L Vanarsdall, Pearson, & Rohrmann, 2007; Zemskov, Kang, & Maeda, 2000). It can be hypothesised that based on the functional role of these eight common ORFs, the composition of the quasispecies will alternate after *in vivo* selection is applied.

From these eight common ORFs, BRO-A and BRO-B mutations had previously been reported in tissue-culture strains, with DNA polymerase and 38.7K mutations occurring within the AC53-T2 strain (Noune & Hauxwell, 2016a). This suggests that ODV-EC27, ORF17, P49 and P74 mutations are *in vivo* derivatisation specific. In addition, six of these eight ORFs were found to be identical in each pressured strain, with DNA polymerase and 38.7K observed to have unique mutations within the F5 and F4 fast strains respectively.

Furthermore, the 10 ORF mutations identified to not be common across all strains could be tied to the applied selection pressure. ORF132, ORF131, ORF128/128a/128b were exclusive to the fast strains, with the ORF128a/128b splitting previously identified in the AC53-T4.2 strain (Noune & Hauxwell, 2016a), and ORF130a/ORF130b was unique to the F4 and F5 maxOB strains. The slow strains did not contain any specific mutations present throughout each generation of selection except for a single non-synonymous *egt* mutant identified in the F5 slow strain.

Recombination analysis identified a significant portion of the AC53 MiSeq genome to be originating from the fast strains with no foreign segments identified in the slow strains. This contrasts with the nucleotide distance analysis which suggested that the slow strains are more closely related to the AC53 MiSeq genome. In addition, Tajima's D and Fay and Wu's H indicated a false-positive bottleneck to be occurring like what was observed in chapter 7, and evolutionary statistics failed to complete due to computational limitations. The recombination result may be incorrect and would need to be repeated with a larger dataset as bratNextGen was designed for hundreds of sequences (Marttinen et al., 2012). However, as it is possible that recombination is playing a significant role in producing genotypes that overcome insect resistance (Okano, Vanarsdall, & Rohrmann, 2007).

Whole-genome and core-SNP phylogenetic results could distinguish and cluster each trait-specific population. This may suggest that mutations occurring within non-coding regions and Hr regions could have greater effect on chapter 8's phenotypic observations. In addition, time-tree analysis estimated divergence time of the selected strains from the AC53 isolate with an overall divergence time of 365 hrs. Fast strains were estimated to have diverged within the first 24 hrs of infection while slow strains and maxOB strains diverged between 40 hrs and 104 hrs of the infection cycle. This result correlates with the ST_{50} results observed in chapter 8.

Polymorphisms identified within the shotgun data of AC53 MiSeq and each selected strain suggests that new mutations are being produced. This is indicative of the higher number of polymorphisms present in each selection pressure strain. Most of these polymorphisms were localised to the following ORFs; P49 in early generation fast strains, 38.7K in late generation fast strains and all slow and maxOB strains, and Hr2 identified as the most variable Hr region. Furthermore, ORF130 polymorphisms were exclusively identified in MaxOB strains suggesting an important role in OB production. IE-1 and ODV-E56 mutations were exclusive to fast strains

and are required for early replication, apoptosis and ODV envelope formation, Mutations in *lef-8*, which encodes a subunit of the baculovirus RNA polymerase, were exclusive to slow strains (Braunagel, Elton, Ma, & Summers, 1996; Passarelli, Todd, & Miller, 1994; Schultz, Wetter, Fiore, & Friesen, 2009; Titterington, Nun, & Passarelli, 2003). However, with the imminent release of MuTect2 and version 4 of the Genome Analysis Toolkit the accuracy in identifying correct polymorphisms and identification of Illumina errors should improve this analysis.

MLE analysis of these polymorphisms indicates that slow and fast strains have significant cross-over suggesting that these polymorphisms may be more prevalent within these two traits. However, computational issues and lack of scalability to many datasets limited the analysis and in-depth comparisons of polymorphic sites.

This lack of scalability was again evident with the *k*-means clustering approach applied to estimate genotype abundance within each analysed dataset. Detailed comparisons were limited by current visualisation techniques and the high amount of variability observed between clusters and strains. Regardless, no obvious trend was observed within the clusters but in most datasets, a high alternative allele abundance was identified. Furthermore, this clustering approach evolves a previously described *k*-means approach (Chateigner et al., 2015) and can be applied as a solution to estimate the abundance of the viral population within an NPV.

However, analysis of the sub-clustered within-strain strait-specific polymorphisms identified a clear, competitive exclusion of genotypes during each generation of selection and compares with previous results observed in 'Vesicular stomatitis virus' (Solé et al., 1999). The slow strains specific polymorphism occurring within *lef-8* was observed to contain the only alternative allele that outcompetes the AC53 allele during selection, whereas the both the fast strains and maxOB strains highlight the AC53 allelic cluster to be outcompeting the alternative allelic cluster. These results contrast chapter 7's results which did not observe competitive exclusion, but instead highlighted niche differentiation to be occurring. This suggests that an unpressured population will differentiate host resources and occupy resource niches, whereas once a pressure is exerted, competitive exclusion takes places due to lower fitness genotypes being unable to adapt to the changing environment.

In addition, these results are indicative of NPVs acting as a viral quasispecies as previously suggested in chapter 7 through mutational robustness. (Domingo et al., 2012; Van Nimwegen et al., 1999; Vignuzzi et al., 2006; Wilke, 2005; Wilke et al., 2001). Rather than recombination influencing population diversity, it may be through the application of *in vivo* selection and mutational robustness that NPVs are able to exploit selection to maintain diversity (Wilke et al., 2001). To put sparingly, if a high mutation rate is maintained throughout the quasispecies, the genotype with the highest fitness would be unable to adapt to its host immunity response (Summers & Litwin, 2006). Furthermore, cooperation of genotypes within the quasispecies may allow variants to infect different host tissues and has been previously

demonstrated with poliovirus (Vignuzzi et al., 2006). However, analysis of the polymorphisms does not support the hypothesis of genotype cooperation as trends in the analysis point to competitive exclusion, but common polymorphisms found in a selection of ORFs does suggest cooperation among genotypes.

In conclusion, this chapter describes the genomic characteristics associated with *in vivo* selection. These results contrasted to the previously described tissue-culture *in vitro* derivatisation where different ORF mutations were observed. In addition, the computational limitations and lack of scalability when analysing polymorphisms within more than two datasets has affected the level of detail when comparing polymorphisms and would need to be addressed. Future studies may also benefit from a transcriptome analysis, as the activation of genes during the infection cycle may shed further light as to how each *in vivo* derived strain behaves. Furthermore, the differences observed between the AC53 MiSeq consensus sequence and the original AC53 sequence has implications in other NPV sequencing studies utilising NGS as reproducibility is an issue.

Chapter 10: Conclusions

10.1 TRENDS & RESULT SUMMARY

The overarching objective of this study was to apply NGS and bioinformatics to improve the understanding of baculovirus dynamics, diversity and evolution. Essentially, the study aimed to develop new bioinformatic techniques to analyse non-model systems, apply these techniques to monitor the infection cycle and identify selection pressure specific mutations and evolutionary rates. This was achieved using the commercial baculovirus isolate AC53, derivatisation of strains undergoing *in vivo* and *in vitro* selection, and development of a new bioinformatic technique which can accurately identify and calculate relative abundance of genotypes within a metapopulation. In addition, this study contributed additional results confirming the reclassification of HaSNPV and HzSNPV isolates to a single species, development of an *in vivo* technique to produce trait-specific strains and identified systematic issues with genome assembly of baculoviruses.

In summary, the overall trends and results of this study in reference to the objectives listed in chapter 1, section 1.2 are as follows:

Objective 1 - Apply NGS and develop bioinformatic techniques to assemble whole-genome sequences and develop and analyse custom meta-barcodes to quantify and describe the strain diversity and abundance within isolates.

a. A baseline reference genome was assembled for AC53 and the automated genome assembly pipeline 'IMG-AP' was developed.
b. A new software pipeline (MetaGaAP and MetaGaAP-Py) was developed to overcome the current challenges with analysing and identifying genotypes and their relative abundance within metapopulations. The results identified a single dominant nucleotide genotype with 28 minor nucleotide genotypes identified above 20x coverage, and were validated by comparison to the AC53-T2 strain and Sanger sequencing.

Objective 2 - Apply *in vivo* and *in vitro* selection to derive strains from the wild-type isolate and identify, characterise and quantify strain variation.

a. Nine *in vitro* strains were derived using a modified plaque purification technique.
b. MetaGaAP-Py was applied to monitor the AC53 infection cycle and identified a statistically significant change in both nucleotide and amino-acid genotype abundance when the inoculum was compared to the final OB products produced

post-infection. In addition, the relative abundance of both amino acid and nucleotide genotypes during the infection was relatively flat even though read count increased before peaking at 120 hrs and 144 hrs P.I. This may be indicative of peak viral load.

c. Three *in vivo* selection pressures (fast speed-of-kill, slow speed-of-kill and maximum OB production) were applied to the AC53 over five generations to produce fifteen trait-specific strains.
d. The IMG-AP was applied to assemble the nine *in vitro* and fifteen *in vivo* strains using the AC53 parent isolate sequence as a reference.
e. A comparison of *in vitro* plaque purified AC53 derived strains identified nine mutant ORFs specific to this technique, and when compared to global isolates, results indicated that HaSNPV and HzSNPV are the same species – confirming previous studies (J. A. Jehle et al., 2006; Rowley et al., 2011), with the species potentially originating in Australia. Genetic analysis of the *in vivo* strains identified 8 mutant ORFs common across these strains in addition to trait-specific mutations. Systematic discrepancies were identified between the AC53 MiSeq genome and the original AC53 reference sequence produced in chapter 3. Furthermore, the results suggest that baculoviruses are a viral quasispecies.

Objective 3 - Characterise the biology of the strains selected by *in vivo* passage.

a. The biological performance of these strains was assessed on the F5 generation and identified significant virulence-transmission trade-offs occurring.
    i. The maxOB strain was observed to have the highest OB production and the heaviest insects during the infection cycle, but with much lower density when compared to the slow strain and the AC53 parent isolate. In addition, the maxOB strain had the longest infection cycle.
    ii. The fast strain was observed to be faster than both the maxOB and slow strains including the AC53 parent isolate. However, the fast strain had lower efficacy and produced the least OBs.
    iii. AC53 had the highest percentage kill and was the most efficient as it could balance virulence-transmission trade-offs effectively. This may be indicative of AC53 containing a full population of cooperating genotypes unbiased towards specific traits.

10.2 SIGNIFICANCE OF KEY FINDINGS

The significance of the study can be summarised in four key-points:

1. NGS and the bioinformatic techniques used to perform genome assembly on metapopulations need to be re-assessed as fundamental systematic errors are present within the consensus genome sequences. This issue was first alluded in chapter 6 in which the BRO-A Sanger sequencing result was identical to the dominant genotype identified by MetaGaAP, and highlighted further in chapter 9 with a brief comparison of a newly assembled AC53 genome sequence to the original sequence produced in chapter 3. In chapter 9, the AC53 MiSeq and original AC53 reference sequence had 99.44% nucleotide similarity, and both BRO-A and DNA polymerase Sanger sequences were 100% identical to the MiSeq genome.

This can be attributed to various factors such as different sequencing platforms utilised, differences in bioinformatic techniques, assembly of a consensus genome from a mixed population and platform-specific bias (AT or GC bias) (Bragg et al., 2013; McElroy et al., 2014; Quail et al., 2012; Steven & Salzberg, 2005). Differences in the multiple methods and platforms used for shotgun sequencing results in difficulty in reproducing results (Nekrutenko & Taylor, 2012). The differences we observed in consensus sequences produced on different platforms were mitigated by use of a single platform in each study (e.g. chapter 9 uses only Illumina MiSeq data). In addition, the 'Invertebrates & Microbiology Groups' Assembly Pipeline' which was developed in chapter 4 and 5 was designed to help alleviate some issues associated with baculovirus genome assembly but is limited by the sequencing platform utilised. The sequencing platform used needs to be selected objectively rather than relying simply on that which is currently in vogue. The advances in third-generation sequencing technologies and the constant improvement in bioinformatic techniques will eventually eliminate many of these issues (Bleidorn, 2016).

NPVs are metapopulations or quasispecies, and it can be hypothesised that during DNA purification and the several clean-up steps used during library preparation that some DNA is lost and can alter the consensus genome. The multiple minor differences resulting from use of different platforms and potential changes in consensus sequence resulting from the different proportions of variants within an isolate calls into question the accuracy of all baculovirus genomes sequence.

2. Most bioinformatic techniques currently available are designed for model systems, struggle with metapopulations, and are unable to scale-up for detailed polymorphic comparisons (da Fonseca et al., 2016; A. D. Johnson, 2009; Nielsen et al., 2011). MetaGaAP has helped to

alleviate the issue for detailed meta-barcoding analyses for non-model systems, but still requires further optimisations and enhancements to deal with the current computational limitations.

Furthermore, limitations with current software and the inability to scale-up polymorphic comparisons were observed in chapter 9. This was highlighted with the limited comparisons of polymorphisms and *k*-means clustering abundance of polymorphisms and, mean evolutionary rate statistics failing with whole genomes. These issues may be overcome if a smaller set of data was compared rather than a population of derived strains, however, this would limit the scope of the study and may miss important polymorphic data.

However, as previously mentioned, with the constant evolution and development of third-generation sequencing technologies and bioinformatic techniques, these issues will be overcome.

3. Trait-specific derivatisation of strains introduces commercial implications with distinct virulence-transmission trade-offs. In chapter 8, strains were derived through the application of three selection pressures which resulted in fifteen trait-specific strains with various virulence-transmission trade-offs observed in the F5 strains. Fast strains were found to be faster than the parent isolate but requires high dosages thus a cost of reduced fitness and reduced transmission. Previous studies have suggested that this is a result of artificial infection removing transmission costs (van Baalen & Sabelis, 1995; White et al., 2012). Slow and maxOB strains experienced similar trade-offs with increased transmission observed at the cost of efficacy and virulence but these effects compounded in the maxOB strains. Commercialisation of these strains and techniques is possible, however, trade-offs need to be taken into consideration but could be offset through strain mixing or continuously applied selection until negative traits are diminished (Bull & Lauring, 2014)

In addition, no in-depth genetic analysis of *in vivo* selected baculoviruses strains has been completed and the study presented in chapter 9 could be considered one of the first. Previous studies have focused on either *in vitro* derivatisation or biological characterisations (Arrizubieta et al., 2015b; Arrizubieta, Williams, Caballero, & Simón, 2014; Graillot et al., 2014; Noune & Hauxwell, 2016a; Elizabeth M Redman et al., 2016; White et al., 2012). Both chapters 5 and 9 have identified core ORFs which contained mutations that occurred during either *in vivo* or *in vitro* derivatisation. However, it can be hypothesised that mutations which alter the phenotypic properties may not be associated with any ORF, and instead caused by polymorphisms and consensus sequence variations within Hr and non-coding regions as most mutations were identified in those regions. This hypothesis can be supported by previous studies which have noted phenotypic affects caused by mutant Hr regions (Carstens & Wu, 2007; Guarino, Gonzalez, & Summers, 1986; Landais et al., 2006; Lin, Chen, Yi, & Zhang, 2010). Furthermore, the validation of HzSNPV and HaSNPV as a single species should ease product

registration of these baculoviruses as it has been shown that they have a global distribution (Buerger et al., 2007).

4. NPVs are constantly adapting to the non-static environment of the insect host causing changes in genotypic abundance during replication and virion occlusion in genotype abundance during and post infection. Furthermore, this study has presented evidence that NPVs fit the 'viral quasispecies' model and should be extended to include these viruses (Domingo et al., 1998; Domingo et al., 2012; Summers & Litwin, 2006; Van Nimwegen et al., 1999; Vignuzzi et al., 2006; Wilke, 2005; Wilke et al., 2001).

Chapter 7 is integral to this key-point, as it applies MetaGaAP to monitor the change in genotype abundance during the infection cycle. The results showed a clear reduction in relative abundance of the dominant genotype and increase minor genotype abundance when the initial AC53 inoculum was compared to the final OB products produced post infection. Furthermore, presence-absence analysis of genotypes during the infection cycle highlighted that up to 87% of the nucleotide genotypes and 81% of the amino-acid genotypes identified within the OB were not present in amplimers derived from non-occluded viral DNA. Instead, two genotype clusters were identified: a cluster of OB-specific genotypes and a cluster of core-genotypes present in all analysed datasets.

However, this could be an artificial result caused by two factors: 1) BV samples were densely packed onto a single 318v2 chip whereas OB samples were sequenced on a less-dense chip with much higher coverage per samples. 2) BV and OB samples underwent different DNA extraction protocols, BV samples undergoing a LN_2 whole-insect extraction without alkaline lysis and OB samples undergoing an SDS extraction from insect cadavers with alkaline lysis. Therefore, OB DNA in the BV samples were not extracted causing a loss of genotypes. In the eventuality that these results are not artificial it is possible that the OB specific genotypes were not detected as the concentration of these OB genotypes may be too low for current second-generation NGS platforms, thus occupying an undetectable resource niche during the BV infection cycle.

Regardless, the differences in genotype abundance observed in the OB samples is a cause for concern regarding commercialisation of these viruses as the produced OB is different from the starting stock. However, safe measures are in place to limit changes in biological activity such as re-assessment of biological performance after each production cycle and completing production cycles using the original viral isolation stock (Buerger et al., 2007). In addition, MetaGaAP-Py can be adapted as an added quality control measure during production runs to monitor and compare genotypes and relative abundance within the initial stock used for production and the final product.

As observed in chapter 7, genotypes had periods of significantly slight increases and decreases of relative abundance during the infection cycle which may be due to NGS variances

or indicative of a viral quasispecies. In addition, this result may be explained by competitive exclusion principle or niche differentiation as the dominant genotype outcompetes the minor genotypes for host-resources which limits the minor genotypes to a resource niche. Or, the most likely explanation for these results may be attributed to mutational robustness and genotype cooperation as part of the viral quasispecies model. However, analyses of trait-specific polymorphisms in chapter 9 contrasts what was observed in chapter 7 as competitive exclusion was clearly observed during each round of selection and compares with previous quasispecies research in 'Vesicular stomatitis virus' (Solé et al., 1999). These results suggest two outcomes; unpressured baculoviruses exhibit niche differentiation of host resources whereas baculoviruses under selection pressure undergo competitive exclusion of lower-fitness genotypes. Both outcomes are indicative of a viral quasispecies (Domingo et al., 1998; Domingo et al., 2012; Solé et al., 1999; Summers & Litwin, 2006; Vignuzzi et al., 2006; Wilke, 2005)

Furthermore, evolutionary rate analysis identified hotspots within the targeted BRO-A fragment where mutation rates were higher than neutral which is indicative of a quasispecies. However, mutation rates were higher in hotspots and not the entire genome and this point is integral to applying the quasispecies model. If mutational rates were high across the entire quasispecies, the dominant genotype would be unable to adapt to its environment and would have deleterious effects on the population (Summers & Litwin, 2006).

Applying selection to the parent isolate highlighted this phenomenon further as chapter 6 and 9 demonstrated. In chapter 6, the AC53-T2 strain had a clear reduction in identified genotypes with two competing genotypes identified, signifying that mutations observed in chapter 5 are caused by selection acting on clouds of mutants. This is again demonstrated with chapter 9 in more detail as *k*-means clustering identified various amounts of genotype clusters within each derived strain that had alternative relative abundances.

The quasispecies model was demonstrated with the bratNextGen recombination analysis, which produced a result at odds with nucleotide distance and ORF analysis in chapter 9. This could be explained through mutational robustness as quasispecies exploit selection to maintain diversity rather than recombination (Van Nimwegen et al., 1999; Wilke et al., 2001). However, the bratNextGen software requires hundreds of sequences for an accurate analysis and result may change when more derived sequences are produced and included in the analysis (Marttinen et al., 2012).

Studies of viral quasispecies are normally applied to RNA viruses, which have a high mutation rate due to low fidelity during RNA replication, however, the application of quasispecies theory to DNA viruses has been previously suggested to result from lower fidelity and error correction in endogenous viral DNA polymerases compared to those involved in cellular replication (Andino & Domingo, 2015; Domingo et al., 1998; Domingo et al., 2012; Lauring & Andino, 2010). The possibility that baculoviruses are quasispecies has been suggested

(Chateigner et al., 2015; Cory et al., 2005). This thesis is the first clear evidence that validates the quasispecies hypothesis in baculoviruses.

10.3 FUTURE DIRECTIONS & FINAL THOUGHTS

As with many research projects many questions continue to linger after concluding. The presented research has open many pathways to continue this study by analysing phenotypic, ecological and evolutionary aspects of derived viruses, and overcoming computational and NGS limitations when analysing NPVs.

As previously mentioned, most bioinformatic techniques are designed for model organisms and this project required extensive research and development to produce the two software pipelines and various R scripts described in this thesis (da Fonseca et al., 2016; Noune, 2016). This is an area that requires active development. Future development should focus on current second-generation techniques to enable accurate genome assembly of each individual genotype within quasispecies i.e. expansion of MetaGaAP-Py to whole-genomes rather than small fragments and scalable in-depth comparisons and visualisation of polymorphisms with large numbers of datasets.

Due to limited time constraints, biological characterisation of *in vitro* derived strains was never completed, and this is key in assigning ORFs to phenotypic traits, in addition to an extensive comparison of *in vitro* and *in vivo* derived strains. Currently, no study has been completed providing an in-depth analysis and comparison of *in vivo* and *in vitro* phenotypes and genotypes. In addition, continuous passage of selection should be completed to validate the previously mentioned hypothesis, that continuous selection will eventually offset negative traits. This can be followed with further genetic analysis and a genome-wide association study. Furthermore, the limited time-constraints impacted some of the statistical analysis undertaken in chapters 7 and 8 as linear regressions were applied rather than a GLM with an appropriate distribution and would need to be implemented to satisfy publication requirements.

Chapter 7 revealed an important aspect with genotype abundance changes during the infection cycle and should be investigated further. This could be achieved by targeting multiple highly variable regions within the AC53 genome using MetaGaAP-Py, and assessing phenotypic and genotypic properties after serial passaging without applying a selection pressure. This could enhance commercial production runs by providing additional information regarding quality control such as viral mutations occurring during an infection cycle. In addition, applying MetaGaAP-Py to several different ORFs within each derived strain and monitoring the infection cycle would provide in-depth detail as to when genotype abundance is shifting and can use this analysis as a basis to derive strains at a time point containing a trait or traits.

Unfortunately, sequencing and designing primers for Hr regions within baculoviruses is a difficult task due to the AT-richness and high variability which introduces primer bias and

platform specific errors (Bragg et al., 2013; Clark & Whittam, 1992; Hoff, 2009; P. L. Johnson & Slatkin, 2008; McElroy et al., 2014; Nakamura et al., 2011; Noune & Hauxwell, 2016a; Quail et al., 2012; Schirmer et al., 2016; Schirmer et al., 2015). This issue will eventually resolve with third-generation sequencing platforms as will short-read limitations, eventually improving genome assembly as single reads can encompass a whole genome and rendering MetaGaAP-Py obsolete (Bleidorn, 2016; Carneiro et al., 2012; Jain et al., 2016; Olasagasti et al., 2010).

Commercially, this study has implicated the production of baculoviruses as pesticides since selection, including inadvertent selection, will impact the final OB produced as the composition of the quasispecies adapts to the sudden change in its host environment. Furthermore, the advent of third-generation sequencing platforms and constant development of bioinformatic techniques will help to resolve many issues that were present throughout this study.

The final thought for this thesis, is that baculoviruses are a highly diverse viral quasispecies that still require extensive research. Some questions have been answered, but many more still need answering, especially as baculoviruses can be used as a model system to answer core virological questions which cannot be answered in a non-insect environment.

Bibliography

Abbott, W. S. (1925). A Method of Computing the Effectiveness of an Insecticide. *Journal of economic entomology, 18*(2), 265-267. doi:10.1093/jee/18.2.265a

Afonso, C. L., Tulman, E. R., Lu, Z., Balinsky, C. A., Moser, B. A., Becnel, J. J., Rock, D. L., & Kutish, G. F. (2001). Genome Sequence of a Baculovirus Pathogenic for *Culex nigripalpus*. *Journal of virology, 75*(22), 11157-11165. doi:10.1128/jvi.75.22.11157-11165.2001

AgBiTech. (2013). Brazil Turning to Australian Agbiotech to Fight Insect Plague [Press release]

AgBiTech. (2014). AgBiTech Secures US EPA Approval for Biological Insect Control [Press release]

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

. Anaconda Software Distribution. (2017). https://continuum.io: Continuum Analytics.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic acids research, 9*(13), 3015-3027.

Andino, R., & Domingo, E. (2015). Viral quasispecies. *Virology, 479-480*(Supplement C), 46-51. doi:https://doi.org/10.1016/j.virol.2015.03.022

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/: Babraham Bioinformatics.

Arbiza, J., Mirazo, S., & Fort, H. (2010). Viral quasispecies profiles as the result of the interplay of competition and cooperation. *BMC evolutionary biology, 10*(1), 137.

Arif, B. M. (2005). A brief journey with insect viruses with emphasis on baculoviruses. *Journal of Invertebrate Pathology, 89*(1), 39-45.

Arrizubieta, M., Simón, O., Williams, T., & Caballero, P. (2015a). Genomic sequences of five Helicoverpa armigera nucleopolyhedrovirus genotypes from Spain that differ in their insecticidal properties. *Genome announcements, 3*(3), e00548-00515.

Arrizubieta, M., Simón, O., Williams, T., & Caballero, P. (2015b). A Novel Binary Mixture of Helicoverpa armigera Single Nucleopolyhedrovirus Genotypic Variants Has Improved Insecticidal Characteristics for Control of Cotton Bollworms. *Applied and Environmental Microbiology, 81*(12), 3984-3993.

Arrizubieta, M., Williams, T., Caballero, P., & Simón, O. (2014). Selection of a nucleopolyhedrovirus isolate from Helicoverpa armigera as the basis for a biological insecticide. *Pest Management Science, 70*(6), 967-976.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology, 71*(12), 7724-7736.

Asser-Kaiser, S., Fritsch, E., Undorf-Spahn, K., Kienzle, J., Eberle, K., Gund, N., Reineke, A., Zebitz, C., Heckel, D., & Huber, J. (2007). Rapid emergence of baculovirus resistance in codling moth due to dominant, sex-linked inheritance. *Science, 317*(5846), 1916-1918.

Asser-Kaiser, S., Fritsch, E., Undorf-Spahn, K., Kienzle, J., Eberle, K., Gund, N., Reineke, A., Zebitz, C., Heckel, D. G., & Huber, J. (2007). Rapid emergence of baculovirus resistance in codling moth due to dominant, sex-linked inheritance. *Science, 317*(5846), 1916-1918.

Asser-Kaiser, S., Heckel, D. G., & Jehle, J. A. (2010). Sex linkage of CpGV resistance in a heterogeneous field strain of the codling moth Cydia pomonella (L.). *Journal of Invertebrate Pathology, 103*(1), 59-64. doi:http://dx.doi.org/10.1016/j.jip.2009.10.005

Baillie, V. L., & Bouwer, G. (2011). Development of highly sensitive assays for detection of genetic variation in key Helicoverpa armigera nucleopolyhedrovirus genes. *Journal of Virological Methods*.

Baillie, V. L., & Bouwer, G. (2012a). High levels of genetic variation within core *Helicoverpa armigera* nucleopolyhedrovirus genes. *Virus Genes, 44*(1), 149-162.

Baillie, V. L., & Bouwer, G. (2012b). High levels of genetic variation within Helicoverpa armigera nucleopolyhedrovirus populations in individual host insects. *Archives of Virology, 157*(12), 2281-2289.

Ball, C. L., Gilchrist, M. A., & Coombs, D. (2007). Modeling within-host evolution of HIV: mutation, competition and strain replacement. *Bulletin of mathematical biology, 69*(7), 2361-2385.

Beerenwinkel, N., Gunthard, H. F., Roth, V., & Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol, 3*, 329. doi:10.3389/fmicb.2012.00329

Beerenwinkel, N., Günthard, H. F., Roth, V., & Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol, 3*. doi:10.3389/fmicb.2012.00329

Belaich, M. N., Rodríguez, V. A., Bilen, M. F., Pilloff, M. G., Romanowski, V., Sciocco-Cap, A., & Ghiringhelli, P. D. (2006). Sequencing and Characterisation of p74 Gene in Two Isolates of Anticarsia Gemmatalis MNPV. *Virus Genes, 32*(1), 59-70. doi:10.1007/s11262-005-5846-z

Belyavskyi, M., Braunagel, S. C., & Summers, M. D. (1998). The structural protein ODV-EC27 of Autographa californica nucleopolyhedrovirus is a multifunctional viral cyclin. *Proceedings of the National Academy of Sciences, 95*(19), 11205-11210.

Berling, M., Blachere-Lopez, C., Soubabere, O., Lery, X., Bonhomme, A., Sauphanor, B., & Lopez-Ferber, M. (2009). Cydia pomonella granulovirus Genotypes Overcome Virus Resistance in the Codling Moth and Improve Virus Efficiency by Selection against Resistant Hosts. *Applied and Environmental Microbiology, 75*(4), 925-930. doi:10.1128/aem.01998-08

Berling, M., Rey, J.-B., Ondet, S.-J., Tallot, Y., Soubabère, O., Bonhomme, A., Sauphanor, B., & Lopez-Ferber, M. (2009). Field trials of CpGV virus isolates overcoming resistance to CpGV-M. *Virologica Sinica, 24*(5), 470-477.

Bernal, A., Simón, O., Williams, T., Muñoz, D., & Caballero, P. (2013). A Chrysodeixis chalcites Single-Nucleocapsid Nucleopolyhedrovirus Population from the Canary Islands Is Genotypically Structured To Maximize Survival. *Applied and Environmental Microbiology, 79*(24), 7709-7718. doi:10.1128/aem.02409-13

Bézier, A., Annaheim, M., Herbinière, J., Wetterwald, C., Gyapay, G., Bernard-Samain, S., Wincker, P., Roditi, I., Heller, M., & Belghazi, M. (2009). Polydnaviruses of braconid wasps derive from an ancestral nudivirus. *Science, 323*(5916), 926-930.

Bideshi, D. K., Renault, S., Stasiak, K., Federici, B. A., & Bigot, Y. (2003). Phylogenetic analysis and possible function of bro-like genes, a multigene family widespread among large double-stranded DNA viruses of invertebrates and bacteria. *Journal of general virology, 84*(9), 2531-2544.

Black, B. C., Brennan, L. A., Dierks, P. M., & Gard, I. (1997). *Commercialization of baculoviral insecticides* (Vol. 341): Plenum Press, New York.

Black, B. C., Brennan, L. A., Dierks, P. M., & Gard, I. E. (1997). Commercialization of baculoviral insecticides *The baculoviruses* (pp. 341-387): Springer.

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity, 14*(1), 1-8.

Blissard, G. W., & Rohrmann, G. F. (1990). Baculovirus diversity and molecular biology. *Annual Review of Entomology, 35*(1), 127-155.

Bonning, B. C., & Nusawardani, T. (2007). Introduction to the use of baculoviruses as biological insecticides. *Methods in Molecular Biology, 388*, 359.

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., & Tyson, G. W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol, 9*(4), e1003031.

Braunagel, S., Elton, D., Ma, H., & Summers, M. (1996). Identification and analysis of an *Autographa californica* nuclear polyhedrosis virus structural protein of the occlusion-derived virus envelope: ODV-E56. *Virology, 217*(1), 97-110.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs, 27*(4), 325-349.

Brittnacher, M. J., Heltshe, S. L., Hayden, H. S., Radey, M. C., Weiss, E. J., Damman, C. J., Zisman, T. L., Suskind, D. L., & Miller, S. I. (2016). GUTSS: An Alignment-Free Sequence Comparison Method for Use in Human Intestinal Microbiome and Fecal Microbiota Transplantation Analysis. *PLoS ONE, 11*(7), e0158897.

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., & Girerd, P. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC microbiology, 15*(1), 66.

Brown, M., & Faulkner, P. (1977). A plaque assay for nuclear polyhedrosis viruses using a solid overlay. *Journal of general virology, 36*(2), 361-364.

Brown, M., & Faulkner, P. (1978). Plaque assay of nuclear polyhedrosis viruses in cell culture. *Applied and Environmental Microbiology, 36*(1), 31-35.

Buerger, P., Hauxwell, C., & Murray, D. (2007). Nucleopolyhedrovirus introduction in Australia. *Virologica Sinica, 22*(2), 173-179.

Bull, J. J., & Lauring, A. S. (2014). Theory and empiricism in virulence evolution. *PLoS Pathog, 10*(10), e1004387.

Burand, J. P., Nakai, M., & Smith, I. (2009). Host-Virus Interactions. In S. P. Stock (Ed.), *Insect pathogens: molecular approaches and techniques*: CABI.

Burden, J. P., Nixon, C. P., Hodgkinson, A. E., Possee, R. D., Sait, S. M., King, L. A., & Hails, R. S. (2003). Covert infections as a mechanism for long□term persistence of baculoviruses. *Ecology Letters, 6*(6), 524-531.

Burges, H. D. (1998). *Formulation of microbial biopesticides: beneficial microorganisms, nematodes, and seed treatments*: Springer.

Burke, M. K. (2012). How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20120799.

Burrows, M., & Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi:10.1186/1471-2105-10-421

Capobianchi, M. R., Giombini, E., & Rozera, G. (2013). Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection, 19*(1), 15-22. doi:10.1111/1469-0691.12056

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., & Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods, 7*(5), 335-336. doi:http://www.nature.com/nmeth/journal/v7/n5/suppinfo/nmeth.f.303_S1.html

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S., Nusbaum, C., & Depristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics, 13*. doi:10.1186/1471-2164-13-375

Carstens, E. B., & Wu, Y. (2007). No single homologous repeat region is essential for DNA replication of the baculovirus Autographa californica multiple nucleopolyhedrovirus. *Journal of general virology, 88*(1), 114-122.

Cary, L. C., Goebel, M., Corsaro, B. G., Wang, H. G., Rosen, E., & Fraser, M. J. (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology, 172*(1), 156-169.

Chang, S., Zhang, J., Liao, X., Zhu, X., Wang, D., Zhu, J., Feng, T., Zhu, B., Gao, G. F., & Wang, J. (2007). Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. *Nucleic acids research, 35*(suppl 1), D376-D380.

Chaston, T. B., & Lidbury, B. A. (2001). Genetic 'budget'of viruses and the cost to the infected host: a theory on the relationship between the genetic capacity of viruses, immune evasion, persistence and disease. *Immunology and cell biology, 79*(1), 62-66.

Chateigner, A., Bézier, A., Labrousse, C., Jiolle, D., Barbe, V., & Herniou, E. A. (2015). Ultra Deep Sequencing of a Baculovirus Population Reveals Widespread Genomic Variations. *Viruses, 7*(7), 3625-3646.

Chen, E. Z., Bushman, F. D., & Li, H. (2016). A Model-Based Approach for Species Abundance Quantification Based on Shotgun Metagenomic Data. *Statistics in Biosciences*, 1-15.

Chen, X., IJkel, W. F., Tarchini, R., Sun, X., Sandbrink, H., Wang, H., Peters, S., Zuidema, D., Lankhorst, R. K., & Vlak, J. M. (2001). The sequence of the Helicoverpa armigera single nucleocapsid nucleopolyhedrovirus genome. *Journal of general virology, 82*(1), 241-257.

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics, 30*(1), 31-37. doi:10.1093/bioinformatics/btt310

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., & Eichler, E. E. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods, 10*(6), 563-569.

Christian, P. D., Gibb, N., Kasprzak, A. B., & Richards, A. (2001). A rapid method for the identification and differentiation of *Helicoverpa* nucleopolyhedroviruses (NPV *Baculoviridae*) isolated from the environment. *Journal of Virological Methods, 96*(1), 51-65.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly, 6*(2), 80-92. doi:10.4161/fly.19695

Clark, A. G., & Whittam, T. S. (1992). Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution, 9*(4), 744-752.

Clarke, D. K., Duarte, E. A., Elena, S. F., Moya, A., Domingo, E., & Holland, J. (1994). The red queen reigns in the kingdom of RNA viruses. *Proceedings of the National Academy of Sciences, 91*(11), 4821-4824.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular ecology resources, 14*(6), 1160-1170.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews, 17*(4), 840-862.

Clavijo, G., Williams, T., Muñoz, D., Caballero, P., & López-Ferber, M. (2010). Mixed genotype transmission bodies and virions contribute to the maintenance of diversity in an insect virus. *Proceedings of the Royal Society B: Biological Sciences, 277*(1683), 943-951.

Clavijo, G., Williams, T., Muñoz, D., López-Ferber, M., & Caballero, P. (2009). Entry into midgut epithelial cells is a key step in the selection of genotypes in a nucleopolyhedrovirus. *Virologica Sinica, 24*(4), 350-358.

Cock, P. J. (2010). BioPython Redundant Fasta Sequence Removal Function. http://lists.open-bio.org/pipermail/biopython/2010-April/012615.html.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., & Wilczynski, B. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422-1423.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic acids research*, gkt1244.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., & Walters, L. (1998). New goals for the US human genome project: 1998-2003. *Science, 282*(5389), 682-689.

Copping, L. G., & Menn, J. J. (2000). Biopesticides: a review of their action, applications and efficacy. *Pest Management Science, 56*(8), 651-676. doi:10.1002/1526-4998(200008)56:8<651::aid-ps201>3.0.co;2-u

Corsaro, B. G., & Fraser, M. J. (1987). Characterization of Genotypic and Phenotypic Variation in Plaque-Purified Strains of HzSNPV Elkar Isolate. *Intervirology, 28*(4), 185-198.

Cory, J. S., Green, B. M., Paul, R. K., & Hunter-Fujita, F. (2005). Genotypic and phenotypic diversity of a baculovirus population within an individual insect host. *Journal of Invertebrate Pathology, 89*(2), 101-111.

Crotty, S., Cameron, C. E., & Andino, R. (2001). RNA virus error catastrophe: direct molecular test by using ribavirin. *Proceedings of the National Academy of Sciences, 98*(12), 6895-6900.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., Zepeda-Mendoza, M. L., Campos, P. F., Heller, R., & Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine genomics, 30*, 3-13.

Daly, J. C., & Murrary, D. A. H. (1988). Evolution of resistance to pyrethroids in Heliothis armigera (Hubner)(Lepidoptera: Noctuidae) in Australia. *Journal of economic entomology, 81*(4), 984-988.

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). A Model of Evolutionary Change in Proteins *Atlas of protein sequence and structure* (Vol. 5, pp. 345-352): National Biomedical Research Foundation Silver Spring, MD.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi:10.1038/ng.806

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology, 72*(7), 5069-5072.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297-302.

Djira, G. D., Hasler, M., Gerhard, D., Schaarschmidt, F., & Schaarschmidt, M. F. (2011). Package 'mratios'.

Domingo, E., Baranowski, E., Ruiz-Jarabo, C. M., Martín-Hernández, A. M., Sáiz, J. C., & Escarmís, C. (1998). Quasispecies structure and persistence of RNA viruses. *Emerging infectious diseases, 4*(4), 521.

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews, 76*(2), 159-216.

Dowle, M., Short, T., & Lianoglou, S. (2013). data. table: Extension of data. frame for fast indexing, fast ordered joins, fast assignment, fast grouping and list columns. *R package version, 1*(8).

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics, 9*(4), 267-276.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics, 20*(3), 482-491.

Eberle, K. E., & Jehle, J. A. (2006). Field resistance of codling moth against Cydia pomonella granulovirus (CpGV) is autosomal and incompletely dominant inherited. *Journal of Invertebrate Pathology, 93*(3), 201-206. doi:http://dx.doi.org/10.1016/j.jip.2006.07.001

Eigen, M. (1978). The hypercycle: A principle of natural self-organization. *International Journal of Quantum Chemistry, 14*(S5), 219-219. doi:10.1002/qua.560140722

Eigen, M., & Biebricher, C. K. (1988). Sequence space and quasispecies distribution. *RNA genetics, 3*, 211-245.

Ekblom, R., & Wolf, J. B. (2014). A field guide to whole□genome sequencing, assembly and annotation. *Evolutionary applications, 7*(9), 1026-1042.

Endrullat, C., Glökler, J., Franke, P., & Frohme, M. (2016). Standardization and quality management in next-generation sequencing. *Applied & Translational Genomics, 10*, 2-9.

Erlandson, M. (2008). Insect Pest Control by Viruses. In B. W. J. M. Editors-in-Chief: & M. H. V. v. Regenmortel (Eds.), *Encyclopedia of Virology (Third Edition)* (pp. 125-133). Oxford: Academic Press.

Evans, P. O., & O'Reilly, D. R. (1998). Purification and kinetic analysis of a baculovirus ecdysteroid UDP-glucosyltransferase. *Biochemical Journal, 330*(3), 1265-1270.

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics, 155*(3), 1405-1413.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 175-185.

Fitt, G., Cotter, S., & Sharma, H. (2005). The Helicoverpa problem in Australia: biology and management. *Heliothis/Helicoverpa management: emerging trends and strategies for future research*, 45-61.

Fitt, G. P. (2000). An Australian approach to IPM in cotton: integrating new technologies to minimise insecticide dependence. *Crop Protection, 19*(8–10), 793-800. doi:10.1016/s0261-2194(00)00106-x

Fitt, G. P. (2003). Deployment and impact of transgenic Bt cotton in Australia. *The economic and environmental impacts of agbiotech: A global perspective*, 141-164.

Fleming-Davies, A. E., Dwyer, G., Rohani, P., & Kalisz, S. (2015). Phenotypic Variation in Overwinter Environmental Transmission of a Baculovirus and the Cost of Virulence. *The American Naturalist, 186*(6), 797-806.

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications, 1*.

Fraley, C., & Raftery, A. E. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification, 16*(2), 297-306.

Fraley, C., & Raftery, A. E. (2006). *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. Retrieved from

Fraser, M., Smith, G. E., & Summers, M. D. (1983). Acquisition of host cell DNA sequences by baculoviruses: relationship between host DNA insertions and FP mutants of

Autographa californica and Galleria mellonella nuclear polyhedrosis viruses. *Journal of virology, 47*(2), 287-300.

Fraser, M. J., Cary, L., Boonvisudhi, K., & Wang, H.-G. H. (1995). Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. *Virology, 211*(2), 397-407.

Fullwood, M. J., Wei, C.-L., Liu, E. T., & Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res, 19*(4), 521-532.

Fuxa, J., & Richter, A. (1989). Reversion of resistance by *Spodoptera frugiperda* to nuclear polyhedrosis virus. *Journal of Invertebrate Pathology, 53*(1), 52-56.

Fuxa, J., Weidner, E., & Richter, A. (1992). Polyhedra without virions in a vertically transmitted nuclear polyhedrosis virus. *Journal of Invertebrate Pathology, 60*(1), 53-58.

Fuxa, J. R., & Richter, A. R. (1992). Virulence and multigeneration passage of a nuclear polyhedrosis virus selected for an increased rate of vertical transmission. *Biological Control, 2*(3), 171-175.

Garavaglia, M. J., Miele, S. A. B., Iserte, J. A., Belaich, M. N., & Ghiringhelli, P. D. (2012). The ac53, ac78, ac101, and ac103 genes are newly discovered core genes in the family Baculoviridae. *Journal of virology, 86*(22), 12069-12079.

Garcia-Maruniak, A., Maruniak, J. E., Zanotto, P. M., Doumbouya, A. E., Liu, J.-C., Merritt, T. M., & Lanoie, J. S. (2004). Sequence analysis of the genome of the Neodiprion sertifer nucleopolyhedrovirus. *Journal of virology, 78*(13), 7036-7051.

Gardner, R. C., Howarth, A. J., Hahn, P., Brown-Luedi, M., Shepherd, R. J., & Messing, J. (1981). The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic acids research, 9*(12), 2871-2888.

Gebhardt, M. M., Eberle, K. E., Radtke, P., & Jehle, J. A. (2014). Baculovirus resistance in codling moth is virus isolate-dependent and the consequence of a mutation in viral gene pe38. *Proceedings of the National Academy of Sciences, 111*(44), 15711-15716.

Gershburg, E., Stockholm, D., Froy, O., Rashi, S., Gurevitz, M., & Chejanovsky, N. (1998). Baculovirus□mediated expression of a scorpion depressant toxin improves the insecticidal efficacy achieved with excitatory toxins. *FEBS letters, 422*(2), 132-136.

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E. A., & Cordaux, R. (2014). Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. *Nat Commun, 5*. doi:10.1038/ncomms4348

Gomi, S., Majima, K., & Maeda, S. (1999). Sequence analysis of the genome of Bombyx mori nucleopolyhedrovirus. *Journal of general virology, 80*(5), 1323-1337. doi:doi:10.1099/0022-1317-80-5-1323

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351.

Gordon, A., & Hannon, G. (2010). Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished)* http://hannonlab.cshl.edu/fastx_toolkit.

Goulson, D., & Cory, J. S. (1995). Sublethal effects of baculovirus in the cabbage moth, Mamestra brassicae. *Biological Control, 5*(3), 361-367.

Goulson, D., & Hauxwell, C. (1995). Resistance or covert infection; baculovirus studies re-examined. *Functional ecology*.

Graillot, B., Berling, M., Blachere-López, C., Siegwart, M., Besse, S., & López-Ferber, M. (2014). Progressive adaptation of a CpGV isolate to codling moth populations resistant to CpGV-M. *Viruses, 6*(12), 5135-5144.

Guarino, L. A., Gonzalez, M. A., & Summers, M. D. (1986). Complete sequence and enhancer function of the homologous DNA regions of Autographa californica nuclear polyhedrosis virus. *Journal of virology, 60*(1), 224-229.

Haight, F. A. (1967). Handbook of the Poisson distribution.

Hails, R. S., Hernandez-Crespo, P., Sait, S. M., Donnelly, C. A., Green, B. M., & Cory, J. S. (2002). Transmission patterns of natural and recombinant baculoviruses. *Ecology, 83*(4), 906-916.

Hamblin, M., Van Beek, N. A. M., Hughes, P. R., & Wood, H. A. (1990). Co-Occlusion and Persistence of a Baculovirus Mutant Lacking the Polyhedrin Gene. *Applied and Environmental Microbiology, 56*(10), 3057-3062.

Hardin, G. (1960). The competitive exclusion principle. *Science, 131*(3409), 1292-1297.

Harrison, R. L. (2009a). Genomic sequence analysis of the Illinois strain of the Agrotis ipsilon multiple nucleopolyhedrovirus. *Virus Genes, 38*(1), 155-170. doi:10.1007/s11262-008-0297-y

Harrison, R. L. (2009b). Structural divergence among genomes of closely related baculoviruses and its implications for baculovirus evolution. *Journal of Invertebrate Pathology, 101*(3), 181-186.

Harrison, R. L. (2013). Concentration-and time-response characteristics of plaque isolates of Agrotis ipsilon multiple nucleopolyhedrovirus derived from a field isolate. *Journal of Invertebrate Pathology, 112*(2), 159-161.

Hauxwell, C. (1999). *Evaluation of potential baculovirus insecticides : studies of the infection process and host susceptibility.* Imperial College London (University of London).

Hauxwell, C. (2008a). Against the one hundredth locust: the commercial use of insect pathogens. *Bioindustry, 1*, 45.

Hauxwell, C. (2008b). Against the one hundredth locust: the commercial use of insect pathogens. *Microbiology Australia, 29*(1), 45-47.

Hayakawa, T., Ko, R., Okano, K., Seong, S.-I., Goto, C., & Maeda, S. (1999). Sequence Analysis of the Xestia c-nigrum Granulovirus Genome. *Virology, 262*(2), 277-297. doi:https://doi.org/10.1006/viro.1999.9894

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, 89*(22), 10915-10919.

Herniou, E. A., & Jehle, J. A. (2007). Baculovirus phylogeny and evolution. *Current drug targets, 8*(10), 1043-1050.

Herniou, E. A., Olszewski, J. A., Cory, J. S., & O'Reilly, D. R. (2003). The genome sequence and evolution of baculoviruses. *Annual Review of Entomology, 48*(1), 211-234.

Herniou, E. A., Olszewski, J. A., O'reilly, D. R., & Cory, J. S. (2004). Ancient coevolution of baculoviruses and their insect hosts. *Journal of virology, 78*(7), 3244-3251.

Highnam, G., Wang, J. J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., & Mittelman, D. (2015). An analytical framework for optimizing variant discovery from personal genomes. *Nature communications, 6*.

Hodgson, D. J., Hitchman, R. B., Vanbergen, A. J., Hails, R. S., Possee, R. D., & Cory, J. S. (2004). Host ecology determines the relative fitness of virus genotypes in mixed□ genotype nucleopolyhedrovirus infections. *Journal of evolutionary biology, 17*(5), 1018-1025.

Hodgson, D. J., Vanbergen, A. J., Hartley, S. E., Hails, R. S., & Cory, J. S. (2002). Differential selection of baculovirus genotypes mediated by different species of host food plant. *Ecology Letters, 5*(4), 512-518. doi:10.1046/j.1461-0248.2002.00338.x

Hoff, K. J. (2009). The effect of sequencing errors on metagenomic gene prediction. *BMC genomics, 10*(1), 1.

Holland, J. J. d., De La Torre, J., & Steinhauer, D. (1992). RNA virus populations as quasispecies *Genetic Diversity of RNA Viruses* (pp. 1-20): Springer.

Huang, J., Hu, R., Pray, C., Qiao, F., & Rozelle, S. (2003). Biotechnology as an alternative to chemical pesticides: a case study of Bt cotton in China. *Agricultural Economics, 29*(1), 55-67. doi:10.1111/j.1574-0862.2003.tb00147.x

Hughes, D. S., Possee, R. D., & King, L. A. (1993). Activation and Detection of a Latent Baculovirus Resembling *Mamestra brassicae* Nuclear Polyhedrosis Virus in *M. brassicae* Insects. *Virology, 194*(2), 608-615.

Hughes, D. S., Possee, R. D., & King, L. A. (1997). Evidence for the presence of a low-level, persistent baculovirus infection of Mamestra brassicae insects. *Journal of general virology, 78*(7), 1801-1805.

Hughes, P. R., & Wood, H. A. (1981). A synchronous peroral technique for the bioassay of insect viruses. *Journal of Invertebrate Pathology, 37*(2), 154-159. doi:http://dx.doi.org/10.1016/0022-2011(81)90069-0

Hunt, M., Newbold, C., Berriman, M., & Otto, T. D. (2014). A comprehensive evaluation of assembly scaffolding tools. *Genome Biology, 15*(3), R42. doi:10.1186/gb-2014-15-3-r42

Hunter-Fujita, F. R., Entwistle, P., Evans, H., & Crook, N. (1998). *Insect viruses and pest management*: John Wiley & Sons Ltd.

Hutter, S., Vilella, A. J., & Rozas, J. (2006). Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics, 7*(1), 409.

Ikeda, M., Hamajima, R., & Kobayashi, M. (2015). Baculoviruses: diversity, evolution and manipulation of insects. *Entomological Science, 18*(1), 1-20.

Institute, B. Picard. Retrieved from http://broadinstitute.github.io/picard/

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology, 17*(1), 239.

Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology, 72*(3), 1719-1728.

Javed, M. A., Biswas, S., Willis, L. G., Harris, S., Pritchard, C., van Oers, M. M., Donly, B. C., Erlandson, M. A., Hegedus, D. D., & Theilmann, D. A. (2017). Autographa californica multiple nucleopolyhedrovirus AC83 is a per os infectivity factor (PIF) protein required for occlusion-derived virus (ODV) and budded virus nucleocapsid assembly as well as assembly of the PIF complex in ODV envelopes. *Journal of virology, 91*(5), e02115-02116.

Jehle, J., Eberle, K., Asser-Kaiser, S., Schulze-Bopp, S., & Schmitt, A. (2010). *Resistance of codling moth against Cydia pomonella granulovirus (CpGV): state of knowledge.* Paper presented at the 14th international conference on organic fruit-growing, Hohenheim, Germany.

Jehle, J. A., Blissard, G. W., Bonning, B. C., Cory, J. S., Herniou, E. A., Rohrmann, G. F., Theilmann, D. A., Thiem, S. M., & Vlak, J. M. (2006). On the classification and nomenclature of baculoviruses: A proposal for revision. *Archives of Virology, 151*(7), 1257-1266. doi:10.1007/s00705-006-0763-6

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., & Bustamante, C. D. (2005). Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. *Genetics, 170*(3), 1401-1410. doi:10.1534/genetics.104.038224

Johnson, A. D. (2009). SNP bioinformatics: a comprehensive review of resources. *Circulation. Cardiovascular genetics, 2*(5), 530-536. doi:10.1161/CIRCGENETICS.109.872010

Johnson, M.-L., Pearce, S., Wade, M., Davies, A., Silberbauer, L., Gregg, P., & Zalucki, M. (2000). Review of beneficials in cotton farming systems. *Cotton Research and Development Corporation, Narrabri, Australia*.

Johnson, P. L., & Slatkin, M. (2008). Accounting for bias from sequencing error in population genetic estimates. *Molecular Biology and Evolution, 25*(1), 199-206.

Jones, M. B., Highlander, S. K., Anderson, E. L., Li, W., Dayrit, M., Klitgord, N., Fabani, M. M., Seguritan, V., Green, J., & Pride, D. T. (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences, 112*(45), 14024-14029.

Kang, W., Suzuki, M., Zemskov, E., Okano, K., & Maeda, S. (1999). Characterization of Baculovirus Repeated Open Reading Frames (bro) in Bombyx mori Nucleopolyhedrovirus. *Journal of virology, 73*(12), 10339-10345.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution, 30*(4), 772-780. doi:10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., & Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res, 22*(3), 568-576. doi:10.1101/gr.129684.111

Kolde, R. (2012). Pheatmap: pretty heatmaps. *R package version, 61*.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., Bates, S. T., Bruns, T. D., Bengtsson□Palme, J., & Callaghan, T. M. (2013). Towards a unified paradigm for sequence□based identification of fungi. *Molecular ecology, 22*(21), 5271-5277.

Krell, P. J. (2008). *Baculoviruses: General Features, in Encyclopedia of Virology (Third Edition)*: Academic Press: Oxford.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, msw054.

Kuzio, J., Jaques, R. t., & Faulkner, P. (1989). Identification of p74, a gene essential for virulence of baculovirus occlusion bodies. *Virology, 173*(2), 759-763.

Kuzio, J., Pearson, M. N., Harwood, S. H., Funk, C. J., Evans, J. T., Slavicek, J. M., & Rohrmann, G. F. (1999). Sequence and Analysis of the Genome of a Baculovirus Pathogenic for *Lymantria dispar*. *Virology, 253*(1), 17-34.

Landais, I., Vincent, R., Bouton, M., Devauchelle, G., Duonor-Cerutti, M., & Ogliastro, M. (2006). Functional analysis of evolutionary conserved clustering of bZIP binding sites in the baculovirus homologous regions (hrs) suggests a cooperativity between host and viral transcription factors. *Virology, 344*(2), 421-431.

Lauring, A. S., & Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog, 6*(7), e1001005.

Lawrie, D. S., Petrov, D. A., & Messer, P. W. (2011). Faster than Neutral Evolution of Constrained Sequences: The Complex Interplay of Mutational Biases and Weak Selection. *Genome Biology and Evolution, 3*, 383-395. doi:10.1093/gbe/evr032

Lennon, J. T., & Locey, K. J. (2016). The underestimation of global microbial diversity. *mBio, 7*(5), e01298-01216.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, X., Burnight, E. R., Cooney, A. L., Malani, N., Brady, T., Sander, J. D., Staber, J., Wheelan, S. J., Joung, J. K., & McCray, P. B. (2013). piggyBac transposase tools for genome engineering. *Proceedings of the National Academy of Sciences, 110*(25), E2279-E2287.

Li, Z., Pan, L., Yu, H., Li, L., Gong, Y., Yang, K., & Pang, Y. (2006). Characterization of Spodoptera litura multicapsid nucleopolyhedrovirus 38.7 k protein, which contains a conserved BRO domain. *Virus Research, 115*(2), 185-191.

Lin, X. a., Chen, Y., Yi, Y., & Zhang, Z. (2010). Baculovirus immediately early 1, a mediator for homologous regions enhancer function in trans. *Virology journal, 7*(1), 32.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology, 2012*, 11. doi:10.1155/2012/251364

Longworth, J., & Cunningham, J. (1968). The activation of occult nuclear-polyhedrosis viruses by foreign nuclear polyhedra. *Journal of Invertebrate Pathology, 10*(2), 361-367.

Lua, L. H., Pedrini, M. R., Reid, S., Robertson, A., & Tribe, D. E. (2002). Phenotypic and genotypic analysis of Helicoverpa armigera nucleopolyhedrovirus serially passaged in cell culture. *Journal of general virology, 83*(4), 945-955.

Lua, L. H., & Reid, S. (2000). Virus morphogenesis of Helicoverpa armigera nucleopolyhedrovirus in Helicoverpa zea serum-free suspension culture. *Journal of general virology, 81*(10), 2531-2543.

Lueders, T., & Friedrich, M. W. (2003). Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. *Appl Environ Microbiol, 69*. doi:10.1128/aem.69.1.320-326.2003

Lynch, M., & Walsh, B. (2007). *The origins of genome architecture* (Vol. 98): Sinauer Associates Sunderland.

Lynn, D. (2006). Baculovirus Multicapsid Nucleopolyhedrovirus (pp. Diagram of nucleopolyhedrovirus virions drawn by Dwight Lynn, USDA employee on personal time.). https://en.wikipedia.org/wiki/File:Nucleopolyhedrovirus.jpg#filelinks: Wikipedia.

Maghodia, A. B., Jarvis, D. L., & Geisler, C. (2014). Complete genome sequence of the Autographa californica multiple nucleopolyhedrovirus strain E2. *Genome announcements, 2*(6), e01202-01214.

Marchant, A., Mougel, F., Mendonça, V., Quartier, M., Jacquin-Joly, E., da Rosa, J., Petit, E., & Harry, M. (2016). Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus. *Insect biochemistry and molecular biology, 69*, 25-33.

Martin, E. R., Kinnamon, D., Schmidt, M. A., Powell, E., Zuchner, S., & Morris, R. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics, 26*(22), 2803-2810.

Marttinen, P., Hanage, W. P., Croucher, N. J., Connor, T. R., Harris, S. R., Bentley, S. D., & Corander, J. (2012). Detection of recombination events in bacterial genomes from large population samples. *Nucleic acids research, 40*(1), e6-e6.

McCarthy, C. B., & Theilmann, D. A. (2008). AcMNPV ac143 (odv-e18) is essential for mediating budded virus production and is the 30th baculovirus core gene. *Virology, 375*(1), 277-291. doi:10.1016/j.virol.2008.01.039

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research, 16*(3), 285-292.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability: Chapman & Hall.

McElroy, K., Thomas, T., & Luciani, F. (2014). Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microbial Informatics and Experimentation, 4*(1), 1-14. doi:10.1186/2042-5783-4-1

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution, 28*(11), 659-669.

Meynell, G. G., & Stocker, B. A. D. (1957). Some hypotheses on the aetiology of fatal infections in partially resistant hosts and their application to mice challenged with Salmonella paratyphi-B or Salmonella typhimurium by intraperitoneal injection. *Microbiology, 16*(1), 38-58.

Microsoft. (2016). Microsoft R Open. https://mran.revolutionanalytics.com/rro/: Microsoft.
Miele, S. A. B., Garavaglia, M. J., Belaich, M. N., & Ghiringhelli, P. D. (2011a). Baculovirus: Molecular Insights on Their Diversity and Conservation. *International Journal of Evolutionary Biology, 2011*, 379424. doi:10.4061/2011/379424
Miele, S. A. B., Garavaglia, M. J., Belaich, M. N., & Ghiringhelli, P. D. (2011b). Baculovirus: molecular insights on their diversity and conservation. *International Journal of Evolutionary Biology, 2011*.
Mignard, S., & Flandrois, J. (2006). 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of microbiological methods, 67*(3), 574-581.
Monobrullah, M., & Shankar, U. (2008). Sub-lethal effects of Splt MNPV infection on developmental stages of Spodoptera litura (Lepidoptera: Noctuidae). *Biocontrol Science and Technology, 18*(4), 431-437.
Moody, G. (2004). *Digital code of life: how bioinformatics is revolutionizing science, medicine, and business*: John Wiley & Sons.
Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schaffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics, 24*(16), 1757-1764. doi:10.1093/bioinformatics/btn322
Moscardi, F. (1999). Assessment of the application of baculoviruses for control of Lepidoptera. *Annual Review of Entomology, 44*(1), 257-289.
Müller, J., & Müller, K. (2004). TreeGraph: automated drawing of complex tree figures using an extensible tree description format. *Molecular Ecology Notes, 4*. doi:10.1111/j.1471-8286.2004.00813.x
Murphy, C. I., & Piwnica□Worms, H. (2001). Overview of the baculovirus expression system. *Current Protocols in Protein Science*, 5.4. 1-5.4. 4.
Murrell, P. (2002). The grid graphics package. *R News, 2*(2), 14-19.
Myers, J. H., Malakar, R., & Cory, J. S. (2000). Sublethal nucleopolyhedrovirus infection effects on female pupal weight, egg mass size, and vertical transmission in gypsy moth (Lepidoptera: Lymantriidae). *Environmental Entomology, 29*(6), 1268-1272.
Nadalin, F., Vezzi, F., & Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics, 13*(14), S8.
Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., & Takahashi, H. (2011). Sequence-specific error profile of illumina sequencers. *Nucleic Acids Res, 39*. doi:10.1093/nar/gkr344
Nayfach, S., Bradley, P. H., Wyman, S. K., Laurent, T. J., Williams, A., Eisen, J. A., Pollard, K. S., & Sharpton, T. J. (2015). Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLoS Comput Biol, 11*(11), e1004573.
Nealis, V., Turnquist, R., Morin, B., Graham, R., & Lucarotti, C. (2015). Baculoviruses in populations of western spruce budworm. *Journal of Invertebrate Pathology, 127*, 76-80.
Nei, M., & Kumar, S. (2000). *Molecular evolution and phylogenetics*: Oxford university press.
Neilson, J. W., Jordan, F. L., & Maier, R. M. (2013). Analysis of Artifacts Suggests DGGE Should Not Be Used For Quantitative Diversity Analysis. *Journal of microbiological methods, 92*(3), 256-263. doi:10.1016/j.mimet.2012.12.021
Nekrutenko, A., & Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics, 13*(9), 667-672.
Nguyen, Q., Chan, L. C., Nielsen, L. K., & Reid, S. (2013). Genome scale analysis of differential mRNA expression of Helicoverpa zea insect cells infected with a H. armigera baculovirus. *Virology, 444*(1), 158-170.
Nguyen, Q., Qi, Y. M., Wu, Y., Chan, L. C. L., Nielsen, L. K., & Reid, S. (2011). In vitro production of Helicoverpa baculovirus biopesticides—Automated selection of insect cell clones for manufacturing and systems biology studies. *Journal of Virological Methods, 175*(2), 197-205. doi:10.1016/j.jviromet.2011.05.011

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics, 12*(6), 443-451.

Noune, C. (2016). The Invertebrates & Microbiology Group Pipelines. https://github.com/CNoune/IMG_pipelines: GitHub, Queensland University of Technology.

Noune, C., & Hauxwell, C. (2015). Complete Genome Sequences of Helicoverpa armigera Single Nucleopolyhedrovirus Strains AC53 and H25EA1 from Australia. *Genome announcements, 3*(5). doi:10.1128/genomeA.01083-15

Noune, C., & Hauxwell, C. (2016a). Comparative Analysis of HaSNPV-AC53 and Derived Strains. *Viruses, 8*(11), 280.

Noune, C., & Hauxwell, C. (2016b). Complete Genome Sequences of Seven Helicoverpa armigera SNPV-AC53-Derived Strains. *Genome announcements, 4*(3). doi:10.1128/genomeA.00260-16

Noune, C., & Hauxwell, C. (2017a). Enhanced Pipeline 'MetaGaAP-Py' for the Analysis of Quasispecies and Non-Model Microbial Populations using Ultra-Deep 'Meta-barcode' Sequencing. *bioRxiv*. doi:10.1101/171520

Noune, C., & Hauxwell, C. (2017b). MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data. *Biology, 6*(1), 14.

O'Reilly, D. R., & Miller, L. K. (1991). Improvement of a baculovirus pesticide by deletion of the egt gene. *Nature biotechnology, 9*(11), 1086-1089.

Ogembo, J. G., Chaeychomsri, S., Kamiya, K., Ishikawa, H., Katou, Y., Ikeda, M., & Kobayashi, M. (2007). Cloning and comparative characterization of nucleopolyhedroviruses isolated from African bollworm, Helicoverpa armigera, (Lepidoptera: Noctudiae) in different geographic regions. *Journal of Insect Biotechnology and Sericology, 76*(1), 39-49.

Okano, K., Vanarsdall, A. L., & Rohrmann, G. F. (2007). A baculovirus alkaline nuclease knockout construct produces fragmented DNA and aberrant capsids. *Virology, 359*(1), 46-54. doi:https://doi.org/10.1016/j.virol.2006.09.008

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. (2007). The vegan package. *Community ecology package, 10*.

Olasagasti, F., Lieberman, K. R., Benner, S., Cherf, G. M., Dahl, J. M., Deamer, D. W., & Akeson, M. (2010). Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat Nanotechnol, 5*. doi:10.1038/nnano.2010.177

Onwuegbuzie, A. J., Daniel, L., & Leech, N. L. (2007). Pearson product-moment correlation coefficient. *Encyclopedia of measurement and statistics*, 751-756.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., & Iliopoulos, I. (2015). Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and biology insights, 9*, 75.

Passarelli, A. L., Todd, J. W., & Miller, L. K. (1994). A baculovirus gene involved in late gene expression predicts a large polypeptide with a conserved motif of RNA polymerases. *Journal of virology, 68*(7), 4673-4678.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50*(302), 157-175.

Pearson, M. N., & Rohrmann, G. F. (2002). Transfer, incorporation, and substitution of envelope fusion proteins among members of the Baculoviridae, Orthomyxoviridae, and Metaviridae (insect retrovirus) families. *Journal of virology, 76*(11), 5301-5304.

Peng, K., van Oers, M. M., Hu, Z., van Lent, J. W. M., & Vlak, J. M. (2010). Baculovirus Per Os Infectivity Factors Form a Complex on the Surface of Occlusion-Derived Virus. *Journal of virology, 84*(18), 9497-9504. doi:10.1128/jvi.00812-10

Pierre, L. (2015). Linearize a fasta sequence. https://gist.github.com/lindenb/2c0d4e11fd8a96d4c345#file-linearizefasta-awk.

Pierre, R. (2009). Renamed Pydee to Spyder (it changes everything...!). https://github.com/spyder-ide/spyder/commit/78a22a22577bbdde2c879da0429f08ad88dcff29#diff-e5fb0cda12f90dc4341247ddab54d1da: GitHub.

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics, 8*(1), 14.

Pocheville, A. (2015). The ecological niche: history and recent controversies *Handbook of evolutionary thinking in the sciences* (pp. 547-586): Springer.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS ONE, 9*(4), e93827.

Posada, D., Crandall, K. A., & Holmes, E. C. (2002). Recombination in evolutionary genomics. *Annual Review of Genetics, 36*(1), 75-97.

Prosperi, M. C., Prosperi, L., Bruselles, A., Abbate, I., Rozera, G., Vincenti, D., Solmone, M. C., Capobianchi, M. R., & Ulivi, G. (2011). Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics, 12*. doi:10.1186/1471-2105-12-5

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics, 13*, 341. doi:10.1186/1471-2164-13-341

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research, 41*(D1), D590-D596.

Redman, E. M., Wilson, K., & Cory, J. S. (2016). Trade□offs and mixed infections in an obligate□killing insect pathogen. *Journal of Animal Ecology*.

Redman, E. M., Wilson, K., Grzywacz, D., & Cory, J. S. (2010). High levels of genetic diversity in Spodoptera exempta NPV from Tanzania. *Journal of Invertebrate Pathology, 105*(2), 190-193. doi:10.1016/j.jip.2010.06.008

Reeson, A. F., Wilson, K., Gunn, A., Hails, R. S., & Goulson, D. (1998). Baculovirus resistance in the noctuid Spodoptera exempta is phenotypically plastic and responds to population density. *Proceedings of the Royal Society of London B: Biological Sciences, 265*(1407), 1787-1791.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics, 16*(6), 276-277.

Richards, A. R., & Christian, P. D. (1999). A rapid bioassay screen for quantifying nucleopolyhedroviruses (Baculoviridae) in the environment. *Journal of Virological Methods, 82*(1), 63-75. doi:10.1016/s0166-0934(99)00080-4

Rohel, D. Z., & Faulkner, P. (1984). Time course analysis and mapping of Autographa californica nuclear polyhedrosis virus transcripts. *Journal of virology, 50*(3), 739-747.

Rohrmann, G. (2011a). Introduction to the baculoviruses and their taxonomy *Baculovirus Molecular Biology* (2 ed.). Bethesda: National Center for Biotechnology Information.

Rohrmann, G. (2011b). Structural Proteins of Baculovirus Occlusion Bodies and Virions *Baculovirus Molecular Biology* (2 ed.). Bethesda: National Center for Biotechnology Information.

Rohrmann, G., Pearson, M., Bailey, T., Becker, R., & Beaudreau, G. (1981). N-terminal polyhedrin sequences and occluded Baculovirus evolution. *Journal of molecular evolution, 17*(6), 329-333.

Rohrmann, G. F. (1992). Baculovirus structural proteins. *Journal of general virology, 73*(4), 749-761.

Rohrmann, G. F. (2013a). The baculovirus replication cycle: effects on cells and insects.

Rohrmann, G. F. (2013b). Baculoviruses, retroviruses, DNA transposons (piggyBac), and insect cells.

Rohrmann, G. F. (2013c). Introduction to the baculoviruses, their taxonomy, and evolution.

Rohrmann, G. F. (2013d). Structural proteins of baculovirus occlusion bodies and virions.

Rowley, D. L., Popham, H. J. R., & Harrison, R. L. (2011). Genetic variation and virulence of nucleopolyhedroviruses isolated worldwide from the heliothine pests Helicoverpa armigera, Helicoverpa zea, and Heliothis virescens. *Journal of Invertebrate Pathology, 107*(2), 112-126. doi:10.1016/j.jip.2011.03.007

Sait, S., Begon, M., & Thompson, D. (1994). The effects of a sublethal baculovirus infection in the Indian meal moth, Plodia interpunctella. *Journal of Animal Ecology, 63*(3), 541-550.

Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., Harkins, T. T., Cookson, B. T., & Hoffman, N. G. (2014). Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling. *Applied and Environmental Microbiology, 80*(24), 7583-7591. doi:10.1128/AEM.02206-14

Sanschagrin, S., & Yergeau, E. (2014). Next-generation sequencing of 16S ribosomal RNA gene amplicons. *JoVE (Journal of Visualized Experiments)*(90), e51709-e51709.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., & Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics, 17*(1), 125.

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

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE, 6*(12), e27310.

Schultz, K. L., Wetter, J. A., Fiore, D. C., & Friesen, P. D. (2009). Transactivator IE1 is required for baculovirus early replication events that trigger apoptosis in permissive and nonpermissive cells. *Journal of virology, 83*(1), 262-272.

Shafer, R. W., Stevenson, D., & Chan, B. (1999). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research, 27*(1), 348-352.

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science, 5*, 209.

Simon, O., Palma, L., Beperet, I., Munoz, D., Lopez-Ferber, M., Caballero, P., & Williams, T. (2011). Sequence comparison between three geographically distinct Spodoptera frugiperda multiple nucleopolyhedrovirus isolates: Detecting positively selected genes. *J Invertebr Pathol, 107*(1), 33-42. doi:10.1016/j.jip.2011.01.002

Simón, O., Palma, L., Williams, T., López-Ferber, M., & Caballero, P. (2012). Analysis of a naturally-occurring deletion mutant of Spodoptera frugiperda multiple nucleopolyhedrovirus reveals sf58 as a new per os infectivity factor of lepidopteran-infecting baculoviruses. *Journal of Invertebrate Pathology, 109*(1), 117-126. doi:10.1016/j.jip.2011.10.010

Simón, O., Williams, T., Caballero, P., & López-Ferber, M. (2006). Dynamics of deletion genotypes in an experimental insect virus population. *Proceedings of the Royal Society B: Biological Sciences, 273*(1588), 783-790.

Simón, O., Williams, T., López-Ferber, M., & Caballero, P. (2005). Functional importance of deletion mutant genotypes in an insect nucleopolyhedrovirus population. *Applied and Environmental Microbiology, 71*(8), 4254-4262.

Sipos, R., Székely, A., Révész, S., & Márialigeti, K. (2010). Addressing PCR biases in environmental microbiology studies. *Bioremediation: Methods and Protocols*, 37-58.

Smith, I. R., & Crook, N. E. (1988). In vivo isolation of baculovirus genotypes. *Virology, 166*(1), 240-244.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research, 23*(01), 23-35.

Smith, K. M. (2012). The cytoplasmic virus diseases. *Insect pathology, an advanced treatise, 1*, 457-497.

Sokal, R. R., & Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon, 11*(2), 33-40. doi:10.2307/1217208

Solé, R. V., Ferrer, R., González-García, I., Quer, J., & Domingo, E. (1999). Red Queen Dynamics, Competition and Critical Points in a Model of RNA Virus Quasispecies. *Journal of Theoretical Biology, 198*(1), 47-59. doi:http://dx.doi.org/10.1006/jtbi.1999.0901

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics, 14*(1), 91.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr., 5*, 1-34.

Spence, R. J., Noune, C., & Hauxwell, C. (2016). Complete Genome Sequences of Four Isolates of Plutella xylostella Granulovirus. *Genome announcements, 4*(3). doi:10.1128/genomeA.00633-16

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics, 22*. doi:10.1093/bioinformatics/btl446

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033

Steven, L., & Salzberg, J. (2005). Beware of mis—assembled genomes. *Bioinformatics, 21*(4), 320-324.

Stöver, B. C., & Müller, K. F. (2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics, 11*(1), 1-9. doi:10.1186/1471-2105-11-7

Summers, J., & Litwin, S. (2006). Examining the theory of error catastrophe. *Journal of virology, 80*(1), 20-26.

Sun, F., & Xia, L. C. (2015). Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads. *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools*, 21-25.

Sun, X., Wang, H., Sun, X., Chen, X., Peng, C., Pan, D., Jehle, J. A., Van der Werf, W., Vlak, J. M., & Hu, Z. (2004). Biological activity and field efficacy of a genetically modified Helicoverpa armigera SNPV expressing an insect-selective toxin from a chimeric promoter. *Biol. Control, 29*, 124-137.

Sun, X., Wu, D., Jin, L., Ma, Y., Bonning, B. C., Peng, H., & Hu, Z. (2009). Impact of Helicoverpa armigera nucleopolyhedroviruses expressing a cathepsin L-like protease on target and nontarget insect species on cotton. *Biological Control, 49*(1), 77-83.

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., & Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America, 109*(45), 18488-18492. doi:10.1073/pnas.1216223109

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*(3), 585-595.

Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., & Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences, 109*(47), 19333-19338.

Team, R. (2014). RStudio: Integrated Development for R. *RStudio, Inc., Boston, MA. URL* http://www. *RStudio. com/ide*.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., Kõljalg, U., Kisand, V., Nilsson, H., & Hildebrand, F. (2015). Shotgun metagenomes and multiple primer

pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys, 10*, 1.

Thézé, J., Bézier, A., Periquet, G., Drezen, J.-M., & Herniou, E. A. (2011). Paleozoic origin of insect large dsDNA viruses. *Proceedings of the National Academy of Sciences, 108*(38), 15931-15935.

Titterington, J. S., Nun, T. K., & Passarelli, A. L. (2003). Functional dissection of the baculovirus late expression factor-8 gene: sequence requirements for late gene promoter activation. *Journal of general virology, 84*(7), 1817-1826.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36-46.

van Baalen, M., & Sabelis, M. W. (1995). The dynamics of multiple infection and the evolution of virulence. *American Naturalist*, 881-910.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics, 11*(1110), 11 10 11-11 10 33. doi:10.1002/0471250953.bi1110s43

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics, 30*(9), 418-426.

Van Nimwegen, E., Crutchfield, J. P., & Huynen, M. (1999). Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences, 96*(17), 9716-9720.

van Oers, M. M., Pijlman, G. P., & Vlak, J. M. (2015). Thirty years of baculovirus–insect cell protein expression: from dark horse to mainstream technology. *Journal of general virology, 96*(1), 6-23. doi:doi:10.1099/vir.0.067108-0

Vanarsdall, A. L., Okano, K., & Rohrmann, G. F. (2005). Characterization of the replication of a baculovirus mutant lacking the DNA polymerase gene. *Virology, 331*(1), 175-180. doi:http://dx.doi.org/10.1016/j.virol.2004.10.024

Vanarsdall, A. L., Pearson, M. N., & Rohrmann, G. F. (2007). Characterization of baculovirus constructs lacking either the Ac 101, Ac 142, or the Ac 144 open reading frame. *Virology, 367*(1), 187-195.

Vasconcelos, S. D., Cory, J. S., Speight, M. R., & Williams, T. (2002). Host stage structure and baculovirus transmission in Mamestra brassicae L.(Lepidoptera: Noctuidae) larvae: a laboratory examination of small scale epizootics. *Neotropical Entomology, 31*(3), 391-396.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*: Springer Science & Business Media.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., & Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature, 439*(7074), 344-348.

Vilella, A. J., Blanco-Garcia, A., Hutter, S., & Rozas, J. (2005). VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics, 21*(11), 2791-2793.

Virus Taxonomy: 2016 Release. (2016, 2016). August, 2016.

Vlak, J. M. (1979). The proteins of nonoccluded Autographa californica nuclear polyhedrosis virus produced in an established cell line of Spodoptera frugiperda. *Journal of Invertebrate Pathology, 34*(2), 110-118. doi:http://dx.doi.org/10.1016/0022-2011(79)90089-2

Vu, V. Q. (2011). ggbiplot: A ggplot2 based biplot. *R package version*.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society, 54*(3), 426-482.

Wall, J. D., Tang, L. F., Zerbe, B., Kvale, M. N., Kwok, P.-Y., Schaefer, C., & Risch, N. (2014). Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res, 24*(11), 1734-1739. doi:10.1101/gr.168393.113

Wang, Y.-j., Burand, J. P., & Jehle, J. A. (2007). Nudivirus genomics: diversity and classification. *Virologica Sinica, 22*(2), 128-136.

Wang, Y., & Jehle, J. A. (2009). Nudiviruses and other large, double-stranded circular DNA viruses of invertebrates: new insights on an old topic. *Journal of Invertebrate Pathology, 101*(3), 187-193.

Warton, D. I., & Guttorp, P. (2011). Compositional analysis of overdispersed counts using generalized estimating equations. *Environmental and ecological statistics, 18*(3), 427-446.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika, 61*(3), 439-447.

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., Angenent, L. T., Knight, R., & Ley, R. E. (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *The ISME journal, 6*(1), 94-103.

Westenberg, M., Uijtdewilligen, P., & Vlak, J. M. (2007). Baculovirus envelope fusion proteins F and GP64 exploit distinct receptors to gain entry into cultured insect cells. *Journal of general virology, 88*(12), 3302-3306.

White, S. M., Burden, J. P., Maini, P. K., & Hails, R. S. (2012). Modelling the within-host growth of viral infections in insects. *Journal of Theoretical Biology, 312*, 34-43. doi:http://dx.doi.org/10.1016/j.jtbi.2012.07.022

Wickham, H. (2009). plyr: Tools for splitting, applying and combining data. *R package version 0.1, 9*, 651.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1-23.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*: Springer.

Wickham, H., & Chang, W. (2015). devtools: Tools to make developing R code easier. *R package version, 1*(0).

Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC evolutionary biology, 5*(1), 44.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature, 412*(6844), 331-333.

Wilson, L., Mensah, R., & Fitt, G. (2004). Implementing integrated pest management in Australian cotton. *Insect Pest Management. Field and Protected Crops. Berlin*.

Xia, J., Wang, Q., Jia, P., Wang, B., Pao, W., & Zhao, Z. (2012). NGS catalog: a database of next generation sequencing studies in humans. *Human mutation, 33*(6).

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology, 12*(9), 635-645.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution, 3*(4), 613-623.

Yu, X., & Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics, 14*, 274-274. doi:10.1186/1471-2105-14-274

Zagordi, O., Bhattacharya, A., Eriksson, N., & Beerenwinkel, N. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics, 12*(1), 1.

Zemskov, E. A., Kang, W., & Maeda, S. (2000). Evidence for Nucleic Acid Binding Ability and Nucleosome Association of Bombyx mori Nucleopolyhedrovirus BRO Proteins. *Journal of virology, 74*(15), 6784-6789. doi:10.1128/jvi.74.15.6784-6789.2000

Zhang, G. (1989). Commercial viral insecticide-Heliothis armigera viral insecticide in China. *The IPM Practitioner, 11*, 13.

Zhang, G. (1994). Research, development and application of Heliothis viral pesticide in China. *Resources and environment in the Yangtze Valley, 3*, 1-6.

Zhang, G., & Bai, C. (1992). *Research and development of first commercial viral pesticide-Heliothis nuclear polyhedrosis virus pesticide in China*.

Zhou, R., Ling, S., Zhao, W., Osada, N., Chen, S., Zhang, M., He, Z., Bao, H., Zhong, C., & Zhang, B. (2011). Population genetics in nonmodel organisms: II. natural selection in marginal habitats revealed by deep sequencing on dual platforms. *Mol Biol Evol, 28*. doi:10.1093/molbev/msr102

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology, 32*(3), 246-251.

Zwart, M. P., Van Der Werf, W., Van Oers, M. M., Hemerik, L., Van Lent, J., De Visser, J., Vlak, J. M., & Cory, J. S. (2009). Mixed infections and the competitive fitness of faster□acting genetically modified viruses. *Evolutionary applications, 2*(2), 209-221.

Supplementary Material

12.1 COMPARITIVE ANALYSIS OF HASNPV-AC53 AND DERIVED STRAINS

Viruses **2016**, *8*, 280; doi:10.3390/v8110280 S1 of S12

Supplementary Materials: Comparative Analysis of HaSNPV-AC53 and Derived Strains

Christopher Noune and Caroline Hauxwell

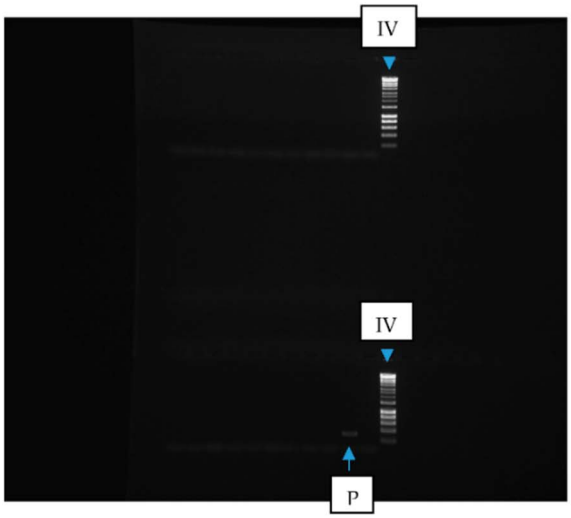


Figure S1. PCR detection for all NPV using rPol primer set. The markers used were the Hyper IV ladder (Bioline) and indicated as 'IV' on the figure. Positive control lane is indicated as "P". The positive control lane (purified HaSNPV-AC53) shows a 400 bp PCR fragment.

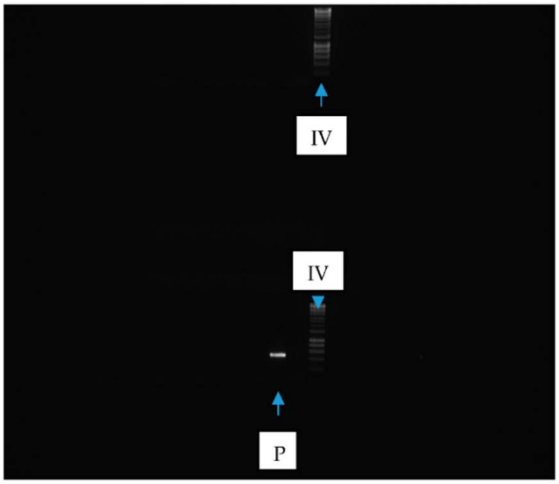


Figure S2. PCR detection of HaSNPV using the A44-RIX primer set. The markers used (were the Hyper IV ladders (Bioline) and indicated as 'IV' on the figure. Positive control lane is indicated as 'P'. The positive control lane (HaSNPV-AC53) shows a 500 bp PCR fragment.

Figure S3. Nucleotide comparison of ORF7 within AC53 and its derived strains. A single substitution and a 16 bp insertion extends the length of the open reading frame (ORF) within the derived strains. Substitutions (multi-colored boxes, green is a T substitution, red is an A substitution, blue is a C substitution and yellow is a G substitution) and deletions (black dotted line) are highlighted

Figure S4. Nucleotide comparison of ORF5 within AC53 and its derived strains. The deletion of an AC-repeat within all of the derived strains (except AC53-C3) results in the truncation of the ORF; CDS, coding DNA sequence. Two G substitutions (yellow) and a deletion (black dotted line) are highlighted.

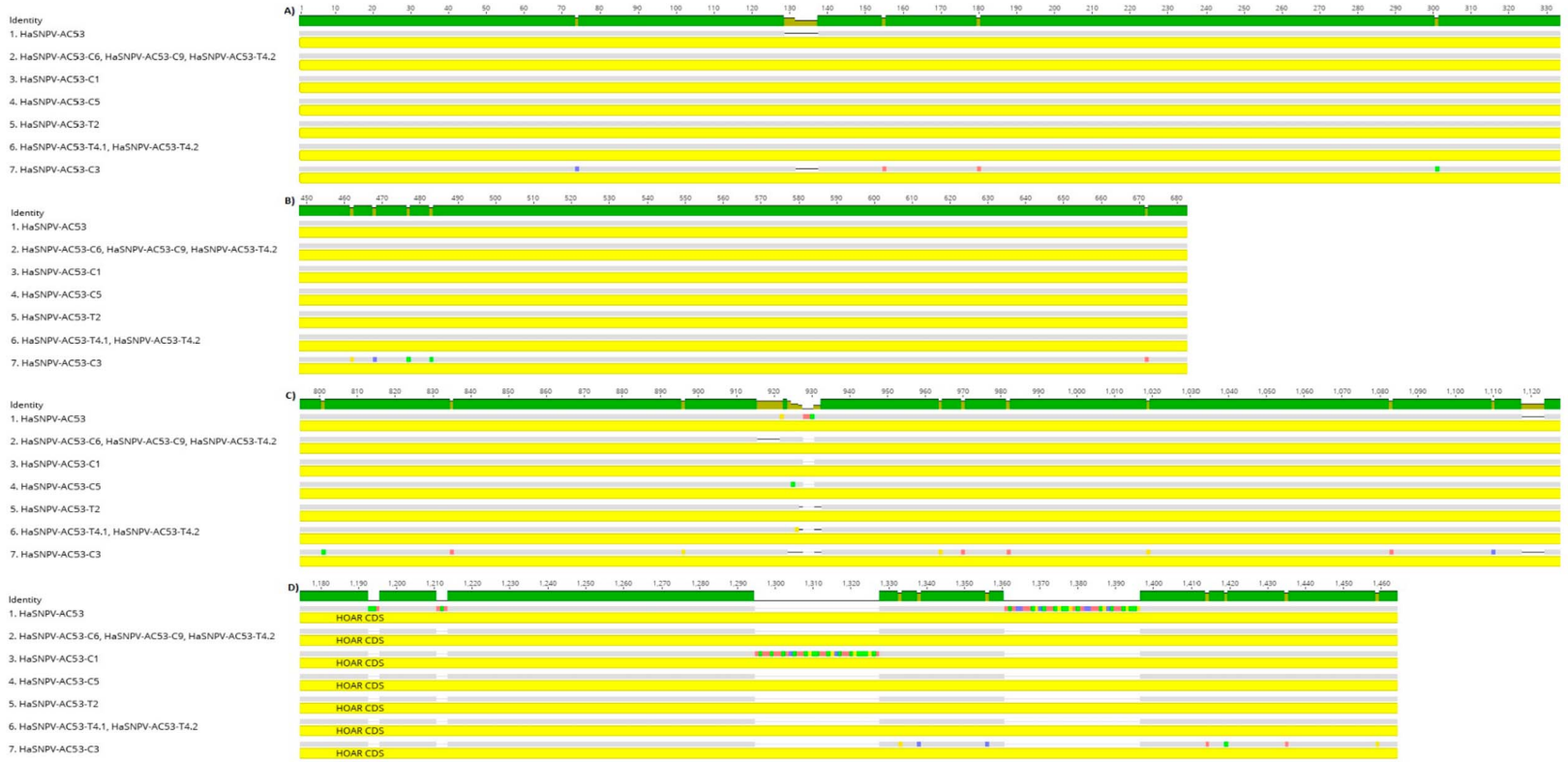


Figure S5. Nucleotide comparison of the four regions (A, B, C and D) containing mutations within the HOAR nucleotide sequence of AC53 and its derived strains. A total of six genotypes have been identified with the derived strains. Substitutions (multi-colored boxes, green is a T substitution, red is an A substitution, blue is a C substitution and yellow is a G substitution) and deletions (thin black line) are highlighted.

Figure S6. Nucleotide comparison of ORF128 with AC53 and its derivatives to the AC53-T4. Exclusive ORF128a and ORF128b, highlighting the substitutions (multi-colored boxes, green is a T substitution, red is an A substitution, blue is a C substitution and yellow is a G substitution) and deletions (thin black line) that have produced the fragmentation.

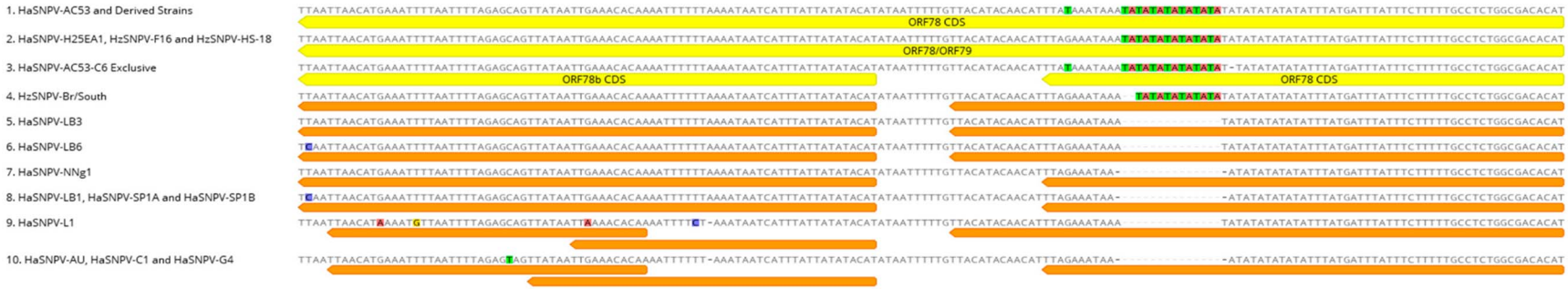


Figure S7. Nucleotide comparison of fragmentation occurring within ORF78/79 with 10 distinct genotypes observed across all *Helicoverpa armigera* Single Nucleopolyhedrovirus (HaSNPV) and *Helicoverpa zea* Single Nucleopolyhedrovirus (HzSNPV) strains. Manually annotated ORFs are underlined with orange. Substitutions and deletions are highlighted in the same manner as Figure S6.

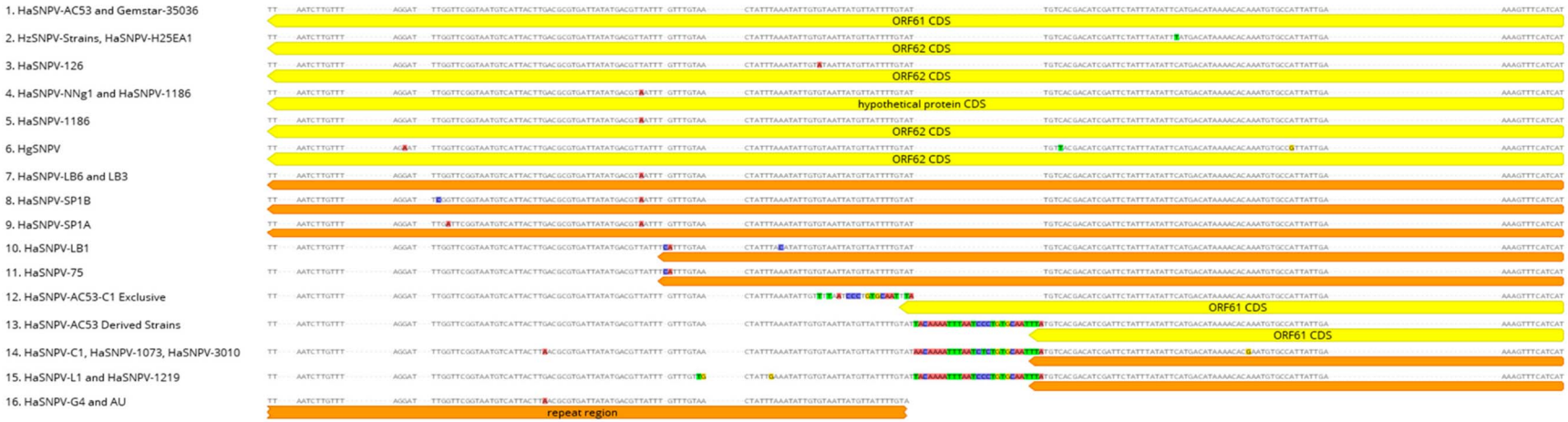


Figure S8. Nucleotide comparison of fragmentation occurring within ORF61/62 with 16 distinct genotypes observed across all HaSNPV and HzSNPV strains. Manually annotated ORFs are underlined with orange. Substitutions and deletions are highlighted in the same manner as Figure S6.

Table S1. Nucleotide and amino acid comparison of the AC53 and H25EA1 strains.

ORF/Homologous Repeat	AC53 Positions		H25EA1 Positions		Direction	Nucleotide Length (bp) (AC53)	Nucleotide Length (bp) (H25EA1)	Nucleotide Identity (%)	Amino Acid Identity (%)
	Start	End	Start	End					
Polyhedrin	1	741	1	741	forward	741	741	99.73	100
ORF2	738	1979	738	1979	reverse	1242	1242	99.6	99.75
PK1	1928	2797	1928	2797	forward	870	870	99.5	99.25
HOAR	2924	5255	2924	5255	reverse	2332	2332	99.43	100
ORF5	5388	5567	5388	5567	forward	180	180	98.89	98.30
ORF6	5717	6595	5717	6595	forward	879	879	99.77	100
ORF7	6807	6962	6807	6962	reverse	156	156	98.08	96.07
***ac141* Homolog**	6950	7807	6950	7807	forward	858	858	99.65	99.64
P49	7824	9230	7824	9230	forward	1407	1407	99.72	100
ODV-E18	9241	9486	9241	9486	forward	246	246	100	100
ODV-EC27	9501	10,355	9501	10,355	forward	855	855	99.65	99.64
ORF12	10,348	10,677	10,348	10,677	forward	330	330	99.64	100
ORF13	10,704	11,315	10,704	11,315	reverse	612	612	99.84	100
IE-1	11,318	13,324	11,318	13,324	forward	2007	2007	99.9	100
ODV-E56	13,377	14,441	13,377	14,441	reverse	1065	1065	99.62	99.71
ME53	14,591	15,670	14,591	15,670	forward	1080	1080	99.54	100
ORF17	15,673	15,840	15,673	15,840	forward	168	168	99.4	98.18
ORF18	15,893	16,174	15,893	16,174	reverse	282	282	100	100
P74	16,180	18,261	16,180	18,261	forward	2082	2082	99.37	99.85
P10	18,314	18,622	18,314	18,622	reverse	309	309	100	100
P26	18,660	19,463	18,660	19,463	reverse	804	804	99.5	99.62
ORF22	19,576	19,779	19,576	19,779	forward	204	204	100	100
***lef*-6**	19,855	20,418	19,855	20,418	reverse	564	564	100	100
DBP1	20,432	21,403	20,432	21,403	reverse	972	972	99.9	100
ORF25	21,547	22,023	21,547	22,023	forward	477	477	99.79	100
Hr1	22,024	23,949	22,024	23,949	forward	1926	1926	99.27	-
ORF26	23,950	24,102	23,950	24,102	forward	153	153	100	100
ORF27	24,045	24,812	24,045	24,812	reverse	768	768	100	100
Ubiquitin	24,652	24,903	24,652	24,903	forward	252	252	100	100
ORF29	24,949	25,473	25,005	25,472	forward	525	468	99.61	72.72
ORF30	25,492	26,064	25,491	26,063	forward	573	573	99.83	100
39K/PP31	26,128	27,063	26,127	27,062	reverse	936	936	99.68	100
***lef*-11**	27,029	27,481	27,028	27,480	reverse	453	453	99.73	100
ORF33	27,381	28,097	27,380	28,096	reverse	717	717	99.58	100

ORF34	28,328	29,407	28,327	29,406	forward	1080	1080	99.81	99.72
P47	29,475	30,713	29,474	30,712	reverse	1239	1239	99.91	100
ORF36	30,786	31,457	30,785	31,456	forward	672	672	99.85	100
ORF37	31,543	31,785	31,542	31,784	forward	243	243	100	100
***lef*-8**	31,782	34,487	31,781	34,486	reverse	2706	2706	99.74	100
ORF39	34,438	35,118	34,437	35,117	forward	681	681	99.48	99.47
ORF40	35,115	35,411	35,258	35,410	forward	297	153	98.32	96.94
Chitinase	35,419	37,146	35,418	37,145	reverse	1728	1728	99.83	99.82
ORF42	37,227	37,772	37,226	37,771	reverse	546	546	100	100
ORF43	37,870	38,298	37,869	38,297	forward	429	429	100	100
ORF44	38,305	39,441	38,304	39,440	reverse	1137	1137	99.82	100
ORF45	39,449	39,688	39,448	39,687	reverse	240	240	100	100
***lef*-10**	39,606	39,851	39,605	39,850	forward	246	246	99.54	100
VP1054	39,724	40,779	39,723	40,778	forward	1056	1056	99.72	99.71
ORF48	40,899	41,105	40,898	41,104	forward	207	207	100	100
ORF49	41,106	41,300	41,105	41,299	forward	195	195	99.49	98.43
ORF50	41,580	42,071	41,579	42,070	forward	492	492	100	100
ORF51	42,150	42,617	42,149	42,616	reverse	468	468	99.79	99.35
ORF52	42,629	42,895	42,628	42,894	reverse	267	267	100	100
FP	43,107	43,817	43,106	43,816	reverse	711	711	100	100
ORF54	43,890	44,117	43,889	44,116	forward	228	228	100	100
hypothetical protein	44,146	44,244	44,145	44,243	reverse	99	99	100	100
***lef*-9**	44,243	45,802	44,242	45,801	forward	1560	1560	100	100
Cathepsin	45,886	46,989	45,885	46,988	reverse	1104	1104	100	100
ORF57	47,030	47,635	47,029	47,634	reverse	606	606	100	100
GP37	47,688	48,527	47,687	48,526	reverse	840	840	99.45	-
Hr2	47,690	50,066	47,689	50,065	forward	2377	2377	100	100
BRO-A	49,989	50,702	49,988	50,701	forward	714	714	89.78	94.78
BRO-B	50,780	51,871	50,779	51,870	forward	1092	1092	96.41	99.70
Hr3	51,872	52,353	51,871	52,352	forward	482	482	94.61	-
ORF61	52,354	52,533	52,353	52,532	reverse	180	180	99.44	100
HE56	52,582	53,310	52,581	53,309	forward	729	729	100	100
IAP-2	53,387	54,139	53,386	54,138	reverse	753	753	99.20	100
ORF64	54,187	55,032	54,186	55,031	reverse	846	846	99.64	100
ORF64	54,980	55,381	54,979	55,380	reverse	402	402	99.00	78.57
***lef*-3**	55,392	56,540	55,391	56,539	forward	1149	1149	99.39	100
ORF67	56,647	59,004	56,646	59,003	reverse	2358	2358	99.96	100
DNA polymerase	59,035	62,097	59,034	62,096	forward	3063	3063	99.74	99.90
ORF69	62,174	62,647	62,173	62,646	reverse	474	474	99.78	100
ORF70	62,698	63,090	62,697	63,089	reverse	393	393	100	100
ORF71	63,087	63,344	63,086	63,343	reverse	258	258	100	100
VLF-1	63,385	64,629	63,384	64,628	reverse	1245	1245	99.92	100
ORF73	64,642	64,986	64,641	64,985	reverse	345	345	100	100

GP41	65,043	66,011	65,042	66,010	reverse	969	969	100	100
ORF75	65,941	66,705	65,940	66,704	reverse	765	765	100	100
ORF76	66,539	67,216	66,538	67,215	reverse	678	678	100	100
VP91	67,146	69,596	67,145	69,595	forward	2451	2451	99.76	99.62
ORF78	69,599	69,775	69,598	69,774	reverse	177	177	99.44	100
CG30	69,741	70,655	69,740	70,654	reverse	915	915	100	100
VP39	70,681	71,562	70,680	71,561	reverse	882	882	100	100
***lef*-4**	71,519	72,946	71,518	72,945	forward	1428	1428	99.93	100
ORF82	72,999	73,763	72,998	73,762	reverse	765	765	99.87	100
ORF83	73,723	74,253	73,722	74,252	forward	531	531	99.59	100
ODV-E25	74,299	74,991	74,298	74,990	forward	693	693	99.86	100
ORF85	75,023	75,520	75,022	75,519	reverse	498	498	99.39	99.39
helicase	75,539	79,300	75,538	79,299	reverse	3762	3762	99.89	99.84
ORF87	79,257	79,778	79,256	79,777	forward	522	522	100	100
ORF88	79,837	80,922	79,836	80,858	reverse	1086	1023	99.59	100
***lef*-5**	80,698	81,645	80,697	81,644	forward	948	948	99.89	100
P6.9	81,639	81,968	81,638	81,967	reverse	330	330	99.89	100
ORF91	82,033	83,142	82,032	83,141	reverse	1110	1110	100	100
ORF92	83,188	83,556	83,187	83,555	reverse	369	369	100	100
ORF93	83,556	84,689	83,555	84,688	reverse	1134	1134	100	100
VP80	84,784	86,601	84,783	86,600	forward	1818	1818	99.67	99.83
ORF95	86,598	86,774	86,597	86,773	forward	177	177	100	100
ORF96	86,789	87,874	86,788	87,873	forward	1086	1086	99.91	100
ORF97	87,919	88,203	87,918	88,202	forward	285	285	99.30	100
ODV-E66	88,270	90,288	88,269	90,287	reverse	2019	2019	99.95	100
ORF99	90,309	91,139	90,308	91,138	reverse	831	831	100	100
Hr4	91,140	93,316	91,139	93,315	forward	2177	2177	98.44	-
ORF100	93,317	93,916	93,316	93,915	forward	600	600	99.83	100
ORF101	93,920	94,276	93,919	94,275	forward	357	357	99.72	100
ORF102	94,371	95,897	94,370	95,896	forward	1527	1527	99.61	100
ORF103	95,976	96,737	95,975	96,736	forward	762	762	99.34	99.20
ORF104	96,752	97,084	96,751	97,083	forward	333	333	99.4	98.18
ORF105	97,143	97,949	97,142	97,948	reverse	807	807	100	100
ORF106	97,946	98,239	97,945	98,238	reverse	294	294	100	100
BRO-C	98,205	99,710	98,204	99,709	reverse	1506	1506	100	100
SOD	99,878	100,357	99,877	100,356	forward	480	480	99.58	100
ORF109	100,364	101,737	100,363	101,736	forward	1374	1374	99.49	99.78
ORF110	101,790	102,368	101,789	102,367	reverse	579	579	99.83	100
ORF111	102,489	102,884	102,488	102,883	forward	396	396	100	100
ORF112	102,862	103,161	102,861	103,160	forward	300	300	99.67	100
ORF113	103,229	104,815	103,228	104,814	forward	1587	1587	99.94	99.80
ORF114	104,812	105,048	104,811	105,047	forward	237	237	100	100
FGF	105,071	105,976	105,070	105,975	reverse	906	906	99.89	100

ALK-EXO	106,103	107,389	106,102	107,388	reverse	1287	1287	99.84	100
ORF117	107,409	107,798	107,408	107,797	reverse	390	390	99.49	99.22
Hr5	107,803	109,187	107,802	109,186	forward	1385	1385	97.04	-
ORF118	109,188	110,114	109,187	110,113	reverse	927	927	99.89	100
ORF119	110,315	110,530	110,314	110,529	forward	216	216	100	100
***lef*-2**	110,646	111,362	110,645	111,361	reverse	717	717	100	100
P24	111,724	112,470	111,723	112,469	forward	747	747	100	100
GP19	112,532	112,816	112,531	112,815	forward	285	285	100	100
CALYX/PEP	112,868	113,890	112,867	113,889	forward	1023	1023	99.51	100
ORF124	113,942	114,433	113,941	114,432	forward	492	492	100	100
ORF125	114,564	115,154	114,563	115,153	forward	591	591	100	100
38.7K protein	115,198	116,376	115,197	116,375	reverse	1179	1179	99.58	100
***lef*-1**	116,378	117,115	116,377	117,114	reverse	738	738	100	100
ORF128	117,090	117,524	117,089	117,523	reverse	435	435	98.85	99.30
EGT	117,669	119,216	117,668	119,215	forward	1548	1548	99.61	99.80
ORF130	119,374	119,994	119,373	119,993	forward	621	621	100	100
ORF131	119,945	120,745	119,944	120,744	forward	801	801	99.88	100
ORF132	120,828	123,671	120,827	123,670	reverse	2844	2844	99.65	99.78
PKIP-1	124,012	124,521	124,011	124,520	forward	510	510	99.41	100
ARIF-1	124,588	125,385	124,587	125,384	reverse	798	798	100	100
ORF135	125,647	126,798	125,646	126,797	forward	1152	1152	99.39	99.17
ORF136	126,839	128,872	127,269	128,870	reverse	2034	1602	99.31	78.52
ORF137	129,014	129,556	129,012	129,554	reverse	543	543	99.26	98.88
ORF138	129,749	130,336	129,747	130,334	forward	588	588	100	100

Table S2. Nucleotide distance matrix of AC53 and its derived strains. All derived strains when compared to each other have between 99.82% and 99.99% sequence identity.

Genome	HaSNPV-AC53	HaSNPV-AC53-C1	HaSNPV-AC53-C5	HaSNPV-AC53-C6	HaSNPV-AC53-T4.1	HaSNPV-AC53-T5	HaSNPV-AC53-C9	HaSNPV-AC53-T2	HaSNPV-AC53-T4.2	HaSNPV-AC53-C3
HaSNPV-AC53	-	99.624	99.600	99.601	99.602	99.603	99.599	99.596	99.530	99.595
HaSNPV-AC53-C1	99.624	-	99.929	99.922	99.926	99.926	99.925	99.921	99.856	99.877
HaSNPV-AC53-C5	99.600	99.929	-	99.986	99.989	99.990	99.989	99.988	99.922	99.947
HaSNPV-AC53-C6	99.601	99.922	99.986	-	99.995	99.993	99.995	99.982	99.922	99.946
HaSNPV-AC53-T4.1	99.602	99.926	99.989	99.995	-	99.997	99.993	99.985	99.921	99.945
HaSNPV-AC53-T5	99.603	99.926	99.990	99.993	99.997	-	99.992	99.985	99.920	99.945
HaSNPV-AC53-C9	99.599	99.925	99.989	99.995	99.993	99.992	-	99.985	99.925	99.949
HaSNPV-AC53-T2	99.596	99.921	99.988	99.982	99.985	99.985	99.985	-	99.920	99.941
HaSNPV-AC53-T4.2	99.530	99.856	99.922	99.922	99.921	99.920	99.925	99.920	-	99.878
HaSNPV-AC53-C3	99.595	99.877	99.947	99.946	99.945	99.945	99.949	99.941	99.878	-

Table S3. *Lef-8* analysed strains.

Strain	Accession No.	Country of Origin
HaSNPV-G4	AF271059	China
HaSNPV-C1	AF303045	China
HzSNPV-F16	AF334030	USA
HaSNPV NNg1	AP010907	Kenya
HaSNPV-South Africa	AY118080	South Africa
Busseola fusca NPV isolate A2-4	AY519223	Unknown
HzSNPV-Gemstar-35022	HQ246097	USA
HaSNPV-75	HQ246098	Sudan
HaSNPV-126	HQ246099	India
HzSNPV-566	HQ246103	Unknown
HzSNPV-668	HQ246104	Unknown
HzSNPV-1013	HQ246105	Unknown
HzSNPV-1073	HQ246108	China
HaSNPV-1115	HQ246110	India
HzSNPV-1180	HQ246111	Unknown
HaSNPV-1186	HQ246112	South Africa
HaSNPV-1240	HQ246114	India
HaSNPV-1623	HQ246116	India
HzSNPV-3010	HQ246121	China
HaSNPV-3104	HQ246122	Unknown
HzSNPV-3108	HQ246123	Unknown
HaSNPV-AU	JN584482	Australia—Sequenced in China
HzSNPV-HS-18	KJ004000	Unknown—Sequenced in Russia
HaSNPV-LB1	KJ701029	Iberian
HaSNPV-LB3	KJ701030	Iberian
HaSNPV-LB6	KJ701031	Iberian
HaSNPV-SP1A	KJ701032	Iberian
HaSNPV-SP1B	KJ701033	Iberian
HaSNPV-AC53	KJ909666	Australia
HaSNPV-H25EA1	KJ922128	Australia
HaSNPV-Faridkot	KM357512	India
HzSNPV-Br/South	KM596835	Brazil
Helicoverpa gelotopoeon SNPV (HgSNPV)	KP340515	Argentina
HaSNPV-L1	KT013224	India
HaSNPV AC53-AC53-C1	KU738896	Australia
HaSNPV AC53-AC53-C3	KU738897	Australia
HaSNPV AC53-AC53-C5	KU738898	Australia
HaSNPV AC53-AC53-C6	KU738899	Australia
HaSNPV AC53-AC53-C9	KU738900	Australia
HaSNPV AC53-T2	KU738901	Australia
HaSNPV AC53-T4.1	KU738902	Australia
HaSNPV AC53-T4.2	KU738903	Australia
HaSNPV AC53-T5	KU738904	Australia
HzSNPV-Elcar	U67265	USA

Table S4. *Lef-9* analysed strains.

Strain	Accession No.	Country of Origin
Busseola fusca NPV isolate A2-4	AY519224	Unknown
HzSNPV-543	HQ246129	Unknown
HaSNPV-G4	AF271059	China
HaSNPV-C1	AF303045	China
HaSNPV-1073	HQ246135	China
HzSNPV-F16	AF334030	USA
HaSNPV-NNg1	AP010907	Kenya

HzSNPV-Gemstar-35022	HQ246124	USA
HaSNPV-126	HQ246126	India
HaSNPV- 138	HQ246127	Poland
HaSNPV-AU	JN584482	Australia—Sequenced in China
HzSNPV-HS-18	KJ004000	Unknown—Sequenced in Russia
HaSNPV-LB1	KJ701029	Iberian
HaSNPV-LB3	KJ701030	Iberian
HaSNPV-LB6	KJ701031	Iberian
HaSNPV-SP1A	KJ701032	Iberian
HaSNPV-SP1B	KJ701033	Iberian
HaSNPV-AC53	KJ909666	Australia
HaSNPV-H25EA1	KJ922128	Australia
HzSNPV-Br/South	KM596835	Brazil
Helicoverpa gelotopoeon SNPV (HgSNPV)	KP340516	Argentina
HaSNPV-L1	KT013224	India
HaSNPV AC53-C1	KU738896	Australia
HaSNPV AC53-C3	KU738897	Australia
HaSNPV AC53-C5	KU738898	Australia
HaSNPV AC53-C6	KU738899	Australia
HaSNPV AC53-C9	KU738900	Australia
HaSNPV AC53-T2	KU738901	Australia
HaSNPV AC53-T4.1	KU738902	Australia
HaSNPV AC53-T4.2	KU738903	Australia
HaSNPV AC53-T5	KU738904	Australia
HaSNPV-1115	HQ246137	India
HaSNPV-Faridkot	KM357515	India

Table S5. *Polh* analysed strains.

Strain	Accession No.	Country of Origin
HaSNPV-RI-G	AF157012	South Africa
HaSNPV-G4	AF271059	China
HaSNPV-C1	AF303045	China
HzSNPV-F16	AF334030	USA
HaSNPV-NNg1	AP010907	Kenya
Busseola fusca SNPV isolate A2-4	AY519223	Unknown
Helicoverpa assulta NPV	DQ157735	South Korea
HaSNPV-PAU	FJ157291	India
HaSNPV-Bathinda	FJ157292	India
HaSNPV-PDBC	FJ157293	India
HaSNPV-Jodhan	FJ157294	India
HzSNPV-Gemstar-35022	HQ246070	USA
HaSNPV-75	HQ246071	Sudan
HaSNPV-138	HQ246073	Poland
HaSNPV-141	HQ246074	Poland
HzSNPV-1024	HQ246079	Unknown
HaSNPV-1073	HQ246081	China
HaSNPV-1113	HQ246082	India
HaSNPV-1186	HQ246085	South Africa
HzSNPV-1578	HQ246088	USA
HaSNPV-1625	HQ246090	China
HaSNPV-1825	HQ246091	Unknown
HaSNPV-2066	HQ246092	Unknown
HaSNPV-3010	HQ246094	China
HaSNPV-3104	HQ246095	Unknown
HaSNPV-AU	JN584482	Australia—Sequenced in China
HaSNPV-Bangalore	JQ612524	India
HaSNPV-Faridkot	KC174715	India

HzSNPV-HS-18	KJ004000	Unknown—Sequenced in Russia
HaSNPV-LB1	KJ701029	Iberian
HaSNPV-LB3	KJ701030	Iberian
HaSNPV-LB6	KJ701031	Iberian
HaSNPV-SP1A	KJ701032	Iberian
HaSNPV-SP1B	KJ701033	Iberian
HaSNPV-AC53	KJ909666	Australia
HaSNPV-H25EA1	KJ922128	Australia
HaSNPV-Ludhiana	KM268536	India
HaSNPV-Faridkot	KM357499	India
HzSNPV-Br/South	KM596835	Brazil
Helicoverpa gelotopoeon SNPV (HgSNPV)	KP340517	Argentina
HaSNPV-L1	KT013224	India
HaSNPV AC53-C1	KU738896	Australia
HaSNPV AC53-C3	KU738897	Australia
HaSNPV AC53-C5	KU738898	Australia
HaSNPV AC53-C6	KU738899	Australia
HaSNPV AC53-C9	KU738900	Australia
HaSNPV AC53-T2	KU738901	Australia
HaSNPV AC53-T4.1	KU738902	Australia
HaSNPV AC53-T4.2	KU738903	Australia
HaSNPV AC53-T5	KU738904	Australia
HaSNPV-Palmpur	LK031772	India
HaSNPV-F29	U67255	Australia
HaSNPV-E17	U67256	Australia
HaSNPV-AE20	U67257	Australia
HzSNPV-Elcar	U67264	USA
HaSNPV-U95055	U95055	China
HaSNPV-U97657	U97657	Unknown

Table S6. BRO-A and BRO-B analysed strains.

Strain	Country of Origin	BRO-A (Accession No.)	BRO-B (Accession No.)
HzSNPV-F16	USA	AF334030	AF334030
Heliothis virescens Ascovirus 3e	Australia	EF133465	EF133465
HzSNPV-HS-18	Unknown—Sequenced in Russia	KJ004000	KJ004000
HaSNPV-AC53	Australia	KJ909666	KJ909666
HaSNPV-H25EA1	Australia	KJ922128	KJ922128
HzSNPV-Br/South	Brazil	KM596835	KM596835
HaSNPV AC53-C1	Australia	KU738896	KU738896
HaSNPV AC53-C3	Australia	KU738897	KU738897
HaSNPV AC53-C5	Australia	KU738898	KU738898
HaSNPV AC53-C6	Australia	KU738899	KU738899
HaSNPV AC53-C9	Australia	KU738900	KU738900
HaSNPV AC53-T2	Australia	KU738901	KU738901
HaSNPV AC53-T4.1	Australia	KU738902	KU738902
HaSNPV AC53-T4.2	Australia	KU738903	KU738903
HaSNPV AC53-T5	Australia	KU738904	KU738904
HaSNPV-G4	China	NA	AF303045
HaSNPV-C1	China	NA	AF271059
HaSNPV-NNg1	Kenya	NA	AP010907
HaSNPV-LB1	Iberian Peninsula	NA	KJ701029
HaSNPV-LB3	Iberian Peninsula	NA	KJ701030
HaSNPV-LB6	Iberian Peninsula	NA	KJ701031

NA, not applicable.

Table S7. ORF42, ORF61 and ORF78 analysed strains.

Strain	Country of Origin	ORF42 (Accession No.)	ORF61 (Accession No.)	ORF78 (Accession No.)
HaSNPV-AC53	Australia	KJ909666	KJ909666	KJ909666
HaSNPV AC53-C1	Australia	KU738896	KU738896	KU738896
HaSNPV AC53-C3	Australia	KU738897	KU738897	KU738897
HaSNPV AC53-C5	Australia	KU738898	KU738898	KU738898
HaSNPV AC53-C6	Australia	KU738899	KU738899	KU738899
HaSNPV AC53-C9	Australia	KU738900	KU738900	KU738900
HaSNPV AC53-T2	Australia	KU738901	KU738901	KU738901
HaSNPV AC53-T4.1	Australia	KU738902	KU738902	KU738902
HaSNPV AC53-T4.2	Australia	KU738903	KU738903	KU738903
HaSNPV AC53-T5	Australia	KU738904	KU738904	KU738904
HzSNPV-F16	USA	AF334030	AF334030	AF334030
HzSNPV-HS-18	Unknown—Sequenced in Russia	KJ004000	KJ004000	KJ004000
HaSNPV-H25EA1	Australia	KJ922128	KJ922128	KJ922128
HzSNPV-Br/South	Brazil	KM596835	KM596835	KM596835
HaSNPV-G4	China	AF303045	AF303045	AF303045
HaSNPV-C1	China	AF271059	AF271059	AF271059
HaSNPV-NNg1	Kenya	AP010907	AP010907	AP010907
HaSNPV-LB1	Iberian Peninsula	KJ701029	KJ701029	KJ701029
HaSNPV-LB3	Iberian Peninsula	KJ701030	KJ701030	KJ701030
HaSNPV-LB6	Iberian Peninsula	KJ701031	KJ701031	KJ701031
HaSNPV-AU	Australia—Sequenced in China	JN584482	JN584482	JN584482
HaSNPV-SP1A	Iberian Peninsula	KJ701032	KJ701032	KJ701032
HaSNPV-SP1B	Iberian Peninsula	KJ701033	KJ701033	KJ701033
HaSNPV-Faridkot	India	KM357465	N.A	NA
HaSNPV-1186	South Africa	NA	HQ246054	NA
HaSNPV-1073	China	NA	HQ246052	NA
HaSNPV-3010	China	NA	HQ246056	NA
HaSNPV-126	India	NA	HQ246051	NA
HaSNPV-L1	India	KT013224	KT013224	NA
HaSNPV-75	Sudan	NA	HQ246050	NA
HaSNPV-1219	India	NA	HQ246055	NA
HaSNPV-1113	India	NA	HQ246053	NA
HzSNPV-Gemstar 35022	USA	NA	HQ246048	NA
HzSNPV-Gemstar 35036	USA	NA	HQ246049	NA
HaSNPV-3104	Unknown	NA	HQ246057	NA
HgSNPV	Argentina	NA	KP340518	NA

NA, not applicable.

12.2 METAGAAP: A NOVEL PIPELINE TO ESTIMATE COMMUNITY COMPOSITION AND ABUNDANCE FROM NON-MODEL SEQUENCE DATA

Biology **2017**, *5*, 14

Supplementary Materials: MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data

Christopher Noune, Caroline Hauxwell

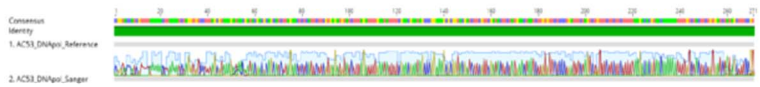


Figure S1. Comparison of the AC53 DNA polymerase Sanger sequence and the AC53 DNA polymerase reference sequence showing 100% nucleotide identity and no polymorphisms identified.

Table S1. Polymorphisms detected within ORFs. BRO-A has the highest number of polymorphisms (30) and HOAR and P74 have the second highest (13).

ORF	Polymorphisms
Exons and Intergenic Regions	45
BRO-A	30
Hr 4	21
Hr 5	19
Hr 2	16
HOAR	13
P74	13
Hr 1	12
Helicase	9
ODV-E66	9
Lef-8	8
Cathepsin	7
ORF82	7
ORF91	7
ORF105	7
Chitinase	6
ORF132	6
VP80	6
DNA polymerase	5
ORF93	5
ORF102	5
P49	5
VP39	5
EGT	4
IE-1	4
Lef-4	4
ORF64	4
ORF67	4
ORF88	4

ORF136	4
ORF137	4
ORF138	4
P26	4
Hr3	4
ALK-EXO	3
BRO-B	3
HE56	3
IAP-2	3
lef-3	3
ME53	3
ODV-EC27	3
ORF6	3
ORF13	3
ORF25	3
ORF33	3
ORF44	3
ORF76	3
ORF96	3
ORF125	3
P6.9	3
P47	3
VP91	3
FP	2
GP41	2
Hypothetical ORF	2
Lef-1	2
ODV-E56	2
ORF2	2
ORF5	2
ORF26	2
ORF27	2
ORF34	2
ORF71	2
ORF73	2
ORF75	2
ORF99	2
ORF124	2
PKIP-1	2
Polyhedrin	2
VP1054	2
38.7K protein	1
39K/PP31	1

BRO-C	1
CALYX/PEP	1
CG30	1
DBP1	1
FGF	1
GP19	1
GP37	1
IE-0	1
Lef-5	1
Lef-9	1
ODV-E25	1
ORF18	1
ORF29	1
ORF36	1
ORF37	1
ORF40	1
ORF50	1
ORF52	1
ORF61	1
ORF78	1
ORF85	1
ORF92	1
ORF97	1
ORF100	1
ORF103	1
ORF104	1
ORF106	1
ORF109	1
ORF110	1
ORF112	1
ORF118	1
ORF135	1
P10	1
PK1	1
Ubiquitin	1
VLF-1	1
ORF113	0
ORF131	0
ARIF-1	0
P24	0
Lef-2	0
ORF57	0
ORF39	0

ORF130	0
ORF30	0
Lef-6	0
ORF42	0
ORF87	0
ORF83	0
SOD	0
ORF51	0
ORF69	0
ORF128	0
ORF43	0
ORF117	0
Lef-11	0
ORF70	0
ORF101	0
ORF111	0
ORF12	0
ORF114	0
ORF45	0
Lef-10	0
ORF119	0
ORF48	0
ORF22	0
ORF49	0
ORF54	0
ODV-E18	0
ORF95	0
ORF17	0
ORF7	0

12.3 TIME-COURSE ANALYSIS OF BRO-A DURING THE HASNPV-AC53 INFECTION CYCLE

Table 12-1: Identified amino-acid genotypes encoding either a predicted functional BRO-A protein, or a predicted non-functional protein caused by a truncation of the BRO-A ORF.

Amino-acid Genotype	Functional/Non-Functional
A.A_1	Functional
A.A_2	Functional
A.A_3	Functional
A.A_4	Functional
A.A_5	Functional
A.A_6	Functional
A.A_7	Functional
A.A_8	Functional
A.A_9	Functional
A.A_12	Functional
A.A_13	Functional
A.A_26	Functional
A.A_37	Functional
A.A_39	Functional
A.A_40	Functional
A.A_45	Functional
A.A_58	Functional
A.A_61	Functional
A.A_67	Functional
A.A_81	Functional
A.A_83	Functional
A.A_88	Functional
A.A_91	Functional
A.A_92	Functional
A.A_95	Functional
A.A_103	Functional
A.A_106	Functional
A.A_107	Functional
A.A_108	Functional
A.A_86	Non-Functional
A.A_10	Non-Functional
A.A_11	Non-Functional
A.A_14	Non-Functional
A.A_15	Non-Functional
A.A_16	Non-Functional
A.A_17	Non-Functional
A.A_18	Non-Functional
A.A_19	Non-Functional

A.A_20	Non-Functional
A.A_21	Non-Functional
A.A_22	Non-Functional
A.A_24	Non-Functional
A.A_25	Non-Functional
A.A_27	Non-Functional
A.A_28	Non-Functional
A.A_29	Non-Functional
A.A_30	Non-Functional
A.A_31	Non-Functional
A.A_32	Non-Functional
A.A_33	Non-Functional
A.A_34	Non-Functional
A.A_35	Non-Functional
A.A_36	Non-Functional
A.A_38	Non-Functional
A.A_41	Non-Functional
A.A_42	Non-Functional
A.A_43	Non-Functional
A.A_44	Non-Functional
A.A_46	Non-Functional
A.A_47	Non-Functional
A.A_48	Non-Functional
A.A_49	Non-Functional
A.A_50	Non-Functional
A.A_51	Non-Functional
A.A_52	Non-Functional
A.A_53	Non-Functional
A.A_54	Non-Functional
A.A_55	Non-Functional
A.A_56	Non-Functional
A.A_57	Non-Functional
A.A_59	Non-Functional
A.A_60	Non-Functional
A.A_62	Non-Functional
A.A_63	Non-Functional
A.A_64	Non-Functional
A.A_65	Non-Functional
A.A_66	Non-Functional
A.A_68	Non-Functional
A.A_69	Non-Functional
A.A_70	Non-Functional
A.A_71	Non-Functional
A.A_72	Non-Functional

A.A_73	Non-Functional
A.A_74	Non-Functional
A.A_75	Non-Functional
A.A_76	Non-Functional
A.A_77	Non-Functional
A.A_78	Non-Functional
A.A_79	Non-Functional
A.A_80	Non-Functional
A.A_82	Non-Functional
A.A_84	Non-Functional
A.A_85	Non-Functional
A.A_87	Non-Functional
A.A_89	Non-Functional
A.A_90	Non-Functional
A.A_93	Non-Functional
A.A_94	Non-Functional
A.A_96	Non-Functional
A.A_97	Non-Functional
A.A_98	Non-Functional
A.A_99	Non-Functional
A.A_100	Non-Functional
A.A_101	Non-Functional
A.A_102	Non-Functional
A.A_104	Non-Functional
A.A_105	Non-Functional

Table 12-2: Model output for read count during the infection cycle. Results are indicating a significant increase in read count over time, or as a proxy for virus titre, a significant increase in virus titre over time.

Parameters	**Estimate**	**Standard Error**	**t value (Wald test (Wald, 1943))**	**Pr(>|t|)**	**Dispersion**
β_0	9.731	0.143	68.181	$<2x10^{-16}$	3626.227
β_1	0.008	0.001	6.205	$4.08x10^{-8}$	
β_2	4.310	0.144	29.883	$<2x10^{-16}$	
β_3	-0.009	0.001	-6.298	$2.81x10^{-8}$	

Table 12-3: Model summary for both the nucleotide and amino-acid genotypes within the BV datasets. Both dominant genotypes (G_33554431 and A.A_1) had a significant, non-linear reduction in relative abundance. A significant, non-linear increase in the abundance of reads of the minor genotypes G_33554303, G_33552383 and G_33554423 was observed with the exception of G_16777215 for which no significant change in abundance was observed. The amino-acid genotypes A.A_2, A.A_3 and A.A_4 were found to have non-significant results, with the exception of A.A_8 for which a significant, non-linear increase in abundance was observed.

Nucleotide/ Amino-Acid	Genotype	Parameters	Estimate	Standard Error	t value (Wald test (Wald, 1943))	Pr(>|t|)	Dispersion
Nucleotide	G_33554431	β_0	9.6842	0.1447	66.917	$<2x10^{-16}$	3553.156
		β_1	0.0083	0.0014	6.108	$6x10^{-8}$	
		β_2	-3.0280	0.6682	-4.532	$2.52x10^{-5}$	
		β_3	0.0003	0.0063	0.049	0.961	
	G_33554303	β_0	4.7479	1.6547	2.869	0.0055	3550.89
		β_1	0.0095	0.0154	0.616	0.54	
		β_2	4.9768	1.6608	2.997	0.0038	
		β_3	-0.0011	0.0154	-0.074	0.9416	
	G_33552383	β_0	4.7289	1.7051	2.773	0.0072	3549.646
		β_1	0.0087	0.0160	0.545	0.5875	
		β_2	4.9958	1.7110	2.920	0.0048	
		β_3	-0.0004	0.0161	-0.023	0.9814	
	G_16777215	β_0	3.4831	3.2487	1.072	0.2876	3549.524
		β_1	0.0079	0.0308	0.257	0.7978	
		β_2	6.2464	3.2518	1.921	0.0591	
		β_3	0.0004	0.0308	0.014	0.9890	
	G_33554423	β_0	4.3524	2.0713	2.101	0.0394	3552.083
		β_1	0.0085	0.0195	0.436	0.6640	
		β_2	5.3744	2.0761	2.589	0.0118	
		β_3	-0.0002	0.0195	-0.008	0.9937	
Amino-Acid	A.A_1	β_0	9.9683	0.1429	67.838	$<2x10^{-16}$	3515.797
		β_1	0.0083	0.0013	6.178	$4.54x10^{-8}$	
		β_2	-3.7733	0.9377	-4.024	0.00015	
		β_3	0.0006	0.0088	0.073	0.9417	
	A.A_2	β_0	1.9369	6.6275	0.292	0.771	3488.318
		β_1	0.0098	0.0614	0.160	0.874	
		β_2	7.7770	6.6290	1.173	0.245	
		β_3	-0.0014	0.0614	-0.024	0.981	
	A.A_3	β_0	-0.5340	23.6586	-0.0230	0.9820	3511.363
		β_1	0.0086	0.2224	0.0390	0.9690	
		β_2	10.2549	23.6590	0.4330	0.6660	
		β_3	-0.0002	0.2225	-0.0010	0.9990	
	A.A_4	β_0	3.3436	3.5910	0.9310	0.3552	3511.704
		β_1	0.0066	0.0346	0.1900	0.8496	
		β_2	6.3757	3.5938	1.7740	0.0807	
		β_3	0.0018	0.0347	0.0510	0.9598	
	A.A_8	β_0	4.7649	1.6691	2.8550	0.0058	3512.215
		β_1	0.0087	0.0157	0.5520	0.5830	
		β_2	4.9490	1.6752	2.9540	0.0043	
		β_3	-0.0003	0.0157	-0.0200	0.9845	

Table 12-4: Model summary for both the nucleotide and amino-acid genotypes within the BV datasets. Both dominant genotypes (G_33554431 and A.A_1) had a significant, linear decrease in relative abundance. However, all minor amino-acid genotypes and three of the four minor nucleotide genotypes had significant, linear increases in relative abundance except for G_16777215 which was found to have no significant changes in abundance.

Nucleotide/ Amino-Acid	Genotype	Parameters	Estimate	Standard Error	t value (Wald test (Wald, 1943))	Pr(>|t|)	Dispersion
Nucleotide	G_33554431	δ_0	12.46	0.1098	113.534	3.61×10^{-8}	3108.962
		δ_1	0.0028	0.0034	0.823	0.4570	
		δ_2	-3.3×10^{-5}	2.3×10^{-5}	-1.448	0.2211	
		δ_3	-3.498	0.6404	-5.462	0.0055	
		δ_4	0.0363	0.0129	2.82	0.0479	
		δ_5	-0.0002	6.86×10^{-5}	-2.307	0.08234	
	G_33554303	δ_0	12.48	0.1053	118.525	3.04×10^{-8}	3550.89
		δ_1	0.0059	0.0031	1.9180	0.1276	
		δ_2	-4.73×10^{-5}	2.05×10^{-5}	-2.3040	0.0826	
		δ_3	-5.0800	1.3400	-3.7910	0.0192	
		δ_4	0.0167	0.0310	0.5380	0.6188	
		δ_5	-0.0001	0.0002	-0.4240	0.6936	
	G_33552383	δ_0	12.49	0.1053	118.585	3.03×10^{-8}	3549.646
		δ_1	0.0059	0.0031	1.9300	0.1258	
		δ_2	-4.75×10^{-5}	2.05×10^{-5}	-2.3160	0.0815	
		δ_3	-5.8200	1.9360	-3.0070	0.0397	
		δ_4	0.0225	0.0415	0.5420	0.6164	
		δ_5	-0.0001	0.0002	-0.4180	0.6971	
	G_16777215	δ_0	12.49	0.1050	118.91	3.0×10^{-8}	3549.524
		δ_1	0.0060	0.0030	1.981	0.1187	
		δ_2	-4.81×10^{-5}	2.40×10^{-5}	-2.356	0.0780	
		δ_3	-5.987	2.098	-2.853	0.0463	
		δ_4	0.0037	0.05857	0.063	0.9526	
		δ_5	-2.38×10^{-5}	3.904×10^{-4}	-0.061	0.9544	
	G_33554423	δ_0	12.49	0.1052	118.722	3.02×10^{-8}	3552.083
		δ_1	0.0059	0.0031	1.9340	0.1253	
		δ_2	-4.76×10^{-5}	2.05×10^{-5}	-2.3210	0.0810	
		δ_3	-6.2100	2.3500	-2.6430	0.0574	
		δ_4	0.0271	0.0480	0.5650	0.6021	
		δ_5	-0.0001	0.0003	-0.4400	0.6825	
Amino-Acid	A.A_1	δ_0	12.4400	0.1526	81.5490	5.25×10^{-9}	5974.11
		δ_1	0.0056	0.0044	1.2850	0.2552	
		δ_2	-0.0001	2.96×10^{-5}	-1.7600	0.1388	
		δ_3	-3.1750	0.6388	-4.9700	0.0042	
		δ_4	0.0107	0.0054	1.9880	0.1036	
		δ_5	12.4400	0.1526	81.5490	5.25×10^{-9}	
	A.A_2	δ_0	6.8160	1.4000	4.8670	0.0046	2716.122
		δ_1	0.0167	0.0120	1.3860	0.2243	
		δ_2	4.79×10^{-5}	1.98×10^{-5}	-2.4200	0.0601	
		δ_3	5.6620	1.4020	4.0400	0.0099	
		δ_4	-0.0108	0.0116	-0.9300	0.3950	
		δ_5	6.8160	1.4000	4.8670	0.0046	
	A.A_3	δ_0	6.1540	1.8800	3.2730	0.0221	2553.557
		δ_1	0.0169	0.0158	1.0660	0.3352	
		δ_2	-4.76×10^{-5}	1.92×10^{-5}	-2.4840	0.0556	
		δ_3	6.3280	1.8810	3.3640	0.0200	
		δ_4	-0.0110	0.0155	-0.7080	0.5104	
		δ_5	6.1540	1.8800	3.2730	0.0221	
	A.A_4	δ_0	6.3890	1.6580	3.8540	0.0120	2387.159
		δ_1	0.0156	0.0142	1.1020	0.3207	

		δ_2	-4.87×10^{-5}	1.85×10^{-5}	-2.6320	0.0464	
		δ_3	6.0940	1.6590	3.6740	0.0144	
		δ_4	-0.0096	0.0138	-0.6920	0.5198	
		δ_5	6.3890	1.6580	3.8540	0.0120	
	A.A_8	δ_0	7.2930	1.1220	6.4990	0.0013	2547.059
		δ_1	0.0137	0.0100	1.3730	0.2281	
		δ_2	-4.77×10^{-5}	1.91×10^{-5}	-2.490	0.0552	
		δ_3	5.1850	1.1240	4.6130	0.0058	
		δ_4	-0.0079	0.0095	-0.8240	0.4476	
		δ_5	7.2930	1.1220	6.4990	0.0013	

Table 12-5: Model output for the presence-absence data indicating a significant, linear increase in present genotypes over the course of the infection within the host.

Nucleotide/Amino-acid	**Parameters**	**Estimate**	**Standard Error**	**t value (Wald test (Wald, 1943))**	**Pr(>|t|)**	**Dispersion**
Nucleotide	β_0	3.7453	0.0932	40.176	$<2\times10^{-16}$	3.19
	β_1	0.0046	0.0009	5.0320	3.99×10^{-6}	
	β_2	1.7801	0.1033	17.226	$<2\times10^{-16}$	
	β_3	-0.0059	0.0010	-5.704	2.98×10^{-7}	
Amino-Acid	β_0	3.1598	0.1048	30.1640	$<2\times10^{-16}$	2.23
	β_1	0.0045	0.0010	4.3300	5.18×10^{-5}	
	β_2	1.3068	0.1231	10.6180	6.42×10^{-16}	
	β_3	-0.0066	0.0013	-5.2740	1.59×10^{-6}	

Table 12-6: Frequencies of present-absent nucleotide genotypes in each analysed BRO-A dataset.

Dataset	**Presence-Absence**	**Frequencies**	**Proportions (%)**
Inoculum	Present	289	100
BV_24.1	Absent	272	94.12
	Present	17	5.88
BV_24.2	Absent	264	91.35
	Present	25	8.65
BV_24.3	Absent	281	97.23
	Present	8	2.77
BV_24.4	Absent	248	85.81
	Present	41	14.19
BV_24.5	Absent	217	75.09
	Present	72	24.91
BV_24.6	Absent	225	77.86
	Present	64	22.15
BV_48.1	Absent	220	76.13
	Present	69	23.88
BV_48.2	Absent	233	80.62
	Present	56	19.38
BV_48.3	Absent	276	95.50
	Present	13	4.50

BV_48.4	Absent	221	76.47
	Present	68	23.53
BV_48.5	Absent	202	69.90
	Present	87	30.10
BV_48.6	Absent	227	78.55
	Present	62	21.45
BV_72.1	Absent	224	77.51
	Present	65	22.49
BV_72.2	Absent	215	74.39
	Present	74	25.61
BV_72.3	Absent	221	76.47
	Present	68	23.53
BV_72.4	Absent	221	76.47
	Present	68	23.53
BV_72.5	Absent	243	84.08
	Present	46	15.92
BV_96.1	Absent	234	80.97
	Present	55	19.03
BV_96.2	Absent	226	78.20
	Present	63	21.80
BV_96.3	Absent	220	76.13
	Present	69	23.88
BV_96.4	Absent	214	74.05
	Present	75	25.95
BV_96.5	Absent	213	73.70
	Present	76	26.30
BV_96.6	Absent	210	72.66
	Present	79	27.34
BV_120.1	Absent	215	74.39
	Present	74	25.61
BV_120.2	Absent	206	71.28
	Present	83	28.72
BV_120.3	Absent	210	2.12
	Present	79	97.88
BV_120.4	Absent	206	94.12
	Present	83	5.88
BV_120.5	Absent	221	91.35
	Present	68	8.65
BV_120.6	Absent	215	97.23
	Present	74	2.77
BV_144.1	Absent	225	85.81
	Present	64	14.19
BV_144.2	Absent	213	75.09
	Present	76	24.91

BV_144.3	Absent	212	77.86
	Present	77	22.15
BV_144.4	Absent	221	76.13
	Present	68	23.88
BV_144.5	Absent	203	80.62
	Present	86	19.38
BV_144.6	Absent	208	95.50
	Present	81	4.50
OB_96.3	Present	289	100
OB_96.4	Present	289	100
OB_120.1	Present	289	100
OB_120.2	Present	289	100

Table 12-7: Frequencies of present-absent amino-acid genotypes in each analysed BRO-A dataset.

Dataset	**Presence-Absence**	**Frequencies**	**Proportions (%)**
Inoculum	Present	100	93.458
BV_24.1	Absent	100	93.46
	Present	7	6.54
BV_24.2	Absent	93	86.92
	Present	14	13.08
BV_24.3	Absent	103	96.26
	Present	4	3.74
BV_24.4	Absent	82	76.64
	Present	25	23.36
BV_24.5	Absent	65	60.75
	Present	42	39.25
BV_24.6	Absent	74	69.16
	Present	33	30.84
BV_48.1	Absent	66	61.68
	Present	41	38.32
BV_48.2	Absent	78	72.90
	Present	29	27.10
BV_48.3	Absent	100	93.46
	Present	7	6.54
BV_48.4	Absent	68	63.55
	Present	39	36.45
BV_48.5	Absent	64	59.81
	Present	43	40.19
BV_48.6	Absent	72	67.29
	Present	35	32.71
BV_72.1	Absent	73	68.22
	Present	34	31.78
BV_72.2	Absent	63	58.88

	Present	44	41.12
BV_72.3	Absent	67	62.62
	Present	40	37.38
BV_72.4	Absent	69	64.49
	Present	38	35.51
BV_72.5	Absent	82	76.64
	Present	25	23.36
BV_96.1	Absent	80	74.77
	Present	27	25.23
BV_96.2	Absent	75	70.09
	Present	32	29.91
BV_96.3	Absent	70	65.42
	Present	37	34.58
BV_96.4	Absent	61	57.01
	Present	46	42.99
BV_96.5	Absent	63	58.88
	Present	44	41.12
BV_96.6	Absent	63	58.88
	Present	44	41.12
BV_120.1	Absent	64	59.81
	Present	43	40.19
BV_120.2	Absent	65	60.75
	Present	42	39.25
BV_120.3	Absent	67	62.62
	Present	40	37.38
BV_120.4	Absent	55	51.40
	Present	52	48.60
BV_120.5	Absent	69	64.49
	Present	38	35.51
BV_120.6	Absent	69	64.49
	Present	38	35.51
BV_144.1	Absent	71	66.36
	Present	36	33.64
BV_144.2	Absent	66	61.68
	Present	41	38.32
BV_144.3	Absent	67	62.62
	Present	40	37.38
BV_144.4	Absent	70	65.42
	Present	37	34.58
BV_144.5	Absent	63	58.88
	Present	44	41.12
BV_144.6	Absent	61	57.01
	Present	46	42.99
OB_96.3	Present	107	100

OB_96.4	Present	107	100
OB_120.1	Present	107	100
OB_120.2	Present	107	100

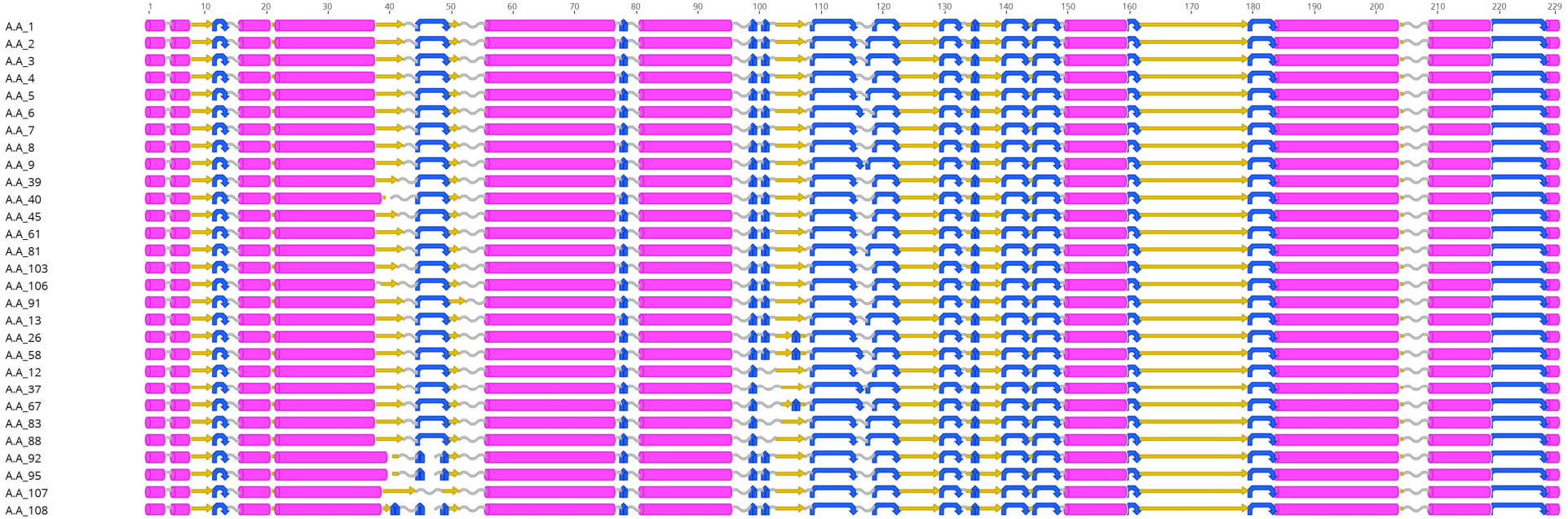


Figure 12-1: Aligned BRO-A predicted protein structures for each amino-acid genotype encoding a functional protein. Protein structures are depicted as follows: Alpha helix (purple cylinder), beta strands (yellow arrows), coils (grey wavy lines), and the turns (blue curved arrows). Major structural differences occur between positions 40 to 50, in which a single turn has been shown to split, or replaced with a beta strand, and between positions 100 to 110, in which the beta strand has split and had a turn inserted, or has lost one turn.

12.4 *IN VIVO* SELECTION & VIRULENCE-TRANSMISSION TRADE-OFFS IN HASNPV-AC53

Table 12-8: Principal Component Analysis of performance metrics from the generational selection of the fast strains.

Variable	PC1	PC2	PC3	PC4	PC5
ST_{50} (Hrs)	0.42	0.34	-0.35	0.60	-0.15
Dosage (OB/mL)	-0.30	-0.67	-0.13	0.54	-0.36
Total Deaths (Up to 72hrs)	-0.47	-0.19	-0.02	-0.14	0.21
Mean Density (OB/µg)	0.45	-0.31	0.02	-0.48	-0.63
Mean Weight (µg)	0.40	-0.29	0.75	0.28	0.30
Mean Total Yield (OB/mL)	0.38	-0.47	-0.55	-0.14	0.56

Table 12-9: Principal Component Analysis of performance metrics from the generational selection of the slow strains.

Variable	PC1	PC2	PC3	PC4	PC5
ST_{50} (Hrs)	-0.11	0.53	-0.49	0.65	0.20
Dosage (OB/mL)	-0.34	0.14	-0.67	-0.60	-0.20
Final Death Time Point (Hrs)	0.47	0.46	0.06	-0.45	0.58
Selected Density (OB/µg)	0.49	-0.31	-0.39	0.04	-0.25
Selected Weight (µg)	0.16	0.62	0.29	-0.04	-0.71
Selected Total Yield (OB/mL)	0.61	-0.09	-0.27	0.11	-0.14

Table 12-10: Principal Component Analysis of performance metrics from the generational selection of the MaxOB strains.

Variable	PC1	PC2	PC3	PC4	PC5
ST_{50} (Hrs)	-0.40	0.52	0.21	0.60	0.40
Dosage (OB/mL)	0.37	0.57	-0.56	-0.26	0.34
Mean Density (OB/µg)	-0.44	-0.21	-0.25	0.08	-0.05
Mean Weight (µg)	-0.40	-0.32	-0.66	0.05	0.24
Mean Total Yield (OB/mL)	-0.43	0.05	0.35	-0.72	0.42
Time Point (Hrs) with Maximum Density (OB/µg)	-0.41	0.51	-0.16	-0.20	-0.70

Table 12-11: Correlation statistics for the fast strains showing dosage and total deaths to be positively correlated whereas dosage is negatively correlated to all other performance metrics.

Variable	**ST_{50} (Hrs)**	**Dosage (OB/mL)**	**Total Deaths (Up to 72hrs)**	**Mean Density (OB/µg)**	**Mean Weight (µg)**	**Mean Total Yield (OB/mL)**
ST_{50} (Hrs)	1.00	-0.72	-0.97	0.63	0.59	0.54
Dosage (OB/mL)	-0.72	1.00	0.83	-0.40	-0.28	-0.17
Total Deaths (Up to 72hrs)	-0.97	0.83	1.00	-0.70	-0.62	-0.57
Mean (OB/µg)	0.63	-0.40	-0.70	1.00	0.95	0.89
Mean Weight (µg)	0.59	-0.28	-0.62	0.95	1.00	0.77
Mean Total Yield (OB/mL)	0.54	-0.17	-0.57	0.89	0.77	1.00

Table 12-12: Covariance statistics for the fast strains mirroring the result from Table 12-11.

Variable	ST_{50} (Hrs)	Dosage (OB/mL)	Total Deaths (Up to 72hrs)	Mean Density (OB/µg)	Mean Weight (µg)	Mean Total Yield (OB/mL)
ST_{50} (Hrs)	$4.92x10^{2}$	$-1.03x10^{8}$	$-3.22x10^{2}$	$1.35x10^{5}$	$1.41x10^{3}$	$7.88x10^{5}$
Dosage (OB/mL)	$-1.03x10^{8}$	$4.2x10^{13}$	$8.06x10^{7}$	$-2.51x10^{10}$	$-1.95x10^{8}$	$-7.39x10^{10}$
Total Deaths (Up to 72hrs)	$-3.22x10^{2}$	$8.06x10^{7}$	$2.23x10^{2}$	$-1.01x10^{5}$	$-9.92x10^{2}$	$-5.66x10^{5}$
Mean Density (OB/µg)	$1.35x10^{5}$	$-2.51x10^{10}$	$-1.01x10^{5}$	$9.45x10^{7}$	$9.92x10^{5}$	$5.74x10^{8}$
Mean Weight (µg)	$1.41x10^{3}$	$-1.95x10^{8}$	$-9.92x10^{2}$	$9.92x10^{5}$	$1.16x10^{4}$	$5.47x10^{6}$
Mean Total Yield (OB/mL)	$7.88x10^{5}$	$-7.39x10^{10}$	$-5.66x10^{5}$	$5.74x10^{8}$	$5.47x10^{6}$	$4.39x10^{9}$

Table 12-13: Correlation statistics for the slow strains showing that dosage is positively correlated with ST_{50}, but negatively correlated with all other metrics. This result suggests that as dosage increases, density, weight and yield decrease.

Variable	ST_{50} (Hrs)	Dosage (OB/mL)	Final Death Time Point (Hrs)	Selected Density (OB/µg)	Selected Weight (µg)	Selected Total Yield (OB/mL)
ST_{50} (Hrs)	1.00	0.89	0.30	-0.46	0.33	-0.43
Dosage (OB/mL)	0.89	1.00	-0.12	-0.26	-0.13	-0.30
Final Death Time Point (Hrs)	0.30	-0.12	1.00	-0.09	0.90	0.08
Selected Density (OB/µg)	-0.46	-0.26	-0.09	1.00	-0.46	0.98
Selected Weight (µg)	0.33	-0.13	0.90	-0.46	1.00	-0.29
Selected Total Yield (OB/mL)	-0.43	-0.30	0.08	0.98	-0.29	1.00

Table 12-14: Covariance statistics for the slow strains mirroring the result from the Table 12-13.

Variable	ST_{50} (Hrs)	Dosage (OB/mL)	Final Death Time Point (Hrs)	Selected Density (OB/µg)	Selected Weight (µg)	Selected Total Yield (OB/mL)
ST_{50} (Hrs)	$5.94x10^{1}$	$4.17x10^{8}$	$6.02x10^{1}$	$-1.23x10^{6}$	$7.88x10^{2}$	$-3.3x10^{8}$
Dosage (OB/mL)	$4.17x10^{8}$	$3.68x10^{15}$	$-1.9x10^{8}$	$-5.50x10^{12}$	$-2.4x10^{9}$	$-1.86x10^{15}$
Final Death Time Point (Hrs)	$6.02x10^{1}$	$-1.9x10^{8}$	$6.91x10^{2}$	$-7.89x10^{5}$	$7.33x10^{3}$	$2.15x10^{8}$
Selected Density (OB/µg)	$-1.23x10^{6}$	$-5.50x10^{12}$	$-7.89x10^{5}$	$1.23x10^{11}$	$-4.9x10^{7}$	$3.49x10^{13}$
Selected Weight (µg)	$7.88x10^{2}$	$-2.4x10^{9}$	$7.33x10^{3}$	$-4.9x10^{7}$	$9.53x10^{4}$	$-9.2x10^{9}$
Selected Total Yield (OB/mL)	$-3.3x10^{8}$	$-1.86x10^{15}$	$2.15x10^{8}$	$3.49x10^{13}$	$-9.2x10^{9}$	$1.02x10^{16}$

Table 12-15: Correlation statistics for the maxOB strains showing dosage to be the main source of variance in the data i.e. increased dosage lead to decreased density, weight, yield, time point with maximum density and ST_{50}

Variable	ST_{50} (Hrs)	Dosage (OB/mL)	Mean Density (OB/µg)	Mean Weight (µg)	Mean Total Yield OB/mL	Time Point (Hrs) with Maximum Density (OB/µg)
ST_{50} (Hrs)	1.00	-0.07	0.56	0.50	0.85	0.95
Dosage (OB/mL)	-0.07	1.00	-0.33	-0.30	-0.40	-0.04
Mean Density (OB/µg)	0.56	-0.33	1.00	1.00	0.37	0.68
Mean Weight (µg)	0.50	-0.30	1.00	1.00	0.30	0.63
Mean Total Yield OB/mL	0.85	-0.40	0.37	0.30	1.00	0.82
Time Point (Hrs) with Maximum Density (OB/µg)	0.95	-0.04	0.68	0.63	0.82	1.00

Table 12-16: Covariance statistics for the maxOB strains mirroring the results from Table 12-15.

Variable	**ST_{50} (Hrs)**	**Dosage (OB/mL)**	**Mean Density (OB/µg)**	**Mean Weight (µg)**	**Mean Total Yield OB/mL**	**Time Point (Hrs) with Maximum Density (OB/µg)**
ST_{50} (Hrs)	$1.71x10^{2}$	$-1.5x10^{9}$	$3.68x10^{5}$	$5.98x10^{3}$	$3.9x10^{8}$	$2.99x10^{2}$
Dosage (OB/mL)	$-1.5x10^{9}$	$2.66x10^{18}$	$-2.69x10^{13}$	$-4.43x10^{11}$	$-2.26x10^{16}$	$-1.4x10^{9}$
Mean Density (OB/µg)	$3.68x10^{5}$	$-2.69x10^{13}$	$2.54x10^{9}$	$4.56x10^{7}$	$6.50x10^{11}$	$8.19x10^{5}$
Mean Weight (µg)	$5.98x10^{3}$	$-4.43x10^{11}$	$4.56x10^{7}$	$8.21x10^{5}$	$9.52x10^{9}$	$1.37x10^{4}$
Mean Total Yield OB/mL	$3.9x10^{8}$	$-2.26x10^{16}$	$6.50x10^{11}$	$9.52x10^{9}$	$1.23x10^{15}$	$6.88x10^{8}$
Time Point (Hrs) with Maximum Density (OB/µg)	$2.99x10^{2}$	$-1.4x10^{9}$	$8.19x10^{5}$	$1.37x10^{4}$	$6.88x10^{8}$	$5.76x10^{2}$

Table 12-17: Standardised LC_{50} bioassay with Abbotts corrected mortality per dose.

Virus	Dose (OB/mL)	Effective Number Treated	Number Control	Observed control mortality	Corrected Mortality	Corrected Mortality (proportion Abbots)
AC53	$1.52x10^5 \pm 1\%$	40	40	0	38	0.95
	$1.44x10^5 \pm 1\%$	43	40	0	41	0.95
	$1.12x10^5 \pm 1\%$	42	40	0	39	0.92
	$8.0x10^4 \pm 1\%$	42	40	0	37	0.88
	$4.8x10^4 \pm 1\%$	42	40	0	37	0.88
	$1.6x10^4 \pm 1\%$	42	40	0	19	0.45
	$8.0x10^3 \pm 1\%$	42	40	0	18	0.42
	$1.6x10^3 \pm 1\%$	42	40	0	4	0.09
Slow	$1.52x10^5 \pm 1\%$	41	40	0	39	0.95
	$1.44x10^5 \pm 1\%$	41	40	0	38	0.93
	$1.12x10^5 \pm 1\%$	42	40	0	37	0.88
	$8.0x10^4 \pm 1\%$	42	40	0	35	0.83
	$4.8x10^4 \pm 1\%$	42	40	0	29	0.69
	$1.6x10^4 \pm 1\%$	41	40	0	24	0.58
	$8.0x10^3 \pm 1\%$	40	40	0	6	0.15
	$1.6x10^3 \pm 1\%$	41	40	0	2	0.04
MaxOB	$1.52x10^5 \pm 1\%$	42	40	0	22	0.52
	$1.44x10^5 \pm 1\%$	42	40	0	15	0.36
	$1.12x10^5 \pm 1\%$	42	40	0	12	0.28
	$8.0x10^4 \pm 1\%$	41	40	0	8	0.19
	$4.8x10^4 \pm 1\%$	42	40	0	6	0.14
	$1.6x10^4 \pm 1\%$	42	40	0	5	0.11
	$8.0x10^3 \pm 1\%$	42	40	0	0	0.00
	$1.6x10^3 \pm 1\%$	42	40	0	0	0.00
Fast	$1.52x10^5 \pm 1\%$	42	40	0	13	0.31
	$1.44x10^5 \pm 1\%$	42	40	0	12	0.29
	$1.12x10^5 \pm 1\%$	40	40	0	10	0.25
	$8.0x10^4 \pm 1\%$	42	40	0	6	0.14
	$4.8x10^4 \pm 1\%$	42	40	0	5	0.11
	$1.6x10^4 \pm 1\%$	40	40	0	4	0.10
	$8.0x10^3 \pm 1\%$	40	40	0	3	0.07
	$1.6x10^3 \pm 1\%$	42	40	0	0	0.00

Table 12-18: Standardised ST_{50} bioassay Abbotts corrected mortality at observed time-points for AC53 and the F5 selected fast strain at a dosage of $1.8x10^6$ ± 1% OB/mL.

Virus	Time (Hrs)	Effective Number Treated	Number Control	Observed Control Mortality	Corrected Mortality	Corrected Mortality (proportion Abbots)
AC53	24	42	42	0	0	0.00
	36	42	42	0	0	0.00
	48	42	42	0	1	0.02
	56	42	42	0	3	0.07
	64	42	42	0	9	0.21
	72	42	42	0	10	0.24
	80	42	42	0	23	0.55
	88	42	42	0	32	0.76
	96	42	42	0	35	0.83
	104	42	42	0	37	0.88
	112	42	42	0	41	0.98
	120	42	42	0	42	1.00
Fast	24	42	42	0	0	0.00
	36	32	42	0	6	0.19
	48	32	42	0	10	0.31
	56	32	42	0	14	0.44
	64	32	42	0	20	0.63
	72	32	42	0	23	0.72
	80	32	42	0	24	0.75
	88	32	42	0	28	0.88
	96	32	42	0	30	0.94
	104	32	42	0	31	0.97
	112	32	42	0	31	0.97
	120	32	42	0	32	1.00

Table 12-19: Standardised ST_{50} bioassay Abbotts corrected mortality at observed time-points for F5 selected strains and AC53 at a dosage of $1.52x10^5$ ± 1% OB/mL (dose 1).

Virus	Time (Hours)	Effective Number Treated	Number Control	Observed Control Mortality	Corrected Mortality	Corrected Mortality (proportion Abbots)
AC53	24	38	40	0	0	0.00
	36	38	40	0	0	0.00
	48	38	40	0	0	0.00
	56	38	40	0	0	0.00
	64	38	40	0	1	0.03
	72	38	40	0	1	0.03
	80	38	40	0	9	0.24
	88	38	40	0	17	0.45
	96	38	40	0	24	0.63
	104	38	40	0	29	0.76
	112	38	40	0	36	0.95
	120	38	40	0	36	0.95
	128	38	40	0	37	0.97
	136	38	40	0	37	0.97
	144	38	40	0	38	1.00
Slow	24	39	40	0	0	0.00
	36	39	40	0	0	0.00
	48	39	40	0	0	0.00
	56	39	40	0	0	0.00
	64	39	40	0	1	0.03
	72	39	40	0	1	0.03
	80	39	40	0	2	0.05
	88	39	40	0	6	0.15
	96	39	40	0	9	0.23
	104	39	40	0	21	0.54
	112	39	40	0	33	0.85
	120	39	40	0	36	0.92
	128	39	40	0	38	0.97
	136	39	40	0	38	0.97
	144	39	40	0	38	0.97
	152	39	40	0	39	1.00
MaxOB	24	22	40	0	0	0.00
	36	22	40	0	0	0.00
	48	22	40	0	0	0.00
	56	22	40	0	0	0.00
	64	22	40	0	2	0.09
	72	22	40	0	2	0.09
	80	22	40	0	4	0.18
	88	22	40	0	6	0.27
	96	22	40	0	6	0.27
	104	22	40	0	7	0.32
	112	22	40	0	7	0.32
	120	22	40	0	7	0.32
	128	22	40	0	11	0.50
	136	22	40	0	12	0.55

	144	22	40	0	12	0.55
	152	22	40	0	14	0.64
	160	22	40	0	15	0.68
	168	22	40	0	16	0.73
	176	22	40	0	19	0.86
	184	22	40	0	20	0.91
	192	22	40	0	20	0.91
	200	22	40	0	20	0.91
	208	22	40	0	21	0.95
	216	22	40	0	21	0.95
	224	22	40	0	21	0.95
	232	22	40	0	21	0.95
	240	22	40	0	22	1.00
Fast	24	13	40	0	0	0.00
	36	13	40	0	2	0.15
	48	13	40	0	5	0.38
	56	13	40	0	7	0.54
	64	13	40	0	7	0.54
	72	13	40	0	8	0.62
	80	13	40	0	10	0.77
	88	13	40	0	10	0.77
	96	13	40	0	10	0.77
	104	13	40	0	10	0.77
	112	13	40	0	10	0.77
	120	13	40	0	10	0.77
	128	13	40	0	10	0.77
	136	13	40	0	10	0.77
	144	13	40	0	10	0.77
	152	13	40	0	11	0.85
	160	13	40	0	11	0.85
	168	13	40	0	11	0.85
	176	13	40	0	11	0.85
	184	13	40	0	11	0.85
	192	13	40	0	11	0.85
	200	13	40	0	12	0.92
	208	13	40	0	12	0.92
	216	13	40	0	12	0.92
	224	13	40	0	12	0.92
	232	13	40	0	13	1.00

Table 12-20: Standardised ST_{50} bioassay Abbotts corrected mortality at observed time-points for F5 selected strains and AC53 at a dosage of $1.44x10^5$ ± 1% OB/mL (dose 2).

Virus	Time (Hours)	Effective Number Treated	Number Control	Observed Control Mortality	Corrected Mortality	Corrected Mortality (proportion Abbots)
AC53	24	41	40	0	0	0.00
	36	41	40	0	0	0.00
	48	41	40	0	0	0.00
	56	41	40	0	0	0.00
	64	41	40	0	0	0.00
	72	41	40	0	0	0.00
	80	41	40	0	3	0.07
	88	41	40	0	15	0.37
	96	41	40	0	19	0.46
	104	41	40	0	26	0.63
	112	41	40	0	32	0.78
	120	41	40	0	35	0.85
	128	41	40	0	38	0.93
	136	41	40	0	38	0.93
	144	41	40	0	38	0.93
	152	41	40	0	39	0.95
	160	41	40	0	39	0.95
	168	41	40	0	41	1.00
Slow	24	38	40	0	0	0.00
	36	38	40	0	0	0.00
	48	38	40	0	0	0.00
	56	38	40	0	0	0.00
	64	38	40	0	1	0.03
	72	38	40	0	1	0.03
	80	38	40	0	1	0.03
	88	38	40	0	5	0.13
	96	38	40	0	11	0.29
	104	38	40	0	18	0.47
	112	38	40	0	28	0.74
	120	38	40	0	31	0.82
	128	38	40	0	34	0.89
	136	38	40	0	34	0.89
	144	38	40	0	34	0.89
	152	38	40	0	35	0.92
	160	38	40	0	35	0.92
	168	38	40	0	36	0.95
	176	38	40	0	37	0.97
	184	38	40	0	38	1.00
MaxOB	24	14	40	0	0	0.00
	36	14	40	0	1	0.07
	48	14	40	0	1	0.07
	56	14	40	0	1	0.07
	64	14	40	0	3	0.21
	72	14	40	0	3	0.21
	80	14	40	0	4	0.29

	88	14	40	0	6	0.43
	96	14	40	0	7	0.50
	104	14	40	0	7	0.50
	112	14	40	0	7	0.50
	120	14	40	0	7	0.50
	128	14	40	0	8	0.57
	136	14	40	0	11	0.79
	144	14	40	0	11	0.79
	152	14	40	0	11	0.79
	160	14	40	0	11	0.79
	168	14	40	0	11	0.79
	176	14	40	0	12	0.86
	184	14	40	0	12	0.86
	192	14	40	0	12	0.86
	200	14	40	0	13	0.93
	208	14	40	0	13	0.93
	216	14	40	0	13	0.93
	224	14	40	0	13	0.93
	232	14	40	0	14	1.00
Fast	24	12	40	0	0	0.00
	36	12	40	0	0	0.00
	48	12	40	0	1	0.08
	56	12	40	0	5	0.42
	64	12	40	0	5	0.42
	72	12	40	0	8	0.67
	80	12	40	0	8	0.67
	88	12	40	0	8	0.67
	96	12	40	0	8	0.67
	104	12	40	0	8	0.67
	112	12	40	0	8	0.67
	120	12	40	0	8	0.67
	128	12	40	0	8	0.67
	136	12	40	0	8	0.67
	144	12	40	0	8	0.67
	152	12	40	0	8	0.67
	160	12	40	0	8	0.67
	168	12	40	0	12	1.00

Table 12-21: OB counts for all F5 selected strains and the AC53 parent strain collected during the dose 1 ST_{50} bioassay.

Strain	Time (Hrs)	Mortality	Cumulative Mortality (%)	Cumulative Total Virus (OB/mL)	Total Virus/Insect (OB/mL)	Cumulative Total Viral Density (OB/µg)	Total Viral Density/Insect (OB/µg)	Cumulative Total Viral Capacity (µg)	Total Viral Capacity/Insect (µg)
AC53	56	1	3	$1.20x10^5$	$1.20x10^5$	$6.00x10^2$	$6.00x10^2$	$2.00x10^2$	$2.00x10^2$
	80	9	24	$5.20x10^6$	$5.78x10^5$	$6.90x10^4$	$7.67x10^3$	$1.05x10^3$	$1.17x10^2$
	88	17	45	$1.12x10^7$	$6.59x10^5$	$1.57x10^5$	$9.26x10^3$	$1.75x10^3$	$1.03x10^2$
	96	24	63	$3.35x10^8$	$1.40x10^7$	$9.62x10^5$	$4.01x10^4$	$4.35x10^3$	$1.81x10^2$
	104	29	76	$4.99x10^8$	$1.72x10^7$	$1.51x10^6$	$5.20x10^4$	$5.55x10^3$	$1.91x10^2$
	112	36	95	$9.07x10^8$	$2.52x10^7$	$2.90x10^6$	$8.06x10^4$	$7.75x10^3$	$2.15x10^2$
	128	37	97	$1.02x10^9$	$2.75x10^7$	$3.12x10^6$	$8.43x10^4$	$8.25x10^3$	$2.23x10^2$
	144	38	100	$1.38x10^9$	$3.62x10^7$	$3.13x10^6$	$8.25x10^4$	$3.35x10^4$	$8.81x10^2$
Slow	56	1	3	$4.00x10^4$	$4.00x10^4$	$1.33x10^2$	$1.33x10^2$	$3.00x10^2$	$3.00x10^2$
	88	5	13	$4.16x10^6$	$8.32x10^5$	$5.11x10^4$	$1.02x10^4$	$8.50x10^2$	$1.70x10^2$
	96	11	29	$5.68x10^7$	$5.17x10^6$	$1.22x10^6$	$1.11x10^5$	$1.40x10^3$	$1.27x10^2$
	104	18	47	$1.96x10^8$	$1.09x10^7$	$2.24x10^6$	$1.24x10^5$	$2.20x10^3$	$1.22x10^2$
	112	28	74	$4.60x10^8$	$1.64x10^7$	$4.09x10^6$	$1.46x10^5$	$4.40x10^3$	$1.57x10^2$
	120	31	82	$7.80x10^8$	$2.52x10^7$	$5.10x10^6$	$1.64x10^5$	$5.40x10^3$	$1.74x10^2$
	128	34	89	$9.43x10^8$	$2.77x10^7$	$5.81x10^6$	$1.71x10^5$	$6.00x10^3$	$1.76x10^2$
	152	35	92	$1.02x10^9$	$2.92x10^7$	$5.97x10^6$	$1.71x10^5$	$6.50x10^3$	$1.86x10^2$
	168	36	95	$1.26x10^9$	$3.51x10^7$	$5.98x10^6$	$1.66x10^5$	$2.61x10^4$	$7.26x10^2$
	176	37	97	$1.53x10^9$	$4.14x10^7$	$6.00x10^6$	$1.62x10^5$	$4.50x10^4$	$1.22x10^3$
	184	38	100	$1.53x10^9$	$4.04x10^7$	$6.00x10^6$	$1.58x10^5$	$4.92x10^4$	$1.29x10^3$
MaxOB	36	1	5	$4.00x10^4$	$4.00x10^4$	$4.00x10^2$	$4.00x10^2$	$1.00x10^2$	$1.00x10^2$
	48	2	9	$3.60x10^5$	$1.80x10^5$	$6.80x10^3$	$3.40x10^3$	$1.50x10^2$	$7.50x10^1$
	64	4	18	$4.40x10^5$	$1.10x10^5$	$7.07x10^3$	$1.77x10^3$	$7.50x10^2$	$1.88x10^2$
	80	6	27	$1.44x10^6$	$2.40x10^5$	$9.72x10^3$	$1.62x10^3$	$1.50x10^3$	$2.50x10^2$
	88	7	32	$2.24x10^6$	$3.20x10^5$	$1.77x10^4$	$2.53x10^3$	$1.60x10^3$	$2.29x10^2$
	104	8	36	$8.24x10^6$	$1.03x10^6$	$4.77x10^4$	$5.97x10^3$	$1.80x10^3$	$2.25x10^2$

	128	13	59	$1.04x10^{8}$	$7.97x10^{6}$	$2.67x10^{5}$	$2.05x10^{4}$	$3.70x10^{3}$	$2.85x10^{2}$
	136	14	64	$1.18x10^{8}$	$8.40x10^{6}$	$2.84x10^{5}$	$2.03x10^{4}$	$4.50x10^{3}$	$3.21x10^{2}$
	152	15	68	$2.78x10^{8}$	$1.85x10^{7}$	$4.84x10^{5}$	$3.23x10^{4}$	$6.10x10^{3}$	$4.07x10^{2}$
	160	17	77	$3.03x10^{8}$	$1.78x10^{7}$	$4.85x10^{5}$	$2.85x10^{4}$	$2.91x10^{4}$	$1.71x10^{3}$
	168	18	82	$5.03x10^{8}$	$2.79x10^{7}$	$4.92x10^{5}$	$2.73x10^{4}$	$6.01x10^{4}$	$3.34x10^{3}$
	176	20	91	$7.03x10^{8}$	$3.51x10^{7}$	$5.01x10^{5}$	$2.51x10^{4}$	$1.01x10^{5}$	$5.06x10^{3}$
	184	21	95	$7.13x10^{8}$	$3.39x10^{7}$	$5.02x10^{5}$	$2.39x10^{4}$	$1.25x10^{5}$	$5.95x10^{3}$
	240	22	100	$1.21x10^{9}$	$5.51x10^{7}$	$5.08x10^{5}$	$2.31x10^{4}$	$2.06x10^{5}$	$9.36x10^{3}$
Fast	36	2	15	$1.20x10^{5}$	$6.00x10^{4}$	$4.02x10^{2}$	$2.01x10^{2}$	$6.50x10^{2}$	$3.25x10^{2}$
	48	5	38	$2.02x10^{5}$	$4.03x10^{4}$	$2.02x10^{5}$	$4.04x10^{4}$	$1.93x10^{3}$	$3.86x10^{2}$
	56	7	54	$3.22x10^{5}$	$4.59x10^{4}$	$3.23x10^{5}$	$4.61x10^{4}$	$2.43x10^{3}$	$3.47x10^{2}$
	72	8	62	$3.62x10^{5}$	$4.52x10^{4}$	$3.63x10^{5}$	$4.54x10^{4}$	$2.53x10^{3}$	$3.16x10^{2}$
	80	10	77	$4.03x10^{5}$	$4.03x10^{4}$	$4.05x10^{5}$	$4.05x10^{4}$	$2.98x10^{3}$	$2.98x10^{2}$
	192	11	85	$4.83x10^{5}$	$4.39x10^{4}$	$4.85x10^{5}$	$4.41x10^{4}$	$3.68x10^{3}$	$3.35x10^{2}$
	224	12	92	$1.16x10^{7}$	$9.64x10^{5}$	$1.16x10^{7}$	$9.64x10^{5}$	$1.37x10^{4}$	$1.14x10^{3}$
	232	13	100	$1.16x10^{7}$	$8.89x10^{5}$	$1.16x10^{7}$	$8.90x10^{5}$	$2.37x10^{4}$	$1.82x10^{3}$

Table 12-22: GLM of F5 strains and AC53 OB counts indicating results are statistically significant.

Performance Metric	**Virus**	**δ_1 Estimate (Standard Error)**	**δ_1 Pr(>|t|)**	**Dispersion**
Viral Yield (OB/mL)	AC53	3.9664 (0.8881)	0.0043	3380361
	Slow	2.6365 (0.4305)	1.74×10^{4}	2411443
	MaxOB	3.4177 (0.3346)	2.85×10^{-6}	1523471
	Fast	-0.3925 (0.1988)	0.143	766.86
Viral Capacity (µg)	AC53	2.289 (1.048)	0.072	114.44
	Slow	2.599 (0.995)	0.028	260.51
	MaxOB	3.8803 (0.5885)	2.56×10^{-5}	556.05
	Fast	-0.1555 (0.1571)	0.395	3.4127
Viral Density (OB/µg)	AC53	3.2675 (0.7721)	0.006	8638.69
	Slow	1.4142 (0.4373)	0.010	21041.98
	MaxOB	1.7409 (0.3384)	2.43×10^{-4}	3076.26
	Fast	1.483 (1.019)	0.241	11734.09

12.5 GENETIC ANALYSIS OF TRAIT-SPECIFIC *IN VIVO* DERIVED STRAINS FROM HASNPV-AC53

Table 12-23: Polymorphic type and total polymorphisms identified within each ORF and Hr region within the AC53 MiSeq genome and each selected strain.

Virus	ORF/Region	Substitution	Insertion	Deletion	Total Polymorphisms
AC53 MiSeq	BRO-A	4	0	0	4
AC53 MiSeq	BRO-B	4	0	0	4
AC53 MiSeq	CALYX/PEP	3	0	0	3
AC53 MiSeq	DNA polymerase	1	0	0	1
AC53 MiSeq	GP19	2	0	0	2
AC53 MiSeq	Hr1	8	0	1	9
AC53 MiSeq	Hr2	11	1	5	17
AC53 MiSeq	Hr3	1	0	0	1
AC53 MiSeq	Hr4	5	2	0	7
AC53 MiSeq	Hr5	8	2	1	11
AC53 MiSeq	ME53	2	0	0	2
AC53 MiSeq	ODV-EC27	5	0	0	5
AC53 MiSeq	ORF131	1	0	0	1
AC53 MiSeq	ORF132	3	0	0	3
AC53 MiSeq	P49	6	0	0	6
F1 Fast	38.7K	8	2	0	10
F1 Fast	ARIF-1	2	0	0	2
F1 Fast	BRO-A	7	0	0	7
F1 Fast	CALYX/PEP	6	0	0	6
F1 Fast	DNA polymerase	2	0	0	2
F1 Fast	GP19	2	0	1	3
F1 Fast	Hr1	2	0	0	2
F1 Fast	Hr2	21	0	2	23
F1 Fast	Hr3	1	0	0	1
F1 Fast	Hr4	5	1	1	7
F1 Fast	Hr5	2	2	0	4
F1 Fast	Hypothetical ORF	0	0	2	2
F1 Fast	IE-1	2	0	0	2
F1 Fast	ME53	7	0	0	7
F1 Fast	ODV-E18	2	0	0	2
F1 Fast	ODV-E56	9	0	0	9
F1 Fast	ODV-EC27	10	0	0	10
F1 Fast	ORF12	3	0	0	3
F1 Fast	ORF13	3	0	0	3
F1 Fast	ORF131	10	0	0	10
F1 Fast	ORF132	3	0	0	3
F1 Fast	ORF136	9	0	0	9
F1 Fast	ORF17	1	0	0	1

F1 Fast	P49	13	0	0	13
F1 Fast	P74	1	0	0	1
F2 Fast	38.7K	8	2	0	10
F2 Fast	BRO-A	7	0	0	7
F2 Fast	CALYX/PEP	1	0	0	1
F2 Fast	DNA polymerase	2	0	0	2
F2 Fast	GP19	2	0	1	3
F2 Fast	Hr1	1	0	0	1
F2 Fast	Hr2	20	1	4	25
F2 Fast	Hr3	1	0	0	1
F2 Fast	Hr4	5	0	1	6
F2 Fast	Hr5	8	4	1	13
F2 Fast	Hypothetical ORF	0	0	2	2
F2 Fast	IE-1	2	0	0	2
F2 Fast	LEF-8	1	0	0	1
F2 Fast	ME53	7	0	0	7
F2 Fast	ODV-E18	2	0	0	2
F2 Fast	ODV-E56	9	0	0	9
F2 Fast	ODV-EC27	10	0	0	10
F2 Fast	ORF12	3	0	0	3
F2 Fast	ORF13	3	0	0	3
F2 Fast	ORF131	10	0	0	10
F2 Fast	ORF132	3	0	0	3
F2 Fast	ORF136	9	0	0	9
F2 Fast	ORF17	1	0	0	1
F2 Fast	P49	13	0	0	13
F2 Fast	P74	1	0	0	1
F3 Fast	38.7K	8	2	0	10
F3 Fast	BRO-A	9	0	0	9
F3 Fast	CALYX/PEP	1	0	0	1
F3 Fast	cathepsin	1	0	0	1
F3 Fast	DNA polymerase	2	0	0	2
F3 Fast	GP19	3	0	0	3
F3 Fast	Hr1	1	0	0	1
F3 Fast	Hr2	23	0	0	23
F3 Fast	Hr3	1	0	0	1
F3 Fast	Hr4	8	2	0	10
F3 Fast	Hr5	13	4	0	17
F3 Fast	Hypothetical ORF	1	0	1	2
F3 Fast	IE-1	2	0	0	2
F3 Fast	ME53	7	0	0	7
F3 Fast	ODV-E18	2	0	0	2
F3 Fast	ODV-E56	9	0	0	9
F3 Fast	ODV-EC27	10	0	0	10
F3 Fast	ORF12	3	0	0	3

F3 Fast	ORF13	3	0	0	3
F3 Fast	ORF131	10	0	0	10
F3 Fast	ORF132	3	0	0	3
F3 Fast	ORF136	9	0	0	9
F3 Fast	ORF17	1	0	0	1
F3 Fast	P49	13	0	0	13
F3 Fast	P74	1	0	0	1
F4 Fast	38.7K	8	2	0	10
F4 Fast	BRO-A	10	0	0	10
F4 Fast	CALYX/PEP	6	0	0	6
F4 Fast	DNA polymerase	2	0	0	2
F4 Fast	GP19	2	0	1	3
F4 Fast	Hr1	4	0	0	4
F4 Fast	Hr2	25	1	5	31
F4 Fast	Hr3	1	0	0	1
F4 Fast	Hr4	9	1	2	12
F4 Fast	Hr5	7	2	3	12
F4 Fast	Hypothetical ORF	0	0	2	2
F4 Fast	IE-1	2	0	0	2
F4 Fast	ME53	5	0	0	5
F4 Fast	ODV-E18	2	0	0	2
F4 Fast	ODV-EC27	9	0	0	9
F4 Fast	ORF12	2	0	0	2
F4 Fast	ORF13	1	0	0	1
F4 Fast	ORF131	10	0	0	10
F4 Fast	ORF132	3	0	0	3
F4 Fast	ORF136	9	0	0	9
F4 Fast	ORF17	1	0	0	1
F4 Fast	P49	10	0	0	10
F4 Fast	P74	1	0	0	1
F5 Fast	38.7K	8	2	0	10
F5 Fast	39K/PP31	1	0	0	1
F5 Fast	BRO-A	9	0	0	9
F5 Fast	CALYX/PEP	1	0	0	1
F5 Fast	DNA polymerase	2	0	0	2
F5 Fast	GP19	1	0	1	2
F5 Fast	Hr1	5	0	0	5
F5 Fast	Hr2	21	0	1	22
F5 Fast	Hr3	1	0	0	1
F5 Fast	Hr4	5	1	1	7
F5 Fast	Hr5	6	2	0	8
F5 Fast	Hypothetical ORF	1	0	1	2
F5 Fast	ME53	5	0	0	5
F5 Fast	ODV-EC27	2	0	0	2
F5 Fast	ORF136	9	0	0	9

F5 Fast	ORF17	1	0	0	1
F5 Fast	ORF67	1	0	0	1
F5 Fast	P49	4	0	0	4
F5 Fast	P74	1	0	0	1
F1 Slow	38.7K	8	2	0	10
F1 Slow	BRO-A	7	0	0	7
F1 Slow	CALYX/PEP	1	0	0	1
F1 Slow	DNA polymerase	2	0	0	2
F1 Slow	GP19	1	0	1	2
F1 Slow	Hr1	2	0	0	2
F1 Slow	Hr2	22	0	2	24
F1 Slow	Hr3	1	0	0	1
F1 Slow	Hr4	7	2	1	10
F1 Slow	Hr5	2	2	0	4
F1 Slow	Hypothetical ORF	0	0	1	1
F1 Slow	ME53	5	0	0	5
F1 Slow	ODV-EC27	2	0	0	2
F1 Slow	ORF128	6	0	0	6
F1 Slow	ORF131	10	0	0	10
F1 Slow	ORF132	3	0	0	3
F1 Slow	ORF136	9	0	0	9
F1 Slow	ORF17	1	0	0	1
F1 Slow	P49	4	0	0	4
F1 Slow	P74	1	0	0	1
F2 Slow	38.7K	8	2	0	10
F2 Slow	BRO-A	7	0	0	7
F2 Slow	CALYX/PEP	1	0	0	1
F2 Slow	DNA polymerase	2	0	0	2
F2 Slow	GP19	1	0	1	2
F2 Slow	Hr1	4	0	0	4
F2 Slow	Hr2	25	0	4	29
F2 Slow	Hr3	1	0	0	1
F2 Slow	Hr4	4	2	1	7
F2 Slow	Hr5	3	3	0	6
F2 Slow	Hypothetical ORF	0	0	2	2
F2 Slow	LEF-8	1	0	0	1
F2 Slow	ME53	5	0	0	5
F2 Slow	ODV-EC27	2	0	0	2
F2 Slow	ORF131	10	0	0	10
F2 Slow	ORF132	3	0	0	3
F2 Slow	ORF136	9	0	0	9
F2 Slow	ORF17	1	0	0	1
F2 Slow	P49	4	0	0	4
F2 Slow	P74	1	0	0	1
F3 Slow	38.7K	8	2	0	10

F3 Slow	BRO-A	13	0	0	13
F3 Slow	CALYX/PEP	1	0	0	1
F3 Slow	DNA polymerase	2	0	0	2
F3 Slow	GP19	1	0	1	2
F3 Slow	Hr1	3	0	0	3
F3 Slow	Hr2	29	0	3	32
F3 Slow	Hr3	1	0	0	1
F3 Slow	Hr4	13	2	1	16
F3 Slow	Hr5	2	2	0	4
F3 Slow	Hypothetical ORF	0	0	1	1
F3 Slow	LEF-8	1	0	0	1
F3 Slow	ME53	5	0	0	5
F3 Slow	ODV-EC27	2	0	0	2
F3 Slow	ORF131	10	0	0	10
F3 Slow	ORF132	3	0	0	3
F3 Slow	ORF136	9	0	0	9
F3 Slow	ORF17	1	0	0	1
F3 Slow	P49	4	0	0	4
F3 Slow	P74	1	0	0	1
F4 Slow	38.7K	8	2	0	10
F4 Slow	BRO-A	9	0	0	9
F4 Slow	CALYX/PEP	1	0	0	1
F4 Slow	DNA polymerase	2	0	0	2
F4 Slow	GP19	1	0	1	2
F4 Slow	Hr1	7	0	0	7
F4 Slow	Hr2	26	1	3	30
F4 Slow	Hr3	1	0	0	1
F4 Slow	Hr4	24	2	3	29
F4 Slow	Hr5	1	1	0	2
F4 Slow	Hypothetical ORF	0	0	1	1
F4 Slow	LEF-8	1	0	0	1
F4 Slow	ME53	5	0	0	5
F4 Slow	ODV-EC27	2	0	0	2
F4 Slow	ORF131	10	0	0	10
F4 Slow	ORF132	3	0	0	3
F4 Slow	ORF136	9	0	0	9
F4 Slow	ORF17	1	0	0	1
F4 Slow	P49	4	0	0	4
F4 Slow	P74	1	0	0	1
F5 Slow	38.7K	8	2	0	10
F5 Slow	BRO-A	7	0	0	7
F5 Slow	CALYX/PEP	1	0	0	1
F5 Slow	DNA polymerase	2	0	0	2
F5 Slow	EGT	1	0	0	1
F5 Slow	GP19	1	0	1	2

F5 Slow	Hr1	4	0	0	4
F5 Slow	Hr2	23	1	7	31
F5 Slow	Hr3	1	0	0	1
F5 Slow	Hr4	12	2	2	16
F5 Slow	Hr5	8	3	2	13
F5 Slow	Hypothetical ORF	1	0	0	1
F5 Slow	LEF-8	1	0	0	1
F5 Slow	ME53	5	0	0	5
F5 Slow	ODV-EC27	2	0	0	2
F5 Slow	ORF131	10	0	0	10
F5 Slow	ORF132	3	0	0	3
F5 Slow	ORF136	9	0	0	9
F5 Slow	ORF17	1	0	0	1
F5 Slow	P49	4	0	0	4
F5 Slow	P74	1	0	0	1
F1 MaxOB	38.7K	8	2	0	10
F1 MaxOB	BRO-A	10	0	0	10
F1 MaxOB	CALYX/PEP	1	0	0	1
F1 MaxOB	DNA polymerase	2	0	0	2
F1 MaxOB	GP19	1	0	1	2
F1 MaxOB	Hr1	4	0	1	5
F1 MaxOB	Hr2	25	1	2	28
F1 MaxOB	Hr3	1	0	0	1
F1 MaxOB	Hr4	10	1	1	12
F1 MaxOB	Hr5	2	2	0	4
F1 MaxOB	Hypothetical ORF	1	0	0	1
F1 MaxOB	ME53	5	0	0	5
F1 MaxOB	ODV-EC27	2	0	0	2
F1 MaxOB	ORF131	10	0	0	10
F1 MaxOB	ORF132	3	0	0	3
F1 MaxOB	ORF136	9	0	0	9
F1 MaxOB	ORF17	1	0	0	1
F1 MaxOB	ORF92	1	0	0	1
F1 MaxOB	P49	4	0	0	4
F1 MaxOB	P74	1	0	0	1
F2 MaxOB	38.7K	8	2	0	10
F2 MaxOB	BRO-A	9	0	0	9
F2 MaxOB	CALYX/PEP	1	0	0	1
F2 MaxOB	DNA polymerase	2	0	0	2
F2 MaxOB	GP19	1	0	1	2
F2 MaxOB	Hr1	1	0	0	1
F2 MaxOB	Hr2	23	1	3	27
F2 MaxOB	Hr3	1	0	0	1
F2 MaxOB	Hr4	10	2	1	13
F2 MaxOB	Hr5	2	2	0	4

F2 MaxOB	Hypothetical ORF	0	0	1	1
F2 MaxOB	IE-1	2	0	0	2
F2 MaxOB	ME53	5	0	0	5
F2 MaxOB	ODV-EC27	2	0	0	2
F2 MaxOB	ORF128	6	0	0	6
F2 MaxOB	ORF13	3	0	0	3
F2 MaxOB	ORF130	1	3	5	9
F2 MaxOB	ORF131	10	0	0	10
F2 MaxOB	ORF132	3	0	0	3
F2 MaxOB	ORF136	9	0	0	9
F2 MaxOB	ORF17	1	0	0	1
F2 MaxOB	P49	4	0	0	4
F2 MaxOB	P74	1	0	0	1
F3 MaxOB	38.7K	8	2	0	10
F3 MaxOB	BRO-A	13	0	0	13
F3 MaxOB	CALYX/PEP	1	0	0	1
F3 MaxOB	DNA polymerase	2	0	0	2
F3 MaxOB	GP19	1	0	1	2
F3 MaxOB	Hr1	2	0	0	2
F3 MaxOB	Hr2	29	1	2	32
F3 MaxOB	Hr3	1	0	0	1
F3 MaxOB	Hr4	27	5	2	34
F3 MaxOB	Hr5	2	2	0	4
F3 MaxOB	Hypothetical ORF	0	0	2	2
F3 MaxOB	IE-1	1	0	0	1
F3 MaxOB	ME53	5	0	0	5
F3 MaxOB	ODV-EC27	2	0	0	2
F3 MaxOB	ORF130	1	0	1	2
F3 MaxOB	ORF131	10	0	0	10
F3 MaxOB	ORF132	3	0	0	3
F3 MaxOB	ORF136	9	0	0	9
F3 MaxOB	ORF17	1	0	0	1
F3 MaxOB	P49	4	0	0	4
F3 MaxOB	P74	1	0	0	1
F4 MaxOB	38.7K	8	2	0	10
F4 MaxOB	BRO-A	7	0	0	7
F4 MaxOB	CALYX/PEP	1	0	0	1
F4 MaxOB	DNA polymerase	2	0	0	2
F4 MaxOB	GP19	1	0	1	2
F4 MaxOB	Hr1	1	0	0	1
F4 MaxOB	Hr2	20	1	2	23
F4 MaxOB	Hr3	1	0	0	1
F4 MaxOB	Hr4	12	2	3	17
F4 MaxOB	Hr5	1	1	0	2
F4 MaxOB	Hypothetical ORF	0	0	1	1

F4 MaxOB	ME53	5	0	0	5
F4 MaxOB	ODV-EC27	2	0	0	2
F4 MaxOB	ORF130	1	3	5	9
F4 MaxOB	ORF131	10	0	0	10
F4 MaxOB	ORF132	3	0	0	3
F4 MaxOB	ORF136	9	0	0	9
F4 MaxOB	ORF17	1	0	0	1
F4 MaxOB	P49	4	0	0	4
F4 MaxOB	P74	1	0	0	1
F5 MaxOB	38.7K	8	2	0	10
F5 MaxOB	BRO-A	9	0	0	9
F5 MaxOB	CALYX/PEP	1	0	0	1
F5 MaxOB	DNA polymerase	2	0	0	2
F5 MaxOB	GP19	1	0	1	2
F5 MaxOB	Hr1	6	0	0	6
F5 MaxOB	Hr2	23	2	4	29
F5 MaxOB	Hr3	1	0	0	1
F5 MaxOB	Hr4	8	1	1	10
F5 MaxOB	Hr5	4	2	4	10
F5 MaxOB	Hypothetical ORF	0	0	1	1
F5 MaxOB	ME53	5	0	0	5
F5 MaxOB	ODV-EC27	2	0	0	2
F5 MaxOB	ORF130	1	3	5	9
F5 MaxOB	ORF131	10	0	0	10
F5 MaxOB	ORF132	3	0	0	3
F5 MaxOB	ORF136	9	0	0	9
F5 MaxOB	ORF17	1	0	0	1
F5 MaxOB	P49	4	0	0	4
F5 MaxOB	P74	1	0	0	1

Table 12-24: *K*-means clustering and mean abundance of reference allele and alternative allele within each cluster identified within each analysed virus.

Virus	Clusters	Reference Allele Mean Cluster Abundance (%)	Alternative Allele Mean Cluster Abundance (%)
AC53 MiSeq	1	77.41	22.59
	2	26.07	73.93
	3	73.47	26.53
	4	1.29	98.71
	5	53.85	46.15
	6	43.66	56.34
F1 Fast	1	29.86	68.34
	2	9.02	90.98
	3	76.16	23.84
F1 MaxOB	1	6.56	93.44
	2	1.24	98.76
	3	60.95	39.05
	4	43.72	56.28
	5	78.53	19.10
F1 Slow	1	54.49	45.51
	2	70.80	29.20
	3	38.37	61.63
	4	84.54	15.46
	5	1.32	98.68
	6	99.27	0.73
F2 Fast	1	71.01	28.99
	2	43.84	56.16
	3	86.98	13.02
	4	3.13	96.87
	5	80.67	19.33
	6	64.50	35.50
	7	89.54	10.46
	8	3.46	96.54
	9	16.79	83.21
F2 MaxOB	1	1.42	98.58
	2	89.65	10.35
	3	40.90	59.10
	4	69.34	30.66
F2 Slow	1	54.70	45.30
	2	37.05	57.48
	3	0.05	99.95
	4	0.37	99.63
	5	4.93	95.07
	6	85.89	14.11
F3 Fast	1	19.65	79.29

	2	78.52	21.48
F3 MaxOB	1	3.07	95.48
	2	86.74	13.26
	3	64.55	35.45
	4	49.93	48.51
F3 Slow	1	73.31	26.69
	2	0.00	100.00
	3	0.15	99.85
	4	43.44	54.89
F4 Fast	1	17.11	81.51
	2	77.79	22.21
F4 MaxOB	1	88.26	11.74
	2	38.18	61.82
	3	0.56	99.44
	4	51.69	48.31
	5	62.84	37.16
F4 Slow	1	88.75	11.25
	2	1.08	98.92
	3	63.21	36.79
	4	41.12	58.88
F5 Fast	1	1.06	98.94
	2	57.50	38.63
F5 MaxOB	1	68.05	30.55
	2	1.80	98.20
F5 Slow	1	4.17	95.83
	2	22.88	77.12
	3	85.81	14.02
	4	0.10	99.90
	5	45.58	54.42

Table 12-25: MaxOB strains ORF130/130a/130b polymorphic abundance changes during each round of selection.

ORF130	**F1 MaxOB**		**F2 MaxOB**		**F3 MaxOB**		**F4 MaxOB**		**F5 MaxOB**	
Position	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**
119,490	100	0	89.46	10.54	85.71	14.29	0	100	1.28	98.72
119,514	100	0	89.68	10.32	100	0	39.66	60.34	71.69	28.31
119,534	100	0	92.14	7.86	100	0	43.22	56.78	72.79	27.21
119,539	100	0	92.17	7.83	100	0	43.22	56.78	72.30	27.70
119,546	100	0	92.16	7.84	92.31	7.69	43.17	56.83	72.79	27.21
119,552	100	0	92.21	7.79	100	0	43.09	56.91	72.30	27.70
119,556	100	0	92.20	7.80	100	0	42.99	57.01	72.00	28.00
119,562	100	0	92.13	7.87	100	0	42.19	57.81	71.05	28.95
119,565	100	0	89.97	10.03	100	0	0.23	99.77	0	100
Mean	**100**	**0**	**91.35**	**8.65**	**89.01**	**10.99**	**33.09**	**66.91**	**63.12**	**43.75**

Table 12-26: Slow strains *lef-8* polymorphic abundance changes during each round of selection.

Lef-8	**F1 Slow**		**F2 Slow**		**F3 Slow**		**F4 Slow**		**F5 Slow**	
Position	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**
33,358	100	0	45.35	54.64	38.32	61.67	30.41	69.85	0.33	99.67

Table 12-27: Fast strains ORF12 polymorphic abundance changes during each round of selection.

ORF12	F1 Fast		F2 Fast		F3 Fast		F4 Fast		F5 Fast	
Position	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**
10,450	79.76	20.24	85.79	14.21	78.55	21.45	92.09	7.91	100	0
10,510	75.05	24.95	88.04	11.96	78.98	21.02	100	0	100	0
10,546	74.73	25.27	87.90	12.10	79.81	20.19	92.23	7.77	100	0
Mean	**76.51**	**23.49**	**87.24**	**12.76**	**79.12**	**20.88**	**94.77**	**5.23**	**100.00**	**0.00**

Table 12-28: Fast strains ORF13 polymorphic abundance changes during each round of selection.

ORF13	F1 Fast		F2 Fast		F3 Fast		F4 Fast		F5 Fast	
Position	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**
10,954	75.88	24.12	88.95	11.05	84.32	15.68	100	0	100	0
11,003	77.76	22.24	88.51	11.49	85.52	14.48	100	0	100	0
11,107	80.42	19.58	89.35	10.65	85.45	14.55	91.83	8.17	100	0
Mean	**78.02**	**21.98**	**88.94**	**11.06**	**85.09**	**14.91**	**97.28**	**2.72**	**100.00**	**0.00**

Table 12-29: Fast strains IE-1 polymorphic abundance changes during each round of selection.

IE-1	F1 Fast		F2 Fast		F3 Fast		F4 Fast		F5 Fast	
Position	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**
11,522	75.17	24.83	86.13	13.87	89.70	10.30	91.18	8.82	100	0
12,043	73.05	26.95	89.59	10.41	90.08	9.92	89.76	10.24	100	0
Mean	**74.11**	**25.89**	**87.86**	**12.14**	**89.89**	**10.11**	**90.47**	**9.53**	**100.00**	**0.00**

Table 12-30: Fast strains ODV-E56 polymorphic abundance changes during each round of selection.

ODV-E56	F1 Fast		F2 Fast		F3 Fast		F4 Fast		F5 Fast	
Position	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**	**AC53 Allele**	**Alt Allele**
13,918	75.58	24.42	88.87	11.13	85.90	14.10	100	0	100	0
13,951	76.42	23.58	89.35	10.65	86.50	13.50	100	0	100	0
13,955	76.73	23.27	89.44	10.56	86.59	13.41	100	0	100	0
13,961	76.18	23.82	89.24	10.76	85.98	14.02	100	0	100	0
14,030	77.18	22.82	88.82	11.18	78.03	21.97	100	0	100	0
14,273	76.34	23.66	87.95	12.05	80.80	19.20	100	0	100	0
14,339	74.96	25.04	88.79	11.21	80.67	19.33	100	0	100	0
14,366	74.84	25.16	88.82	11.18	80.35	19.65	100	0	100	0
14,374	76.63	23.37	89.16	10.84	81.08	18.92	100	0	100	0
Mean	**76.10**	**23.90**	**88.94**	**11.06**	**82.88**	**17.12**	**100.00**	**0.00**	**100.00**	**0.00**

Appendices

13.1 APPENDIX A: AN ADDITIONAL CO-AUTHORED PUBLISHED PAPER UNRELATED TO THE THESIS BUT APPLIES THE 'INVERTIBRATES AND MICROBIOLOGY GROUP ASSEMBLY PIPELINE' TO FOUR GRANULOVIRUSES

Statement of Contribution of Co-Authors for Thesis by Published Paper

The authors listed below have certified* that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

5. **Spence, R. J., Noune, C., & Hauxwell, C. (2016). Complete Genome Sequences of Four Isolates of *Plutella xylostella* Granulovirus. Genome announcements, 4(3). doi:10.1128/genomeA.00633-16**

Contributor	Statement of contribution*
Robert Spence Signature: Date: 27/10/2017	Performed experiments and sequencing. Interpreted results. Assembled and annotated genomes.
Christopher Noune Signature: Date: 27/10/2017	Contributed to experimental design, data analysis, result interpretation and minor manuscript input.
Caroline Hauxwell	Contributed to experimental design, data analysis, result interpretation and reviewed and edited manuscript.

Principal Supervisor Confirmation		
I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
Caroline Hauxwell Name	CHauxwell Signature	27/10/17 Date

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Complete Genome Sequences of Four Isolates of *Plutella xylostella* Granulovirus

Robert J. Spence, Christopher Noune, Caroline Hauxwell

Queensland University of Technology (QUT), Brisbane, Australia

Granuloviruses are widespread pathogens of *Plutella xylostella* L. (diamondback moth) and potential biopesticides for control of this global insect pest. We report the complete genomes of four *Plutella xylostella* granulovirus isolates from China, Malaysia, and Taiwan exhibiting pairs of noncoding, homologous repeat regions with significant sequence variation but equivalent length.

Received 12 May 2016 **Accepted** 17 May 2016 **Published** 30 June 2016

Citation Spence RJ, Noune C, Hauxwell C. 2016. Complete genome sequences of four isolates of *Plutella xylostella* granulovirus. Genome Announc 4(3):e00633-16. doi:10.1128/genomeA.00633-16.

Address correspondence to Caroline Hauxwell, caroline.hauxwell@qut.edu.au.

Plutella xylostella L. (diamondback moth) is a globally important insect pest of *Brassica* crops that is widely resistant to insecticides incurring control costs up to five billion U.S. dollars annually (1–3). *Plutella xylostella* granuloviruses (*Plxy*GV) have potential use as biopesticides to manage insecticide resistance and improve pest management (2, 4, 5). Here, we present complete genome sequences of four isolates of *Plxy*GV from China, Malaysia, and two from Taiwan (6, 7).

Isolates were passaged through *P. xylostella* larvae and occlusion bodies purified by centrifugation, followed by incubations in 1 M sodium carbonate (Na_2CO_3) at room temperature for 5 min and 1% N-lauryl sarcosine at 37°C for 15 min. DNA was then purified using an Isolate II Genomic DNA kit (catalogue no. BIO-52067, Bioline) from step 4 according to manufacturer's instructions.

Genomic DNA was prepared for sequencing using the Nextera XT kit and medium-output flow cell on an Illumina Next-Seq 500 with 150 bp paired-end sequencing. Raw sequence data comprised 12,852,411 reads (*Plxy*GV-C, China); 8,878,618 reads (*Plxy*GV-T, Taiwan); 12,251,888 reads (*Plxy*GV-K, Taiwan); and 16,077,717 reads (*Plxy*GV-M, Malaysia), representing average genome coverage of 19,088×, 13,186×, 18,196×, and 23,878×, respectively.

Read inspection, trimming, and genome assembly used the method of Noune and Hauxwell (8). Reads were analyzed and trimmed using FastQC version 0.11.3 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and FastX trimmer version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit). Genomes were assembled *de novo* using Tadpole (BBMap 35.49) (http://sourceforge.net/projects/bbmap) with *k*mer values generated by Kmergenie (9) and mapping to the NC_002593.1 reference (10) using BWA version 0.7.12 (11) and SAMtools version 1.2 (12). The *de novo* assembled contigs, BWA generated mapping, and trimmed reads were merged into a single FastQ file and then remapped to the NC_002593.1 reference using the Geneious R9 mapper (13) with low to medium sensitivity and 5 iterations. Gaps were filled manually from Sanger sequences. All four genomes showed a high degree of similarity, with 40.7% G+C content and sequence homology of 99.9%. Open reading frames (ORFs) were predicted using the Geneious R9 live annotation tool and compared to the NC_002593.1 reference with the larger of overlapping ORFs selected. One hundred eighteen ORFs were predicted for all four *Plxy*GV isolates, two fewer than the NC_002593.1 reference genome. The completed genomes are 100,980 bp (*Plxy*GV-C), 100,978 bp (*Plxy*GV-T), 101,004 bp (*Plxy*GV-K), and 100,980 bp (*Plxy*GV-M) in length.

Genomic differences arise in the position of ORF73 (which shares 58.6% nucleotide sequence identity with AcMNPV ORF91) and is truncated in PlxyGV-K. The regions of greatest sequence variation occur within two pairs of noncoding regions, which are almost identical in length (pair one, 2,596 bp and 2,516 bp; pair two, 1,340 bp and 1,414 bp) in the four isolates and NC_002593.1 reference genome. These are homologous repeat regions that are common features within baculovirus genomes (14).

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession numbers KU529791 (*Plxy*GV-C), KU529792 (*Plxy*GV-K), KU529793 (*Plxy*GV-M), and KU529794 (*Plxy*GV-T). The versions described in this paper are the first versions, KU529791.1, KU529792.1, KU529793.1, and KU529794.1.

ACKNOWLEDGMENTS

This work was funded by the Grains Research and Development Corporation in partnership with AgBiTech Australia Pty. Ltd. and carried out at Queensland University of Technology (QUT), Australia. The data reported in this paper were obtained at the Central Analytical Research Facility operated by the Institute for Future Environments (QUT).

We would like to acknowledge and thank Helen Hesketh (NERC Centre for Ecology and Hydrology, Wallingford, UK) and Robert Possee (Oxford Brookes University) for their kind assistance in accessing isolates.

FUNDING INFORMATION

This work, including the efforts of Robert James Spence and Caroline Hauxwell, was funded by Grains Research Development Corporation (QUT00004). This work, including the efforts of Christopher Noune, was funded by Cotton Research Development Corporation (QUT1402).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. **Dezianian A, Sajap AS, Lau WH, Omar D, Kadir HA, Mohamed R, Yusoh MRM.** 2010. Morphological characteristics of *P. Xylostella granulovirus* and effects on its larval host diamondback moth *Plutella xylostella L.* (*Lepidoptera*, *Plutellidae*). Am J Agric Biol Sci **5:43–49.** http://dx.doi.org/10.3844/ajabssp.2010.43.49.
2. **Furlong MJ, Wright DJ, Dosdall LM.** 2013. Diamondback moth ecology and management: problems, progress, and prospects. Annu Rev Entomol **58:517–541.** http://dx.doi.org/10.1146/annurev-ento-120811-153605.
3. **Grzywacz D, Rossbach A, Rauf A, Russell DA, Srinivasan R, Shelton AM.** 2010. Current control methods for diamondback moth and other Brassica insect pests and the prospects for improved management with lepidopteran-resistant Bt vegetable brassicas in Asia and Africa. Crop Protect **29:68–79.** http://dx.doi.org/10.1016/j.cropro.2009.08.009.
4. **Talekar NS, Shelton AM.** 1993. Biology, ecology, and management of the diamondback moth. Annu Rev Entomol **38:275–301.**
5. **Raymond BEN, Sayyed AH, Hails RS, Wright DJ.** 2007. Exploiting pathogens and their impact on fitness costs to manage the evolution of resistance to *Bacillus thuringiensis*. J Appl Ecol **44:768–780.** http://dx.doi.org/10.1111/j.1365-2664.2007.01285.x.
6. **Kadir HBA.** 1986. The granulosis virus of *Plutella xylostella*, p 261–273. In Biological control in the tropics. Proceedings of the First Regional Symposium on Biological Control, Malaysia. Malaysian Agricultural Research and Development Institute (MARDI).
7. **Kadir HBA, Payne CC, Crook NE, Winstanley D.** 1999. Characterization and cross-transmission of baculoviruses infectious to the diamondback moth, *Plutella xylostella*, and some other lepidopteran pests of brassica crops. Biocontrol Sci Technol **9:227–238.** http://dx.doi.org/10.1080/09583159929802.
8. **Noune C, Hauxwell C.** 2016. Complete genome sequences of seven HaSNPV-AC53-derived strains. Genome Announc **4(3):e00260-16.** http://dx.doi.org/10.1128/genomeA.00260-16.
9. **Chikhi R, Medvedev P.** 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics **30:31–37.** http://dx.doi.org/10.1093/bioinformatics/btt310.
10. **Hashimoto Y, Hayakawa T, Ueno Y, Fujita T, Sano Y, Matsumoto T.** 2000. Sequence analysis of the *Plutella xylostella* granulovirus genome. Virology **275:358–372.** http://dx.doi.org/10.1006/viro.2000.0530.
11. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:1754–1760.** http://dx.doi.org/10.1093/bioinformatics/btp324.
12. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:2078–2079.** http://dx.doi.org/10.1093/bioinformatics/btp352.
13. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A.** 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:1647–1649.** http://dx.doi.org/10.1093/bioinformatics/bts199.
14. **Ferrelli ML, Salvador R, Biedma ME, Berretta MF, Haase S, Sciocco-Cap A, Ghiringhelli PD, Romanowski V.** 2012. Genome of *Epinotia aporema* granulovirus (EpapGV), a polyorganotropic fast killing betabaculovirus with a novel thymidylate kinase gene. BMC Genomics **13:548.** http://dx.doi.org/10.1186/1471-2164-13-548.

13.2 APPENDIX B: ENHANCED PIPELINE 'METAGAAP-PY' FOR THE ANALYSIS OF QUASISPECIES AND NON-MODEL MICROBIAL POPULATIONS USING ULTRA-DEEP 'META-BARCODE' SEQUENCING

*For source code refer to: https://github.com/CNoune/IMG_pipelines/tree/master/MetaGaAP-Py

Statement of Contribution of Co-Authors for Thesis by Published Paper

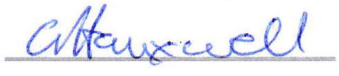
The authors listed below have certified* that:

11. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
12. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
13. there are no other authors of the publication according to these criteria;
14. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
15. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

In the case of this chapter:

6. **Noune, C., & Hauxwell, C. (2017). Enhanced Pipeline 'MetaGaAP-Py' for the Analysis of Quasispecies and Non-Model Microbial Populations using Ultra-Deep 'Meta-barcode' Sequencing. bioRxiv. doi:10.1101/171520**

Contributor	Statement of contribution*
Christopher Noune Signature:  Date: 27/10/17	Performed all programming and contributed equally to the development of concepts and applications, and to the writing of the manuscript.
Caroline Hauxwell	Contributed equally to the development of concepts and applications, and to the writing of the manuscript.

Principal Supervisor Confirmation		
I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.		
Caroline Hauxwell Name	 Signature	27/10/17 Date

Enhanced Pipeline 'MetaGaAP-Py' for the Analysis of Quasispecies and Non-Model Microbial Populations using Ultra-Deep 'Meta-barcode' Sequencing

Christopher Noune [1*], Caroline Hauxwell [2]

[1] Queensland University of Technology, Brisbane, 4000 Australia
[1*] Correspondence: chris.noune@connect.qut.edu.au; ORCID: 0000-0003-3318-5243
[2] Email: caroline.hauxwell@qut.edu.au; ORCID: 0000-0002-1681-9657

Abstract: A pipeline developed to establish sequence identity and estimate abundance of non-model organisms (such as viral quasispecies) using customized ultra-deep sequence 'meta-barcodes' has been modified to improve performance by re-development in the Python programming language. Redundant packages were removed and new features added. RAM and storage usage have been optimized to facilitate the computational speeds though coding optimizations and improved cross-platform compatibility. However, computational limits restrict the approach to barcodes spanning a maximum of 30 polymorphisms. The modified pipeline, MetaGaAP-Py, is available for download here: https://github.com/CNoune/IMG_pipelines

Keywords: Bioinformatics; Python; Baculovirus; Virology; Meta-barcode; Quasispecies;

1. Introduction

The 'Meta-Barcoding Genotyping and Abundance Pipeline' (MetaGaAP) was developed to identify and estimate abundance of strain variants within non-model populations by identification and ultra-deep sequencing of custom 'meta-barcodes' and comparison with a database of all possible polymorphisms generated from the sequence data, which was then validated through analysis of quasispecies within baculovirus isolates [1-3]. This approach facilitates analysis of viral quasispecies for which standard 'barcodes' sequence databases are not readily available. Since the original release on GitHub, the limits on cross-platform compatibility, the large number of dependencies, the high computation capacity required and reliance on Bash and R programming languages were found to limit performance. These were addressed by redevelopment in Python.

2. Method

The Python version 3.6 programming language was selected as a versatile, general purpose, high-level non-compiled language (interpreted language) which is backwards compatible with all versions of Python 3. The pipeline was fully re-coded using the Anaconda 3.6.1 and the Spyder 3.1.4 integrated developer environments [4,5] to ensure no redundant Bash or R code would be carried over.

A core set of dependencies were retained: The Burrows-Wheelers Aligner (version 0.7.15 or above), Samtools (version 1.3 or above), the Genome Analysis Toolkit (version 3.6 or above), fastx-toolkit (0.0.14 or above), Picard-tools (version 2.9 or above), Oracle Java 1.8, mawk (version 1.3.3), Sed (version 4.2 or above) and Biostars175929 that is now included as a pre-compiled version [6-12]. The number of dependencies was reduced: BBmap renamer and duplicate sequence removal tools [13], kentUtils [14], Zenity, and the R coded back-end scripts Subset_Stats.R and Seq_List.R, were discarded and replaced by pure Python implementations coded directly in the source code (Table 1). The implemented Python packages (with the exception of Biopython) are natively-installed with Python 3.6 without the need to write a separate installation script.

Table 1: Python implemented dependencies in the revised pipeline

Package/Library	Function
TKinter	Implements a simple graphical user interface for file selection.
Biopython [15]	A Python library to manipulate fasta sequences.
The Python Data Analysis Library (Pandas)	Removes the need to use R code.
Multiprocessing	A standard Python library to implement multi-threading.
Sys	Captures operating system type i.e. Windows, Linux or Mac. This package was implemented to fix an issue in multi-threading on a Linux system. However, in cases where multi-threading doesn't work, the duplicate removal step will default to a single thread.
Garbage Collection	Optimises RAM utilisation.
Getpass	Captures user information.
OS	Basic Python functions which allow Python to communicate with the operating system.
Subprocess	Python functions allowing for the execution of non-Python packages.

New features were implemented to allow for different analysis types and behind the scenes coding enhancements to improve computational efficiencies (Table 2). To distinguish the original MetaGaAP from the new Python implementation, the original pipeline was renamed as MetaGaAP-Legacy and the new Python implementation named MetaGaAP-Python (MetaGaAP-Py)

Table 2: New features added and coding enhancements

Feature/Enhancement	Function
Multi-reference, multi-sample analysis	Multiple samples with different reference sequences can be analysed at the same time, e.g. two different 'barcodes' or viruses.
Single-reference, multi-sample analysis	Multiple samples using the same reference sequence can be analysed, e.g. time-course analysis
Automatic directory creation	Directories are created at the same time as processing to reduce user interactions needed to select output directories.
RAM optimised, multi-threaded processing, Python-native duplicate sequence removal [16]	Multi-threaded duplicate sequence removal has been implemented to reduce computational time. In some-cases it may default to a single thread. Optimises RAM usage by creating a new database at the same time as duplicate removal facilitates use on systems with lower than the recommended 8 gigabytes.
Sequence combinations database compression by converting multi-line fasta sequences to a single-line fasta sequence [17]	Internal testing has found that the database memory footprint reduces when converted from a multi-line fasta to a single-line fasta file i.e. a single fasta sequence per line rather than a wrapped fasta sequence.
Automatic average sequence length and maximum read depth calculation	Automatically tells the HaplotypeCaller the maximum read depth for identification of polymorphisms and the Biostars175929 tool the sequence length required to produce the combinations database.

3. Results and Conclusions

Porting to Python and improved package selection has resulted in a highly-refined pipeline with an optimized workflow. Furthermore, the reduction in required dependencies and coding in a cross-

platform compatible language enables execution in Mac OS X, the Microsoft Windows 10 Linux Subsystem and all Linux distributions.

The introduction of multi-reference, multi-sample analysis expands the application to analyse sequence sets from multiple samples with different reference sequences at the same time, e.g. multiple samples using two different 'barcodes'. Single-reference, multi-sample analysis enables analysis of sequences from multiple samples using the same reference sequence such as time-course analysis of viral quasispecies.

Some weaknesses persist. The implemented duplicate sequence removal is dependent on the number of cores in the user's central processing unit, on whether the storage unit is a solid-state drive or a mechanical hard-disk drive, and on the size of dataset: the pipeline is functionally limited to analysis of a maximum of 30 polymorphisms across the barcode region due to the lack-of multi-threading within the Biostars175929 tool and memory requirements to store large databases. In addition, fastq files and sample names need to be re-specified when completing the final mapping stage and calculating abundance result as part of a single-reference, multi-sample analysis.

Overall the optimisations, newly implemented features, and reduced dependency requirements facilitate the use of MetaGaAP-Py, resulting in a less computationally demanding and more streamlined user interface that can be applied to generation and application of customised sequence 'barcodes' and libraries for identification and quantification of quasispecies variants in non-model populations, for which standard sequence 'barcodes' and public sequence databases are not available.

Acknowledgments: This work was funded by the Queensland University of Technology (QUT), the Cotton Research Development Corporation and an Australian Government Research Training Program Scholarship. We would like to acknowledge the support of the Invertebrate & Microbiology Group at QUT for their assistance. Some of the data reported in this paper was obtained at the Central Analytical Research Facility (CARF) operated by the Institute for Future Environments (QUT). Access to CARF is supported by generous funding from the Science and Engineering Faculty (QUT).

Author Contributions: Christopher Noune conducted the programming. Both authors contributed equally to the development of concepts and applications, and to the writing of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Software Availability: MetaGaAP-Py is available for download at https://github.com/CNoune/IMG_pipelines.

5. References

1. Noune, C.; Hauxwell, C. MetaGaAP: A Novel Pipeline to Estimate Community Composition and Abundance from Non-Model Sequence Data. *Biology* **2017**, *6*, 14.
2. Noune, C. *The Invertebrates & Microbiology Group Pipelines*, GitHub, Queensland University of Technology: https://github.com/CNoune/IMG_pipelines, 2016.
3. Noune, C.; Hauxwell, C. Comparative Analysis of HaSNPV-AC53 and Derived Strains. *Viruses* **2016**, *8*, 280.
4. Pierre, R. *Renamed Pydee to Spyder (it changes everything...!)*, GitHub: https://github.com/spyder-ide/spyder/commit/78a22a22577bbdde2c879da0429f08ad88dcff29#diff-e5fb0cda12f90dc4341247ddab54d1da, 2009.
5. *Anaconda Software Distribution*, Continuum Analytics: https://continuum.io, 2017.
6. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; Banks, E.; Garimella, K.V.; Altshuler, D.; Gabriel, S.; DePristo, M.A. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **2013**, *11*, 11 10 11-11 10 33.

7. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* **2013**.
8. Institute, B. Picard. http://broadinstitute.github.io/picard/
9. Gordon, A.; Hannon, G. Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished)* http://hannonlab.cshl.edu/fastx_toolkit **2010**.
10. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **2009**, *25*, 2078-2079.
11. Pierre, L. *JVarkit: java-based utilities for Bioinformatics*.
12. Aho, A.V.; Kernighan, B.W.; Weinberger, P.J. *The AWK programming language*. Addison-Wesley Longman Publishing Co., Inc.: 1987.
13. Bushnell, B. BBMap short read aligner. *URL* http://sourceforge.net/projects/bbmap.
14. Kent, J. *kentUtils*, GitHub: https://github.com/ENCODE-DCC/kentUtils, 2014.
15. Cock, P.J.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* **2009**, *25*, 1422-1423.
16. Cock, P.J. *BioPython Redundant Fasta Sequence Removal Function*, http://lists.open-bio.org/pipermail/biopython/2010-April/012615.html, 2010.
17. Pierre, L. *Linearize a fasta sequence*, https://gist.github.com/lindenb/2c0d4e11fd8a96d4c345#file-linearizefasta-awk, 2015.

13.3 APPENDIX C: CONFERENCE ABSTRACTS

13.3.1 Conference 1: Microbiology at QUT and Beyond Workshop, 29 October 2014

Genome Sequence of *Helicoverpa armigera* SNPV Strain AC53 and H25EA1, Isolated from Brookstead, Queensland, Australia

Christopher Noune [a], Caroline Hauxwell [a]
[a] Earth, Environmental and Biological Sciences School, Queensland University of Technology, Brisbane, Queensland, Australia

The *Helicoverpa armigera* single nucleopolyhedrovirus (HaSNPV) infects and kills *H. armigera* larvae, an important pest of cotton in Australia and worldwide. One strain, isolated from Brookstead, Queensland, Australia, and commercialised as the 'Vivus' strain, was sequenced using the next generation sequencing (NGS) platform, Ion Torrent Personal Genome Machine (PGM). An additional virus sample obtained from the University of Queensland (UQ), and believed to be a variant of this virus strain, was also sequenced. The viruses have been designated as the original accessioned nomenclature by the Queensland Department of Primary Industries (QDPI) and CSIRO.

13.3.2 Conference 2: B^3: Big Biology and Bioinformatics Symposium, 24-25 November 2014

Comparative Analysis of two Australian Baculovirus Sequences using Ion Torrent

Noune. C[1], Corander. J[2], Hauxwell. C[3]
[1] Queensland University of Technology (QUT) - Invertebrate Microbiology Group
[2] University of Helsinki, Department of Mathematics and Statistics, FIN-00014
[3] Queensland University of Technology (QUT) - Invertebrate Microbiology Group
(caroline.hauxwell@qut.edu.au)

Two single nucleopolyhedroviruses (SNPV) from *Helicoverpa spp*. with one (HaSNPV-AC53) originating in Brookstead (Queensland, Australia) and the other (HaSNPV-H25EA1) a derivative of the original isolate, were characterised. The aims of this project were to analyse a full baculovirus genome using next generation sequencing (NGS) and to determine whether the Ion Torrent Personal Genome Machine (PGM) is a viable NGS platform for baculovirus sequencing. Ion Torrent PGM™ semiconductor sequencing was used to generate *de novo* sequences of the virus genome and to identify regions of genetic diversity in the population of strains within each isolate. Sequencing identified the HaSNPV-AC53 (KJ909666) and HaSNPV-H25EA1 (KJ922128) isolates as more similar to the *Helicoverpa zea* SNPV reference genome (NC_003349) than to the four HaSNPV genomes NNg1 (NC_011354), G4 (NC_002654), C1 (NC_002654) and an additional Australian isolate (JN584482). This supports the classification of Heliothine SNPVs as a single species. Analysis of the NGS data identified hypervariable regions, recombinant DNA, deletions and insertions. The relative utility of next generation sequencing in detecting strain variation within a baculovirus isolate was assessed. Furthermore, using in-house methods which have been described in the paper, we show that the Ion Torrent is a viable NGS platform for baculovirus analysis.

13.3.3 Conference 3: 49th Annual Meeting of the Society for Invertebrate Pathology, 24-28 July 2016

Genotype Detection and Abundance within Baculoviruses using Next Generation Sequencing

Christopher Noune[1], Caroline Hauxwell[1]
[1] Queensland University of Technology (QUT), Brisbane 4001, AUSTRALIA
(chris.noune@connect.qut.edu.au)

Next Generation Sequencing (NGS) generates short 'reads' of viral sequences between 75 and 500 base pairs (bp). We have developed a method that uses the Ion Torrent PGM to detect genotypes and relative abundance within a baculovirus isolate. The software pipeline was developed to quantify baculovirus genotypes within *Helicoverpa armigera* SNPV isolate AC53. Genotypes and their relative abundance within isolates were determined from nucleotide based polymorphisms within NGS data of amplicons from BRO-A, HOAR and DNA polymerase open reading frames (ORFs). The method was subjected to a two-step validation using Sanger sequencing and analysis of relative abundance of strains derived from the parent strain by passage and plaque selection in tissue culture. The technique was then applied to determine changes in relative abundance of viral variants under selection *in vivo* and *in vitro*. Using the NGS data and an open-source software pipeline we can identify and determine genotype abundance within baculovirus populations with application to wider microbial metagenomic studies.

Genomic Analysis of Four *Plutella xylostella* Granulovirus Isolates

Robert J Spence[1], Christopher Noune[1], Caroline Hauxwell[1]
[1] Queensland University of Technology (QUT), Brisbane 4001, AUSTRALIA
(r1.spence@qut.edu.au)

Diamondback moth (*Plutella xylostella*) is considered the most destructive insect pest of cruciferous crops worldwide for which control is estimated to cost up to five billion US dollars annually. *Plutella xylostella* Granulovirus (PxGV) is a highly virulent pathogen that has co-evolved with diamondback moth and is considered a potential biopesticide. We sequenced four isolates of PxGV from Taiwan, China and Malaysia using the Illumina NextSeq 500 platform resulting in 13,186 to 23,878 X coverage. All four genomes share 99.9% sequence homology and 40.7% GC content. Sequence variability is greatest within the four homologous repeat regions which exhibit 59.8% pairwise identity. All four genomes encode two fewer putative ORFs in conserved order compared to the RefSeq Japanese isolate (NC_002593). Phylogenetic analysis shows the Japanese isolate is more homologous to a Taiwan isolate while the Chinese isolate is most divergent. Genomes of PxGV isolates share very high sequence conservation and order despite geographical separation.

13.3.4 Conference 4: AB^{3}ACBS: The Australian Big Biology, Bioinformatics and Computational Biology Conference, 1-2 November 2016

High-Resolution Termite Metagenomics

Boyd Tarlinton, Christopher Noune, Caroline Hauxwell

Queensland University of Technology

The use of meta-barcoding enables high-resolution insights into the metagenomes of various organisms that cannot be accomplished with conventional wet-lab techniques or shot-gun 'next generation sequencing' (NGS). Targeted NGS of the taxonomically significant 16S rRNA meta-barcode was used to identify and determine relative abundance of bacterial strains within the termite microbiome. A recently developed meta-barcoding software pipeline called 'IMGGAP' was used to produce a high-resolution map of these bacterial strains. This pipeline produces a sequence database based on polymorphisms identified within a dataset, enabling the production of meta-barcodes that are more highly resolved than those produced with pipelines that cluster reads based on sequence similarity. It was revealed by this analysis that the samples were dominated by bacteria of the phylum Fibrobacteres. Also, revealed by the analysis was the greater degree of bacterial diversity that exists within termites of the worker caste compared to soldier termites of the same species.

13.4 APPENDIX D: AWARDS AND GRANTS

1. The Australian Research Training Program Scholarship (2014 – 2017)
2. Cotton Research Development Corporation Post-Graduate Top-Up Scholarship (2014 – 2017)
3. Cotton Research Development Corporation International Travel Grant: The money from this grant was used to travel to the 49th Annual Meeting of the Society for Invertebrate Pathology in Tours, France – valued at AUD $2450
4. Science and Engineering Faulty Higher Degree Research Student High Achievement Award for February - 2017

13.5 APPENDIX E: MEMBERSHIP OF PROFESSIONAL SOCIETIES

1. The American Society of Microbiology
2. The Society for Invertebrate Pathology

13.6 APPENDIX F: CONTINUOUS PROFESSIONAL EDUCATION COMPLETED

1. Introduction to R – September 2015
    a. Hosted by Microsoft and Data Camp
2. Introduction to Python for Data Science – December 2016
    a. Hosted by Microsoft and Data Camp
3. Programming with Python for Data Science – June 2016
    a. Hosted by Microsoft and Coding Dojo
4. 11th Annual Winter School in Mathematical and Computational Biology – 7-11 July 2014
    a. Hosted by the Institute for Molecular Bioscience (University of Queensland) and the Australian Research Council Centre of Excellence in Bioinformatics
5. Personal Bioinformatics Tutoring – 2014-2015
    a. Tutored by Dr Stephen Rudd from the Queensland Facility for Advanced Bioinformatics (QFAB)

