Genomics of West Nile viruses from South Africa

By

Cornéll Kortenhoeven

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

(Zoology)

Department of Zoology and Entomology

Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

South Africa

Supervisors: Prof ADS Bastos

Prof C Abolnik

April 2013

ACKNOWLEDGEMENTS

My supervisors, Prof ADS Bastos and Prof C Abolnik, for their untiring support, effort, advice and continued guidance that saw to the successful completion of this work.

Dr C Potgieter, Deltamune, for his continued advice, insight, time and expertise in the planning and execution of various components of this project.

Prof F Joubert for his unreserved effort, persistence and time spent in overcoming the many technical hurdles associated with the bioinformatics analyses in this project.

The staff of the Applied Biotechnology Department, Agricultural Research Council (ARC) Onderstepoort Veterinary Institute (OVI), especially Ms Y Lotter, Dr C Ellis, Ms N Dutja for their moral support and friendship.

Dr M Romito, Applied Biotechnology Department, ARC-OVI, for his insight, technical assistance, guidance and support throughout the duration of this project.

The staff of the Virology Department, ARC-OVI, especially Ms H Tshabalala, Ms T Mlingo and Mr F Pila for providing technical assistance and after-hours support during virus isolation.

Dr A Lubisi, Virology Department, ARC-OVI, for providing laboratory facilities for virus isolation and the WNV strains used in this study.

Ms B Khoza, Virology Department, ARC-OVI, for her time, patience, expertise and support in teaching me all there is to know about cell culture.

Ms de Castro and Ms E Viljoen, Biotechnology Platform, ARC-OVI, for their insight, technical assistance and training in all aspects of Illumina sequencing.

Dr W van Wyngaardt, Immunology Department, ARC-OVI, for advice on aspects of virus isolation and cell culture.

Ms S Smith and Ms I Wright, Deltamune, for providing training in virus isolation in mice and the preparation of cell cultures during the final stages of this study.

My mother, father, sister and SB Coetzee for their unconditional love, support, consideration, encouragement and sacrifices throughout the duration of this project.

H Fysh, S & H Steenkamp, N & A Louw, J du Toit, L Snyman, M Muller, H Grobbelaar, A la Grange, RS Julius, A Lithole and the Coetzee family for their continued support and encouragement.

The ARC-OVI for providing research funding and facilities.

The South African Poultry Association (SAPA) and AgriSeta for financial support throughout the duration of this project.

Haiku (The low yellow)

The low yellow
moon above the
Quiet lamplit house.

Jack Kerouac

SUMMARY

Genomics of West Nile viruses from South Africa

by

Cornéll Kortenhoeven

Supervisor: Prof ADS Bastos

Department of Zoology

Faculty of Natural and Agricultural Science

University of Pretoria

Co-Supervisor: Prof C Abolnik

Department of Production Animal Studies

Faculty of Veterinary Science

University of Pretoria

In partial fulfilment of the degree Magister Scientiae: Zoology

West Nile Virus (WNV) forms part of the Japanese encephalitis serocomplex in the genus *Flavivirus*, family *Flaviviridae*. This enveloped positive single-stranded RNA (+ssRNA) virus is the etiological agent of West Nile fever, and in more severe cases WNV neuroinvasive disease, in both humans and animals. WNV is distributed worldwide and is phylogenetically classified into five distinct lineages. The WNV genome is ~11 Kb in length and encodes a single open reading frame (ORF) that is post-translationally cleaved into three structural proteins and seven non-structural proteins. In this study, two contemporary and two historic South African WNV strains were genetically characterised as lineage 2 strains based on complete genome sequences. Genetic change as a result of passage number and propagation system was quantified on both the consensus genome- and quasispecies level. A lack of variation was observed amongst the consensus genome sequences of WNV strains subject to changes in propagation system from BHK-21 cell culture to mouse brain and *vice versa*. In contrast, variation amongst the latter was observed on the quasispecies level. Genome-wide single nucleotide polymorphism (SNP) profiles as well as full-length haplotype sequences

reconstructed from ultra deep sequence data indicated that high levels of quasispecies diversity persists, particularly in the capsid gene region, during changes in propagation environment. The changes in frequency of variants were consistent throughout isolates propagated in different systems. The increased variation in the capsid gene region may result from selective pressures brought about by differences in host cell type between propagation systems. This study is the first to demonstrate quasispecies dynamics resulting from changes in propagation system of a lineage 2 WNV based on the reconstruction of full-length haplotype sequences from ultra deep sequence data. The approach demonstrates a cost-effective alternative to the estimation of viral population structure in light of viral evolutionary dynamics, which may in turn be assessed by the single plasmid reverse genetic system designed in this study. Although early attempts at rescuing an infectious WNV clone were unsuccessful, the system shows promise in the application of future studies concerning vaccine and diagnostic development, virulence studies and disease control.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

anchC	-	Nascent Capsid Protein
BI	-	Bayesian Inference
bp	-	Base Pairs
BHK-21	-	Baby Hamster Kidney Cells
bPCR	-	Bridge PCR
BSL 3	-	Biological Security Level 3 Laboratory
C	-	Capsid Protein
°C	-	Degrees Celsius
cDNA	-	Complementary DNA
CO_2	-	Carbon Dioxide
CS	-	Conserved Sequence Element
CRT	-	Cyclic Reversible Termination
DENV1	-	Dengue Virus Type 1
DENV2	-	Dengue Virus Type 2
DMSO	-	Dimethyl Sulfoxide
DNA	-	Deoxyribonucleic Acid
E	-	Envelope Protein
EGFP	-	Enhanced Green Fluorescent Protein
EM	-	Expectation Maximization Algorithm
EMEM	-	Eagle's Minimum Essential Medium
emPCR	-	Emulsion PCR
ER	-	Endoplasmic Reticulum
FCS	-	Fetal Calf Serum
FMDV	-	Foot and Mouth Disease Virus
HMW	-	High Molecular Weight
H_2O	-	Water
JEV	-	Japanese Encephalitis Virus
IFN	-	Interferon
kbp	-	Kilo Base Pair
kDa	-	Kilo Dalton
min	-	Minute
ML	-	Maximum Likelihood
mL	-	Milliliter
mM	-	Millimolar

MSA	-	Multiple Sequence Alignment
MTase	-	Methyl Transferase
MVE	-	Murray Valley Encephalitis Virus
NCR	-	Non-Coding Region
NEAA	-	Non-Essential Amino Acids
NJ	-	Neighbour-Joining
NQS	-	Neighbourhood Quality Standard
NS	-	Non-Structural
NTPase	-	Nucleoside Triphosphatase
ORF	-	Open Reading Frame
PCR	-	Polymerase Chain Reaction
PBS	-	Phosphate Buffered Saline
prM	-	Pre-Membrane Protein
rRT-PCR	-	Real-Time Reverse Transcriptase PCR
s	-	Second
SL	-	Stem Loop
SNA	-	Single Nucleotide Addition
SNP	-	Single Nucleotide Polymorphism
+ssRNA	-	Positive Single Stranded RNA
R_0	-	Basic Reproduction Number
RdRP	-	RNA-Dependent RNA Polymerase
RTPase	-	RNA Triphosphatase
U	-	Units
VEEV	-	Venezuelan Equine Encephalitis Virus
VSV	-	Vesicular Stomatitis Virus
WGS	-	Whole Genome Sequencing
WTA	-	Whole Transcriptome Amplification
WNV	-	West Nile Virus
YF	-	Yellow Fever Virus
µg	-	Microgram
µL	-	Microliter

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Literature Review

1.1 INTRODUCTION

Since its introduction into the Western Hemisphere during the 1990's, West Nile Virus (WNV) has been a recognised cause for concern worldwide. WNV, a member of the genus *Flavivirus,* is the aetiological agent of West Nile Fever; and in more severe cases WNV neuroinvasive disease (Hayes and Gubler, 2006). The positive-sense single-stranded RNA (+ssRNA) genome of WNV is ~11 Kb in length and encodes a single open reading frame (ORF) that is post-translationally cleaved into three structural proteins and seven non-structural proteins (Campbell *et al*., 2002). WNV is a zoonotic arthropod-borne virus transmitted primarily by *Culex sp.* mosquitoes in enzootic/epizootic cycles where migratory birds act as the reservoir host (Campbell *et al*., 2002). Even so, the incidental infection of vertebrate hosts such as humans and equines frequently occur (Nicholas, 2003).

The global increase in frequency and severity of WNV infection is consistent with the epidemiological pattern of an emerging pathogen. Genetic change in a pathogen is, amongst many factors, considered a fundamental driver of disease emergence (Woolhouse *et al*.2005). The interactions between population variants, termed quasispecies dynamics, dictate the influence of genetic change on the fitness of a consensus- or master genotype of a virus (Ciota *et al*.2007). Large population sizes coupled with high mutation rates during viral replication result in increased genetic diversity within a quasispecies (Domingo *et al*.2012).

The evolutionary change that shapes emergence is seldom observed in the consensus genotype of an arbovirus such as WNV (Domingo *et al*.2012). Population variants are subject to the adaptive constraints associated with host cycling, resulting in a reduced rate of consensus level evolution (Woolhouse *et al*.2001). However, minority variants that contribute to quasispecies dynamics without reflecting in the consensus genotype of a virus are more than often overlooked (Domingo *et al*.2012).

Recent advances in next generation sequencing and data analysis has led to an increased understanding of, amongst others, viral diversity in light of pathogenesis, virus evolution and selection pathways. The increased availability of complete genome sequences has enabled comparative studies of genome wide genetic diversity amongst viruses. Computational methods that recognise genetic diversity beyond that of the consensus genome sequence are developing at a rapid pace, enabling the reconstruction of quasispecies from next generation sequence data.

1.2.1 AETIOLOGY

1.2.1 CLASSIFICATION OF WEST NILE VIRUS *(TAXONOMY)*

West Nile Virus (WNV) is classified as a Flavivirus (family *Flaviviridae,* genus *Flavivirus*) (Van Regenmortel *et al*., 2000). Whether compared according to classical serological criteria or molecular phylogenetic techniques, a good parallel exists between the relationships of Flaviviruses (Calisher and Gould, 2003). Serological criteria more than often reflect only a small portion of the genetic

properties of a virus, whereas phylogenetic relationships based on partial or complete genome sequences provide resolution as to the true genetic relationships amongst viruses (Calisher and Gould, 2003). In theory, viruses that cross-react in neutralization tests are likely to be phylogenetically closely related (Calisher and Gould, 2003). With some exceptions, this principle holds for most Flaviviruses including West Nile Virus (WNV) (Calisher and Gould, 2003).

1.2.1.1 SEROLOGICAL CLASSIFICATION

According to classical serological criteria, members of the family *Flaviviridae* can be assigned to eight distinct antigenic complexes (Calisher *et al*., 1989). The latter includes Tickborne encephalitis, Rio Bravo, Japanese encephalitis, Tyuleniy, Ntaya, Uganda S, Dengue and Modoc (Calisher *et al*., 1989). WNV is serologically classified as a member of the Japanese encephalitis antigenic complex that includes Japanese encephalitis, Murray Valley encephalitis, Kokobera, Alfuy, Stratford, St Louis encephalitis, Usutu, Kunjin and Koutango (Calisher *et al*., 1989). The genus *Flavivirus* is furthermore divided into mosquito-borne viruses, tick-borne viruses and viruses with no known arthropod vector (De Madrid and Porterfield, 1974; Calisher *et al*., 1989). Members of the Japanese encephalitis antigenic group, and therefore WNV, are mosquito-borne viruses (De Madrid and Porterfield, 1974; Calisher *et al*., 1989).

1.2.1.2 PHYLOGENETIC CLASSIFICATION

Phylogenetic analyses demonstrated two distinct lineages of WNV strains based on complete genome sequence data (Campbell *et al*., 2002). WNV strains that form part of lineage 1 have a worldwide distribution and can be subdivided into at least three clades (Campbell *et al*., 2002, Bakonyi *et al*., 2005). WNV strains that belong to clade 1a have been isolated in Europe, Africa, United States of America and Israel (Bakonyi *et al*., 2005), whilst clades 1b and 1c consist of isolates from Australia (Kunjin) (Bakonyi *et al*., 2005) and India (Bakonyi *et al*., 2005), respectively. In contrast, WNV strains that form part of lineage 2 have been isolated in the sub-Saharan region of Africa and in Madagascar, exclusively (Lanciotti *et al*., 2002). Recent studies suggest, however, that more than two lineages of WNV can be distinguished (Bakonyi *et al*., 2005; Bondre *et al*., 2007). A third WNV lineage was proposed following the isolation of Rabensburg virus strain 97-103 (RabV 97-103) and Rabensburg virus strain 99-222 (RabV 99-22) in South Monrovia, Czech Republic in 1977 and 1999, respectively (Bakonyi *et al*., 2005). These two isolates share >99% nucleotide identity, 75%-77% nucleotide identity with representative strains from lineage 1, and 89%-90% nucleotide identity with representative strains from lineage 2, based on complete genome sequences (Bakonyi *et al*., 2005). According to a study by Bakonyi *et al*.(2005) these RabV isolates represent either a novel flavivirus of the Japanese encephalitis antigenic group, or a third lineage of WNV. In the same instance, a fourth distinct WNV lineage is proposed following the isolation of strain LEIV-Krnd88-190 in the northwest Caucasus Mountain valley, Russia in 1998 (Prilipov *et al*., 2002; L'Vov *et al*., 2004), and more

recently a fifth lineage was proposed by Bondre *et al.*(2007) following phylogenetic analysis of fifteen WNV strains isolated in India. Thirteen of these isolates formed a distinct genetic lineage differing by 20%-22% from lineage 1 and 2 strains , and by 24%-25% from lineage 3 and 4 strains, based on complete genomic sequences (Bondre *et al*., 2007). The remaining two strains were closely related to members of lineage 1, in accordance with the previously proposed lineage 1c consisting of Indian WNV strains (Bakonyi *et al*., 2005; Bondre *et al*., 2007).

1.2.2 GENOME ORGANIZATION

The mature WNV particle is enveloped, spherical and approximately 50 nm in diameter (Campbell *et al*., 2002). It consists of a host-derived lipid bilayer membrane surrounding a nucleocapsid core (Campbell *et al*., 2002). The latter contains a positive-sense single-stranded RNA genome ~11Kb in length (Campbell *et al*., 2002). The WNV genome encodes a single open reading frame (ORF) which is flanked by 5' and 3' untranslated regions (UTR) (Rossi *et al*., 2010). The ~3000 amino acid polyprotein is cleaved into three structural proteins and seven non-structural proteins (Rossi *et al*., 2010). The structural proteins are required for virion formation and include the capsid protein (C), premembrane protein (prM) and envelope protein (E) (Rossi *et al*., 2010). The non-structural (NS) proteins are required for viral genome replication and include NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5 (Rossi *et al*., 2010).

1.2.2.1 NON-CODING REGION

5' Non-coding Region (NCR)

The 5' NCR is a 95-132 nucleotide region capped with a type 1 5' cap, $m^7GpppAmpN_s$ (Brinton and Dispoto, 1988). Although not highly conserved amongst different flaviviruses, the 5'NCR contains common secondary structures influencing genome translation (Brinton and Dispoto, 1988). The latter includes a 5' stem-loop structure and a complementary region in the negative strand (Brinton and Dispoto, 1988). The 5' stem-loop is an important determinant of genome replication; while the complementary region in the negative strand of the 5' NCR provides an initiation site for positive strand synthesis during RNA replication (Lindenbach and Rice, 2003).

3'Non-coding Region

The 3' NCR which can range from 114 to 624 nucleotides in length and is highly variable amongst different flaviviruses. The latter differs significantly between tick-borne and mosquito-borne viruses (Lindenbach and Rice, 2003). Despite the high level of variability, the 3' NCR contains a number of conserved features, including a 3' stem-loop (3'SL) structure and a 3'-proximal conserved sequence element (CS1) (Hahn *et al*., 1987; Brinton and Dispoto, 1988; Lindenbach and Rice, 2003).

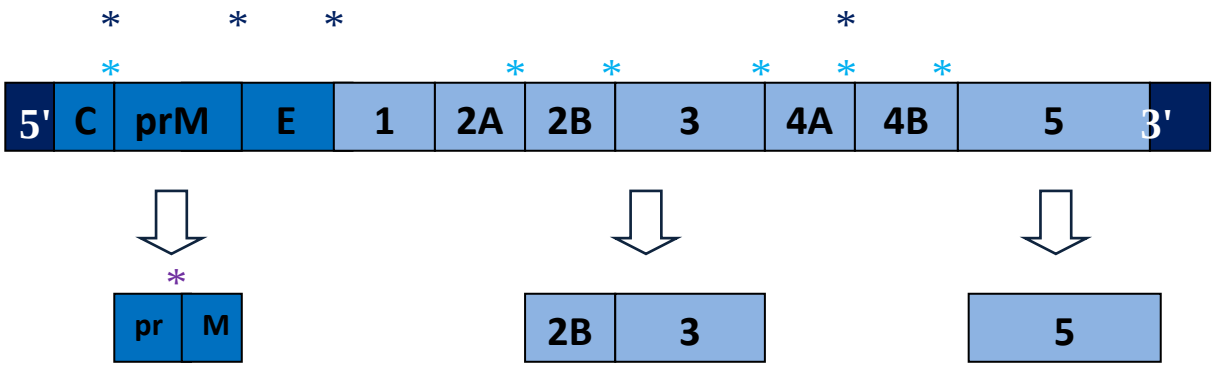

Figure 1.1 Schematic representation of the West Nile virus (WNV) genome and associated polyprotein processing. The three structural proteins and seven non-structural proteins situated on the 5' and 3' end of the genome, respectively are colour coded. The pre-membrane protein (prM) is posttranslationally cleaved resulting in mature membrane protein. Following posttranslational cleavage, non-structural (NS) proteins 2B and 3 exhibit serine protease, helicase and RNA triphosphatase activity. Similarly, NS protein 5 exhibits methyl transferase and RNA dependent RNA polymerase activity. The viral serine protease cleavage sites*, signal peptidase cleavage sites* and furin enzyme cleavage sites* are indicated accordingly.

The 3' SL structure consists of 90-120 nucleotides and is essential for virus replication through interaction with proteins of functional relevance, including NS3 and NS5 (Lindenbach and Rice, 2003). A 3'-proximal conserved sequence element (CS1) adjacent to the 3' stem-loop has been identified exclusively in mosquito-borne flaviviruses, including yellow fever (YF), dengue fever (DEN2), West Nile virus (WN) and Murray Valley encephalitis (MVE) viruses (Hahn *et al*., 1987). This sequence element has been shown to basepair with a complementary domain in the capsid protein (C), namely 5' CS, resulting in cyclization (Hahn *et al*., 1987). It has been proposed that the complementary cyclizations due to long-distance basepairing of complementary domains are essential for virus replication and the selection of RNA templates for replication. An alternative hypothesisis that genome cyclization is involved in regulating the use of RNA as templates for translation rather than RNA replication (Lindenbach *et al*., 2007). It has been shown that CS1 can basepair with an internal loop in the stalk of 3' SL forming a pseudoknot (Shi *et al*., 1996). The formation of 3' SL structures and genome cyclization acts as a conformational switch between various uses of the RNA template (Lindenbach et al., 2007).

1.2.2.2 CODING REGION

Capsid Protein (C)

The capsid protein (C) is a highly basic protein shaping the viral nucleocapsid (Lindenbach et al., 2007). It presents in two forms in infected cells, namely mature C protein (C) and nascent C protein

(anchC) (Lindenbach et al., 2007). The latter is a membrane-anchored protein with C-terminal hydrophobic residues that serve as a signal peptide for translocation of membrane protein (prM) to the endoplasmic reticulum (ER) (Lindenbach et al., 2007). Mature C protein results from cleavage of these hydrophobic residues by viral serine proteases (Lobigs, 1993). It is ~11 kDa in size and is composed of a central hydrophobic region that facilitates membrane association (Ma *et al*., 2004). The latter is surrounded by charged residues grouped at the N- and C- termini that mediate RNA interaction (Ma *et al*., 2004).

Membrane Protein (prM)

The glycoprotein precursor of M protein is translocated into the ER by the C-terminal hydrophobic domain of the C protein (Lindenbach et al., 2007). Signal peptidase cleavage of prM is delayed until viral serine protease cleaves the hydrophobic residues from anchC to form mature C protein (Lobigs, 1993). The prM protein is ~26 kDa in size and the N-terminal region contains one to three N-linked glycosylation sites (Chambers *et al*., 1990) and six cysteine residues (Nowak *et al*., 1989). The prM folds rapidly after synthesis, forming a heterodimeric complex with the E-protein (Lin and Wu, 2005). The maturation of viral particles in the secretory pathways coincides with the cleavage of prM into pr and M by host protease furin (Stadler *et al*., 1997). The early association of prM and the E protein prevents the E protein from undergoing rearrangement into the fusogenic form when transported through the secretory pathway (Guirakhoo *et al*., 1992). The co-expression of both results in the release of subviral particles from infected cells without the presence of other viral elements (Lorenz *et al*., 2002). The prM protein is cleaved by furin-like proteases during virus release; resulting in the dimerization of E proteins on the virion surface (Lindenbach and Rice, 1999).

Envelope Protein (E)

The most conserved of flavivirus structural proteins, the envelope protein (E) is the major protein found on the virion surface (Lindenbach et al., 2007). The E protein is ~53 kDa in size and mediates receptor binding and membrane fusion (Lindenbach et al., 2007). The E protein is a type I membrane protein and consists of 12 conserved cysteines forming disulfide bonds that contain up to three glycosylation motifs (Nowak and Wengler, 1987). As mentioned earlier, the folding of the E protein is dependent on co-expression with prM (Chambers *et al*., 1990).

The flavivirus envelope consists of an anchor domain and an ectodomain (Lindenbach et al., 2007). The anchor domain contains two predicted α-helical segments involved in stabilization of the prM/E interactions and functions in anchoring the E protein. The anchor domain can furthermore undergo pH-induced conformational changes (Allison *et al*., 1999). In addition to the latter, the anchor domain has two transmembrane segments that function as signal sequences for the transloaction of NS1 into the ER lumen (Mandl *et al*., 1989). The ectodomain is sub-divided into three domains. Domain I

consists of 120 residues forming a β-barrel. Domain II, or the dimerization domain, consists of 180 residues and contains flavivirus cross-reactive epitopes (Rey *et al*., 1995). Domain III consists of 92 residues and contains putative receptor binding regions (Bork *et al*., 1994).

Non-Structural Protein 1 (NS1)

The non-structural protein 1 (NS1) is ~46 kDa in size and contains 12 conserved cysteines forming disulfide bonds and 2 N-linked glycosylation sites (Lee *et al*., 1989). This glycoprotein is translocated to the ER prior to cleavage from the E protein by host signal peptidase; whereas the NS1/NS2A junction is cleaved by an unknown ER resident host enzyme (Falgout *et al*., 1991). Despite the largely hydrophilic amino acid content and lack of transmembrane domains, NS1 forms stable homodimers with a high membrane affinity (Winkler *et al*., 1989).

Although the function of NS1 is unclear, it is thought that it plays a pivotal role in RNA replication and virus production (Muylaert *et al*., 1996). It has been shown that mutation in the N-glycosylation sites of NS1 negates RNA replication and virus reproduction (Muylaert *et al*., 1996). In the same instance trans-complementation studies indicate that NS1 is involved in the earliest stages of RNA replication (Khromykh *et al*., 1999) and that correct replicase function requires an interaction between NS1 and NS4A (Lindenbach and Rice, 1999).

NS1 is mostly maintained within infected cells, but has been shown to relocate to the cell surface prior to being secreted from mammalian cells (Lindenbach and Rice, 2003). This secreted NS1 assembles into hexameric particles consisting of three dimers held together by hydrophobic interactions, approximately 11 nm in size (Crooks *et al*., 1994). Although the function of this form of NS1 is currently unknown, it has been established that secreted NS1 is compartmentalized to late endosomes by hepatocytes (Alcon-LePoder *et al*., 2005).

Non-Structural Protein 2A (NS2A)

The non-structural protein 2A (NS2A) is a poorly conserved hydrophobic protein, ~22 kDa in size. This protein results from cleavage at the NS1/NS2A junction by an unknown ER-resident host enzyme and cytosolic cleavage at the NS2A/NS2B junction by serine protease (Falgout and Markoff, 1995). NS2A interacts with replicase components in subcellular sites of virus RNA replication and is thought to coordinate the shift between RNA packaging and RNA replication (Khromykh *et al*., 2001). It has furthermore been shown that WNV NS2A inhibits interferon (IFN) signalling by acting as an IFN antagonist (Liu *et al*., 2006).

Non-Structural Protein 2B (NS2B)

The non-structural protein 2B (NS2B) is a small membrane-associated protein, ~15 kDa in size. This

protein acts as a cofactor for the NS2B-NS3 serine protease by forming a stable complex with NS3 (Falgout *et al*., 1991). Interaction between NS2B and NS3 is regulated by conserved hydrophilic domains flanked by hydrophobic regions in NS2B; leading to co-translational insertion of the NS2B-NS3 precursor into the ER membranes (Clum *et al*., 1997). It is thought that NS2B might be involved in adjusting membrane permeability during infection (Chang *et al*., 1999).

Non-Structural Protein 3 (NS3)

The non-structural protein 3 (NS3) is a highly conserved multifunctional protein ~70 kDa in size. This protein is tri-functional, exhibiting protease, helicase and RNA triphosphatase (RTPase) activites (Chambers *et al*., 1991; Chambers *et al*., 1993). The latter activities are required for both polyprotein processing and RNA replication (Lindenbach et al., 2007).

The N-terminal of NS3 is the catalytic domain of the NS2B-NS3 serine protease complex. This catalytic domain is not only responsible for cleaving NS2A/NS2B, NS2B/NS3, NS3/NS4A and NS4B/NS5 junctions, but also produces the C-termini of the mature C protein as well as NS4A (Amberg *et al*., 1994) (Lin *et al*., 1993).

The C-terminal of NS3 encodes for a protein that is homologous to the supergroup 2 RNA helicases (Gorbalenya *et al*., 1989) and plays an essential role in viral replication (Matusan *et al*., 2001). The helicase structure is comprised of three subdomains, two conserved domains that are involved in nucleoside triphosphate hydrolysis and a C-terminal domain involved in virus-specific RNA and protein recognition (Lindenbach et al., 2007). It has been shown that RNA unwinding activity and RNA-stimulated nucleoside triphosphatase (NTPase) activity of NS3 is essential for viral replication (Matusan *et al*., 2001). The NS3 RNA triphosphatase (RTPase) activity is thought to dephosphorylate the 5' end of the flavivirus genome prior to cap addition (Wengler and Wengler, 1993).

It was recently suggested that RTPase requires NS3 C-terminal sequences; but also makes use of a motif in the helicase-NTPase catalytic core for phosphodiester bond hydrolysis (Bartelma and Padmanabhan, 2002). Suffice to say that all three functions of NS3 (NTPase, helicase and RTPase) require a common active centre (Lindenbach et al., 2007). It has been shown that interactions between NS3 and NS5 coordinates all three functions, where NS3 binds to the 3' SL together with NS5 resulting in enhanced NTPase activity (Chen *et al*., 1997).

Non-Structural Protein 4A (NS4A) and Non-Structural Protein 4B (NS4B)

Both non-structural proteins 4A (NS4A) and 4B (NS4B) are hydrophobic proteins with a size of ~16 kDa and ~27 kDa, respectively. While NS4A has been shown to co-localize with replication complexes, NS4B co-localizes with NS3 in sites of RNA replication (Mackenzie *et al*., 1998; Miller *et al*., 2006). Both NS4A and NS4B are therefore associated with the viral polymerase complex

(Lindenbach and Rice, 2003).

Non-Structural Protein 5 (NS5)

The non-structural protein 5 (NS5) is a highly conserved multifunctional protein ~103 kDa in size. It exhibits methyltransferase (MTase) and RNA-dependent RNA polymerase (RdRP) activities (Lindenbach et al., 2007). The structure of the N-terminal region of NS5 is homologous with other MTase proteins, suggesting that NS5 is involved in methylation of the 5'cap (Koonin, 1993). Alternatively, the C-terminal region of NS5 is homologous with other RdRPs (Rice, 1985).

1.2.3 REPLICATION CYCLE

The flavivirus replication cycle can be described in three intermediate steps, including binding and entry into the host cell, viral genome replication, and the assembly and release of viral particles from infected host cells (Lindenbach et al., 2007). The replication cycle is initiated when viral glycoproteins interact with cellular receptors, resulting in receptor mediated endocytosis (Lindenbach and Rice, 2003). Fusion between viral particles and host-cell membranes results in the uncoating of the nucleocapsid and release of the viral RNA genome into the host cytoplasm (Gollins and Porterfield, 1985; Gollins and Porterfield, 1986). The viral genome is translated into a single polyprotein and processed by viral serine protease and host cell proteases resulting in mature viral proteins (Lindenbach and Rice, 2003). RNA synthesis is semi-conservative and asymmetric, and a negative strand intermediate serves as a template for the generation of positive strand genomic RNA copies (Lindenbach and Rice, 2003). The latter are produced by viral RNA-dependent polymerase, NS5 and other cellular proteins (Lindenbach and Rice, 2003).Virus particles are assembled and released by budding into the rough endoplasmic reticulum (ER), where nascent particles pass through the host secretory pathway (Lindenbach and Rice, 2003). Virion maturation occurs during exocytotic release, completing the flavivirus replication cycle (Lindenbach and Rice, 2003).

1.2.4 NATURAL TRANSMISSION AND HOSTS

The genetic and antigenic relatedness among flavivirues reflects vector preference, where viruses of the Japanese encephalitis antigenic group are classified as mosquito-borne arboviruses (Calisher *et al*., 1989; Billoir *et al*., 2000). WNV, in turn, is transmitted primarily by *Culex* spp. mosquitoes; and the transmission occurs in enzootic/epizootic cycles (Campbell *et al*., 2002). *Culex* spp. mosquitoes acquire WNV from migratory birds that act as amplifying- or reservoir hosts (Gubler *et al*., 2007). Mosquitoes are infected by ingestion of WNV with a blood meal from a viraemic bird; followed by an extrinsic incubation period of eight to fourteen days (Gubler et al., 2007). The amplification cycle continues as infected mosquitoes feed on the avian host (Gubler et al., 2007).

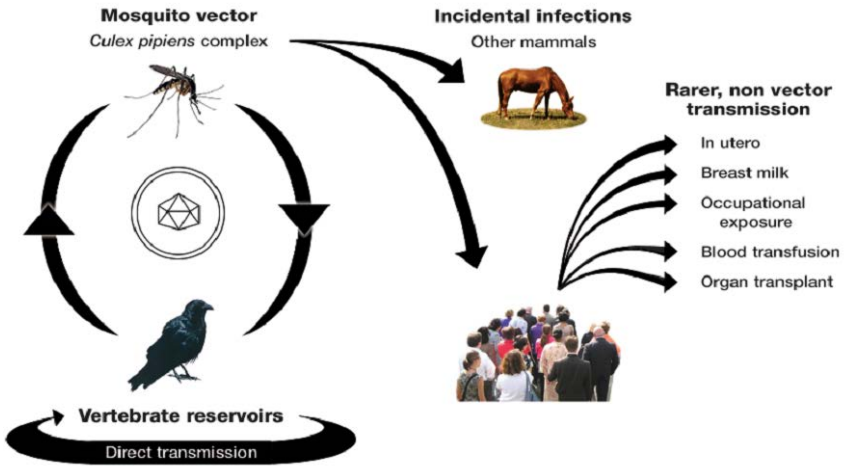

Figure 2.2 Natural transmission cycle of West Nile Virus (WNV). Primary transmission takes place between *Culex spp.* mosquitoes and the avian host, where migratory birds serve as the reservoir host. Incidental infections of non-avian vertebrate vectors commonly occur, including the infection of humans and horses. Non-vector transmission of WNV includes, amongst others, infection *via* occupational exposure and blood transfusion. *Figure adapted from Gould and Fikrig, 2004.*

Although non-avian vertebrates do not participate in the primary transmission cycle, infection of incidental hosts frequently occurs. WNV infection has been documented in a number of vertebrate hosts, including humans, canines, felines, equines, ungulates, bats, rodents, sea mammals and alligators (Nicholas, 2003). The latter are considered "dead-end'' hosts since the viral concentration in the infected vertebrate is inadequate to infect a feeding naive mosquito with WNV (Rossi *et al*., 2010). Although rare, various alternative non-vector transmission routes have been reported. These include *in-utero* infection, infection through breast milk, occupational exposure, blood transfusions and organ transplants (Gould and Fikrig, 2004).

1.3 EPIDEMIOLOGY OF WNV

1.3.1 GEOGRAPHIC DISTRIBUTION OF WNV

WNV is considered the most widespread of flaviviruses, with a geographic distribution across Africa and Eurasia (Hubálek and Halouzka, 1999). WNV strains that form part of lineage 1 have a worldwide distribution (Lanciotti *et al*., 2002; Bakonyi *et al*., 2005). Strains that belong to clade 1a have been isolated in Israel, the United States, Europe and Africa (Bakonyi *et al*., 2005). Strains that belong to clade 1b have been isolated in Australia, and those of clade 1c have been isolated in India (Bakonyi *et al*., 2005). Lineage 2 WNV strains have been isolated in the sub-Saharan region of Africa and Madagascar exclusively (Lanciotti *et al*., 2002).

Recently distinguished lineage 3 isolates have been isolated in the Czech Republic, and those of lineage 4 have been isolated in Russia (Prilipov *et al*., 2002; L'Vov *et al*., 2004; Bakonyi *et al*., 2005). WNV strains that represent lineage 5 have been isolated in India, and can be distinguished from Indian isolates that belong to lineage 1 clade c (Bakonyi *et al*., 2005; Bondre *et al*., 2007).

1.3.2 EMERGING INFECTIOUS DISEASE

Besides epidemics of West Nile fever in Israel during the 1950's (Marberg *et al*., 1956), and two large epidemics in South Africa in 1974 and 1984 (Jupp *et al*., 1986; Burt *et al*., 2002); little WNV activity was reported prior to the 1990's worldwide. Although WNV is considered endemic to South Africa, epidemics have been restricted to episodes of unusually high rainfall, high summer temperatures and an increased density in *Culex* sp. mosquitoes (Hayes, 2001).

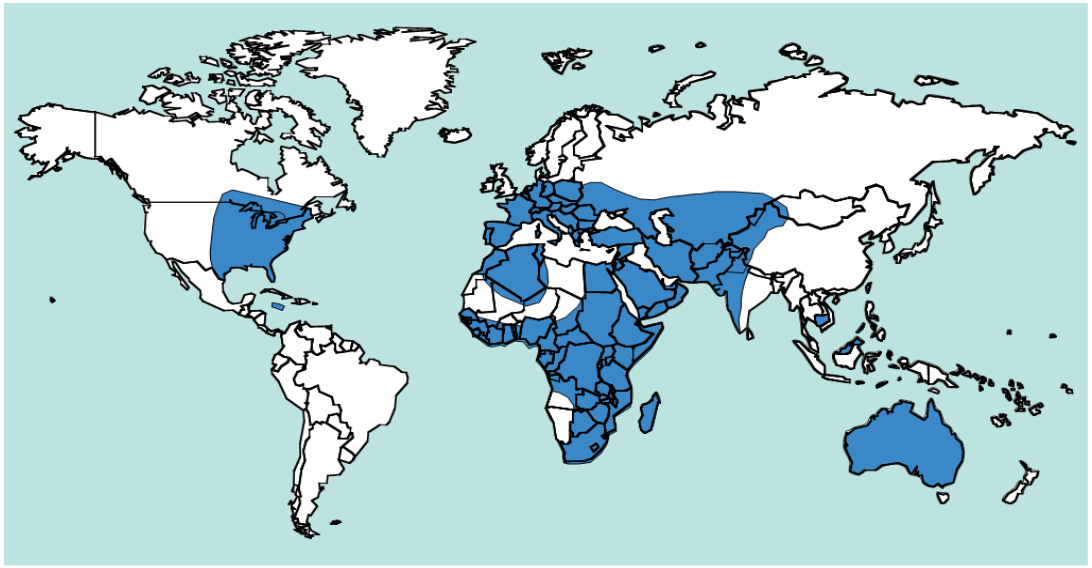

Figure 1.3 Worldwide geographic distribution of West Nile Virus (WNV). *Image adapted from Campbell et al.2002*

One of the largest WNV epidemics ever to be reported occurred in the Karoo region of the Northern Cape Province, South Africa in 1974 (Jupp *et al*., 1986; Burt *et al*., 2002). A second epidemic occurred in 1984 in the Witwatersrand area of the Gauteng province (Jupp *et al*., 1986; Burt *et al*., 2002). In both incidences patients presented with mild febrile disease and no cases of neurological disease or associated fatalities were recorded (Jupp *et al*., 1986; Burt *et al*., 2002). Suffice to say that the early transmission of WNV in the Old World was associated with infrequent outbreaks of mild dengue-like illness and reports of neurological disease and fatalities were rare (Hayes and Gubler, 2006).

During the 1990's, however, this epidemiological pattern changed (Hayes and Gubler, 2006). WNV epidemics worldwide became associated with an increased incidence of severe and fatal neurologic disease in both humans and horses, as well as birds (Hayes and Gubler, 2006). Disease caused by WNV infection in South Africa since the 1990's has been recognized as potentially severe and fatal (Venter *et al*., 2009). A number of human infections resulted in severe nonfatal encephalitis and fatal hepatitis (Venter *et al*., 2009). The global distribution of WNV was compatible with the north-south bird migratory pathways; and outbreaks occurred concurrently with epizootics in birds (Hubalek, 2000). Epidemics started to occur in Algeria (1994), Morocco and Romania (1996), Tunisia and Israel (1997), Italy and Israel (1998), and Israel, Russia and the United States (1999) (Gubler et al., 2007). Following introduction into the Western Hemisphere in 1999, WNV has spread rapidly throughout the United States, South America, the Caribbean, Canada and Mexico (Wertheimer, 2012).

The global increase in frequency and severity of WNV infections corresponds to the epidemiological pattern followed by emerging pathogens (CDC, 2012). According to Woolhouse *et al.*(2005), an emerging pathogen can be defined as the etiological agent of an infectious disease of which there is an increased incidence following its introduction to a new host population. Alternatively, an increased incidence in infectious disease in an existing host population may be the result of long-term changes in the emerging pathogen's epidemiology (Woolhouse, 2002). In order to understand the epidemiology and evolutionary biology that underlies the emergence of infectious disease, a brief overview of pathogen population dynamics is necessary.

1.3.3 PATHOGEN POPULATION DYNAMICS

In order to evaluate the spread of infection through a host population, epidemiological theory quantifies the transmission potential of a disease by means of the basic reproduction number (R_0) (Woolhouse *et al*., 2001). R_0 can be described as the number of secondary cases of infection generated by a single primary case that was introduced into a population of unexposed individuals (Combes and Théron, 2000). The expected size of an outbreak is therefore dependent firstly; on the number of primary cases of infection and secondly; on the transmission potential expressed in terms of R_0 (Woolhouse *et al*., 2005). The latter, in turn, defines an important epidemic threshold (Woolhouse *et al*., 2005).

If R_0 is <1 in the new host population, each primary case of infection will fall short of replacing itself, resulting in nothing more than a minor outbreak (Woolhouse *et al*., 2005). If R_0 is >1, however, each primary infection is capable of generating more than one secondary infection; increasing the probability of a major epidemic to occur (Woolhouse *et al*., 2005). A transition between these

scenarios is set to occur when $R_0 \approx 1$, where the expected size of the outbreak is highly responsive to minor changes in R_0 (Woolhouse *et al*., 2005).

This observation, in light of emerging pathogens, implies that minor changes in R_0 may influence the incidence of infection to a great extent (Woolhouse *et al*., 2005). This change in R_0 can be attributed to a variety of different factors. According to Woolhouse *et al*.(2005) these factors include, firstly; genetic changes in the pathogen, secondly; immunocompromised hosts, and thirdly; changes in host-pathogen ecology. The latter includes changes in host demography and movement; natural environment; and climate and land use (Woolhouse *et al.*, 2002). Results from various independent studies more than often reflect these factors as drivers of the recent WNV emergence. For the purpose of this discussion, however, genetic changes affecting the emergence of WNV will be considered.

Genetic change is mostly studied on the level of the consensus genome. Few studies account for the underlying population variation that contributes to changes in the consensus genome. This underlying population variation, termed quasispecies, ultimately determines the biological attributes of a virus.
In order to comprehend the genetic change that underlies changes in the epidemiological pattern of WNV, it is necessary to review quasispecies theory.

1.3.4 QUASISPECIES THEORY

A viral quasispecies is defined as an RNA population composed of a diverse mutant spectrum surrounding a consensus- or master genotype that confers the highest fitness (Domingo *et al*., 1998). Quasispecies evolution is the result of evolutionary events that target the components of the mutant spectrum instead of the consensus genotype (Domingo, 2002). As a result, changes in quasispecies composition may take place without modification of the consensus genotype. Large population sizes coupled with high mutation rates during viral replication result in increased genetic diversity within the mutant spectrum (Domingo, 2002). The genetic variation generated in this manner more than often determines the biological behaviour of a viral population. The continuity of this genetic variation, however, is subject to various evolutionary processes and constraints. It is these evolutionary processes and constraints will be discussed in the sections to follow.

Genetic Variation and Error Threshold

In light of genetic variation, a viral quasispecies can be described as distributions of non-identical but related genomes continuously subject to genetic change (Domingo *et al*., 2006). The generation of genetic variation can be ascribed to mutation, molecular recombination, gene duplication and genome segment reassortment (Domingo *et al*., 2012). For WNV, in particular, there is minimal

recombination (Pickett & Lefkowitz 2009; Taucher *et al.*, 2010) and the feature of a single open-reading frame makes it incapable of reassortment. Mutation is the key mechanism for the generation of genetic variation in WNV populations (Taucher *et al.*, 2010), and will be the focus of this discussion. RNA viruses have a mutation rate of approximately one mutation for every 10^4 bp of nucleotides replicated (Drake and Holland, 1999). Mutational change of RNA viruses can be attributed to high error rates and the lack of proofreading ability of RNA-dependent RNA polymerase (Holland *et al.*, 1982a). When coupled with large population sizes, the high rate of mutation results in ample genetic variation.

Despite the high genetic variation potential, quasispecies theory accommodates the stable conservation of genetic information in the form of an error threshold. The latter stipulates that for any given amount of genetic information conveyed during replication, there is a maximum error rate that accompanies the maintenance of that genetic information (Eigen and Schuster, 1979). This error threshold is dependent on, firstly, the replication accuracy; and secondly, the fitness of the consensus genotype relative to the mean fitness of the error copies (Eigen and Schuster, 1979).

Viral Fitness and Selection

Viral fitness is a measure of the relative replication capacity of a virus (Domingo and Holland, 1997). Consider an actively replicating virus as a population, where variant genomes represent individual components of this population. In comparing the consensus nucleotide sequence of the virus population with that of the individual components of the population, it becomes apparent that the frequency of individual components varies. It is this variation in frequency that dictates relative fitness amongst individual components, or variant genomes. Variant genomes, or collectively the mutant spectrum, are therefore considered the unit of selection (Domingo *et al.*, 2006).

The continuity of newly acquired genetic variation is determined by the outcome of selective pressures on variant genomes, or components of the mutant spectrum. Positive selection refers to the process where a variant genome becomes dominant in an evolving population as a result of a higher relative fitness value in comparison with other variant genomes. Conversely, negative selection dictates the maintenance of a variant genome at a low frequency in a population or the elimination thereof due to a reduced fitness value. Low frequency or minority variants are often subject to negative selection at different intensities, emphasising the need to quantify the selective advantage of variant genomes through fitness measurements (Domingo *et al.*, 2012).

Viral quasispecies interact competitively towards attaining a mutation-selection equilibrium in a given environment (Wilke, 2005). That is, the frequency at which a variant reaches an equilibrium value

that reflects a balance between the effects of mutation and selection. An infected host is essentially a mosaic of different environments including cells, tissues, organs and associated physiological conditions (Domingo and Holland, 1997). Although in need of additional research, tissue-specific viruses are expected to show an increase in fitness when replicating in the respective tissue as infection advances (Domingo *et al*., 2012). In contrast, the replication of viruses that are not tissue-specific is thought to result in compartmentalization of fitness values (Domingo *et al*., 2012). In addition to being influenced by environment, viral fitness is inherently dependent on population size (Domingo and Holland, 1997; Quinones-Mateu and Arts, 2006).

Environment

A drastic reduction in population size such as successive bottleneck events stochastically reduces variability resulting in a pattern of decreased viral fitness (Brackney *et al*., 2011). As suggested by Muller's ratchet theory, a small population size and high mutation rate results in the irreversible incorporation of deleterious mutations unless compensatory mechanisms restore the variant genomes to a mutation free state (Muller, 1932, Muller, 1964). Although studies of vesicular stomatitis virus (VSV) confirm the consequences of Muller's ratchet (Duarte *et al*., 1993), studies of foot and mouth disease virus (FMDV) illustrate the resistance of extinction as a result of compensatory mutations (Lázaro *et al*., 2003). In contrast, it was shown that dengue virus type 1 (DENV1) (Aaskov *et al*., 2006) and Venezuelan equine encephalitis virus (VEEV) (Smith *et al*., 2008) is not subject to population bottlenecks during natural transmission due to the persistence of putatively defective genomes through complementation. According to Brackney *et al*.(2011), WNV populations are not prone to undergoing bottleneck events during the enzootic transmission cycle. It was furthermore shown that although WNV is subject to compartmentalization, viral populations do not undergo bottleneck events due to anatomical barriers (Brackney *et al*., 2011).

Population Equilibrium

Classical population genetics refers to variation and constancy in terms of consensus genomic sequences (Domingo et al., 2012). The majority of standard sequencing techniques routinely used in virology result in the consensus nucleotide sequence of a viral population (Domingo, 2007). The latter is commonly used in virus identification, phylogenetic studies and the interpretation of evolutionary pressures including selection and drift (Domingo, 2007). The lack of variation in consensus nucleotide sequences is often interpreted as an absence of mutations, and the underlying variation in the mutant spectrum that contributes to the consensus nucleotide sequence is neglected (Domingo *et al*., 2012). The absence of variation does not equate to evolutionary stasis, but rather a state of population equilibrium resulting in a constant consensus genome sequence (Holland *et al*., 1982b; Domingo and

Holland, 2005). The rate of evolution for a consensus nucleotide genome of a virus that is in population equilibrium is approximately 10^{-4} substitutions per site per year or lower (Domingo *et al*., 2012).

In contrast, quasispecies evolution can be seen as the disequilibrium of a mutant spectrum (Domingo *et al*., 2012). Positive selection and stochastic events change the frequency of variant genomes in the population, thereby replacing previous distributions of the mutant spectrum (Domingo *et al*., 2012). The latter reflects as a change in the consensus nucleotide sequence of the viral population, and the continuity thereof depends on the ability of the new distribution to again reach population equilibrium (Domingo *et al*., 2012). The ability to reach population equilibrium is a function of the degree of variance in relative fitness between the parental and progeny distributions (Domingo *et al*., 2012). This concept is well illustrated in the adaptation of a natural virus isolate to cell culture. The growth of subpopulations in a new environment results in a modification of the viral population regardless of the adaptation time (Domingo et al., 2012). Rapid selection of variant genomes takes place following a sudden increase in variation in a mutant swarm *in vivo*, and rates of evolution of up to 10^{-2} substitutions per site per year can be attained (Brown *et al*., 1999; Gebauer *et al*., 1988).

The mutant spectrum therefore has the capability of changing the consensus genome at a rapid rate of evolution of 10^{-2} substitutions per site per year during positive selection; or to keep the consensus genome constant at a reduced rate of evolution of 10^{-4} substitutions per site per year during population equilibrium (Domingo *et al*., 2012). It should be noted that a state of population equilibrium does not imply mutation rates lower than mutation rates operating in genome replication. Similarly, rapid rates of evolution do not imply above-average mutation rates.

Adaptability and Sequence Space

Sequence space is defined as the theoretical representation of all variant possibilities of a sequence, i.e. nucleotide or amino acid sequences (Eigen, 1971; Eigen and Schuster, 1979). The theoretical sequence space of a viral genome of 10 kbp equates to approximately $4^{10\,000}$ (Eigen, 1971; Domingo *et al*., 2001). In reality, the sequence space occupied by a viral genome is decisively less than theoretically predicted due to functional and biological restrictions (Domingo *et al*., 2012). The latter include sequence-dependent regulatory signals encoded by the genome, secondary and tertiary structures in RNA and open reading frames encoding functional proteins which interact with viral and host proteins (Domingo *et al*., 2012). The amount of sequence space occupied by any virus at any time during replication is dependent on the environment in which the virus is replicating, the number of mutations necessary for relevant phenotypic traits and virus fecundity and turnover (Schneider and Roossinck, 2001).

The mutant spectrum of a viral population can be described as a cloud within sequence space that either shifts position in response to selective forces or establishes a unique cloud following a bottleneck event (Domingo *et al*., 2012). Sequence space is highly connected and genotypes are separated by a variable number of point mutations that never surpasses the length of the consensus genome (Domingo *et al*., 2012). Movements within sequence space are either guided by a fitness gradient when two points in sequence space are separated by few mutations, or by mere mutational pressure when separated by a single mutation (Eigen *et al*., 1988).

The assignment of a fitness value to positions in sequence space results in a fitness landscape (Wright, 1931). With regards to viral quasispecies, it is important to emphasise that fitness landscapes are brief and rugged. These features are a result of the rarity of neutral mutations in compact viral genomes, as well as the variability of environment (Domingo *et al*., 2001). The breadth of a mutant spectrum is dependent on the mutation rate, tolerance to mutations and population size (Domingo et al., 2012). In addition, the latter determines the landscape of minority variants and their relative frequencies (Domingo, 2000). A broad mutant spectrum is associated with higher accessibility to points in sequence space (Domingo, 2000). With higher accessibility comes higher adaptability as a result of fitness gain (Domingo, 2000).

A broad mutant spectrum furthermore contributes to overcoming genetic and phenotypic barriers during selection (Domingo and Holland, 1997). The limitations that a virus must defeat in order to attain a required phenotype are collectively termed ‘genetic barriers’ (Domingo and Holland, 1997). An amino acid replacement, for example, has a low genetic barrier should only one mutation be required to complete the replacement (Domingo and Holland, 1997). Should two mutations be required, the replacement would be considered to have a high genetic barrier (Domingo and Holland, 1997). Conversely, the term phenotypic barrier refers to the fitness cost of an amino acid substitution to the viral population (Domingo and Holland, 1997).

Molecular Memory

The fitness gains associated with positive selection during replication permits the maintenance of viral population components in the form of memory genomes (Ruiz-Jarabo *et al*., 2000). The latter can remain present in the viral population at higher frequencies than permitted by basal mutation rates without contributing to the consensus genome (Ruiz-Jarabo *et al*., 2000). Memory genomes are thought to represent a dominant component of the viral populations at earlier phases of the same evolutionary lineage (Domingo *et al*., 2012).

When considering the mutant spectrum as an entity, molecular memory is the outcome of quasispecies dynamics. Memory is eliminated subsequent to population bottleneck events (Domingo, 2000; Ruiz-Jarabo *et al*., 2000); and the frequency at which a memory genome is maintained is dependent on the relative fitness of the population to which it belongs (Ruíz-Jarabo *et al*., 2002). Studies have shown that memory genomes gain fitness similar to genomes that frequently present in the viral population (Arias *et al*., 2001). The latter confirms the Red Queen hypothesis where the most fit genomes are subject to continuous selection irrespective of whether it is an artefact of molecular memory or part of the viral population (Van , 1973).

Cell Tropism and Host Range

Mutational change in structural and non-structural regions of the viral genome markedly influences cell tropism and host range (Domingo and Holland, 1997). The ability of a virus to cause infection depends on the recognition of cell surface receptors and intracellular host factors that permit virion multiplication and release (Domingo and Holland, 1997). Although genetic and phenotypic barriers provide the parameters associated with changes in cell tropism or host range, the evolutionary dynamics of viral subpopulations that accompany these changes are poorly understood (Domingo *et al*., 2012).

Mutations that lead to amino acid substitutions at external residues of surface proteins persist in the form of minority genomes (Domingo, 2000). The latter encode alternative receptor recognition sites capable of transforming cell tropism or host range upon selection (Domingo, 2000). A large number of mutations and amino acid replacements are required to alter receptor recognition sites, affording changes in cell tropism and host range a high genetic barrier (Domingo and Holland, 1997). Regardless of this high genetic barrier, the impending positive effects on viral fitness accommodate for the subsistence of these minority variants in the viral population (Domingo and Holland, 1997).

However, the successful establishment of a receptor-virus interaction depends on the relative fitness beyond that which alternative receptor recognition affords the viral population (Domingo and Holland, 1997). The amino acid substitutions required to change a receptor recognition site might negatively affect other traits such as surface protein or capsid stability (Domingo and Holland, 1997). As such, epistatic gene interactions may result in large phenotypic barriers that impede positive selection of these minority variants (Domingo and Holland, 1997).

Antigenic Drift and Escape Mutants

The term antigenic drift describes the accumulation of mutations in the antibody binding sites of a virus. The latter permits mutants to escape the selective constraints of the immune response and continue to replicate in a host system. Population variants that continue to replicate in this manner are referred to as viral escape mutants and play an important role in the establishment and maintenance of viral persistence (Aebischer *et al.*, 1991) (Ciurea *et al.*, 2000)Gebauer et al., 1988).

Long-Term Virus Evolution

Quasispecies dynamics is a collective term that describes the long-term evolutionary pattern of viruses as determined by the formation and composition changes of mutant spectra. Interactions amongst components of the mutant spectrum determine the biological behaviour and phenotypic traits of a virus, and subsequently modulate the genetic diversity that is transmitted from infected hosts to susceptible hosts. Features that depend on interactions between components of the mutant spectrum include, amongst others, molecular memory, the outcome of selective pressures on diversity, complementation and interference. Although this review discussed these features as separate entities in light of short-term evolution, it should be kept in mind that these processes and interactions occur simultaneously and collectively affect the long-term evolution of a virus.

1.4 REVERSE GENETIC SYSTEM

Classical genetics, or forward genetics, aims to characterise the genetic basis of a phenotypic trait. A forward genetic approach would involve the mutagenesis of a population followed by a genetic screen for a specific phenotype. The transmission of the phenotype is confirmed and mapped to a locus through genetic crosses. The gene responsible for the observed phenotype is isolated and sequenced. Conversely, reverse genetics aims to determine the function of a gene or subset of genes by characterising the phenotypic traits that arise subsequent to manipulation. The application of reverse genetic systems in virology has led to considerable insights into viral replication and pathogenesis resulting from the genetic manipulation of functional complementary DNA (cDNA) clones (Ruggli and Rice, 1999). The general strategy for creating a functional cDNA clone for positive-sense single-stranded RNA (+ssRNA) virus depends on the infectious nature of the genome when transfected into susceptible host cells (Baltimore, 1971).

The cDNA template design imitates the structure of the viral genomic RNA with the addition of genetic manipulations to create a recombinant virus (Ruggli and Rice, 1999). The complete cDNA template is propagated as a plasmid and transcribed *in vitro* (Ruggli and Rice, 1999). The RNA

transcribed from the cDNA clone is transfected into host cells and a recombinant infectious virus is rescued from cell culture (Ruggli and Rice, 1999).

Reverse genetics is considered one of the most powerful tools in modern virology (Ruggli and Rice, 1999). The use of functional cDNA clones in the study of a virus of interest has been applied to many facets of research in virology. More specifically, infectious cDNA clones of WNV have been applied to studies of virus replication and infectivity (Fayzulin *et al*., 2006; Yamshchikov *et al*., 2001), virulence determinants (Audsley *et al*., 2011), virus immune regulation and pathogenesis (Suthar *et al*., 2012), pathogenesis (Shi *et al*., 2002), drug discovery and vaccine development (Puig-Basagoiti *et al*., 2005), and diagnostic assay development (Davis *et al*., 2001). Moreover, infectious WNV cDNA clones have been used extensively in the study of quasispecies dynamics (Ciota *et al*., 2007; Jerzak *et al*., 2008; Fitzpatrick *et al*., 2010).

2.5 NEXT GENERATION SEQUENCING

Conventional Sanger sequencing methods have almost exclusively been implemented for DNA sequence generation since the 1990's (Shendure and Ji, 2008). High-throughput production pipelines that enable the generation of whole genome sequence data follow one of two approaches, *viz.* shotgun *de novo* sequencing and targeted resequencing (Shendure and Ji, 2008). Despite contributing to numerous monumental accomplishments in the generation of genome sequence data over the past two decades, the limitations of Sanger sequencing emphasised the need for improved sequencing technologies (Metzker, 2010). The time constraints and high costs associated with generating whole genome sequence data with Sanger sequencing methods led to the development of next-generation sequencing strategies, enabling the production of large volumes of sequence data at a reduced cost (Metzker, 2010).

Next-generation sequencing technologies is a collective term that describes methods of template preparation, sequencing and imaging, and data analysis (Metzker, 2010). Each sequencing platform is distinguished based on the unique combination of protocols used, resulting in the production of different types of data (Metzker, 2010). The latter dictates the quality and cost associated with each commercial next-generation sequencing platform, as well as the relevance of the data generated to a specific study (Metzker, 2010). Commercially available next-generation sequencing platform technologies include Roche 454, Illumina, Life Technologies SOLiD, Helicos Biosciences Heliscope and the Pacific Bioscience RS system (Metzker, 2010). For the purposes of this discussion, the Illumina platform will be emphasized.

The numerous applications of next-generation sequencing technologies illustrate an unprecedented

capacity to cater for a broad range of research interests (Shendure and Ji, 2008). Complete genome sequencing and resequencing, reduced representation sequencing and targeted genomic resequencing made way for comprehensive genome-wide or targeted polymorphism and mutation discovery (Albert *et al.*, 2007; Dahl et al., 2007; Hodges *et al.*, 2007; Okou *et al.*, 2007; Porreca *et al.*, 2007; Van Tassell *et al.*, 2008; Wheeler *et al.*, 2008). Paired end sequencing enabled the discovery of inherited and acquired structural variation (Campbell *et al.*, 2008; Chen *et al.*, 2008). Metagenomic sequencing led to the discovery of infectious and commensal flora (Cox-Foster *et al.*, 2007). The discovery of transcribed SNPs, quantification of gene expression and alternative splicing and transcript annotation was made possible by transcriptome sequencing (Bainbridge *et al.*, 2006; Kim *et al.*, 2007; , Cloonan *et al.*, 2008; Lister *et al.*, 2008; Mortazavi *et al.*, 2008; Sugarbaker *et al.*, 2008; Wilhelm *et al.*, 2008). Small RNA sequencing, Chromatin immunoprecipiation-sequencing (ChIP-Seq) and nuclease fragmentation sequencing enabled microRNA profiling, the determination of patterns of DNA methylation and genome-wide mapping of protein-DNA interactions respectively (Johnson *et al.*, 2007; Robertson *et al.*, 2007; Morin *et al.*, 2008; Schones *et al.*, 2008; Wold and Myers, 2008). Large scale molecular barcoding was made possible by mulitplex sequencing of samples from multiple individuals (Kim et al., 2007, Meyer *et al.*, 2008). Finally, ultra-deep sequencing enables the discovery of minority variants and rare mutations in a single experiment (Mardis, 2008).

With regards to virology, the inference of genetic variation in viral populations has, until recently, relied on Sanger sequencing of individually cloned variants (Beerenwinkel and Zagordi, 2011). The estimation of underlying quasispecies structure proved time-consuming, laborious and expensive (Beerenwinkel and Zagordi, 2011). As a result, few studies have focused on the underlying quasispecies dynamics of viral populations in considerable detail (Beerenwinkel and Zagordi, 2011). However, the advent of next-generation sequencing, specifically the application of ultra-deep sequencing, has revolutionised the capacity for studying the underlying genetic diversity of viral populations (Mardis, 2008, Beerenwinkel and Zagordi, 2011). The limitations of clonal Sanger sequencing can easily be overcome with the application of next-generation sequencing approaches for detecting low-frequency mutations in viral populations (Beerenwinkel and Zagordi, 2011). Sequencing directly from a mixed sample at high coverage of approximately 10,000 reads per base pair reduces the time-constraints, labour intensity and costs previously associated with the estimation of population variants and their relative frequencies (Beerenwinkel and Zagordi, 2011).

Template Preparation

Two principal methods of template preparation are applied across all commercial next-generation sequencing platforms: single DNA-molecule templates, and clonal amplification originating from DNA molecules (Metzker, 2010). The two most common methods of clonal amplification include

emulsion PCR (emPCR) (Dressman *et al*., 2003) and solid-phase amplification (Fedurco *et al*., 2006). The Illumina platform utilises clonal amplification of single DNA molecules through solid-phase amplification (Metzker, 2010).

Template library preparation commences with fragmentation and adenylation of the DNA template. Adapter oligonucleotides are ligated to both ends of the adenylated DNA fragments, after which fragments are size-selected and purified. Single molecules composed of individual DNA fragments are isothermally amplified in a flow cell in preparation for high throughput sequencing. The surface of the flow cell is densely grafted with large amounts of oligonucleotides which bind to the adapters ligated to the DNA fragments. The DNA fragments are extended and fragment copies are covalently bound to the flow cell surface. Clonal amplification of the fragments takes place through isothermal bridge amplifications, resulting in clusters of DNA fragment copies. Reverse strands are removed and blocked, and sequencing primers are hybridized to the DNA template fragments for sequencing.

Sequencing and Imaging

The process of sequencing is classified into many subtypes including single-nucleotide addition (SNA), sequencing by ligation (SBL), real-time sequencing and cyclic reversible termination (CRT). The Illumina platform incorporates CRT technology during the sequencing process. The CRT process makes use of reversible terminators in cycles of nucleotide incorporation, fluorescence imaging and cleavage. The DNA templates are sequenced in parallel, one base at a time. One of four fluorescently labelled reversible terminator nucleotides is incorporated, representing the complement of the template base. Remaining bases are washed away subsequent to incorporation and the clusters are excited by laser to identify the newly added nucleotide. The fluorescent label and terminating group is removed by cleavage, making way for incorporation of the next nucleotide.

1.6 GENOME ASSEMBLY

The large amounts of data generated by the application of next-generation sequencing remain largely uninformative up the point of data analysis (Flicek and Birney, 2009). The processing of next-generation sequence data is fundamental to creating biological understanding to an appreciable level (Flicek and Birney, 2009). Even so, the implementation of analysis algorithms is subject to many limitations and challenges (Alkan *et al*., 2010). Next-generation sequencing data is inherently error-prone, and large scale data processing is computationally expensive and time consuming (Alkan *et al*., 2010). As a result, the trade off between speed and sensitivity dictates the functionality of any given approach to data processing (Flicek and Birney, 2009). Many algorithms and computational strategies are being developed to compensate for these limitations (Miller *et al*., 2010). In context of genomics,

the development of progressive alignment and assembly methods provide a good example (Miller *et al*., 2010).

Sequence alignment and assembly are considered the most elementary of computational analyses in the processing of genomic sequence data (Flicek and Birney, 2009). Alignment is defined as the process by which the genomic regions from which observed sequence reads originated is determined based on a known reference genome (Flicek and Birney, 2009). Assembly, on the other hand, is defined as the process by which sequence reads are organised into a hierarchical data structure that results in a putative reconstruction of the genome of origin (Miller *et al*., 2010).

Alignment

The most commonly implemented methods of alignment of short-read sequence data are hash table-based methods and Burrows Wheeler transform (BWT)-based methods (Flicek and Birney, 2009). Hash table-based methods implement a common data structure that is able to index complex data to facilitate rapid searching for candidate alignment locations (Flicek and Birney, 2009). Hash based-methods are especially applicable to the alignment of next-generation sequence data. Although DNA sequence data is likely to contain duplicates, it is highly unlikely to represent all possible nucleotide combinations (Flicek and Birney, 2009). Burrows-Wheeler transform (BWT)-based methods provide alignment strategies that are equally sensitive but less time-consuming than hash table-based methods (Burrows and Wheeler, 1994). BWT-based methods implement the Full-text index in Minute space (FM) data structure that simultaneously index and compress complex data to facilitate efficient and accurate searching for candidate alignment locations (Burrows and Wheeler, 1994; Ferragina and Manzini, 2000). The FM index data structure is based on the BWT algorithm commonly used in data compression, resulting in a compact alignment method that requires minimal computational effort compared to other alignment methods (Burrows and Wheeler, 1994; Ferragina and Manzini, 2000). When candidate alignment locations have been identified, slower and more accurate algorithms such as Smith-Waterman can be implemented to determine the final alignment.

Assembly

Next generation sequencing assemblers fall into one of two categories, namely Overlap graph methods and de Bruijn Graph methods (de Bruijn and Erdos, 1946; Myers, 1995). Both methods are based on graphs, consisting of a set of nodes and a set of edges between the nodes (Claude, 1962). Each graph is constructed by either a Hamiltonian cycle, where each node in the graph is visited once, or by a Eulerian cycle, where each edge is visited once (Euler, 1741; Hamilton, 1847).

In Overlap graph methods, nodes represent reads and edges represent pairwise alignments between reads (Myers, 1995). The overlap graph is constructed by a Hamiltonian cycle, creating an alignment between each successive node visited (Hamilton, 1847). The genome is assembled by following the edges in numerical order to combine alignments between successive reads (Kececioglu and Myers, 1995). This method is implemented in the assembly of Sanger sequence reads and is most applicable to a limited number of reads with significant overlap (Flicek and Birney, 2009). Although some genome assembly applications incorporate overlap graph methods, this approach is computationally intensive when applied to large numbers of short read sequence data (Flicek and Birney, 2009).

Methods based on De Bruijn graphs effectively reduce the computational effort required to assemble genomes from short read data by implementing K-mers instead of whole sequence reads (Flicek and Birney, 2009). A K-mer is a motif or subsequence of fixed-length which is observed more than once in a genomic sequence (Miller *et al*., 2010). The graph is constructed in either a Hamiltonian cycle, where nodes represent k-mers and edges represent pairwise alignments or a Eulerian cycle, where nodes represent (k-1)-mers and edges represent k-mers (Compeau *et al*., 2011). When the genome is assembled in a Hamiltonian cycle, an alignment is created in which each successive node is shifted by one position. (Compeau *et al*., 2011). In contrast, when the genome is assembled in a Eulerian cylce, an alignment is created in which each successive edge is shifted by one position (Compeau *et al*., 2011).

1.7 QUASISPECIES RECONSTRUCTION

Several pitfalls are associated with the analysis of deep sequencing data obtained from diverse viral populations. Errors introduced by PCR and sequencing, as well as the short nature of sequence reads reduce the accuracy with which viral population diversity is inferred. Errors arising from PCR include, firstly, the misincorporation of nucleotides due to the inaccuracy of DNA polymerases, secondly, biased amplification due to primer mismatches, thirdly, *in vitro* recombination as a result of premature termination of strand elongation and lastly, resampling due to low DNA concentration (Eckert and Kunkel, 1991; Kanagawa, 2003). In addition to errors in sequence data caused by PCR, many errors are introduced during sequencing. The mismatch rate of the Illumina sequencing platform increases with read length, and artificial indels and substitutions have been reported (2011, Kircher *et al*., 2009; Archer *et al*., 2012). Apart from sequence errors, diversity estimation depends on the read distribution across the genome (Beerenwinkel *et al*., 2012). The reasons for difficulty in attaining uniform coverage across the genome is poorly understood, although GC content of the target sequence has been shown to play significant role (Dohm *et al*., 2008).

In order to compensate for these drawbacks in quasispecies reconstruction; methods for data filtering and alignment, read error correction, local and global haplotype inference and haplotype frequency estimation have been developed (Beerenwinkel *et al*., 2012).

Data Filtering and Local Haplotype Inference

Data is filtered to remove reads of low quality and indels that cause frameshifts in coding regions prior to haplotype estimation (Beerenwinkel *et al*., 2012). The remaining reads are aligned by mapping to a reference genome sequence to generate a multiple sequence alignment (MSA) (Wang *et al*., 2007, Beerenwinkel et al., 2012). The inference of haplotypes and their frequencies in the population ultimately relies on the variation present in this alignment. It should be kept in mind that true variation is sampled in proportion to frequency in the population (Beerenwinkel and Zagordi, 2011). In contrast, variation arising from technical errors are randomly distributed and rare (Beerenwinkel and Zagordi, 2011). A true haplotype is therefore likely to be a group of reads more similar to each other than to any other read (Beerenwinkel and Zagordi, 2011). However, this assumption only applies if the error rate is low compared to the population diversity (Eriksson *et al*., 2008a). Both error correction and the identification of haplotypes relate to a clustering problem, where clusters are formulated probabilistically and solved in a Bayesian fashion (Eriksson et al., 2008a; Zagordi *et al*., 2010).

Global Haplotype Inference

Global haplotype inference, or genome-wide quasispecies reconstruction, is comparable to a jigsaw puzzle problem, where local fragments are assembled into haplotype sequences across the genome (Beerenwinkel *et al*., 2012). Various approaches to solving the jigsaw puzzle have been proposed, including graph-based methods and probabilistic clustering methods (Beerenwinkel *et al*., 2012). Graph-based global quasispecies reconstruction aggregates reads into a graph where nodes represent the locally-error corrected reads and edges represent the connection between reads that belong to the same haplotype (Beerenwinkel *et al*., 2012). If every path is equivalent to a potential haplotype, quasispecies reconstruction aims to find the minimum set of paths that explain the reads well (Eriksson et al., 2008a, Zagordi *et al*., 2011). Probabilistic approaches to global quasispecies reconstruction implements an infinite mixture model (Ewens, 1972; Ferguson, 1973; Beerenwinkel *et al*., 2012). The latter acknowledges the presence of subcomponents within a mixed data set without prior knowledge of the subcomponent to which an entity belongs. In terms of global quasispecies reconstruction, a haplotype represents a subcomponent, an alignment of sequence reads represents a mixed data set and an entity represents a sequence read (Ewens, 1972; Ferguson, 1973; Beerenwinkel *et al*., 2012).

Recent advances in next generation sequencing and data analysis are increasing our understanding of viral diversity in light of pathogenesis, virus evolution, fitness, selection pathways and much more. Although genome-wide diversity on the consensus level has been applied extensively throughout comparative studies, the underlying variation in a virus population is seldom accounted for. Computational methods that recognise genetic diversity beyond that of the consensus genome sequence are developing at a rapid pace despite the many challenges that confound such analyses. Imperfect as it may be, the current capacity to reconstruct quasispecies from next generation sequence data offers the first glance at the underlying genetic variation that contributes to the viral genome.

1.8 RESEARCH OBJECTIVES

The research described in this thesis is aimed at: (1) The genetic classification of two historical and two contemporary South African WNV strains, (2) The characterization of genetic change on both the consensus genome- and quasispecies level in relation to propagation system, and (3) The establishment of a single plasmid reverse genetic system for a South African WNV strain.

In achieving these aims, the following research questions were addressed:

1. In which lineages do the WNV strains used in this study reside?

2. Can genetic changes resulting from alterations between Baby Hamster Kidney (BHK-21) cells and mouse brain as propagation system be observed in the consensus genome?

3. What differences can be observed in the genome-wide SNP profiles of WNV strains resulting from alternations in propagation system?

4. How do changes in propagation system influence the relative frequencies of haplotypes?

Chapter 2

Materials and Methods

2.1 RATIONALE

Prior to this study, none of the four reference WNV strains from South Africa had been genetically characterised. The underlying quasispecies structure and interactions amongst population variants of each isolate were unknown. Moreover, the changes in underlying population diversity potentially elicited by differences in the number of passages and in propagation system had not been studied. The experimental design and methodology implemented to achieve the aims of this study is summarized in Figure 2.1. Briefly, selected viruses were passaged in baby hamster kidney (BHK-21) cell culture and by intra-cerebral inoculation of two day old suckling mice. Viral RNA was extracted and the transcriptome of each isolate was amplified prior to whole genome sequencing. Sequencing data was assembled into complete genomes, and each isolate was genetically characterised by phylogenetic analysis of the C-prM-E gene region. A subset of the sequence data was used to reconstruct haplotypes that contribute to the quasispecies structure of each isolate. To enable future studies, a single plasmid reverse genetic system was designed based on the genome sequence of one of the isolates.

2.2 WNV ISOLATES

The WNV isolates used in this study were acquired from the Agricultural Research Council (ARC) Onderstepoort Veterinary Institute (OVI) antigen bank. Four strains were used, two of which were contemporary viruses isolated in 2008 and 2011, and two historic viruses isolated in 1968 and 1977. Both contemporary strains, designated HS 101/08 and HS 87/11 were originally isolated from horses. The geographical data accompanying the latter is unknown. The geographical data and hosts from which WNV 1968 and WNV 349/77 were isolated are unknown. The WNV 1968 strain was donated to the ARC by BM McIntosh (pers. comm. B. Erasmus). WNV 349/77 was isolated in 1977 by the ARC and it is unknown whether the detailed historical information accompanying this isolate was recorded at the time (pers. comm. B. Erasmus).

Historic strains were previously propagated in either baby hamster kidney (BHK-21) cells or by intra-cerebral inoculation of 2 day old suckling mice. Each of the historic viruses underwent varying rounds of passage in each of their respective propagation systems and were stored in lyophilised form. WNV 1968 was propagated intra-cerebrally in suckling mice once and three times respectively. WNV 349/77 was propagated intra-cerebrally in suckling mice eight times and in BHK cells three, five and seven times. Contemporary viruses were continuously propagated in BHK cell cultures at the time of this study (Table 2.1).

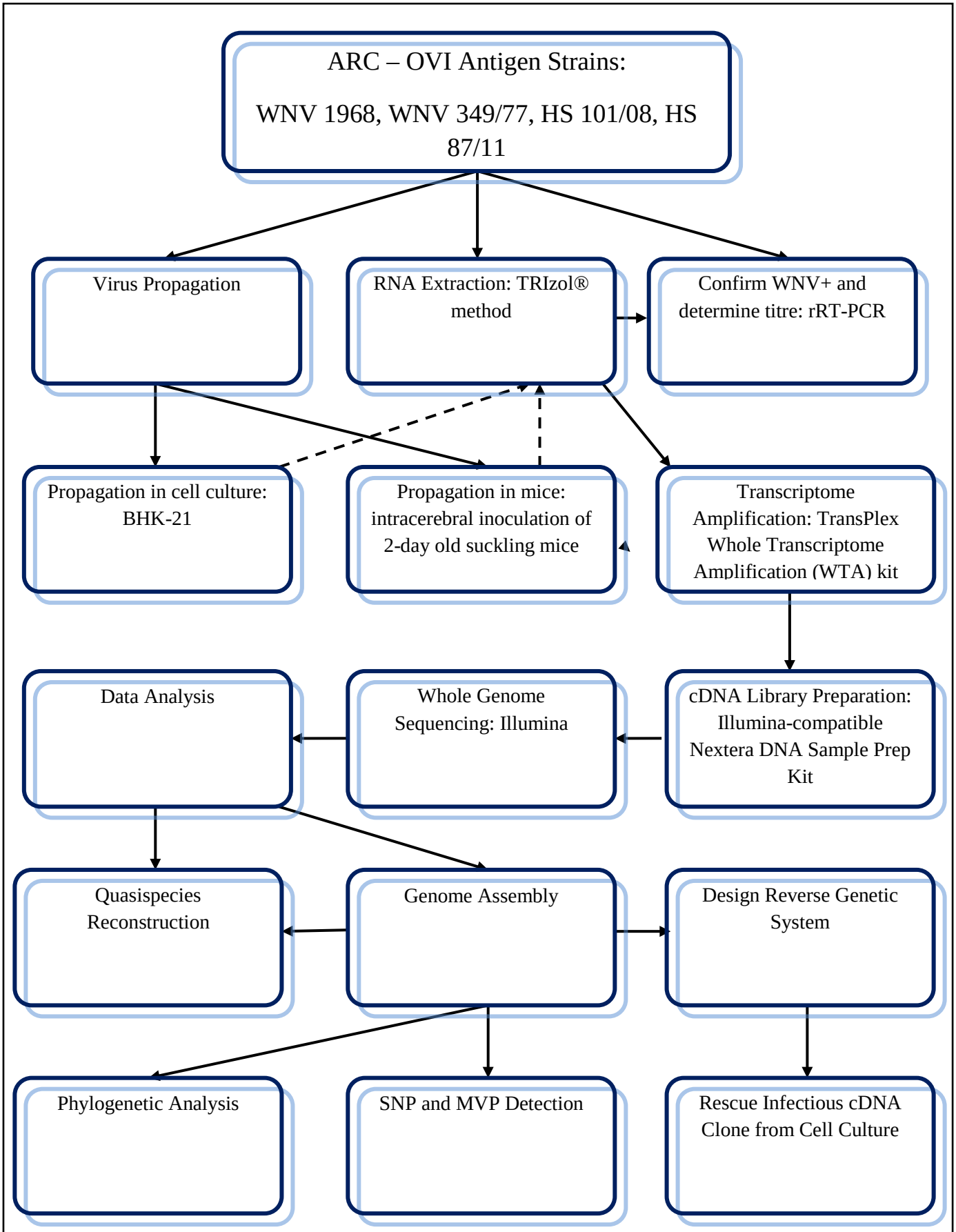

Figure 2.1 Schematic overview of experimental design and methodology implemented in this study.

Table 2.1 West Nile virus (WNV) isolates used in this study

Strain	Isolate	Year of Isolation	Passage Level	Host Source	Location
WNV 1968	A	1968	Mouse Brain #3	Unknown	South Africa
	B	1968	Mouse Brain # 1	Unknown	South Africa
WNV 349/77	C	1977	Mouse Brain # 8	Unknown	South Africa
	D	1977	BHK Cells # 3	Unknown	South Africa
	E	1977	BHK Cells # 5	Unknown	South Africa
	F	1977	Mouse Brain # 8	Unknown	South Africa
	G	1977	BHK Cells # 7	Unknown	South Africa
HS 87/11	H	2011	BHK Cells Ongoing Culture	Horse (*Equus caballus)*	South Africa
HS 101/08	I	2008	BHK Cells Ongoing Culture	Horse (*Equus caballus)*	South Africa

2.3 VIRUS PROPAGATION

2.3.1 Propagation in Mice

In order to investigate the effect of the propagation system and number of passages on the underlying WNV population diversity, selected historical isolates were propagated by intra-cerebral inoculation of two day old suckling mice. One family of mice was dedicated to each of the four isolates, and each isolate was passaged once. Isolates include duplicates of WNV 349/77 previously passed three times and five times in BHK cells respectively, a duplicate of WNV 349/77 previously passaged eight times in mice and a duplicate of WNV 1968 previously passaged once in mice.

Mice families were housed in standard 60cm x20cm x 20cm laboratory mice cages. Cages were kept in a quiet, appropriately lit and air conditioned room at 23°C in the ARC-OVI small animal facility. Mothers were kept with their offspring at all times. Unlimited dried food and clean water were supplied to the lactating mother. Cages were cleaned daily by moving the family to a new cage including clean bedding, water and food with minimal distress. Contents of used cages were discarded in a biohazard bag, autoclaved and incinerated. Cages were washed with F10 detergent and sprayed with 70% ethanol. Water bottles were autoclaved. All animals were observed three times per day to ensure comfort of surroundings, and eliminate first signs of unnecessary distress. Unnecessary handling of all animals were kept to a minimum. Ethical clearance was obtained from the University

of Pretoria Animal Ethics Committee (AEC) (Appendix A).

The lyophilised material of each WNV isolate was resuspended in 1mL phosphate buffered saline (PBS) solution. A further 1:10 dilution was prepared by the addition of 5 mL PBS to 500ul of the original suspension. Suckling mice were inoculated with 50 µl each of the 1:10 dilution by intra-cerebral injection in the occipital region of the skull. Mice were inspected three times daily for neurological- and behavioural symptoms of WNV infection during the seven day incubation period. At first signs of illness, mice were euthanized and virus was harvested from brain tissue in biological security level 3 (BSL 3) conditions. A small amount of infective brain tissue was inactivated by the addition of TRIzol® reagent (Life Technologies) in preparation for RNA extraction. Brain tissue was pooled per isolate, lyophilised and stored at 4 °C.

The remainder of the original suspension of lyophilised material was inactivated by the addition of TRIzol® reagent (Life Technologies) in preparation for RNA extraction. Subsequent to RNA extraction, the viral load of lyophilized material was quantified by rRT-PCR. The remainder of the 1:10 dilution was filter-sterilised and applied to cell culture subsequent to the inoculation of mice.

2.3.2 Propagation in Cell Culture

Both contemporary WNV isolates and all duplicates of historic WNV isolates were propagated in baby hamster kidney (BHK-21) cell cultures. The parent line of BHK-21(C-13) was derived from one- day-old baby Syrian hamster (*Mesocricetus auratus*) kidneys in 1961 (Macpherson, 1963). The BHK-21 cell line used in this study was acquired from the American Type Culture Collection (ATCC) and is the property of the Agricultural Research Council (ARC) Onderstepoort Veterinary Institute (OVI).

Cells were recovered from frozen stocks, maintained routinely and infected with WNV isolates accordingly throughout this study. The methodology concerning all of the aforementioned will be described in the sections to follow. Cells were processed aseptically within a BSL 2 cabinet at all times, and all solutions and equipment that came in contact with the cells were sterile.

Recovery of BHK-21 Cells from Cryopreservation

Vials containing BHK-21 cells were removed from liquid nitrogen and thawed briefly at 37 °C. The thawed cell suspension was transferred to 4 mL complete medium consisting of Eagle's Minimum Essential Medium (EMEM) (Lonza), 20% fetal calf serum (Invitrogen), 1% Penicillin/Streptomycin/Amphotericin B (Gibco), 1% L-glutamine (200 mM) (Gibco) and 1% non-

essential amino acids (NEAA) (Gibco) and centrifuged for ten minutes at 1000 rpm. The supernatant was discarded and cells were resuspended in 1mL complete medium. The cell suspension was transferred to a T-25 flask (Nunc) and made up to a final volume of 5mL. Cells were incubated at 37 °C at 5% CO_2 and observed under a Leica DM IL microscope daily for the formation of a confluent monolayer.

Cell Number and Viability

Cell culture conditions were standardised to attain a viable cell concentration of $2x10^6$ cells/mL. The number of cells and cell viability was determined with the application of Trypan Blue (Gibco) staining and the use of a hemacytometer. The confluent cell monolayer was trypsinized and suspended in 2 mL 10% EMEM (Lonza). A further 1:20 dilution was prepared with the addition of 3 µl 10% EMEM (Lona), 90 µl PBS and 100 µl Trypan Blue (Gibco) to 100 µl of the former cell suspension. The 1:20 dilution of cell suspension was transferred to the hemacytometer. The slide was viewed under the microscope at 100X magnification. Cells were counted and the cell number was calculated according to the following standard equation:

cells/mL = cells/mL(number of cells counted) x (dilution factor)

The percentage viable cells are calculated as follows:

% viable cells = (number of unstained cells)/(total number of cells) x 100

Propagation and Subculture

Cells were propagated in 10% medium consisting oF EMEM (Lonza), 10% FCS (Invitrogen), 1% Penicillin/Streptomycin/Amphotericin B (Gibco), 1% L-glutamine (200 mM) (Gibco) and 1% NEAA (Gibco) at 37 °C in 5% CO_2. Cells were sub-cultured after reaching 80-90% confluency. Medium was removed from the flask and the cell monolayer was rinsed three times with 2 mL 0.25% trypsin. The flask was incubated for approximately 5 to 10 minutes at 37°C and 5% CO_2 to promote detachment of cells. Depending on sub-cultivation ratio, between 1.5 mL and 5.5 mL was added to the detached cell suspension to inactivate trypsin. The cell suspension was transferred to a T-75 flask (Nunc) and made up to a final volume of 15 mL. Excess cells were prepared for cryopreservation and frozen as stock for future use. Cells were maintained to a concentration of $2x10^6$ cells/mL by sub-cultivation at a ratio of 1:2 to 1:6 every two to five days. Cells that were not sub-cultered further were prepared for cryopreservation.

Freezing for Cryopreservation

Freezing medium was prepared prior to processing cells for cryopreservation. Freezing medium comprised of 10% DMSO (Sigma Aldrich), 20% FCS (Invitrogen) and complete growth medium and was prepared to a total volume of 1 mL per cryovial. Cells were trypsinized as per description for subculture of confluent cells. Trypsin was inactivated by the addition of 2 mL complete medium. The cell suspension was centrifuged at 1000 rpm at 4 °C for two minutes. The supernatant was removed and the pelleted cells were resuspended in freezing medium at a density of 3 x 10^6 cells/mL. Cells were transferred to cryovials to a final volume of 1mL per vial. Cryovials containing cells were wrapped in cotton and placed at -20 °C for four to five hours prior to being transferred to -70 °C for overnight storage. The cryovials were then transferred liquid nitrogen for long term storage.

Preparing Cells for Infection

A cell monolayer that had reached 80-90% confluency was sub-cultured as described previously to maintain a cell concentration of 2 x 10^6 cells/mL. Cells were transferred to a 12-well cell culture plate and 10% complete medium. For each culture plate infected with a virus, an uninfected culture plate was prepared as negative control. Cells were incubated at 37°C and 5% CO_2 until 90% confluency was reached. The 10% complete medium was removed and the cell monolayer was rinsed with PBS. Complete medium was replaced with serum-free medium consisting of EMEM growth medium, 1% Penicillin/Streptomycin/Amphotericin B (Gibco), 1% L-glutamine (200 mM) (Gibco) and 1% non-NEAA (Gibco). Cells were transfected with filter sterilised virus originating form resuspended lyophilized material or brain tissue under BSL 3 conditions. Infected cells were incubated at 37 °C in 5% CO2 for 60 to 90 minutes. Medium containing virus was removed from cells and replaced by 2% complete medium. Plates were sealed and incubated at 37 °C in 5% CO_2 for seven days. Cultures were inspected microscopically twice daily for signs of cytopathic effect (CPE) up to seven days post infection. All materials and equipment that came in contact with virus material were disinfected, disposed of in biohazard containers, autoclaved and incinerated.

Virus Isolation from Cell Culture

Virus was isolated from infected cell cultures demonstrating CPE in 50% of the cell monolayer in comparison with the negative control. Cultures were frozen at -70 °C and thawed at room temperature to promote cell lysis and virus release. Thawed culture material containing live virus was redistributed to sterile falcon tubes and centrifuged at 2000 rpm at 4 °C. The supernatant was collected and stored at 4 °C for RNA extraction.

2.4 RNA Extraction

Viral RNA was extracted according to the single-step RNA isolation method using TRIzol® (Life Technologies) reagent (Simms *et al.*, 1993). RNA was extracted from lyophilised virus material, brain tissue from infected mice and cell culture material from infected BHK cell cultures. In each instance, 250 µl of viral material was inactivated with the addition of 750 µl TRIzol® reagent (Life Technologies) under BSL 3 conditions prior to commencing with the extraction protocol.

The reaction was incubated at room temperature for 5 minutes to facilitate the dissociation of nucleoprotein complexes. A volume of 200 µl of chloroform was added to each reaction followed by vigorous mixing and incubation for 10 minutes at room temperature. Reactions were centrifuged for 15 minutes at 13 000 rpm, resulting in separation of biphasic mixtures into an organic and aqueous phase. The RNA was extracted from the aqueous phase with the addition of 500 µl isopropanol and incubation at 4 °C overnight. The reaction was centrifuge for 10 minutes at 13 000 rpm. The supernatant was removed and the pellet was washed with 1 mL 70% ethanol. The pellet was air dried and resuspended in 40 µl elution buffer. RNA quantity and purity was assessed by spectrophotometric measurements of the ratio of absorbance at 260 nm and 280 nm (NanoDrop). RNA was stored at -70 °C.

2.5 REAL-TIME qRT-PCR

A real-time quantitative reverse transcription polymerase chain reaction (qRT-PCR) assay targeting the NS2A region of the WNV genome was used for semi-quantification of WNV isolates in different propagation systems (Eiden *et al.*, 2010). The assay included each WNV isolate following RNA extraction from lyophilised viral material, infected mouse brain and infected BHK-21 cell cultures respectively. Quantification was based on a standard curve derived from 10-fold serial dilutions of a WNV lineage 2 positive control virus.

The LightCycler RNA Amplification Kit Hybridization Probes (Roche) was used according to manufacturer's instructions. Primer and probe sequences are listed in Table 2.2. The reaction mixture consisted of 5.4 µl nuclease free water, 7 mM $MgCl_2$, 20 µM primer FLI-WNF5-F, 20 µM primer FLI-WNF6-R , 10 µM FLI-WNF-Probe 4.0 µl, Hyprobe mix (5x), 0.4 µl enzyme mix and 6 µl RNA. PCR was performed using a LightCycler®Nano according to the following cycling conditions were as follows: one cycle at 50 °C for 30 minutes followed by 95 °C for 15 minutes, 42 cycles at 95 °C for 30 seconds followed by 55 °C for 30 seconds, and 72 °C for 30 seconds.

Table 2.2 Primer and probe sequences

Primer/Probe Name	Sequence	Orientation	Target Genome Position	Tm[1]
FLI-WNF5-F	GGGCCTTCTGGTCGTGTTC	Sense	2558-3576	59.6 ° C
FLI-WNF6-R	GATCTTGGCYGTCCACCTC	Antisense	3621-3603	54.1° C
FLI-WNF-Probe	F-CCACCCAGGAGGTCCTCGCAA--Q	Sense	3581-3602	67.8° C

[1]Melting Temperature

2.6 NEXT GENERATION SEQUENCING

2.6.1 Transcriptome Amplification and cDNA Library Preparation

The transcriptomes of WNV isolates selected for whole genome sequencing (WGS) were amplified with the use of TransPlex Whole Transcriptome Amplification (WTA) kit (Sigma Aldrich) according to manufacturer instructions. The WTA amplification process reverse transcribes RNA by implementing non-self-complementary primers. The resulting cDNA library is composed of random overlapping fragments flanked by a universal end sequence. Fragments were PCR amplified by implementing a universal primer to produce the WTA product. RNA was quantified by spectrophotometry (NanoDrop).

The prescribed amount of 300ng RNA was added to nuclease-free water to a final volume of 19 µl. Equal amounts of WTA library synthesis buffer and WTA library stabilization solution were added to the RNA suspension resulting in a final reaction volume of 24 µl. The reaction mixture was incubated at 70 °C for five minutes to facilitate denaturation and immediately cooled thereafter. The reaction mixture was supplemented with 1ul WTA library synthesis enzyme in preparation for cDNA library synthesis, incubated in a thermocycler and subjected to the following thermal cycling parameters: 24 °C for 15 minutes, 42 °C for two hours and 95 °C for five minutes, after which the reaction was immediately cooled on ice.

The products of cDNA library synthesis were aliquoted in volumes of 5 µl and added to 70 µl of WTA Amplification Mix prior to amplification. The WTA Amplification Mix was composed of 300 µl nuclease-free water, 37.5 µl WTA Amplification Master Mix, 7.5 µl dNTP Mix and 12.5 units of antibody inactivated Hot-Start Taq DNA Polymerase (Qiagen). Reactions were incubated in a thermocycler according to the following parameters: 95 °C for three minutes, followed by 17 cycles of 94 °C for 20 seconds and 64 °C for five minutes. The amplified cDNA libaries were purified using

High Pure PCR Product Purification Kit (Roche) according to manufacturer's instructions and quantified by spectrophotometry (NanoDrop).

2.6.2 Illumina Sequencing

The Illumina-compatible Nextera DNA Sample Prep Kit (EPICENTRE Biotechnologies) was used to prepare genomic cDNA libraries for sequencing according to manufacturer's instructions. Sequencing libraries were prepared by a process of tagmentation which combines DNA fragmentation, end-polishing and adapter-ligation in a single reaction. The DNA product recovered from tagmentation was used as input for bridge PCR (bPCR) and cluster generation as per the standard Illumina protocol. The methods used to prepare genomic cDNA libraries for Illumina sequencing are described briefly in the remainder of this section.

DNA Quantification

The genomic cDNA libraries of WNV isolates were quantified with a fluorescence-based Qubit quantitation assay. DNA was prepared for fluorescence with the use of Qubit dsDNA HS Assay Kit (Invitrogen). The Qubit 2.0 Fluorometer was calibrated according to manufacturer's instructions. The cDNA of each WNV isolate was aliquoted to Qubit assay tubes. Qubit working solution was added to the cDNA to a total volume of 200 µl. The reaction was vortexed and incubated for 2 minutes at room temperature. The cDNA was quantified by fluorescence with the use of the Qubit 2.0 Fluorometer. The Qubit concentration was used to standardise the DNA concentration of each sample to 27.5 ng/µl in a total volume of 5 µl.

Sequencing Library Preparation

Target DNA was fragmented and tagged with Nextera Enzyme Mix to generate Illumina-compatible sequencing libraries during tagmentation. Each tagmentation reaction mixture was composed of 27,5 ng target DNA, 2 µl High Molecular Weight (HMW) buffer and 0.5 µl Nextera Enzyme Mix. Nuclease-free water was added to the reaction mixture to a final volume of 10 µl. The reaction was incubated at 55 °C for five minutes. Products were purified with the QIAGEN MinElute PCR purification kit and eluted in 11 µl nuclease-free water. Tagmentation was followed by nine cycles of limited-cycle PCR to incorporate bridge PCR (bPCR) compatible adapters into the core sequencing library.

The limited-cycle PCR reaction included the following reagents: 12.5 µl 2X Nextera PCR Buffer, 0.5 µl 50X Nextera Primer Cocktail, 2 µl 50X Nextera Adaptor 2,1,75 units Nextera PCR Enzyme, 4.5 µl

nuclease-free water and 5 µl tagmentation reaction product. Reactions were incubated in a thermocycler according to the following parameters: 72 °C for three minutes, 95 °C for 30 seconds followed by nine cycles of 95 °C for 10 seconds, 62 °C for 30 seconds and 72 °C for three minutes.

The DNA product was purified using QIAGEN MinElute PCR purification kit (Qiagen) and eluted in 16 µl Elution Buffer and 0.1% Tween. The DNA concentration of sequencing libraries was quantified with a fluorescence-based Qubit quantitation assay as described previously. Sequencing libraries generated from the tagmentation and limited-cycle PCR reactions were used as input for bridge PCR (bPCR) and cluster generation as per the standard Illumina protocol. Sequencing was performed using a HiScan sytem (Illumina) or MiSeq system (Illumina) at the ARC Biotechnology Platform.

3.7 DATA ANALYSIS

3.7.1 Genome Assembly

Illumina sequence reads were trimmed prior to genome assembly and mapping with the use of CLC Genomics Workbench v5.1.5 (http://www.clcbio.com/). Reads were assembled *de novo* to optimise the paired read lengths of individual data sets. Trimming was repeated using data with read lengths conforming to a normally distributed range within the paired reads distance distribution. Trimmed reads were mapped to the complete genome sequence of SA 93/01 (GenBank accession number: EF429198). The consensus genome sequence was annotated accordingly. The procedure and parameters used to trim and map sequence data is described in the remainder of the section.

Trimming

Sequence reads were trimmed according to quality scores using a modified-Mott algorithm. Quality scores were converted to error probabilities and subtracted from a limit value of 0.01. The running sum of each nucleotide was calculated and regions with a value equal to or below zero were trimmed. The reads resulting from quality trimming were composed of regions between the first positive value of the running sum and the highest value of the running sum. Ambiguous nucleotides at the end of sequence reads were trimmed according to an ambiguity limit of zero. Nextera adapter sequences were aligned to sequence data using a Smith-Waterman alignment (Smith and Waterman, 1981). Default alignment scoring thresholds were specified to include internal matches to a minimum score of 10 and end matches to a minimum score of four. A default cost of two and three was assigned to mismatches and gaps respectively. Sequence data regions matching adapter sequences were trimmed together with downstream nucleotides. Sequence data was trimmed to length by removing five nucleotides from 5’terminal ends and one nucleotide from 3’terminal ends.

Mapping

Mapping was performed by local alignment of paired sequence reads to the reference genome. A distance interval was automatically calculated for paired read data and incorporated during the mapping process. The complete genome sequence of a Lineage 2 WNV isolate (GenBank accession number EF429198.1) was used as reference genome. Mapping parameters were specified as default values with a mismatch cost of two, insertion cost of three and a deletion cost of three. Optimal read alignments achieved during the implementation of mapping parameters were subject to a filtering process. Matched reads were included in the mapping output or discarded as determined by the filtering threshold. Determinants of the filtering threshold were specified as default values with a length fraction of 0.5 and a similarity fraction of 0.8.

2.7.2 Phylogenetic Analysis

The phylogenetic relationships between WNV isolates in this study and viruses representing all five phylogenetic lineages of WNV were determined by Maximum Likelihood (ML) analysis and Bayesian Inference (BI) based on the C-prM-E gene region. The GenBank accession numbers for isolates used in the generation of phylogenetic trees can be found in figure 3.5. Sequences were aligned using MAFFT multiple sequence alignment program (Katoh and Toh, 2008) and adjusted manually using Bio Edit (Hall, 1999). Phylogenetic trees were rooted using Japanese Encephalitis virus strain JEV SP78 (GenBank accession number AF075723) as the outgroup. The best-fit nucleotide substitution model was determined statistically using jModelTest v2.0 (Posada, 2008, Guindon and Gascuel, 2003) prior to BI and ML analysis. The Akaike Information Criterion (AIC) was used for model selection (Akaike, 1974). The proportion of invariable sites (p-invar) as well as the gamma distribution parameter (α) was determined using jModelTest v2.0 (Posada, 2008, Guindon and Gascuel, 2003).

ML analysis was conducted using PhyML v3.0 (Guindon and Gascuel, 2003). ML trees were reconstructed according to General Time Reversible (GTR) substitution model with a fixed gamma-distributed rate variation of 0.25 across sites. Nucleotide equilibrium frequencies were optimised and the number of substitution rate categories was set to four. BI was conducted using MrBayes v2.0 (Huelsenbeck and Ronquist, 2001). A phylogenetic tree was reconstructed using the Markov chain Monte Carlo algorithm (Metropolis *et al.*, 1953) run over 10 000 000 generations under a General Time Reversible (GTR) substitution model (Tavaré, 1986) with gamma-distributed rate variation across sites. A dirichlet prior distribution was specified for the state frequency parameter of the likelihood model. Subsequent to phylogenetic analysis the confidence levels in branching points on

phylogenetic trees generated by ML were determined by bootstrap analysis (1000 replicates) (Felsenstein, 1985).

A Neighbour-Joining (NJ) tree was included to illustrate relative branch lengths (Figure 3.6). The latter was inferred in MEGA5 (Tamura *et al.*, 2011) using uncorrected / p-distances. Confidence levels in branching points were determined by bootstrap analysis (1000 replicates) (Felsenstein, 1985), with support values greater than 70 % from NJ and ML bootstrap resampling and posterior probabilities >94 from Bayesian inference being indicated on the relevant nodes.

2.7.3 Single Nucleotide Polymorphism (SNP) Detection

Single nucleotide polymorphisms (SNPs) were determined in CLC Genomics Workbench v5.1.5 (http://www.clcbio.com/) using the Neighborhood Quality Standard (NQS) algorithm (Altshuler *et al.*, 2000). The complete genome sequence of a Lineage 2 WNV isolate (GenBank accession number EF429198.1) was used as reference genome throughout the study. SNPs were determined by quality and significance assessments on sequence reads mapped to a reference genome. Quality assessments were based on a neighbourhood radius of 11 nucleotides. A minimum average quality score of 15 was allowed for surrounding bases and 20 for the central base. A maximum of two mismatches and gaps were allowed. Significance assessments were based on a minimum coverage threshold of four. A minimum variant frequency of 1 % and 35 % was allowed during SNP and MNV detection respectively. Non-specific and low quality matches were ignored. It is of note that this method does not detect SNPs and insertions and deletions (InDels) simultaneously. The latter was therefore considered subsequent to quasispecies reconstruction described later in this section.

The SNPs detected in sequence data of each isolate were annotated according to gene region. In order to identify gene regions most influenced by changes in propagations system, the diversity of gene regions were compared based on the number of SNPs observed. In each instance, the size of a gene region was scaled to the size of the genome to obtain the fraction of nucleotides comprising that gene region. The number of SNPs observed in each respective gene region was multiplied by the inverse of the aforementioned fraction to obtain the number of SNPs scaled to the size of the genome to accommodate for the difference in size between gene regions.. The number of SNPs were compared per gene region to identify the most variable region of the genome. The standard deviation in the number of SNPs per gene region was calculated between isolates to identify the gene regions most influenced by changes in propagation system and passage number.

In order to study the influence of propagation system on quasispecies variation, the frequencies of SNPs that were shared amongst isolate A (WNV 1968) and isolate B (WNV 1968), as well as

amongst isolate C (WNV 349/77), isolate D (WNV 349/77), isolate E (WNV 349/77), isolate F (WNV 349/77) and isolate G (WNV 349/77) were compared. In studying frequency changes of SNPs shared amongst isolates, all SNPs were considered regardless of codon changes resulting in premature stop codons. Due to the inability of the approach used to detect SNPs to discern the association between respective SNPs, quasispecies were reconstructed to obtain full-length haplotype sequence alignments.

2.7.4 Quasispecies Reconstruction

The haplotypes of WNV 349/77 isolates were reconstructed from ultra deep sequence data in order to estimate the underlying genetic diversity that contributes to the quasispecies of each isolate. The sequence data of each isolate was aligned to its corresponding consensus genome as reference using the Burrows-Wheeler transform-based method (Burrows and Wheeler, 1994) in Bowtie (Langmead *et al.*, 2009). For every isolate, aligned reads were resampled to a total of approximately 50,000 reads to accommodate for the difference in total number of mapped reads amongst isolates using SamTools (Li *et al.*, 2009). Each alignment was sorted and indexed to create a multiple sequence alignment (MSA) using SamTools (Li *et al.*, 2009). The MSA was subject to error correction and local haplotype construction by implementing a model-based probabilistic clustering algorithm (Zagordi *et al.*, 2010) in ShoRAH (Beerenwinkel and Zagordi, 2011). The process was repeated for 5,000 iterations. The quality of the reconstructed haplotypes and corresponding frequencies were estimated in a Bayesian fashion by computing the posterior probability distribution of the latter parameters (Zagordi et al., 2010). Global analysis was performed on the corrected reads using a parsimony principle to compute the minimal set of haplotypes that explains the sequence data (Eriksson *et al.*, 2008b). The frequencies of the haplotypes were estimated by maximum likelihood with the use of an Expectation Maximization (EM) algorithm (Eriksson et al., 2008b). Haplotypes with a posterior probability below 0.8 were discarded.

The relative diversity of gene regions were compared between isolates based on the number of haplotypes recovered for each respective gene region. For each isolate, haplotypes with a posterior probability above 0.8 were annotated according to the gene region in which variation was observed. In order to identify gene regions most influenced by changes in propagations system, the diversity of gene regions were compared based on the number of haplotypes observed. In each instance, the size of a gene region was scaled to the size of the genome to obtain the fraction of nucleotides comprising that gene region. The number of haplotypes observed in each respective gene region was multiplied by the inverse of the aforementioned fraction to obtain the number of hapltoypes scaled to the size of the genome to accommodate for the difference in size between gene regions. The standard deviation in the number of haplotypes per gene region was calculated between isolates to identify gene regions

most influenced by changes in propagation system and passage number.

The frequencies of haplotypes that were present in more than one isolate were compared in order to study the influence of propagation system on the quasispecies composition of WNV 349/77. The sequences of haplotypes with a posterior probability above 0.8 were translated into protein sequences in MEGA5 (Tamura *et al.*, 2011). The haplotype sequences of each isolate were combined in a single sequence alignment. The latter was implemented in DnaSP v5 (Librado and Rozas, 2009) to identify haplotypes that were shared amongst isolates based on nucleotide identity. Due to the computational expenses involved in grouping full-length haplotype sequences, only viable haplotypes containing no stop codons in the WNV open reading frame were considered. The latter were compared based on the gene region in which variation was observed.

2.8 Reverse Genetic System

2.8.1 Infectious cDNA Clone Design

A novel infectious clone of WNV 349/77 was designed to facilitate future viral genetic studies of WNV and its host interactions. The design was based on a single plasmid reverse genetic system and synthesized accordingly (GenScript). The full length WNV genome was flanked by a 5' T7 polymerase promoter and a 3' Hepatitis Delta Virus ribozyme (Figure 2.2). The latter was inserted in a pUC57 vector by *Eco*RI and *Hind*III (Figure 2.2). *Ssp*I restriction sites were engineered at both ends of the WNV 349/77 genome to enable *in vitro* translation and serve as a genetic marker.

2.8.2 Infectious cDNA Clone Rescue

The lyophilised construct was resuspended in nuclease free water to a final concentration of 250 ng/µl and incubated at 37 °C for 30 minutes to dissolve. The resuspended construct was used for *in vitro* transfection of three different cell lines, including BHK-21cells; Vero cells and BSR T7/5 cells. In a further attempt to rescue the infectious WNV clone, the construct was linearised and translated *in vitro* to be used for the transfection of cell cultures. The infectious clone was amplified *in vitro* in *E.coli* for future studies. The strategies implemented in the rescue of the infectious WNV clone are discussed in the remainder of this section

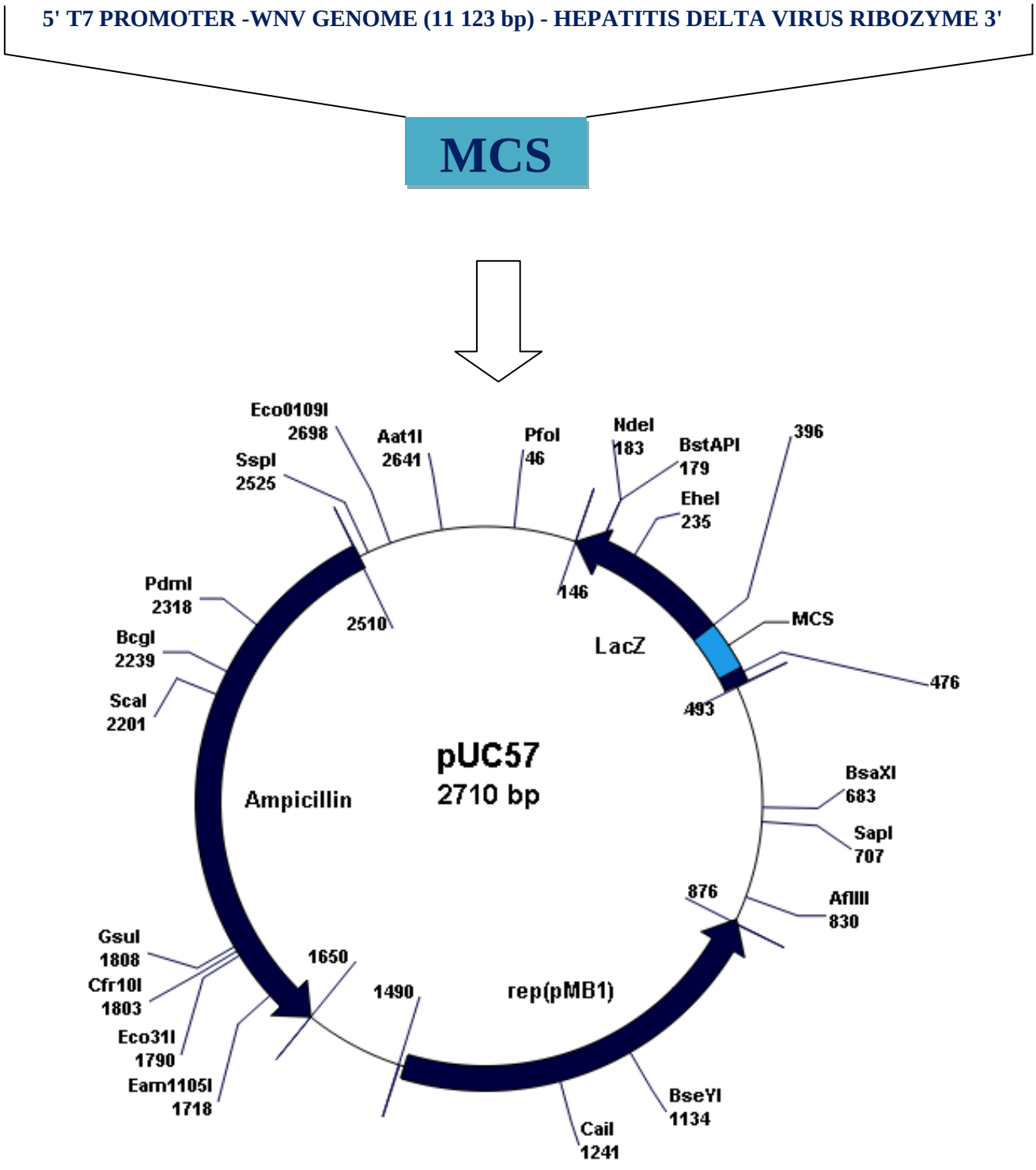

Figure 2.2 Diagrammatic representation of the infectious WNV cDNA clone. The full length WNV genome was flanked by a 5' T7 polymerase promoter and a 3' Hepatitis Delta Virus ribozyme. The latter was inserted in a pUC57 vector by *Eco*RI and *Hind*III in the multiple cloning site (MCS) between positions 396 and 476. The total size of the recombinant plasmid was 13,833 bp.

Cell Cultures

Three different cell lines were transfected with the infectious WNV clone, including BHK-21 cells, Vero cells and BSR T7/5 cells. The BHK-21 cells and Vero cells used in this section of the study was acquired from the American Type Culture Collection (ATCC) and are the property of Deltamune. The BSR T7/5 cells were kindly donated by U. Buchholz (NIH, US). The BSR T7/5 cell line is a modified clone of the BHK-21 cell line that stably expresses the T7 RNA polymerase gene to allow transient expression of genes under the control of a T7 promotor. (Buchholz *et al.*, 1999). The latter was accomplished by transfection of BHK-21 cells with pSC6-T7-NEO that encodes the T7 RNA polymerase gene under control of a Cytomegalovirus (CMV) promoter (Buchholz *et al.*, 1999).

BHK-21, Vero and BSRT7/5 cell cultures were prepared by S. Smith and I. Wright (Deltamune). Both BHK-21 and Vero cell culture were propagated on 6-well culture plates (Nunc) in 10% medium consisting of Dulbecco's Minimum Essential Medium (DMEM) (Gibco), 10% FCS (Invitrogen) and 1% Penicillin/Streptomycin/Amphotericin B (Gibco) at 37 °C in 5% CO_2. BSR-T7 cells were propagated in 6-well culture plates (Nunc), 10% medium consisting of Glasgow MEM (Invitrogen), 1% L-Glutamine (200 mM) (Invitrogen), 2% MEM Amino Acid Solution (Invitrogen), 10% FCS (Invitrogen) and 1% Penicillin/Streptomycin/Amphotericin B (Gibco) at 37 °C in 5% CO_2 (Buchholz *et al.*, 1999). BSR T7/5 cell cultures were treated with Geneticin selection (Invitorgen) to a final concentration of 1 mg/mL every second passage. Cell cultures were transfected at 50% -70% confluency.

***In vitro* Transfection of Cell Cultures**

Prior to transfection with the infectious WNV clone, BHK-21 cell cultures; Vero cell cultures and BSR T7/5 were infected with the recombinant fowlpox virus fpEFLT7pol (Britton *et al.*, 1996). The latter was kindly donated by M. Skinner. The fpEFLT7pol virus stably expresses T7 RNA polymerase in mammalian cells to permit transient expression of the infectious WNV clone under the control of the T7 promoter. The fpEFLT7pol virus was propagated by C. Potgieter (Deltamune) in Chicken Embryo Cells (CER) at a Tissue Culture Infective Dose ($TCID_{50}$) of 10^6. The recombinant virus was harvested from the supernatant of infected CER cell cultures and used for subsequent infections. A single well of each cell culture plate containing BHK-21 cells, Vero cells and BSR T7/5 cells was infected with 750 µl of the fpEFLt7pol virus and incubated at 37 °C and 5% CO_2 for 30 - 60 minutes.

In addition to transfection with the infectious WNV clone, each cell culture was transfected with the pCI Mammalian Expression Vector (Promega) being used as a positive control for transfection

efficiency. The pCI Mammalian Expression Vector (Promega) expresses eGFP under the control of the human Cytomegalovirus (CMV) major immediate-early gene promotor region. The infectious WNV cone and pCI Mammalian Expression Vector (Promega) was prepared for transfection by the addition of 200 µl Optimem medium (Gibco) and 2 µl X-tremeGENE HP DNA Transfection Reagent (Roche) to 2 µg of plasmid DNA respectively and incubated at room temperature for 30 minutes. Following incubation, a single well of each cell culture plate containing BHK-21 cells, Vero cells and BSR T7/5 cells was transfected with the pCI Mammalian Expression Vector (Promega). The single well of each cell culture previously infected with the fpEFLT7pol virus was transfected with the infectious WNV clone. In addition to the latter, a single well of the BSR T7/5 cell culture plate was infected with only the infectious WNV clone. Transfected cell cultures were sealed and incubated at 37 °C and 5% CO_2. Cell cultures were inspected 24 hours after transfection for eGFP fluorescence of the pCI Mammalian Expression Vector (Promega) and for fluorescence of BSR T7/5 cells as an indication of successful transfection (Zeiss Axio Vert.A1). Transfected cell cultures were inspected microscopically daily for cytopathic effect (CPE) up to seven days post transfection (Zeiss Axio Vert.A1). Cells transfected with the infectious cDNA clone were passaged seven days post transfection.

Chapter 3

Results and Discussion

3.1 VIRUS PROPAGATION

3.1.1 WNV Propagation in mice

The following isolates were passaged once in suckling mice: isolate A (strain WNV1968), isolate D (strain WNV349/77), isolate E (strain WNV349/77) and isolate F (strain WNV349/77) (Table 2.1). The progression of neurological symptoms during the 7 day incubation period were recorded on a daily basis. The number of mice euthanized due to onset of neurological symptoms of WNV infection are indicated in Table 3.1. Mice infected with isolate F (strain WNV349/77) were first to display onset of neurological symptoms at three days post-inoculation (DPI), followed by isolate E (strain WNV349/77) and isolate B (strain WNV1968) at four DPI (Table 3.1). Mice infected with isolate D (strain WNV349/77) were last to display onset of neurological symptoms, starting at five DPI (Table 3.1). All mice were euthanized by seven DPI.

Table 3.1 Onset of neurological symptoms of WNV infection in mice

Strain	Isolate	# Mice	0 DPI	1 DPI	2 DPI	3 DPI	4 DPI	5DPI	6 DPI	7 DPI
WNV 1968	B	7	0/7	0/7	0/7	0/7	1/7	3/7	5/7	7/7
WNV349/77	D	16	0/16	0/16	0/16	0/16	0/16	5/16	11/16	16/16
	E	13	0/13	0/13	0/13	0/13	3/13	7/13	10/13	13/13
	F	13	0/13	0/13	0/13	3/13	7/13	8/13	12/13	13/13

DPI: Days Post Inoculation

3.1.2 Propagation in Cell Culture

3.1.2.1 Maintenance and Subculture of BHK-21 Cells

The BHK-21 cell cultures used for virus propagation were maintained and subcultured for WNV infection. The growth of newly seeded BHK-21 cells was recorded on a daily basis. Cultures were photographed two hours after seeding (Figure 3.1a), one day after seeding (Figure 3.1b), two days after seeding (Figure 3.1c) and three days after seeding (Figure 3.1d). Cultures reached 50% confluency one day after seeding, 70% confluency two days after seeding and 100% confluency three days after seeding (Figure 3.1).

3.1.2.2 WNV Propagation in BHK-21 Cell Culture

The following isolates were passaged once in BHK-21 cell cultures: isolate A (strain WNV1968), isolate B (strain WNV1968), isolate C (strain WNV349/77), isolate E (strain WNV349/77), isolate F (strain WNV349/77), isolate H (strain HS87/11) and isolate I (strain HS101/08) (Table 2.1). The progression of CPE in infected BHK-21 cell cultures during the seven day incubation period were recorded. Results are indicated in Table 3.2.

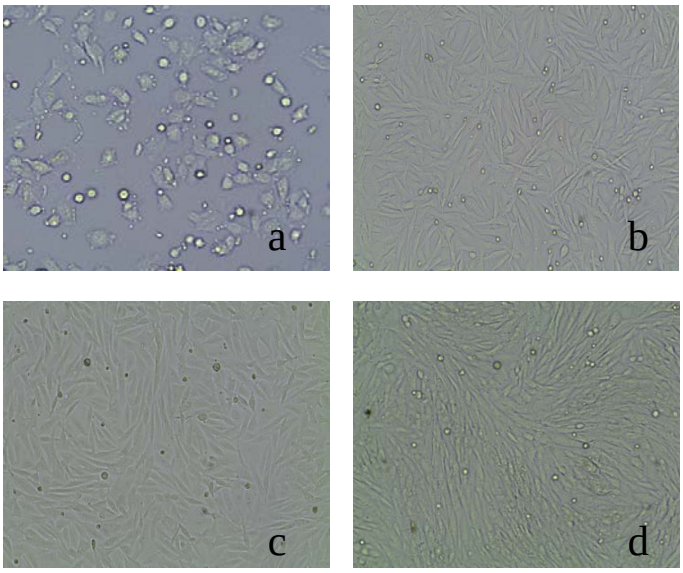

Figure 3.1 Growth of BHK-21 cell cultures under 10X magnification. (a) Two hours after seeding (b) One day after seeding, 50% confluency (c) Two days after seeding, 70% confluency (d) Three days after seeding, 100% confluent.

Cultures infected with isolate F (strain WNV349/77) was first to display onset of cytopathic effects (CPE) characteristic of WNV infection at three days post-inoculation (Table 3.2). Cultures infected with isolate A (strain WNV1968), isolate B (strain WNV1968), isolate C (strain WNV349/77), isolate E (strain WNV349/77) and isolate I (strain HS101/08) displayed onset of CPE four DPI (Table 3.2). Cultures infected with isolate H (strain HS87/11) displayed late onset of CPE attributed to WNV infection six DPI (Table 3.2). Virus was harvested from cell cultures when 70% CPE was reached, or alternatively when the seven day incubation period has been reached.

Table 3.2 Onset of CPE associated with WNV infection in BHK-21 cell culture

Strain	Isolate	0 DPI	1 DPI	2 DPI	3 DPI	4 DPI	5 DPI	6 DPI	7 DPI
WNV 1968	A	0%	0%	0%	0%	20%	40%	70%	-
	B	0%	0%	0%	0%	20%	40%	70%	-
WNV349/77	C	0%	0%	0%	0%	20%	40%	70%	-
	F	0%	0%	0%	20%	40%	70%	-	-
HS87/11	H	0%	0%	0%	0%	0%	0%	10%	20%
HS101/08	I	0%	0%	0%	0%	20%	40%	70%	-

DPI: Days Post Inoculation

3.2 RNA EXTRACTION

Viral RNA was extracted according to the single-step RNA isolation method using TRIzol® (Life Technologies) reagent (Simms *et al.*, 1993). The purity and concentration of RNA was quantified by spectrophotometry (NanoDrop). The RNA concentrations obtained were suitable for real-time qRT-

PCR (Table 3.3) The RNA obtained for each isolate ranged from near-optimal to optimal purity as indicated by the ratio of absorbance at 260 nm and 280 nm (Table 3.3).

Table 3.3 RNA concentration and purity determined by spectrophotometry

Strain	Isolate	Propagation System	Concentration (ng/µl)	$A_{260/280}$ [1]
WNV 1968	A1	Mouse Brain	109.4	2.1
	A2	BHK Cell Culture	137.6	1.89
	A3	Lyophilized	48.1	2.0
	B1	Mouse Brain	1253.1	1.92
	B2	BHK Cell Culture	103.2	1.87
	B3	Lyophilized	256.2	2.04
WNV349/77	C1	Mouse Brain	695.4	2.06
	C2	BHK Cell Culture	28.6	1.89
	C3	Lyophilized	105.4	2.06
	D1	Mouse Brain	347.9	2.01
	D3	Lyophilized	12.9	1.89
	E1	Mouse Brain	529.3	1.96
	E3	Lyophilized	27.2	1.82
	F1	Mouse Brain	771.2	2.01
	F2	BHK Cell Culture	76.8	1.86
	F3	Lyophilized	771.8	1.96
	G2	Lyophilized	68.4	2.04
HS87/11	H2	BHK Cell Culture	132.5	1.92
HS101/08	I2	BHK Cell Culture	114.9	1.95

[1] The ratio of absorbance at 260 nm and 280 nm. A ratio of ~2.0 is accepted as pure for RNA. A value higher or lower than ~2.0 indicates the presence of contaminants that absorb strongly at or near 260 nm or 280 nm respectively.

3.3 REAL-TIME QRT-PCR

A real-time qRT-PCR assay targeting the NS2A region of the WNV genome was used for semi-quantification of WNV isolates in different propagation systems (Eiden *et al.*, 2010). Quantification was based on a standard curve derived from 10-fold serial dilutions of a WNV lineage 2 positive control with a known Tissue Culture Infective Dose ($TCID_{50}$) of 3.1 x 10^5 (Figure 3.2 and 3.3). The cycle threshold (Ct) and associated $TCID_{50}$ of each dilution is indicated in Table 3.4.

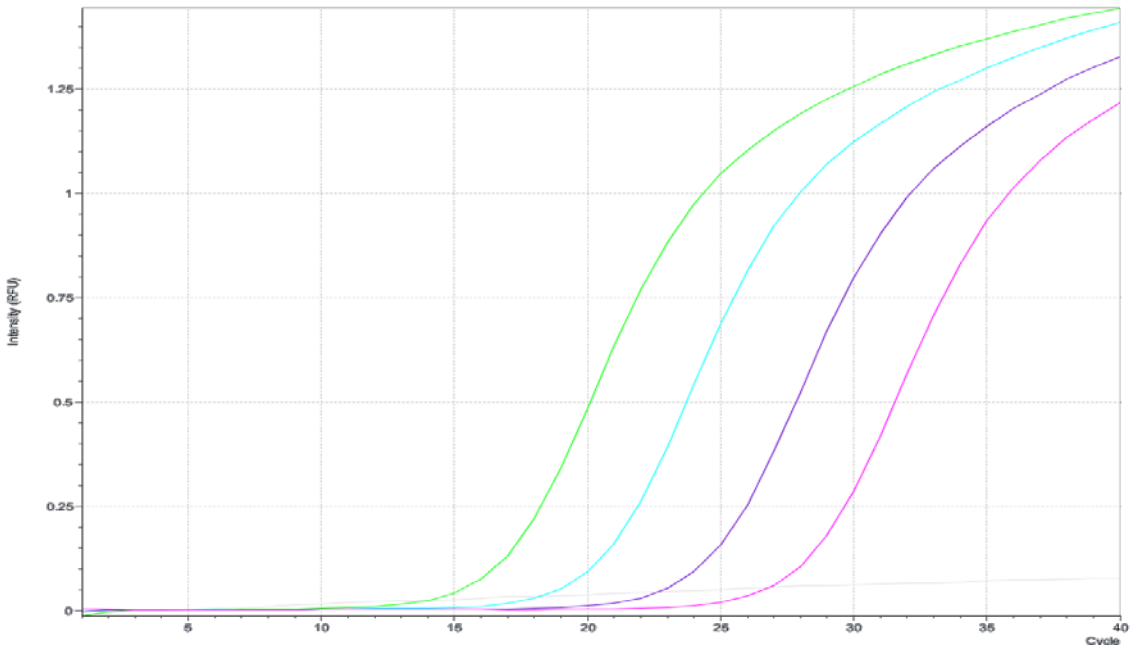

Figure 3.2 Real-Time qRT-PCR plot for serially diluted WNV lineage 2 positive control. The Vertical axis depicts fluorescence detected in relative fluorescence units (RFU). The horizontal axis depicts the number of PCR cycles.

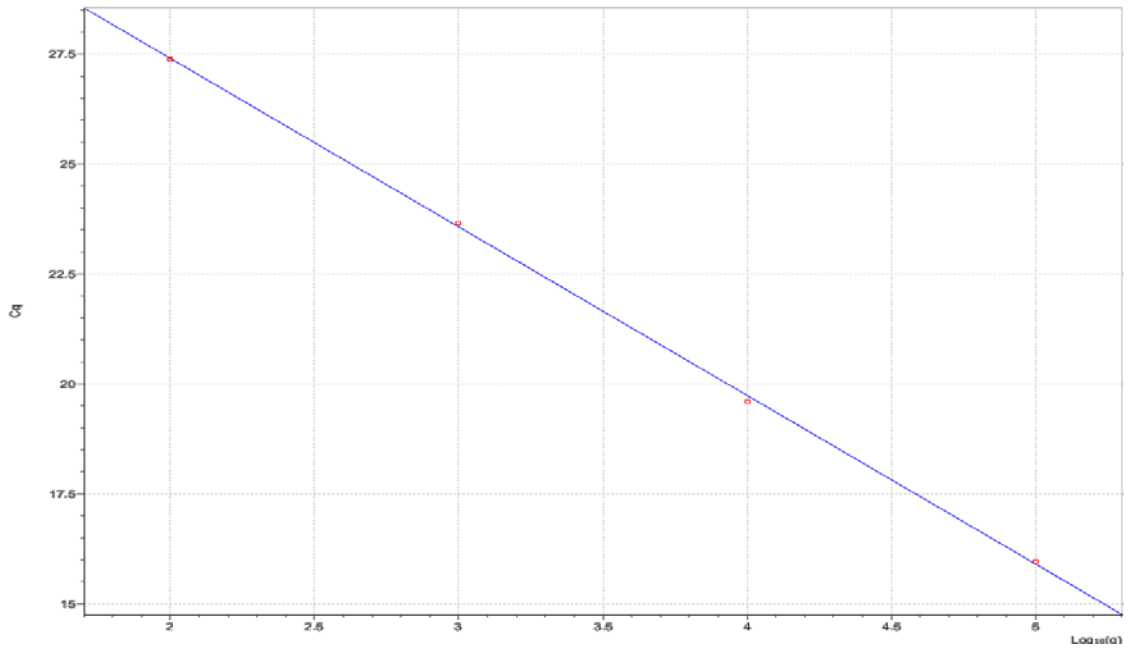

Figure 3.3 Standard curve calculated from Real-Time qRT-PCR plot for serially diluted WNV lineage 2 positive control. The Vertical axis depicts the cycle threshold (Ct). The horizontal axis depicts the starting quality.

Table 3.4 Real-Time qRT-PCR results of serially diluted WNV lineage 2 positive control

Strain	Dilution	Result	Ct[1]	Quantity[2]
WNV Lineage 2 Positive Control	10^5	Positive	15.951	9.6 X 10^4
	10^4	Positive	19.585	1.1 X 10^4
	10^3	Positive	23.639	9.5 X 10^2
	10^2	Positive	27.375	1.0 X 10^2
	10^1	Positive	30.074	2.0 X 10^1

[1]Cycle Threshold is defined as the number of cycles required for the fluorescent signal to cross the threshold or background level. Values are inversely proportional to the amount of target nucleic acid in the sample.

[2]Comparative $TCID_{50}$ per mL

The assay included each WNV isolate following RNA extraction from starting lyophilised viral material, infected mouse brain and infected BHK-21cell cultures respectively. Strong positive reactions were observed for all WNV isolates included in the assay (Figure 3.4). The cycle threshold (Ct) and comparative $TCID_{50}$ per mL of each WNV isolate is indicated in Table 3.5. Cycle Threshold (Ct) levels below 29 were detected for all WNV isolates with the exception of isolate H (HS 87/11), indicating abundant levels of target nucleic acid in each sample (Figure 3.4, Table 3.4).

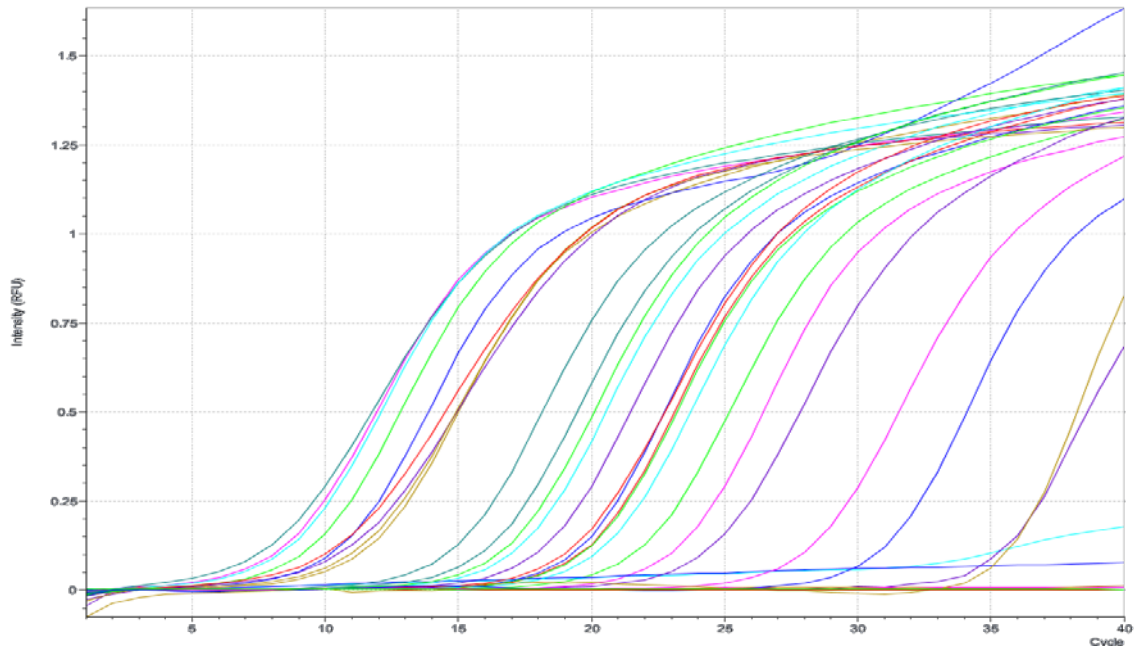

Figure 3.4 Real-Time qRT-PCR plot for WNV isolates propagated in different systems. The Vertical axis depicts fluorescence detected in relative fluorescence units (RFU). The horizontal axis depicts the number of PCR cycles.

Table 3.5 Real-Time qRT-PCR results of WNV isolates propagated in different systems

Strain	Isolate	Propagation System	Result	Ct	Quantity
WNV1968	A1	Mouse Brain	Positive	7.616	1.4×10^7
	A2	BHK Cell Culture	Positive	18.681	1.9×10^4
	A3	Lyophilized	Positive	10.787	2.1×10^6
	B1	Mouse Brain	Positive	10.288	2.9×10^6
	B2	BHK Cell Culture	Positive	22.361	2.0×10^3
	B3	Lyophilized	Positive	9.674	4.2×10^6
WNV249/77	C1	Mouse Brain	Positive	7.2	1.8×10^7
	C2	BHK Cell Culture	Positive	18.994	1.5×10^4
	C3	Lyophilized	Positive	8.619	7.9×10^6
	D1	Mouse Brain	Positive	12.071	1.1×10^6
	D3	Lyophilized	Positive	16.411	7.3×10^4
	E1	Mouse Brain	Positive	10.557	2.5×10^6
	E3	Lyophilized	Positive	14.028	3.1×10^5
	F1	Mouse Brain	Positive	9.8	3.9×10^6
	F2	BHK Cell Culture	Positive	19.069	1.5×10^4
	F3	Lyophilized	Positive	7.831	1.3×10^7
	G2	BHK Cell Culture	Positive	17.342	4.2×10^4
HS87/11	H2	BHK Cell Culture	Positive	34.567	6.8
HS101/08	I2	BHK Cell Culture	Positive	21.024	4.6×10^3
WNV Lineage 2 Positive Control			Positive	15.951	9.6×10^4

[1]Cycle Threshold is defined as the number of cycles required for the fluorescent signal to cross the threshold or background level. Values are inversely proportional to the amount of target nucleic acid in the sample.

[2]Comparative $TCID_{50}$ per mL

The highest comparative $TCID_{50}$ was observed for isolate A (WNV1968), isolate C (WNV249/77), isolate D (WNV249/77) and isolate E (WNV249/77) propagated in mouse brain during this study. Both isolate B (WNV1968) and isolate F (WNV249/77) displayed a higher comparative $TCID_{50}$ when tested from lyophilised virus originally propagated in mouse brain. Isolate G (WNV 349/77), isolate H (HS 87/11) and isolate I (HS 101/08) were only propagated in BHK cell culture in this study and displayed considerably lower comparative $TCID_{50}$ values compared to isolates propagated in mouse brain. Although isolate H (HS 87/11) tested positive, the high cycle threshold (Ct) value and extremely low comparative $TCID_{50}$ value may indicate the absence of live virus at the time of RNA extraction, or alternatively the degradation of RNA at the time of the assay.

3.4 NEXT GENERATION SEQUENCING

The isolates selected for next generation sequencing and their corresponding passage history is indicated in Table 3.6. The genomes of isolate A (WNV 1968) and isolate C (WNV 349/77) was sequenced following RNA extraction from lyophilised material (Table 3.6). In addition, the genomes of isolate B (WNV 1968), isolate D (WNV 349/77), isolate E (WNV 349/77) and isolate F (WNV 349/77) were sequenced following RNA extraction from infected mouse brain (Table 3.6). The genomes of isolate G (WNV 349/77), isolate H (HS 87/11) and isolate I (HS 101/08) were sequenced following RNA extraction from infected BHK-21 cell cultures (Table 3.6).

Table 3.6 Passage history of sequenced isolates

Strain	Isolate	Passage History
WNV 1968	A3	Lyophilised as MB #3 → Sequence
	B1	Lyophilised as MB #1 → Passaged MB #1 → Sequence
WNV349/77	C3	Lyophilised as MB #8 → Sequence
	D1	Lyophilised as BHK #3 → Passaged MB #1→ Sequence
	E1	Lyophilised as BHK #5 → Passaged MB #1→ Sequence
	F1	Lyophilised as MB #8 → Passaged MB #1→ Sequence
	G2	Lyophilised as BHK #7 → Passaged BHK #1→ Sequence
HS87/11	H2	Ongoing Culture BHK→ Sequence
HS101/08	I2	Ongoing Culture BHK→ Sequence

3.4.1 Transcriptome Amplification and cDNA Library Preparation

The transcriptomes of WNV isolates selected for whole genome sequencing (WGS) were amplified with the use of TransPlex Whole Transcriptome Amplification (WTA) kit (Sigma Aldrich). The amplified cDNA libaries were purified using High Pure PCR Product Purification Kit (Roche) and

quantified by with a fluorescence-based Qubit quantitation assay. The genomic cDNA library concentration of each WNV isolate is indicated in Table 3.7.

Table 3.7 Concentration of cDNA libraries generated for each isolate

Strain	Isolate Sequenced	cDNA Library Concentration (ng/µl)
WNV 1968	A3	42.6
	B1	36.3
WNV349/77	C3	45.1
	D1	34.4
	E1	82.3
	F1	32.7
	G2	4.8
HS87/11	H2	48.6
HS101/08	I2	45.4

3.5 DATA ANALYSIS

3.5.1 Genome Assembly

The consensus genome for each WNV isolate was assembled using CLC Genomics Workbench v5.1.5 (http://www.clcbio.com/). Sequence reads were trimmed according to quality scores and length. Adapter sequences were removed prior to assembly. Trimmed reads were mapped to the complete genome sequence of a Lineage 2 WNV isolate (GenBank accession number EF429198.1).

3.5.1.1 Trimming

The read statistics of each WNV isolate that pertains to the trimming process is indicated in Table 3.8. An average of 1,828,179 total reads were obtained, ranging from 1,568,384 for isolate A (WNV 1968) to 2,054,204 for isolate D (WNV 349/77) (Table 3.8). The high number of sequence reads can be ascribed to the presence of non-viral cDNA in the sequencing reaction. The latter is the result of the total RNA extraction method used in this study. The average percentage of trimmed reads were 99.70556, ranging from 99.36 % for isolate A (WNV 1968) to 99.85% for isolate D (WNV 349/77) (Table 3.8). The mean read length after trimming ranged between 79.4 bp for isolate C (WNV 349/77) and 90 bp for isolate H (HS 87/11) (Table 3.8).

Table 3.8 Trimmed sequence read statistics

Strain	Isolate	Total Read Count	Trimmed Read Count	% Trimmed	Mean Read Length After Trim
WNV 1968	A3	1,568,384	1,562,587	99.36	89.1
	B1	2,054,204	2,051,074	99.85	80.7
WNV349/77	C3	1,646,236	1,653,756	99.36	79.4
	D1	2,054,204	2,051,074	99.85	80.7
	E1	1,620,826	1,618,204	99.84	82.5
	F1	1,972,894	1,969,863	99.85	82.1
	G2	1,929,256	1,926,236	99.84	81.4
HS87/11	H2	1,876,898	1,871,364	99.71	90
HS101/08	I2	1,730,712	1,725,291	99.69	87.5

3.5.1.2 Mapping

The trimmed reads of each WNV isolate were mapped to the complete genome sequence of a Lineage 2 WNV isolate (GenBank accession number EF429198.1). The read statistics of each WNV isolate that pertain to the mapping process are indicated in Table 3.9. An average of 1,833,775 reads were mapped, ranging from 311 reads for isolate H (HS 87/11) to 963,103 for isolate E (WNV 349/77) (Table 3.9). The average fraction of the reference genome that was covered during mapping was 0.923333, where the fraction of most isolates was near - or equal to 1.00, with the exception of isolate H (HS 87/11) for which it was 0.36 (Table 3.9). The average coverage level ranged from 2.03 for isolate H (HS 87/11) to 6,923.01 for isolate E (WNV 349/77) (Table 3.9). The mean read length of reads that mapped to the reference genome ranged from 82.43 bp for isolate B (WNV 1968) and 95.79 for isolate A (WNV 1968) (Table 3.9).

Table 3.9 Mapped sequence read statistics

Strain	Isolate	Total Trimmed Read Count	Matched Read Count	Mean Read Length	Fraction Reference Coverage	Average Coverage Level
WNV 1968	A3	1,562,587	709,030	95.79	0.99	5,118.92
	B1	2,051,074	693,868	82.43	1.00	4,749.81
WNV349/77	C3	1,625,850	628,670	90.99	1.00	4,125.35
	D1	2,051,074	693,868	82.43	1.00	5,478.51
	E1	1,618,204	963,103	84.36	1.00	6,923.01
	F1	1,966,832	175,646	82.68	1.00	1,202.94

Strain	Isolate	Total Trimmed Read Count	Matched Read Count	Mean Read Length	Fraction Reference Coverage	Average Coverage Level
	G2	1,923,216	371,000	82.71	1.00	2,543.41
HS87/11	H2	1,871,364	311	86	0.36	2.03
HS101/08	I2	1,720,38	2,509	92.43	0.96	17.07

3.5.2 Phylogenetic Analysis

The phylogenetic relationships between WNV isolates in this study and viruses representing all five phylogenetic lineages of WNV were determined by Maximum Likelihood (ML) and Bayesian Inference (BI) based on the 921 bp C-prM-E gene region. Trees generated by ML and BI were of similar topology, and bootstrap support values obtained from ML analysis are therefore indicated together with the posterior probabilities obtained from Bayesian analysis (Figure 3.5). Bootstrap support values determined for nodes during ML analysis are indicated above branches where applicable (Figure 3.5). Similarly, posterior probabilities determined for nodes during BI are indicated below branches where applicable (Figure 3.5). A Neighbour-Joining (NJ) tree was included to illustrate the relative branch lengths (Figure 3.6).

Nodal support values determined for Bayesian Inference (BI) and Maximum Likelihood (ML) trees supported the separation of WNV isolates into five distinct clades analogous to the five previously defined lineages of WNV (Figure 3.5). WNV isolates AY490240 (China), AY278441 (Russia, AF260967 (Romania), AF404757 (Italy), AY701413 (Morocco), AY660002 (Mexico) and AF260967 (USA) were grouped in a monophyletic clade representative of WNV lineage 1A with nodal support of 89% and 97% from ML and BI, respectively. Kunjin virus isolate D00246 (Australia) represents lineage 1B and is separated from isolates in lineage 1A with bootstrap support of 97% and 100% respectively.

Isolate C3 (WNV 349/77), isolate D1 (WNV 349/77), isolate E1 (WNV 349/77), isolate F1 (WNV 349/77), isolate G2 (WNV 349/77), isolate H2(HS 87/11), M12294 (Uganda), isolate A3 (WNV 1968), isolate B1 (WNV 1968), EF429198 (SA), isolate I2 (HS 101/08) and DQ176636 (Madagascar) were grouped in a monophyletic clade representative of WNV lineage 2 with nodal support of 99% and 91% respectively. Isolate I2 (HS 101/08) and EF4219198 (SA) separate from respective isolates of WNV 349/77, isolate H2 (HS 87/11), respective isolates of WNV 1968 and M12294 (Uganda) with nodal support of 80% determined by ML. Isolate DQ176636 (Madagascar) was separated from the remainder of the isolates grouped within lineage 2 with nodal support of 99% and 100% respectively.

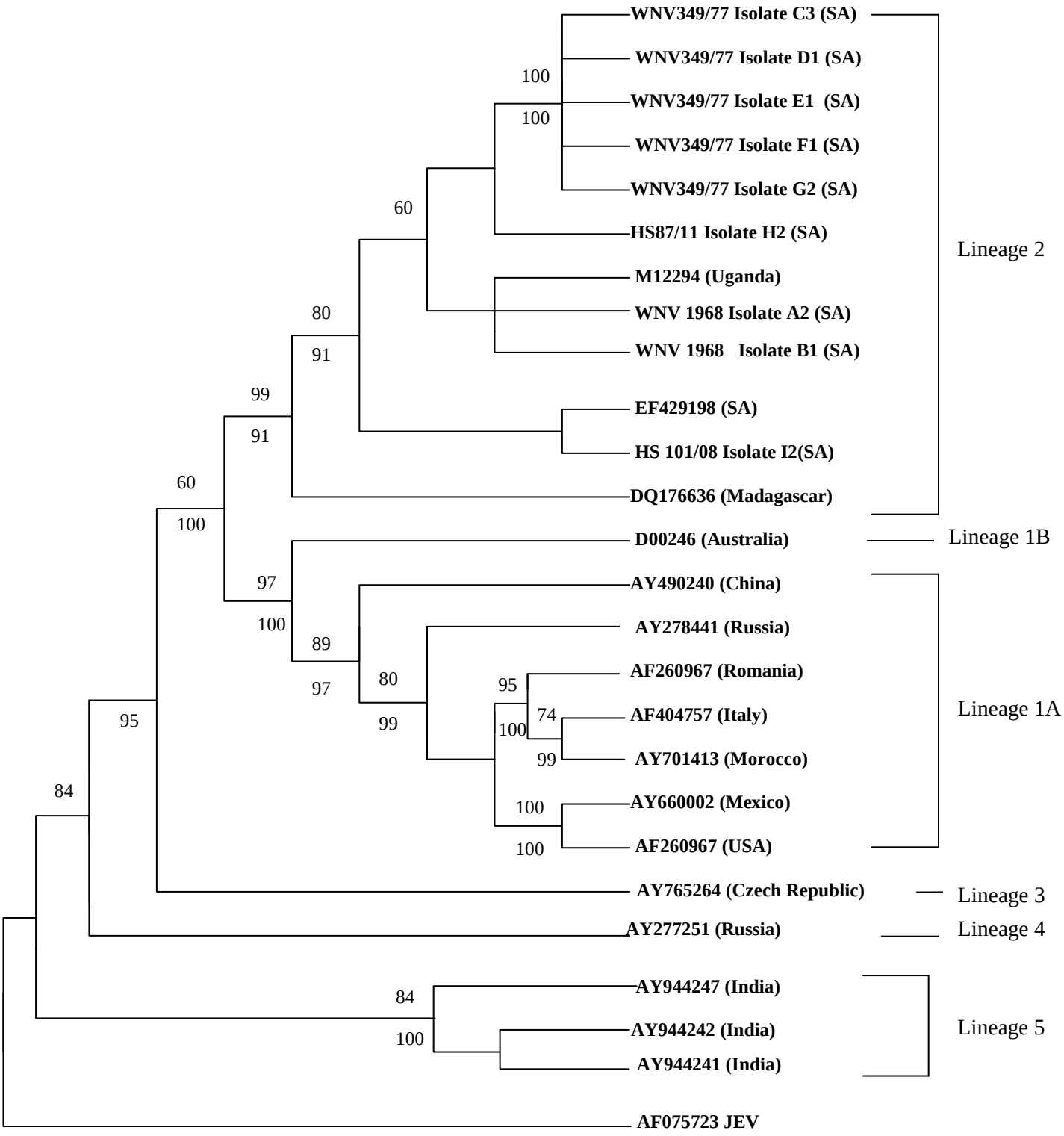

Figure 3.5 Phylogenetic tree of WNV isolates based on the C-prM-E gene region. Trees generated using Maximum Likelihood and Bayesian Inference were identical in topology. Bootstrap support values obtained from Maximum Likelihood analysis are indicated above branches where applicable, and values greater than 60% are shown. Posterior probabilities obtained from Bayesian Inference are indicated below branches where applicable, and values greater than 90% are shown. Strain JEV-GP78 AF075723 was defined as outgroup. GenBank accession numbers for the complete genomic sequences included in phylogenetic analysis are: WNFCG Wengler M12294 (Uganda), HS 93/01 EF429198 (SA), ANMY798 DQ176636 (Madagascar), KUNJIN D00246 (Australia), CHIN-01 AY490240 (China), AST 99-901 AY278441 (Russia), RO 97-50AF260967 (Romania), ITALY 1998EQ AF404757 (Italy), MOROCCO 04.05 AY701413 (Morocco), Mexico TM171-3 AY660002 (Mexico), NY99-EQ AF260967 (USA), RABENSBURG 697-1023 AY765264 (Czech Republic), LEIV-vlg 99-27889 AY277251 (Russia), 821622 AY944247 (India), 80897 AY944242 (India), 622886 AY944241 (India).

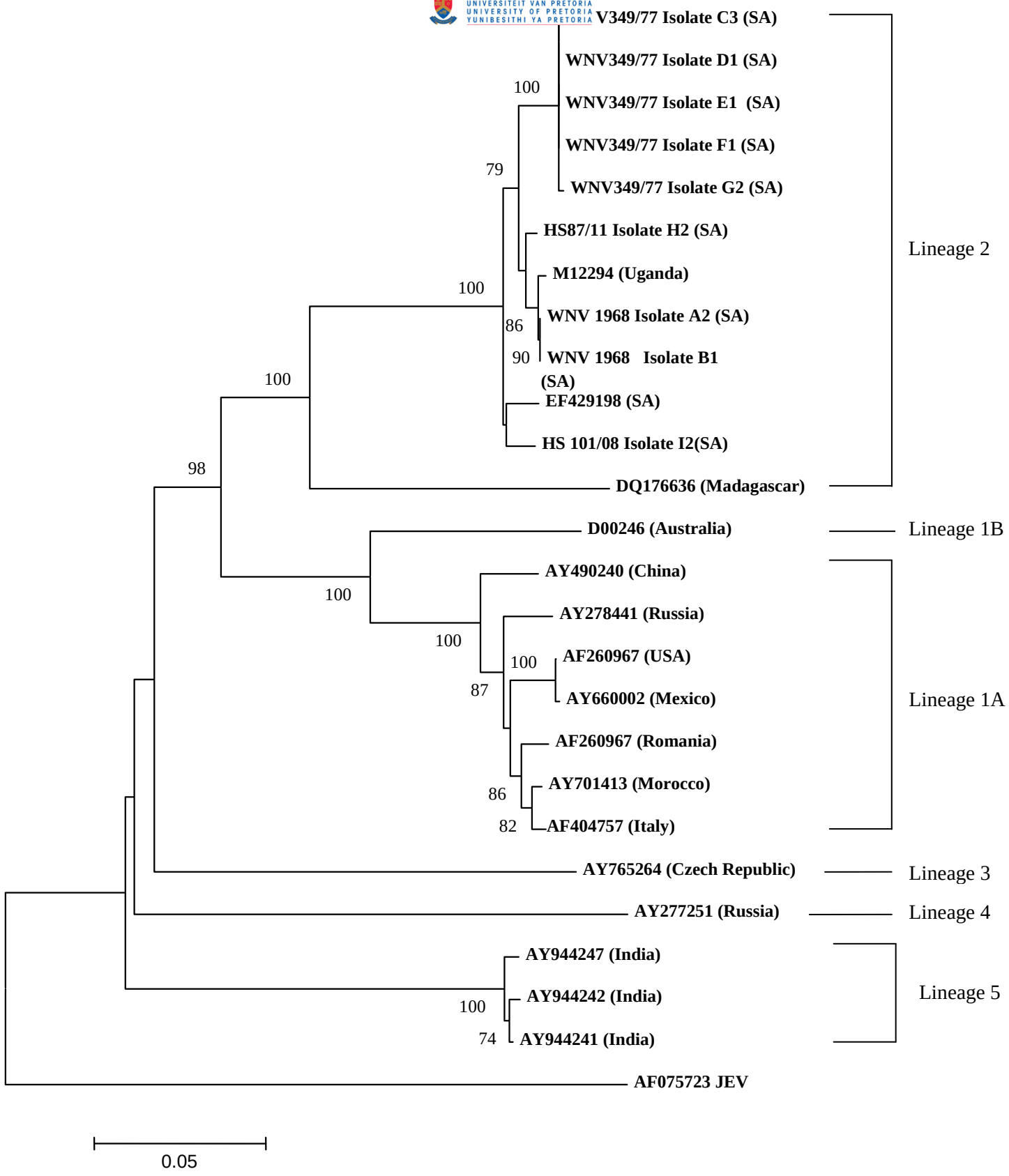

Figure 3.6 Neighbour-Joining tree of WNV isolates based on the C-prM-E gene region. Tree was constructed based on p-distance amongst taxa by neighbour-joining. Bootstrap (1000 replicates) values greater than 60% are indicated on branches. Strain JEV-GP78 AF075723 was defined as outgroup. GenBank accession numbers for the complete genomic sequences included in phylogenetic analysis are: WNFCG Wengler M12294 (Uganda), HS 93/01 EF429198 (SA), ANMY798 DQ176636 (Madagascar), KUNJIN D00246 (Australia), CHIN-01 AY490240 (China), AST 99-901 AY278441 (Russia), RO 97-50AF260967 (Romania), ITALY 1998EQ AF404757 (Italy), MOROCCO 04.05 AY701413 (Morocco), Mexico TM171-3 AY660002 (Mexico), NY99-EQ AF260967 (USA), RABENSBURG 697-1023 AY765264 (Czech Republic), LEIV-vlg 99-27889 AY277251 (Russia), 821622 AY944247 (India), 80897 AY944242 (India), 622886 AY944241 (India).

Isolate AY765264 (Czech Republic), representing WNV lineage 3, was separated from WNV isolates that grouped within lineage 1A, lineage 1B and lineage 2 with nodal support of 100% as determined by BI. Isolate AY277251 (Russia), representing WNV lineage 4, was separated from WNV isolates that grouped within lineage 1A, lineage 1B, lineage 2 and lineage 3 with nodal support of 95% as determined by BI. Isolate AY944247 (India), AY944242 (India) and AY944241 (India) were grouped in a monophyletic group representative of WNV lineage 5, and was separated from WNV isolates grouped within lineage 1A, lineage 1B, lineage 2, lineage 3 and lineage 4 with nodal support of 95% as determined by BI. Trees were rooted to a Japanese Encephalitis virus (JEV) isolate with GenBank accession number AF075723.

The consensus genome sequence obtained for all WNV 349/77 isolates were identical based on pairwise distance calculations as well as phylogenetic analysis (Figure 3.5 and 3.6, Table D1). Similarly, the consensus genome sequence obtained for both WNV 1968 isolates were identical (Figure 3.5 and 3.6, Table D1). The differences in passage number between isolate A3 and isolate B1 did not cause genetic change in the consensus genome of WNV 1968. Regarding WNV 349/77, the consensus genome of isolates propagated in mouse brain, BHK-21 cell cultures, and isolates that were switched from one propagation system to another displayed no genetic changes in the consensus genome of WNV 349/77. The lack of variation observed amongst consensus genome sequences of WNV 1968 and WNV 349/77 isolates is suggestive of a well-maintained mutation-selection equilibrium within the environment in which the respective isolates were propagated.

3.5.3 Single Nucleotide Polymorphism (SNP) Detection

Single nucleotide polymorphisms (SNPs) were determined in CLC Genomics Workbench v5.1.5 (http://www.clcbio.com/). The complete genome sequence of a Lineage 2 WNV isolate (GenBank accession number EF429198.1) was used as reference genome. Genome-wide SNP profiles were compared between WNV 1968 isolates, as well as between WNV 349/77 isolates. Isolate H2 (HS 87/11) and isolate I2 (HS 101/08) were not considered for SNP analysis due to the low average coverage levels obtained during mapping (Table 3.9). The results are discussed in the remainder of this section.

3.5.3.1 SNP number and Position

The SNP profiles of isolate A3 (WNV 1968) and isolate B1 (WNV 1968) were compared based on nucleotide position in the genome. Results were visualized using Circos (Krzywinski *et al.*, 2009) (Figure 3.7). The genome of each isolate is depicted by a band to the right of Figure 3.5, where "a" represents isolate A3 and "b" represents isolate B1. The nucleotide positions where SNPs have been detected are illustrated to the left of Figure 3.5. Genomes are connected to nucleotide positions where

SNPs occur by means of a colour-coded ribbon. In the instance where a SNP was present in both genomes in the same nucleotide position, the nucleotide position is connected to both genomes by two similarly coloured ribbons. As an example, the SNP at position 10281 occurred in the genomes of both isolate A3 and isolate B1 and is connected to both genomes by means of a red ribbon (Figure 3.7). In contrast, a SNP that was present in only one of the two genomes, a single ribbon connects the nucleotide position with the genome. As an example, the SNP occurring at position 2683 occurred in the genome of isolate A3 only and is connected to that isolates' genome by means of a green ribbon (Figure 3.7).

A total of 49 SNPs were detected in the two WNV 1968 isolates sequenced (Figure 3.7). The number of SNPs that occurred in both isolate A3 and isolate B1 corresponded to 51% of the total number of SNPs (Figure 3.7). The number of SNPs that were unique to isolate A3 comprised 24% of the combined total (Figure 3.5). Similarly, 24% of the SNPs identified were unique to isolate B1 (Figure 3.7).

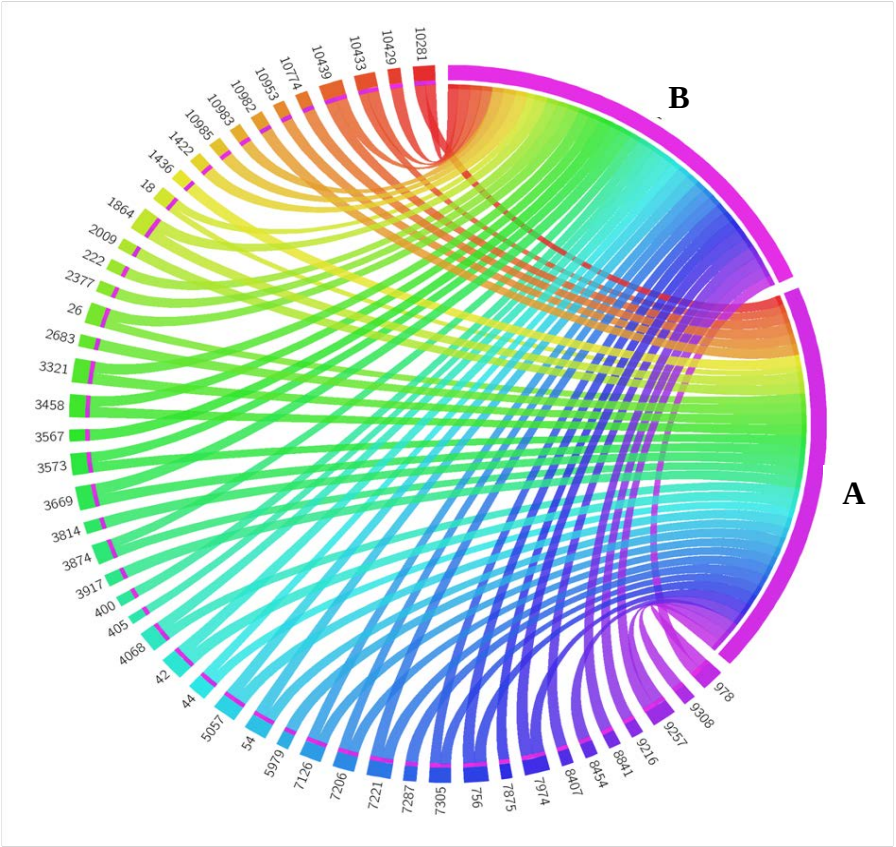

Figure 3.7 Position and number of Single Nucleotide Polymorhpisms (SNPs) of WNV 1968 isolates. The genome of isolate A3 and isolateB1 is depicted by the letters "A" and "B" respectively. The numbers in the diagram depict nucleotide positions where SNPs have been detected**.** Genomes are connected to nucleotide positions where SNPs occur by means of a colour-coded ribbon. In the instance where a SNP was present in both genomes in the same nucleotide position, the nucleotide position is connected to both genomes by two similarly coloured ribbons.

Concerning WNV 349/77 isolates, a combined total of 98 unique SNPs were detected (Figure 3.8).The number of SNPs that occurred in all five isolates comprised 7% of the combined total number of SNPs (Figure 3.8). The number of SNPs shared amongst four isolates, three isolates and two isolates was 4%, 11% and 9% of the total combined number of SNPs each (Figure 3.8). The remaining 69% of the SNPs were unique to just one of the five isolates sequenced (Figure 3.8).

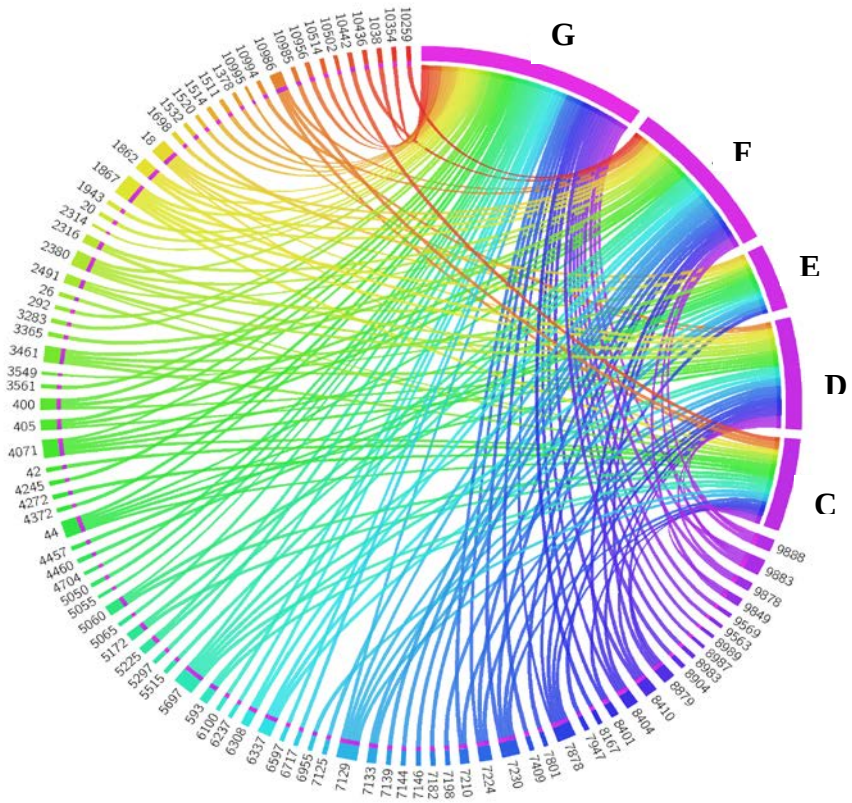

Figure 3.8 Position and number of Single Nucleotide Polymorhpisms (SNPs) of WNV 349/77 isolates. The genomes of isolates C3, D1, E1, F1, and G2 are each depicted by letters C, D, E, F and G. The numbers in the diagram depict the nucleotide position at which SNPs were detected**.** Genomes are connected to nucleotide positions where SNPs occur by means of a colour-coded ribbon. In the instance where a SNP is present in both genomes at the same nucleotide position, the nucleotide position is connected to both genomes by two similarly coloured ribbons.

In order to identify gene regions most influenced by changes in propagations system, the diversity of gene regions were compared based on the number of SNPs observed. The standardised number of SNPs per gene region for WNV 1968 isolates are illustrated in Figure 3.9 and Figure 3.10. The standard deviation in number of SNPs per gene region amongst WNV 1968 isolates is illustrated in Figure 3.11. Similarly, the standardised number of SNPs per gene region for WNV 349/77 isolates is illustrated in Figures 3.12 and 3.13. The standard deviation in number of SNPs per gene region amongst WNV 349/77 isolates is illustrated in Figure 3.14.

The most variable gene region of the WNV 1968 genome was determined to be the 5'UTR region, followed by the 3'UTR region, the NS2A gene region, the NS4B gene region, the membrane gene region, the NS5 gene region, the NS1 gene region, the NS3 gene region and lastly the NS2B and NS4A gene regions (Figure 3.9 and 3.10).

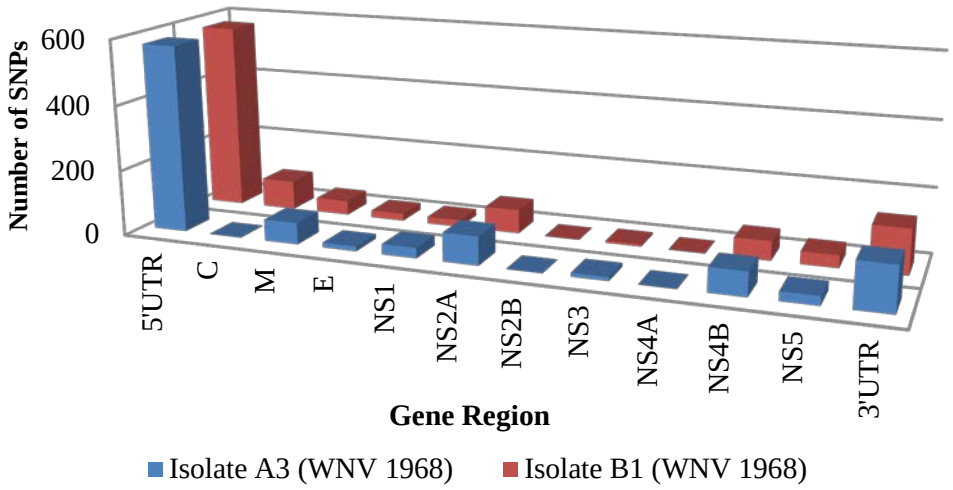

Figure 3.9 Number of Single Nucleotide Polymorphisms (SNPs) per gene region of WNV 1968 isolates. The number of SNPs was standardised to accommodate for the size difference of gene regions. Isolate A3 (WNV 1968) is represented in blue and isolate B1 (WNV 1968) is represented in red. Gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

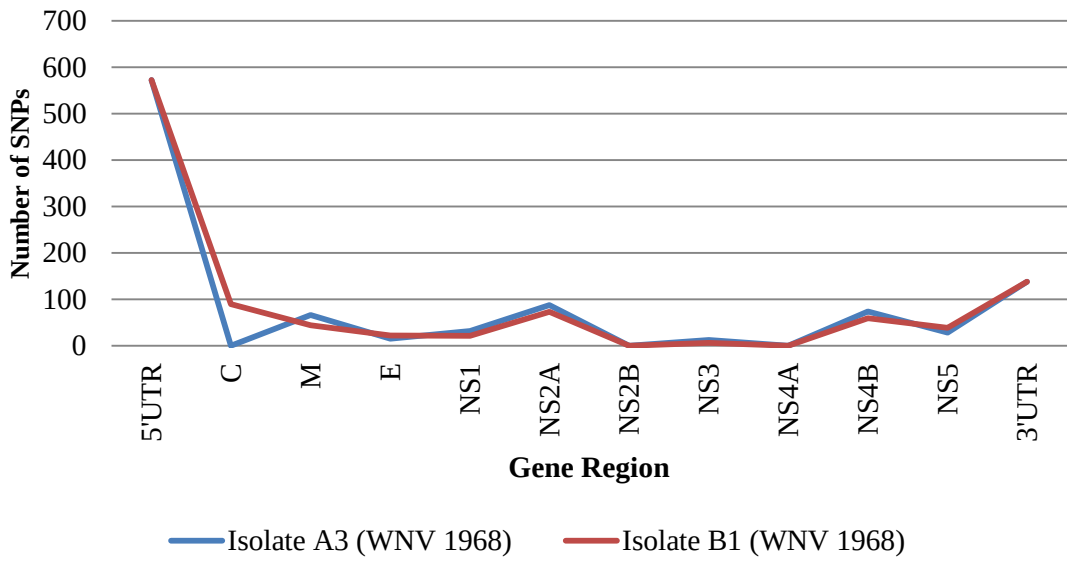

Figure 3.10 Number of Single Nucleotide Polymorphisms (SNPs) per gene region of WNV 1968 isolates. The number of SNPs was standardised to accommodate for the size difference of gene regions. Isolate A3 (WNV 1968) is represented in blue and isolate B1 (WNV 1968) is represented in red. Gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

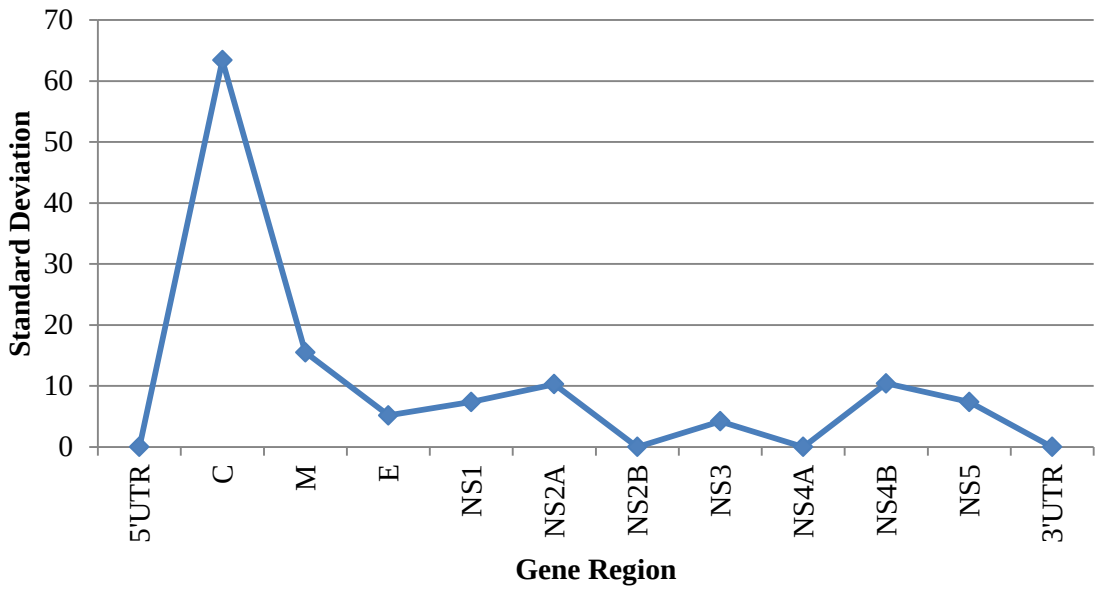

Figure 3.11 Standard deviation in the number of Single Nucleotide Polymorphisms (SNPs) per gene region amongst WNV 1968 isolates. Gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

Concerning the WNV 349/77 genome, the most variable region was estimated as the 5'UTR region, followed by the 3'UTR region, the capsid gene region, the NS4B region, the membrane gene region, the envelope gene region, the NS3 region, the NS5 region, the NS4A region and lastly the NS1 and NS2A regions (Figure 3.12 and 3.13). The highest variation in number of SNPs between the respective isolates of WNV 349/77 was observed for the 5'UTR region, followed by the 3'UTR region, the capsid gene region, the membrane gene region, the envelope gene region, the NS4B region, the NS2A region, the NS5 region, the NS3 region, the NS4A region and lastly the NS1 and NS2A regions (Figure 3.14).

Both the 5'UTR and 3'UTR regions are highly variable amongst flavivirues (Lindenbach and Rice, 2003, Brinton and Dispoto, 1988). The 3'UTR and 5'UTR contain common secondary structures necessary for genome replication (Brinton and Dispoto, 1988). Similarly, the NS2A gene region is poorly conserved (Falgout and Markoff, 1995). The NS2A protein interacts with replicase components of virus replication and coordinates the shift between RNA packaging and RNA replication (Khromykh *et al*., 2001). The high variation observed in the 5'UTR region, the 3'UTR region and the NS2A region in comparison with other gene regions is therefore expected to occur. In contrast, the high variation observed in the capsid protein gene region of WNV 349/77 is significant as the capsid protein forms an integral part in the assembly of infectious virions the capsid protein facilitates membrane association and membrane protein (prM) translocation to the endoplasmic reticulum (Ma *et al*., 2004). The assembly of RNA replication complexes, in turn, is known to occur on intracellular membranes (Ahlquist *et al.*, 2003). As such, the increased variation in the capsid

gene region is suggestive of selection pressures brought about by differences in host cell type between propagation systems.

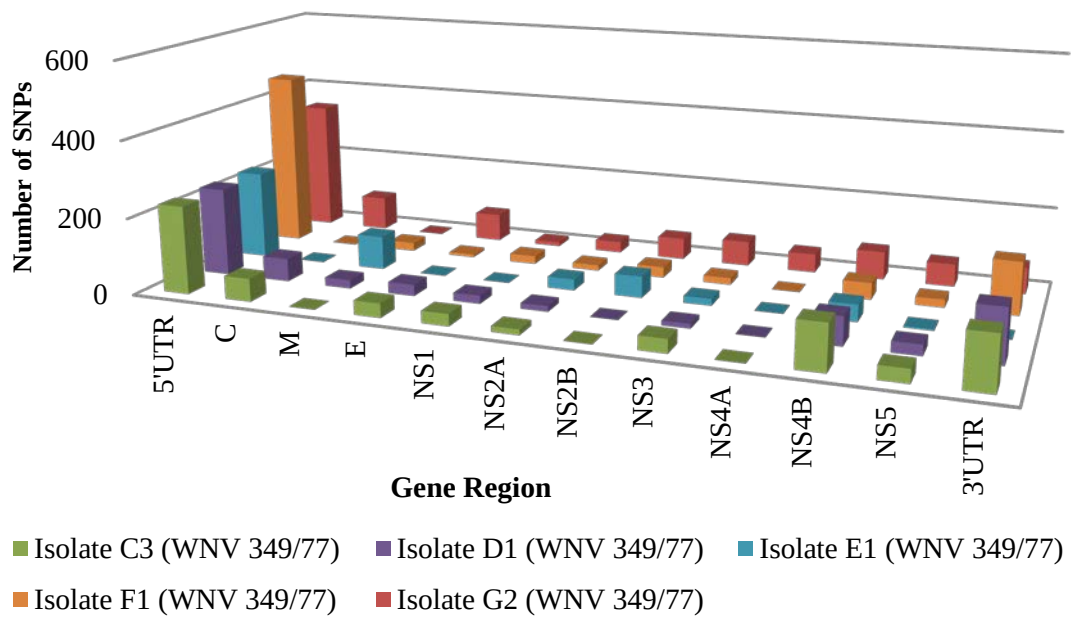

Figure 3.12 Number of Single Nucleotide Polymorphisms (SNPs) per gene region of WNV 349/77 isolates. The number of SNPs was standardised to accommodate for the size difference of gene regions. Isolate C3 (WNV 349/77) is represented in green, isolate D1 (WVN 349/77) in purple, isolate E1 (WNV 349/77) in blue, isolate F1 (WNV 349/77) in orange and isolate G2 (WNV 349/77) in red. Gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

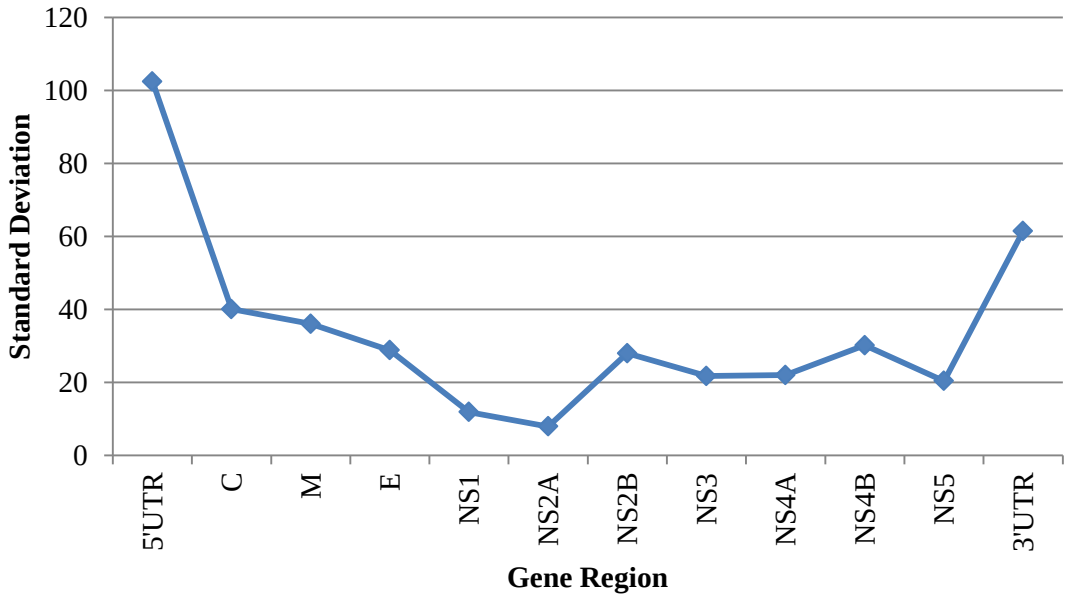

Figure 3.14 Standard deviation in the number of Single Nucleotide Polymorphisms (SNPs) per gene region amongst WNV 349/77 isolates. Gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

3.5.3.1 SNP Frequency

Isolates were compared based on the frequency of the major allele of each SNP in order to study the influence of propagation system and passage number on SNP frequency. For the purpose of this discussion, only SNPs that were present in both WNV 1968 isolates were considered (Figure 3.15). Similarly, SNPs that were present in more than one WNV 349/77 isolate were considered (Figure 3.16).

The SNPs that occurred in both isolate A3 and isolate B1 and their respective frequencies based on the major allele are illustrated in Figure 3.13. A total number of five SNPs were observed in the 5'UTR region, one SNP in the membrane gene region, two SNPs in the envelope gene region, two SNPs in the NS1 region, four SNPs in the NS2A region, 1 SNP in the NS3 region, 4 SNPs in the NS4B region and 5 SNPs in the NS5 region (Figure 3.15). Uniform frequencies were observed for the majority of SNPs present in both isolate A3 and isolate B1. Higher frequencies were observed for SNPs in position 18 (5'UTR), 25 (5'UTR), 3669 (NS2A) and 10433 (NS5) in isolate B1. The major allele at each SNP position remained identical between isolate a and isolate B1, with the exception of the SNP at position 18. The latter was present as adenine at a frequency of 48.9 in isolate A3, and guanine at a frequency of 66.7 in isolate B1. In both cases the codon associated with this nucleotide translates to alanine and the mutation is therefore synonymous.

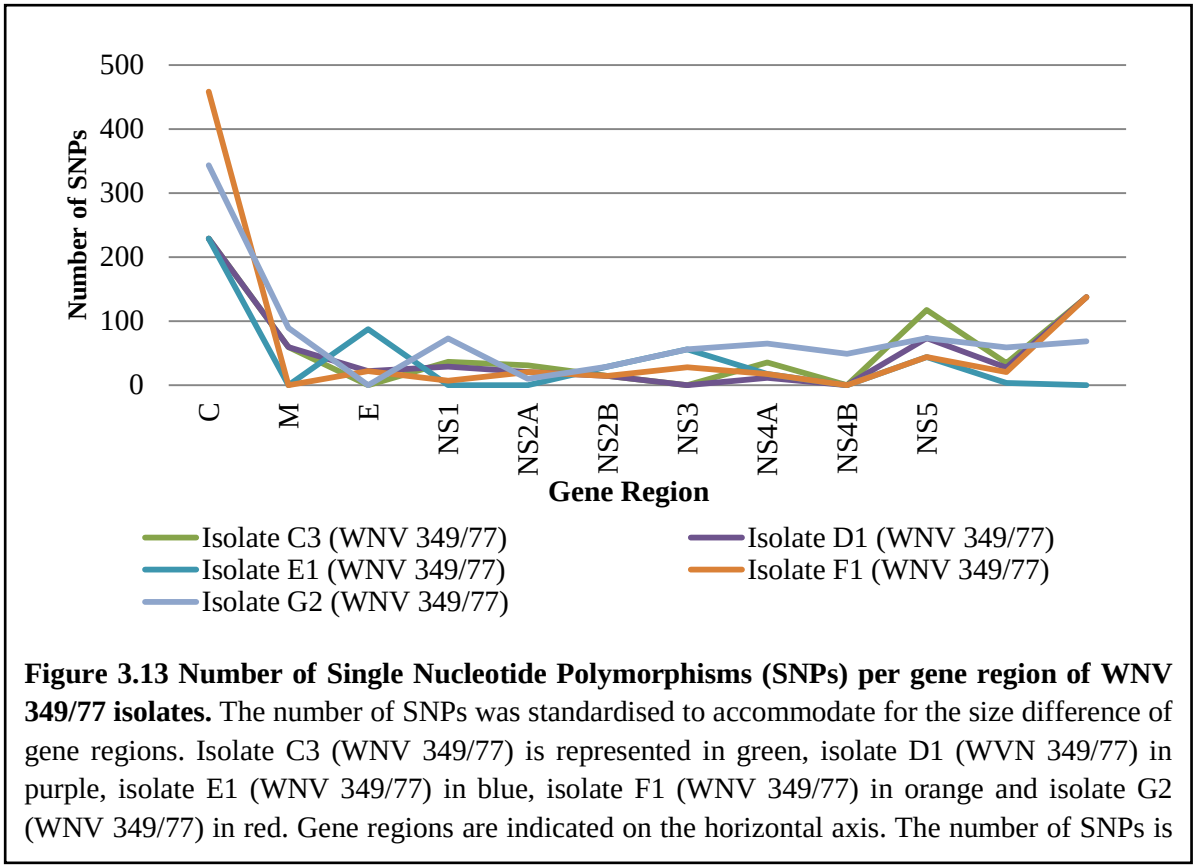

Figure 3.13 Number of Single Nucleotide Polymorphisms (SNPs) per gene region of WNV 349/77 isolates. The number of SNPs was standardised to accommodate for the size difference of gene regions. Isolate C3 (WNV 349/77) is represented in green, isolate D1 (WVN 349/77) in purple, isolate E1 (WNV 349/77) in blue, isolate F1 (WNV 349/77) in orange and isolate G2 (WNV 349/77) in red. Gene regions are indicated on the horizontal axis. The number of SNPs is

Increased frequencies of SNPs were observed in the 5'UTR region, NS2A region and NS5 region in isolate B1. The increased frequencies corresponds to a single passage in mouse brain, as opposed to three passages in mouse brain for isolate A3. In light of this, the variation in SNP frequencies between isolate A3 and isolate B1 may be ascribed to the process of attaining an equilibrium frequency towards the most fit genotype during adaptation to the propagation system.

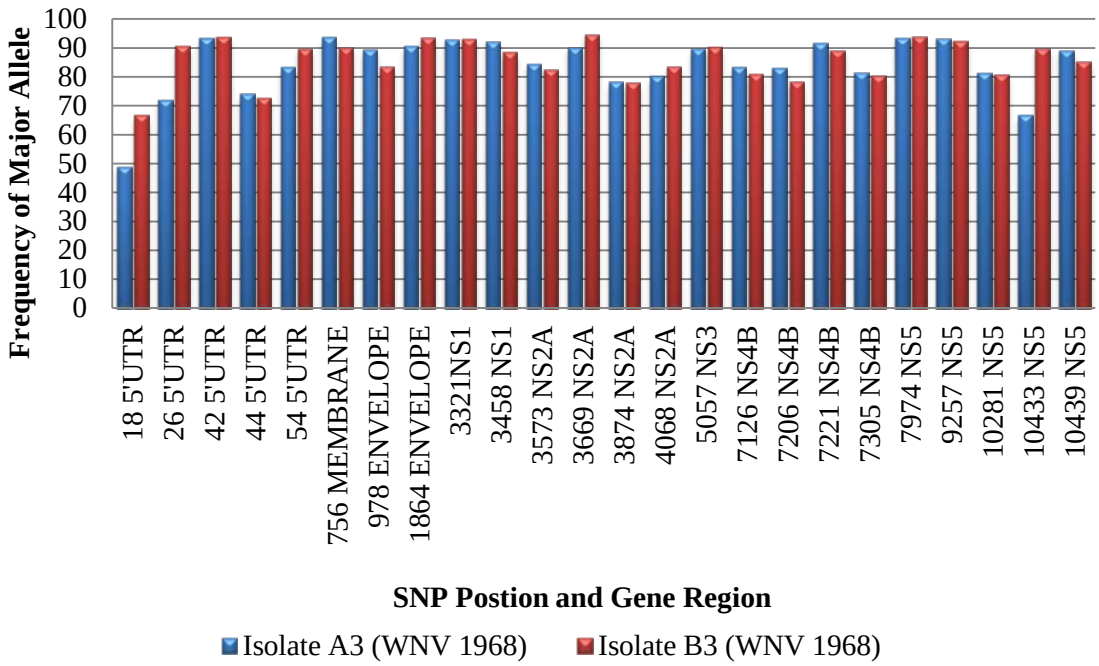

Figure 3.15 Frequency of Single Nucleotide Polymorphisms (SNPs) present in both WNV 1968 isolates. The frequencies that are shown represent the frequency of the major allele. Isolate A3 (WNV 1968) is represented in blue and isolate B1 (WNV 1968) is represented in red.. Frequencies represent the frequency of the major allele in each Gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

The variation in SNP frequencies between WNV 349/77 isolates were substantially higher than that between WNV 1968 isolates (Figure 3.16).The SNPs that occurred in all WNV 349/77 isolates and their respective frequencies based on the major allele are illustrated in Figure 3.14. A total of two SNPs were observed in the 5'UTR region, two SNPs in the capsid gene region, one SNP in the membrane gene region, six SNPs in the envelope gene region, one SNP in the NS1gene region, one SNP in the NS2A gene region, four SNPs in the NS3 gene region, one SNP in the NS4A gene region, five SNPs in the NS4B region, eight SNPs in the NS5 region and one SNP in the 3'UTR region (Figure 3.16). The major allele at each SNP position remained identical between WNV 349/77 isolates, with the exception of SNPs at position 2316 and 2491.

The SNP in position 2316 is situated in the envelope gene region. Whereas thymine is observed in position 2316 in the WNV 349/77 consensus genome, cytosine was observed in isolate C3 at a frequency of 92.5. This non-synonymous substitution resulted in an amino acid change from phenylalanine in the WNV 349/77 consensus genome to an arginine in isolate C3. This suggests that the substitution in position 2316 is under positive selection when the WNV 349/77 strain is passaged in mouse brain, and that the associated changes to the envelope protein confers elevated fitness in this propagation system.

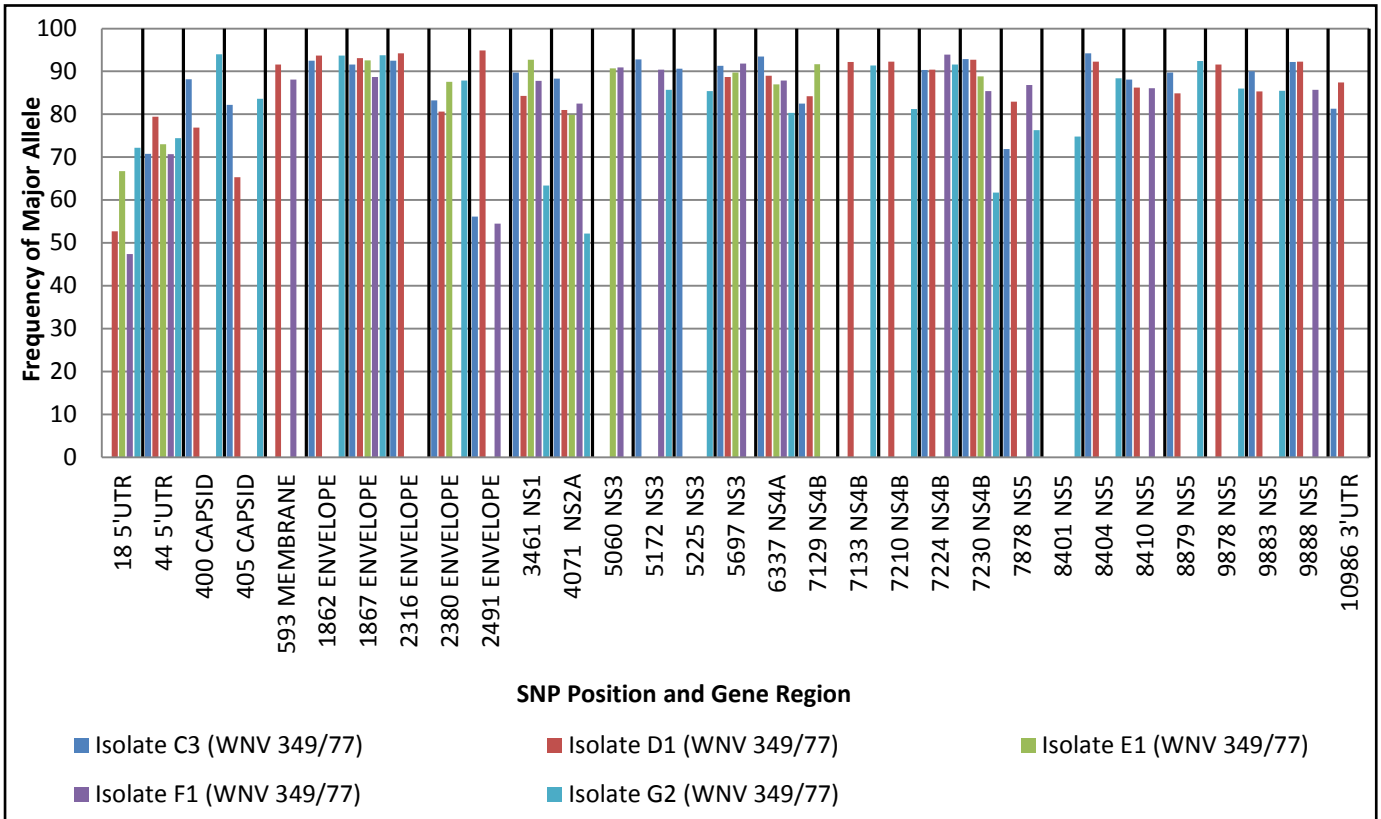

Figure 3.16 Frequency of Single Nucleotide Polymorphisms (SNPs) present in more than one WNV 349/77 isolate. The frequencies that are shown represent the frequency of the major allele. Isolate C3 (WNV 349/77) is represented in blue, isolate D1 (WNV 349/77) in red, isolate E1 (WNV 349/77) in green, isolate F1 (WNV 349/77) in purple and isolate G2 (WNV 349/77) in orange. Frequencies represent the frequency of the major allele in each gene regions are indicated on the horizontal axis. The number of SNPs is indicated on the vertical axis.

Similarly, the SNP in position 2491 is situated in the envelope gene region. The WNV 349/77 consensus genome contains adenine in this position, whereas guanine is observed in isolate C3 and isolate D1 at a frequency of 56.1 and 94.9 in isolate D1. The non-synomous substitution resulted in an amino acid change from isoleucine in the WNV 349/77 consensus genome to valine in isolates C3 and D1. Interestingly, adenine is observed at a frequency of 54.5 in isolate F1. Although the latter is agreement with the WNV 349/77 consensus genome, the frequency shows little deviation from that of

guanine observed in isolate C3. Both isolate C3 and isolate F1 were propagated in mouse brain, with the exception that isolate C3 was passaged eight times and isolate F1 was passaged nine times. The passage history of isolate D1, however, concerns 3 passages in BHK-21 cell culture followed by one passage in mouse brain. When considering the passage history of all three isolates, the significance of the SNP in position 2491 becomes apparent. Firstly, the results suggest that quasispecies variants of WNV 349/77 that contain either adenine or guanine in position 2491 are maintained at a near equal frequency in the population. Secondly, the presence of adenine and the associated amino acid isoleucine in the envelope protein confers elevated fitness when WNV 349/77 is passaged in mouse brain, and the population equilibrium shifts towards adenine in position 2491 with the increase in passage number. In support of the latter, the high frequency at which guanine persists in isolate D1 when switching from 3 passages in BHK-21 cell culture to one passage in mouse brain suggests that the presence of guanine and the associated amino acid change to valine in the envelope protein confers elevated fitness in BHK-21 cell culture. The frequency of guanine declines sharply from 94.9 to 56.1 with increased number of passages in mouse brain as illustrated in isolate C3. It is therefore suggested that adenine in position 2491 is under positive selection when WNV 349/77 is passaged in mouse brain, whereas guanine is under positive selection when passaged in BHK-21 cell culture.

3.5.4 Quasispecies Reconstruction

The haplotypes of WNV 349/77 isolates were reconstructed from ultra deep sequence data in ShoRAH (Beerenwinkel and Zagordi, 2011). The passage history of isolate C3 coincided with that of isolate F1, and isolate C3 was omitted from the analysis. Although the coverage obtained for isolate C3 was higher than that of isolate F1, the sequence data for isolate C3 could not be resampled successfully for unknown reasons. The analysis was resumed based on sequence data for isolate F instead.

The haplotypes obtained consisted of full-length genomes aligned to the consensus genome of the isolate involved (Figure 3.17). The haplotypes obtained for isolate D1, isolate E1, isolate F1 and isolate G2 were compared. The total number of haplotypes, the number of haplotypes with a posterior probability (PSP) above 0.8 and the number of viable haplotypes for each isolate is indicated in Table 3.10. The number of haplotypes of any one isolate that were present in one or more of the other isolates is indicated in Table 3.10.

The highest number of haplotypes with a PSP>0.8 were obtained subsequent to quasispecies reconstruction from sequence data of isolate F1, whereas the total number of haplotypes ranged from 567 for isolate E1 to 738 for isolate F1 (Table 3.10). The number of haplotypes with a posterior probability above 0.8 ranged from 331 (isolate G2) and 386 (isolate F1) (Table 3.10).

Table 3.10 Haplotypes obtained subsequent to quasispecies reconstruction

Isolate	Total # Haplotypes	# Haplotypes with PSP > 0.8	# Viable Haplotypes	# Viable Haplotypes Shared
Isolate D1	659	338	64	20
Isolate E1	567	339	56	19
Isolate F1	572	386	72	21
Isolate G2	738	331	45	15

PSP: Posterior Probability

Figure 3.17 Haplotype sequence alignment obtained from quasispecies reconstruction. Quasispecies were reconstructed in ShoRAH (Zagordi *et al.*, 2011). The haplotypes obtained consisted of full-length genomes aligned to the reference genome. Similarities with the consensus genome of the isolate involved were represented by an indel. Nucleotides that differed from the consensus genome of the isolate involved are shown in scaffolds.

The highest number of viable hapltoypes were obtained for isolate F1, ranging from 45 for isolate G2 to 72 for isolate F1 (Table 3.10). The number of haplotypes that were found to be identical to haplotypes in other isolates ranged from 15 for isolate G2 to 21 for isolate F1 (Table 3.10). Isolates were compared based on the number and frequencies of haplotypes respectively.

3.5.4.1 Variation in Haplotype Number

In order to identify gene regions most influenced by changes in propagations system, the diversity of gene regions were compared based on the number of haplotypes observed. The standardised number

of haplotypes per gene region for isolate D1, isolate E1, isolate F1and isolate G2 are illustrated in Figure 3.18 and 3.19. The standard deviation in number of haplotypes per gene region are illustrated in Figure 3.20.

Based on variation in the number of haplotypes amongst gene regions, the most variable region of the WNV 349/77 genome was estimated as the 5'UTR region, followed by the capsid gene region, the 3'UTR region, the NS3 region, the envelope gene region, the NS2B region, the NS2A and NS4A regions, the NS5 region, the membrane gene region, and the NS1 and NS4B regions (Figure 3.18). According to the variation in number of haplotypes between WNV 349/77 isolates within each gene region, the most variable regions were the 5'UTR region; the capsid gene region and the 3'UTR region. These findings are in agreement with the gene regions found to be most variable according to SNP data.

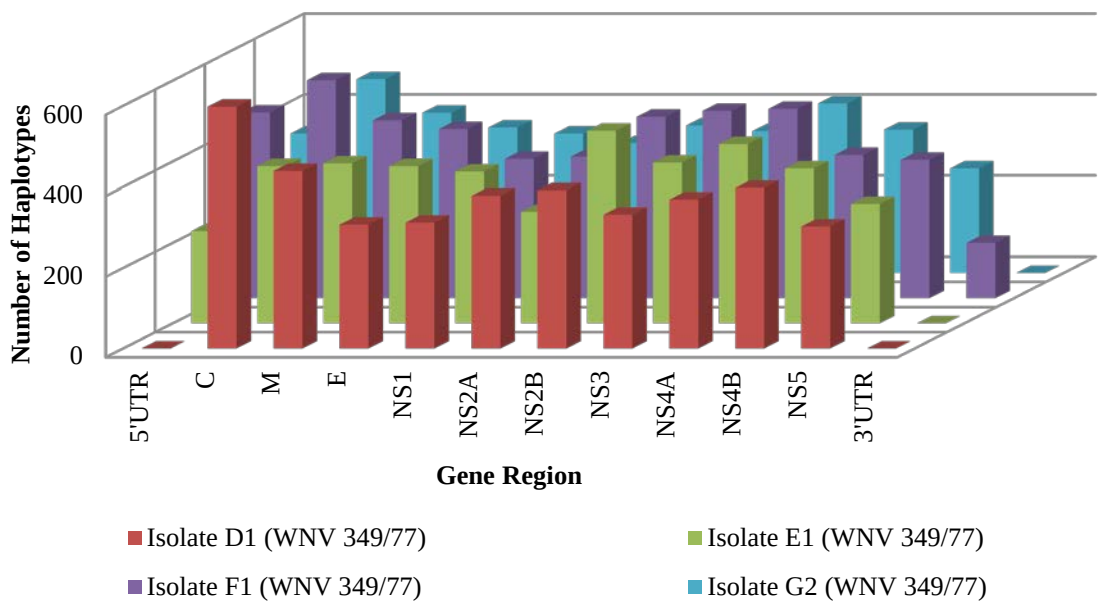

Figure 3.18 Number of haplotypes per gene region of WNV 349/77 isolates. The number of haplotypes was standardised to accommodate for the size difference of gene regions. Isolate D1 (WNV 349/77) is represented in red, isolate E1 (WVN 349/77) in green, isolate F1 (WNV 349/77) in purple and isolate G2 (WNV 349/77) in blue. Gene regions are indicated on the horizontal axis. The number of haplotypes is indicated on the vertical axis.

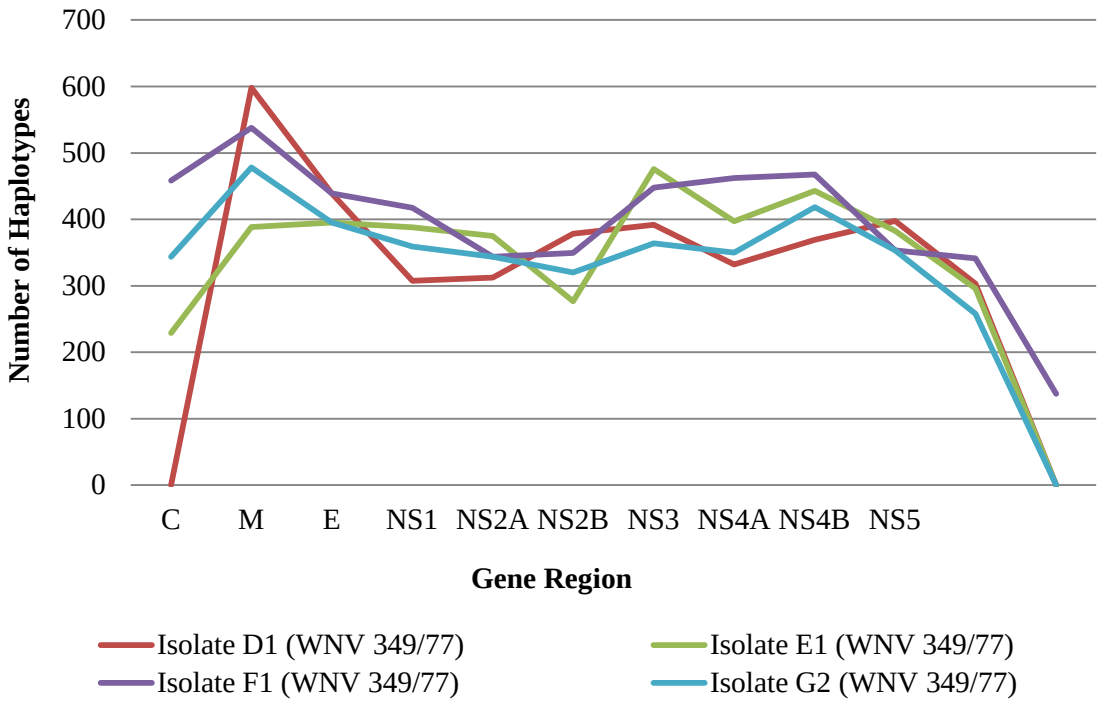

Figure 3.19 Number of haplotypes per gene region of WNV 349/77 isolates. The number of haplotypes was standardised to accommodate for the size difference of gene regions. Isolate D1 (WNV 349/77) is represented in red, isolate E1 (WVN 349/77) in green, isolate F1 (WNV 349/77) in purple and isolate G2 (WNV 349/77) in blue. Gene regions are indicated on the horizontal axis. The number of haplotypes is indicated on the vertical axis.

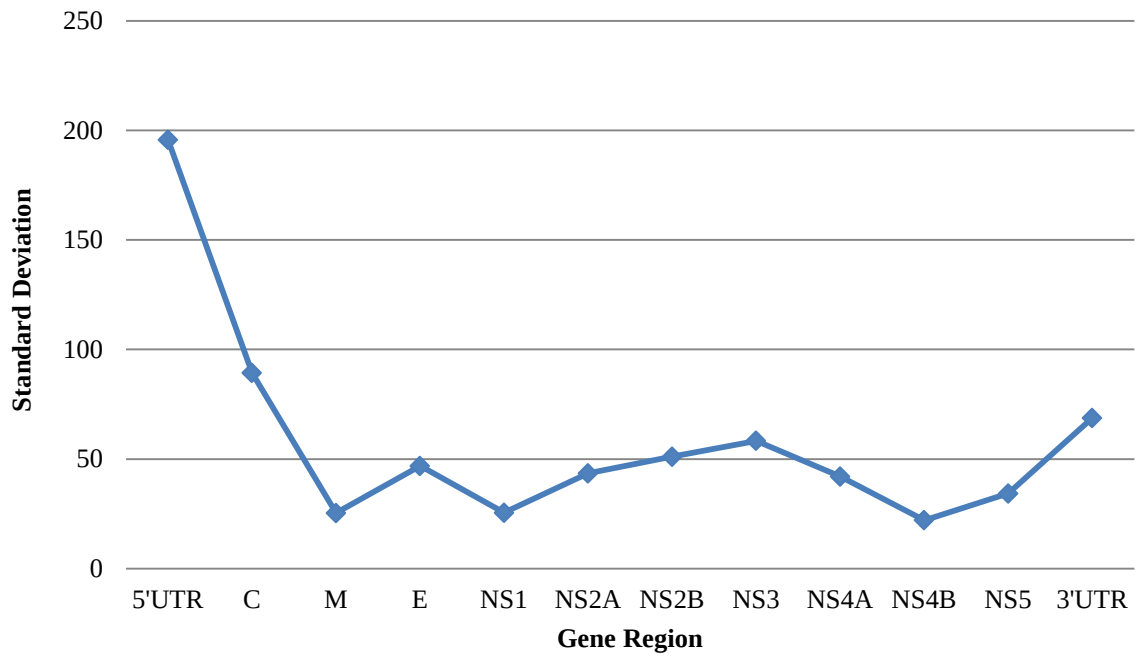

Figure 3.20 Standard deviation in the number of haplotypes per gene region amongst WNV 349/77 isolates. Gene regions are indicated on the horizontal axis. The number of haplotypes is indicated on the vertical axis.

3.5.4.1 Variation in Haplotype Frequency

The frequencies of viable haplotypes that were present in more than one isolate were compared in order to study the influence of propagation system on the quasispecies composition of WNV 349/77. The shared haplotypes were renamed Haplotype 1 through to Haplotype 30 and grouped according to the gene region in which each respective haplotype displayed variation (Appendix C). The variation in haplotype frequency between isolates is illustrated in Figure 3.21. The standard deviation in haplotype frequency between isolates is illustrated in Figure 3. 22.

Haplotypes that varied most in frequency between isolates were observed in the capsid gene region, followed by the envelope gene region and the NS2A region (Figure 3.21 and 3.22). The frequency of Haplotype 20 differed most between isolates, followed by Haplotype 25 and Haplotype 22 (Figure 3.21 and 3.22). For the purposes of this discussion, the emphasis will be placed on the latter.

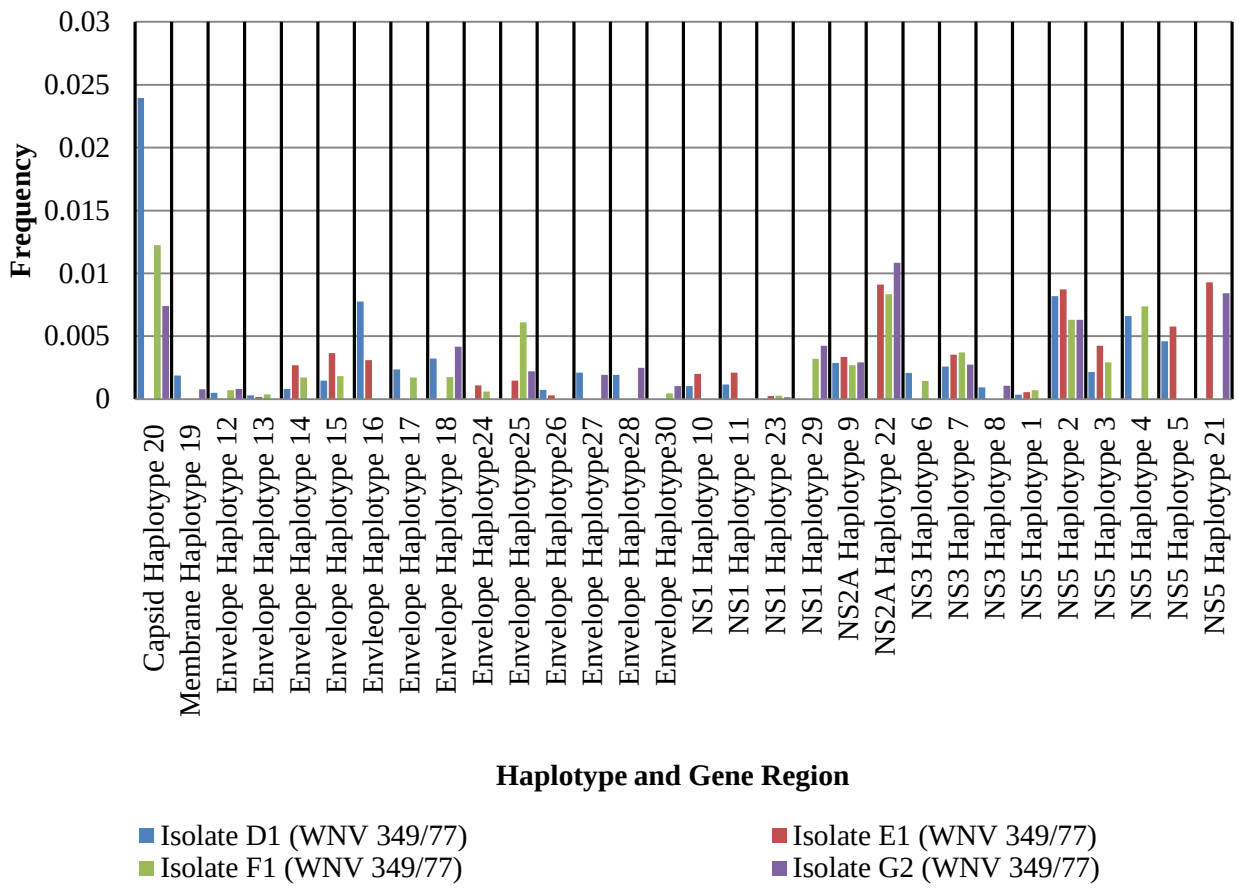

Figure 3.21 Frequency of viable haplotypes present in each respective WNV 349/77 isolate. Isolate D1 (WNV 349/77) is represented in blue, isolate E1 (WNV 349/77) in red, isolate F1 (WNV 349/77) in green and isolate G2 (WNV 349/77) in purple. Haplotype and corresponding gene regions are indicated on the horizontal axis. Haplotype frequency is indicated on the vertical axis.

Haplotype 25 contained variation in the NS2A region between genome position 367 and 3770. The latter can be ascribed to deletion of a single nucleotide (A) in position 2286, causing a frameshift mutation in the envelope gene. A pairwise distance of 0.0081 was observed between the genome sequence of Haplotype 22 and the WNV 349/77 consensus genome. Haplotype 22 was observed at a frequency of 0.01083 in isolate G2, 0.00912 in isolate E1 and 0.00835 in isolate F1. The latter was therefore most prevalent when WNV 349/77 was passaged continuously in BHK-21 cell culture. Haplotype 20 as observed at an intermediate frequency when WNV 349/77 was subject to a change in propagation system from BHK-21 cell culture to mouse brain, and least prevalent when continuously passaged in mouse brain. This suggests that Haplotype 20 confers elevated fitness when WNV 349/77 is propagated in BHK-21 cell culture. Considering the reduction in frequency from continuous propagation in BHK-21 cell culture to that of continuous propagation in mouse brain, it is evident that Haplotype 20 is subject to selection when WNV 349/77 is adapting to BHK-21 cell culture.

Haplotype 22 contained variation in the envelope gene region when compared to the consensus genome of WNV 349/77. The pairwise distance between the genome sequence of Haplotype 25 and the latter was estimated at 0.00667. Variation was observed between genome positions 3672 and 3770 due to a single nucleotide single nucleotide deletion (T) in position 3672, causing a frameshift mutation in the envelope gene. Haplotype 22 was observed at a frequency of 0.01083 in isolate G2, 0.00912 in isolate E1 and 0.00835 in isolate F1. Haplotype 22 was therefore most prevalent when WNV 349/77 was passaged continuously in BHK-21 cell culture. Haplotype 22 was less prevalent when WNV 349/77 was subjected to a change in propagation system, and least prevalent when propagated continuously in mice. Similar to Haplotype 25, results suggest that Haplotype 22 is subject to selection when WVN 349/77 is propagated in BHK-21 cell culture.

The highest frequency of Haplotype 22 was observed in isolate G2, followed by isolate E1 and isolate F1. Haplotype 22 was observed at a frequency of 0.01083 in isolate G2, 0.00912 in isolate E1 and 0.00835 in isolate F1 (Appendix C). A pairwise distance of 0.0081 was observed between the genome sequence of Haplotype 22 and the WNV 349/77 consensus genome. Nucleotide variation was observed in the NS2A region between genome positions 3672 and 3770. The observed variation can be ascribed to a single nucleotide deletion (T) in position 3672, causing a frameshift mutation in the envelope gene of Haplotype 22.

The highest variation amongst WNV 349/77 isolates was observed in Haplotype 20. A pairwise distance of 0.00714 was observed between the genome sequence of Haplotype 20 and the WNV 349/77 genome sequence. With respect to the WNV 349/77 consensus genome, Haplotype 20 contained variation in the capsid gene region between genome position 221 and 318. The variation observed in this genome region can be ascribed to the deletion of a single nucleotide (G) in from

position 221, causing a frameshift mutation in the capsid gene of Haplotype 20. Haplotype 20 was therefore most prevalent when WNV 349/77 was subject to a change in propagation system from BHK-21 cell culture to mouse brain. An intermediate prevalence was observed when WNV 349/77 was continuously passaged in mouse brain, and a low prevalence when passaged continuously in BHK-21 cell culture. The magnitude in frequency variation observed between propagation in a constant environment and that of a change in environment suggests that Haplotype 20 provides a viable wild-type intermediate aiding in the process of adaptation. The latter indicates that the biological properties associated with the capsid protein Haplotype 20 encodes is under positive selection during the transition of BHK-cell culture to mouse brain, without ultimately contributing to the capsid region consensus genome.

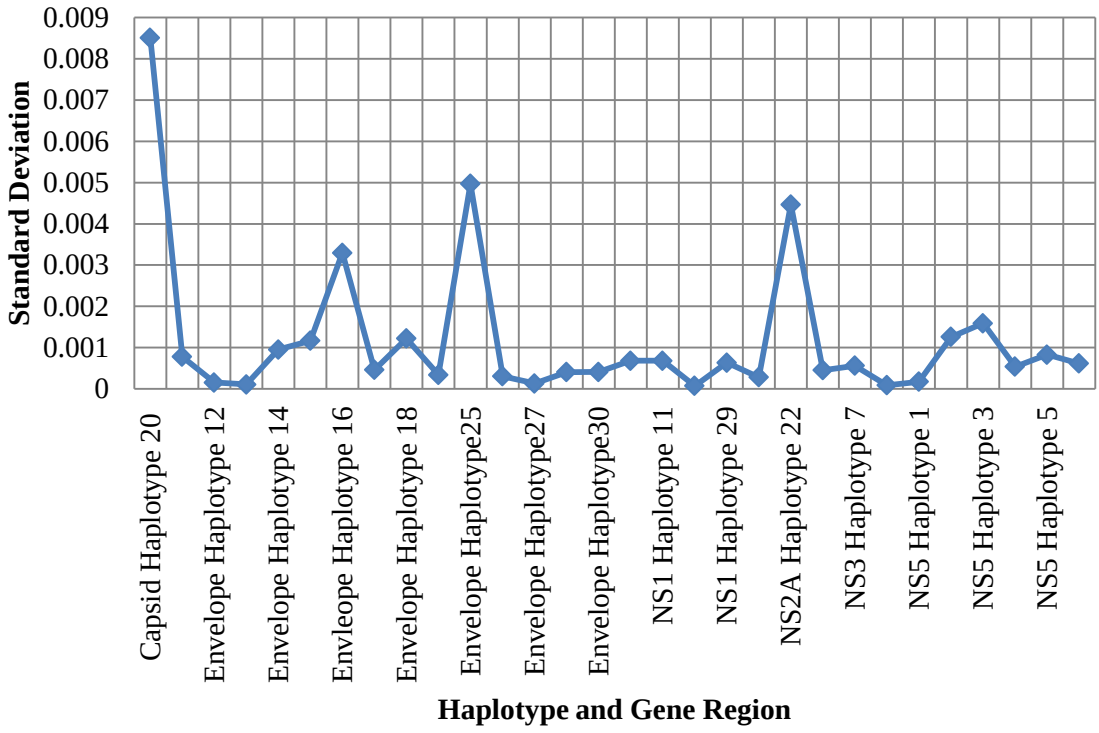

Figure 3.22 Standard deviation in haplotype frequency amongst WNV 349/77 isolates. Haplotypes and corresponding gene regions are indicated on the horizontal axis. Haplotype frequency is indicated on the vertical axis.

3.5.5 Functional Significance of Quasispecies Diversity

Based on observations of the number of SNPs and number of haplotypes for WNV 349/77, the capsid gene region demonstrated the highest degree of variation among isolates based on propagation system and passage number. When considering the variation in haplotype frequency contributing to the quasispecies dynamics of WNV 349/77, the most variation in frequency was observed for Haplotype

20 based on propagation system. In agreement with the observations of the most variable gene region, Haplotype 20 demonstrated variation in the capsid region. As mentioned earlier, the capsid protein plays an integral role in the assembly of infectious virions. Amongst the various important functions of the capsid protein, its role in membrane association is most prominent in light of the results obtained from this study.

The capsid protein is ~11 kDa in size and is composed of a central hydrophobic region that facilitates membrane association (Ma *et al*., 2004). The latter is surrounded by charged residues grouped at the N- and C- termini that mediate RNA interaction (Ma *et al*., 2004). Concerning protein structure, the capsid protein is an alpha-helical protein that associates as a dimer (Ma *et al*., 2004). Each monomer consists of four helices, α1 - α4; where helices α2 and α4 of one monomer are anti-parallel to helices α2 and α4 of the adjacent monomer (Ma *et al*., 2004). The latter form the contact surfaces for dimerization (Schlick *et al*., 2009). Subsequent to dimerization, the interacting α2 helices form a hydrophobic cleft that functions in membrane attachment (Ma *et al*., 2004). Alternatively, the majority of highly positively charged residues remain on the surfaces of the interacting α4 helices, functioning in the interaction with the RNA genome (Ma *et al*., 2004). The central hydrophobic region associated with membrane association spans helix α2 and part of the loop intervening helices α2 and α3 (Schlick et al., 2009). Although highly conserved, it has been demonstrated that the functionality of the hydrophobic domain can compensated for when large structural alterations occur (Schlick et al., 2009).

A study by Schlick *et al*.(2009) demonstrated the structural and functional flexibility of the WNV capsid protein. Deletions of up to a third of the capsid protein gene was shown to produce functional capsid protein (Schlick et al., 2009). Large deletions in the conserved hydrophobic region resulted in the formation of fusion helices, compensating for the functionality of the capsid protein (Schlick et al., 2009). Although the influence of the observed variation on the functionality of the capsid protein of WNV 349/77 is beyond the scope of this study, it is clear that genetic change in the capsid protein is of significance when changes in propagation system or passage number occur.

3.6 REVERSE GENETIC SYSTEM

A novel infectious clone of WNV 349/77 was designed based on a single plasmid reverse genetic system and synthesized accordingly (GenScript). Three different cell lines were transfected with the infectious WNV clone, including BHK-21 cells, Vero cells and BSR T7/5. Cell cultures transfected with the infectious WNV clone were co-infected with the recombinant fowlpox virus fpEFLT7pol to facilitate T7 polymerase gene expression in the WNV construct. In addition, each cell culture was

transfected with the pCI Mammalian Expression Vector (Promega) which served as positive control for transfection efficiency.

Cell cultures were inspected 24 hours after transfection for eGFP fluorescence of the pCI Mammalian Expression Vector (Promega) and for fluorescence of BSR T7/5 under a Zeiss Axio Vert.A1 microscope. Results indicated that the transfection was successful (Figure 3.23). Fluorescence was observed in BHK-21 cell cultures (Figure 3.23a), Vero cell cultures (Figure 3.21b) and BSR T7/5 cell cultures (Figure 2.23c) transfected with the pCI Mammalian Expression Vector (Promega). Fluorescence was also observed in BSR T7/5 cell cultures that were infected with the recombinant fowlpox virus fpEFLT7pol, indicating that BSR T7/5 express T7 polymerase efficiently (Figure 3.23d).

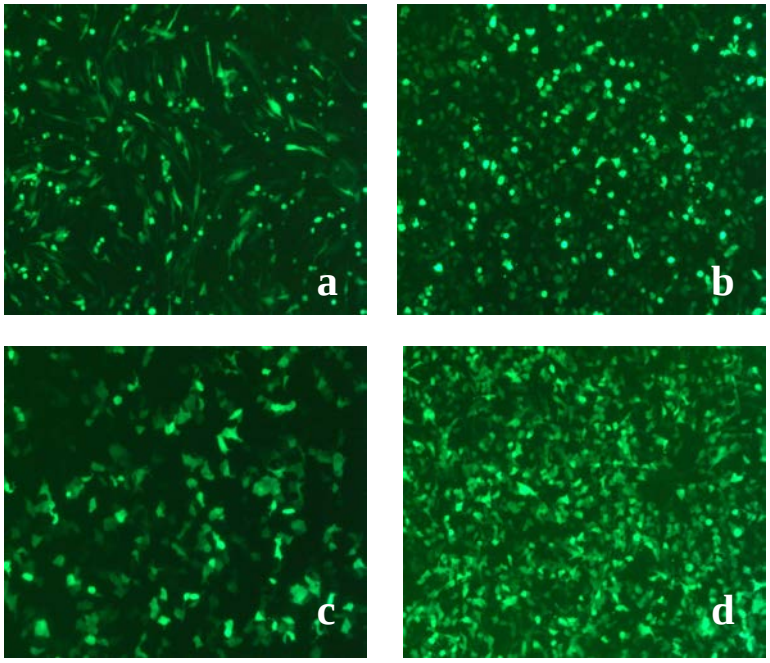

Figure 3.23 Fluorescence due to eGFP expression in cell cultures 24 hours post transfection (a) BHK-21 cells transfected with pCI Mammalian Expression Vector (b) BSR T7/5 cells transfected with pCI Mammalian Expression Vector (c) Vero cells transfected with pCI Mammalian Expression Vector (d) BSR T7/5 cells transfected with fpEFLT7pol

Transfected cell cultures were inspected under the microscope daily for cytopathic effect (CPE) up to seven days post-transfection (Zeiss Axio Vert.A1). Cells transfected with the infectious cDNA clone was passaged blindly seven days post-transfection. No CPE was observed in cell cultures infected with the WNV clone during this time. The attempt to rescue an infectious WNV clone was unsuccessful, although the rescue of full length cDNA clones were successful for other positive strand RNA viruses including hepatitis C virus (Yanagi *et al.*, 1997) and dengue type 4 virus (Lai *et al.*, 1991). Unsuccessful attempts may be ascribed to possible low-yield plasmid preparations and instability during plasmid propagation in bacteria. The presence of endotoxins in the preparation

received upon synthesis of the recombinant plasmid cannot be excluded. Lastly, unsuccessful attempts may be due to sequence error in the consensus genome of WNV 349/77.

Chapter 4

Conclusion and Recommendations

Two contemporary WNV strains, including HS 101/8 and HS 87/11; as well as two historic WNV strains, including WNV 1968 and WNV 349/77 were genetically characterised during this study. In addition, the genetic change associated with propagation system and passage number was investigated at both the consensus genome- and quasispecies level for isolates WNV 1968 and WNV 349/77. In light of future research, a single plasmid reverse genetic system was designed for WNV 349/77.

Both contemporary strains, HS 87/11 and HS 101/08; as well as historical strains, WNV 1968 and WNV 349/77, were shown to cluster within WNV lineage 2 (Figure 3.4). The genetic changes associated with passage number and changes in propagation system from BHK-21 cell culture to mouse brain and *vice versa* was not found to influence the consensus genome sequence of WNV 349/77 and WNV 1968. The abundance of Single Nucleotide Polymorphisms (SNPs) and haplotypes of both WNV 349/77 and WNV 1968, however, provided valuable insights into the genetic changes on the quasispecies level.

Although the detection of SNPs from next generation sequencing data is the most frequent approach to study viral quasipecies, this study is the first to reconstruct full-length haplotypes reflecting the quasispecies structure of WNV. It was found that the capsid gene region is largely subject to variation resulting from passage number and changes in propagation system. The majority of SNPs and haplotypes were found to contribute to variation in the capsid gene region, and the frequency changes associated with these variants were consistent throughout isolates propagated in different systems. Considering the importance of the capsid protein in membrane association, RNA interaction and ultimately the assembly of RNA replication complexes, the increased variation observed in the capsid gene region is likely due to selection pressure brought about by differences in host cell type between propagation systems.

A single plasmid reverse genetic system was designed for WNV 349/77 to complement studies of the functional significance of the quasispecies diversity demonstrated in this study. Based on the observation that the most recent WNV strain circulating South Africa is genetically most similar to the WNV 349/77 strain, the latter was chosen as a candidate for a reverse genetic system with vaccine development in mind. This is the first attempt to obtain a single plasmid reverse genetic system for WNV, as well as the first attempt to establishing a functional reverse genetic system for a South African lineage 2 WNV strain. Although the first attempt to rescue the infectious cDNA clone was unsuccessful, future approaches will aim to overcome the many caveats associated with a full-length infectious WNV clone. The latter includes the restriction digestion of the construct, followed by *in vitro* transcription and transfection of cell cultures using the approach followed in this study.

Briefly, this study is the first to demonstrate quasispecies dynamics resulting from changes in

propagation system of a lineage 2 WNV based on the reconstruction of full-length haplotype sequence data from ultra deep sequencing results. The approach envisages a cost-effective alternative to the estimation of viral population structure in light of viral evolutionary dynamics. The latter, in turn, may be complemented by the single plasmid reverse genetic system designed in this study. Although initial attempts at rescuing the infectious WNV clone were unsuccessful, once successful, the system hold promise for future studies focussing on vaccine development and improved disease control strategies.

REFERENCES

AASKOV, J., BUZACOTT, K., THU, H. M., LOWRY, K. & HOLMES, E. C. 2006. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. *Science,* 311**,** 236-238.

AEBISCHER, T., MOSKOPHIDIS, D., ROHRER, U. H., ZINKERNAGEL, R. M. & HENGARTNER, H. 1991. In vitro selection of lymphocytic choriomeningitis virus escape mutants by cytotoxic T lymphocytes. *Proceedings of the National Academy of Sciences,* 88**,** 11047-11051.

AHLQUIST, P., NOUEIRY, A. O., LEE, W.-M., KUSHNER, D. B. & DYE, B. T. 2003. Host Factors in Positive-Strand RNA Virus Genome Replication. *Journal of Virology,* 77**,** 8181-8186.

AKAIKE, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on,* 19**,** 716-723.

ALBERT, T. J., MOLLA, M. N., MUZNY, D. M., NAZARETH, L., WHEELER, D., SONG, X., RICHMOND, T. A., MIDDLE, C. M., RODESCH, M. J. & PACKARD, C. J. 2007. Direct selection of human genomic loci by microarray hybridization. *Nature Methods,* 4**,** 903-905.

ALCON-LEPODER, S., DROUET, M.-T., ROUX, P., FRENKIEL, M.-P., ARBORIO, M., DURAND-SCHNEIDER, A.-M., MAURICE, M., LE BLANC, I., GRUENBERG, J. & FLAMAND, M. 2005. The Secreted Form of Dengue Virus Nonstructural Protein NS1 Is Endocytosed by Hepatocytes and Accumulates in Late Endosomes: Implications for Viral Infectivity. *Journal of Virology,* 79**,** 11403-11411.

ALKAN, C., SAJJADIAN, S. & EICHLER, E. E. 2010. Limitations of next-generation genome sequence assembly. *Nature Methods,* 8**,** 61-65.

ALLISON, S. L., STIASNY, K., STADLER, K., MANDL, C. W. & HEINZ, F. X. 1999. Mapping of Functional Elements in the Stem-Anchor Region of Tick-Borne Encephalitis Virus Envelope Protein E. *Journal of Virology,* 73**,** 5605-5612.

ALTSHULER, D., POLLARA, V. J., COWLES, C. R., VAN ETTEN, W. J., BALDWIN, J., LINTON, L. & LANDER, E. S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature,* 407**,** 513-516.

AMBERG, S. M., NESTOROWICZ, A., MCCOURT, D. W. & RICE, C. M. 1994. NS2B-3 proteinase-mediated processing in the yellow fever virus structural region: in vitro and in vivo studies. *Journal of Virology,* 68**,** 3794-3802.

ARCHER, J., BAILLIE, G., WATSON, S., KELLAM, P., RAMBAUT, A. & ROBERTSON, D. 2012. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics,* 13**,** 47.

ARIAS, A., LÁZARO, E., ESCARMíS, C. & DOMINGO, E. 2001. Molecular intermediates of fitness gain of an RNA virus: characterization of a mutant spectrum by biological and molecular cloning. *Journal of General Virology,* 82**,** 1049-1060.

AUDSLEY, M., EDMONDS, J., LIU, W., MOKHONOV, V., MOKHONOVA, E., MELIAN, E. B., PROW, N., HALL, R. A. & KHROMYKH, A. A. 2011. Virulence determinants between New York 99 and Kunjin strains of West Nile virus. *Virology,* 414**,** 63-73.

BAINBRIDGE, M. N., WARREN, R. L., HIRST, M., ROMANUIK, T., ZENG, T., GO, A., DELANEY, A., GRIFFITH, M., HICKENBOTHAM, M. & MAGRINI, V. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics,* 7**,** 246.

BAKONYI, T., HUBALEK, Z., RUDOLF, I. & NOWOTNY, N. 2005. Novel flavivirus or new lineage of West Nile virus, central Europe. *Emerging Infectious Diseases,* 11**,** 225-31.

BALTIMORE, D. 1971. Expression of animal virus genomes. *Bacteriological Reviews,* 35**,** 235.

BARTELMA, G. & PADMANABHAN, R. 2002. Expression, Purification, and Characterization of the RNA 5′-Triphosphatase Activity of Dengue Virus Type 2 Nonstructural Protein 3. *Virology,* 299**,** 122-132.

BEERENWINKEL, N., GÜNTHARD, H. F., ROTH, V. & METZNER, K. J. 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology,* 3.

BEERENWINKEL, N. & ZAGORDI, O. 2011. Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol,* 1**,** 413-418.

BILLOIR, F., DE CHESSE, R., TOLOU, H., DE MICCO, P., GOULD, E. A. & DE LAMBALLERIE, X. 2000. Phylogeny of the genus Flavivirus using complete coding sequences of arthropod-borne viruses and viruses with no known vector. *Journal of General Virology,* 81**,** 781-790.

BONDRE, V. P., JADI, R. S., MISHRA, A. C., YERGOLKAR, P. N. & ARANKALLE, V. A. 2007. West Nile virus isolates from India: evidence for a distinct genetic lineage. *Journal of General Virology,* 88**,** 875-884.

BORK, P., HOLM, L. & SANDER, C. 1994. *J. molec. Biol.,* 242**,** 309-320.

BOTHA, E. M. M. W. M. J. T. S. R. G. L. H. V. M. 2008. Genetic Determinants of Virulence in Pathogenic Lineage 2 West Nile Virus Strains. *Emerging Infectious Diseases,* 14**,** 222.

BRACKNEY, D. E., PESKO, K. N., BROWN, I. K., DEARDORFF, E. R., KAWATACHI, J. & EBEL, G. D. 2011. West Nile Virus Genetic Diversity is Maintained during Transmission by <italic>Culex pipiens quinquefasciatus</italic> Mosquitoes. *PLoS ONE,* 6**,** e24466.

BRINTON, M. A. & DISPOTO, J. H. 1988. Sequence and secondary structure analysis of the 5′-terminal region of flavivirus genome RNA. *Virology,* 162**,** 290-299.

BRITTON, P., GREEN, P., KOTTIER, S., MAWDITT, K. L., PENZES, Z., CAVANAGH, D. & SKINNER, M. A. 1996. Expression of bacteriophage T7 RNA polymerase in avian and mammalian cells by a recombinant fowlpox virus. *Journal of General Virology,* 77**,** 963-967.

BROWN, B. A., OBERSTE, M. S., ALEXANDER, J. P., KENNETT, M. L. & PALLANSCH, M. A. 1999. Molecular epidemiology and evolution of enterovirus 71 strains isolated from 1970 to 1998. *Journal of Virology,* 73**,** 9969-9975.

BUCHHOLZ, U. J., FINKE, S. & CONZELMANN, K.-K. 1999. Generation of bovine respiratory syncytial virus (BRSV) from cDNA: BRSV NS2 is not essential for virus replication in tissue culture, and the human RSV leader region acts as a functional BRSV genome promoter. *Journal of Virology,* 73**,** 251-259.

BURROWS, M. & WHEELER, D. J. 1994. A block-sorting lossless data compression algorithm.

BURT, F. J., GROBBELAAR, A. A., LEMAN, P. A., ANTHONY, F. S., GIBSON, G. V. F. & SWANEPOEL, R. 2002. Phylogenetic Relationships of Southern African West Nile Virus Isolates. Centers for Disease Control and Prevention.

CALISHER, C. H. & GOULD, E. A. 2003. Taxonomy of the virus family Flaviviridae. *Advances in Virus Research.* Academic Press.

CALISHER, C. H., KARABATSOS, N., DALRYMPLE, J. M., SHOPE, R. E., PORTERFIELD, J. S., WESTAWAY, E. G. & BRANDT, W. E. 1989. Antigenic Relationships between Flaviviruses as Determined by Cross-neutralization Tests with Polyclonal Antisera. *Journal of General Virology,* 70**,** 37-43.

CAMPBELL, G. L., MARFIN, A. A., LANCIOTTI, R. S. & GUBLER, D. J. 2002. West Nile virus. *The Lancet Infectious Diseases,* 2**,** 519-529.

CAMPBELL, P. J., STEPHENS, P. J., PLEASANCE, E. D., O'MEARA, S., LI, H., SANTARIUS, T., STEBBINGS, L. A., LEROY, C., EDKINS, S. & HARDY, C. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics,* 40**,** 722-729.

CHAMBERS, T. J., GRAKOUI, A. & RICE, C. M. 1991. Processing of the yellow fever virus nonstructural polyprotein: a catalytically active NS3 proteinase domain and NS2B are required for cleavages at dibasic sites. *Journal of Virology,* 65**,** 6042-6050.

CHAMBERS, T. J., HAHN, C. S., GALLER, R. & RICE, C. M. 1990. Flavivirus Genome Organization, Expression, and Replication. *Annual Review of Microbiology,* 44**,** 649-688.

CHAMBERS, T. J., NESTOROWICZ, A., AMBERG, S. M. & RICE, C. M. 1993. Mutagenesis of the yellow fever virus NS2B protein: effects on proteolytic processing, NS2B-NS3 complex formation, and viral replication. *Journal of Virology,* 67**,** 6797-6807.

CHANG, Y.-S., LIAO, C.-L., TSAO, C.-H., CHEN, M.-C., LIU, C.-I., CHEN, L.-K. & LIN, Y.-L. 1999. Membrane Permeabilization by Small Hydrophobic Nonstructural Proteins of Japanese Encephalitis Virus. *Journal of Virology,* 73**,** 6257-6264.

CHASE, M. & DOERMANN, A. 1958. High negative interference over short segments of the genetic structure of bacteriophage T4. *Genetics*, 43**,** 332.

CHEN, C. J., KUO, M. D., CHIEN, L. J., HSU, S. L., WANG, Y. M. & LIN, J. H. 1997. RNA-protein interactions: involvement of NS3, NS5, and 3' noncoding regions of Japanese encephalitis virus genomic RNA. *Journal of Virology*, 71**,** 3466-73.

CHEN, W., KALSCHEUER, V., TZSCHACH, A., MENZEL, C., ULLMANN, R., SCHULZ, M. H., ERDOGAN, F., LI, N., KIJAS, Z. & ARKESTEIJN, G. 2008. Mapping translocation breakpoints by next-generation sequencing. *Genome Research*, 18**,** 1143-1149.

CIOTA, A. T., NGO, K. A., LOVELACE, A. O., PAYNE, A. F., ZHOU, Y., SHI, P.-Y. & KRAMER, L. D. 2007. Role of the mutant spectrum in adaptation and replication of West Nile virus. *Journal of General Virology*, 88**,** 865-874.

CIUREA, A., KLENERMAN, P., HUNZIKER, L., HORVATH, E., SENN, B. M., OCHSENBEIN, A. F., HENGARTNER, H. & ZINKERNAGEL, R. M. 2000. Viral persistence in vivo through selection of neutralizing antibody-escape variants. *Proceedings of the National Academy of Sciences*, 97**,** 2749-2754.

CLAUDE, B. 1962. Théorie des graphes et ses applications. *Russian translation, Moscow*.

CLOONAN, N., FORREST, A. R., KOLLE, G., GARDINER, B. B., FAULKNER, G. J., BROWN, M. K., TAYLOR, D. F., STEPTOE, A. L., WANI, S. & BETHEL, G. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5**,** 613-619.

CLUM, S., EBNER, K. E. & PADMANABHAN, R. 1997. Cotranslational Membrane Insertion of the Serine Proteinase Precursor NS2B-NS3(Pro) of Dengue Virus Type 2 Is Required for Efficient in Vitro Processing and Is Mediated through the Hydrophobic Regions of NS2B. *Journal of Biological Chemistry*, 272**,** 30715-30723.

COMBES, C. & THÉRON, A. 2000. Metazoan parasites and resource heterogeneity: constraints and benefits. *International Journal for Parasitology*, 30**,** 299-304.

COMPEAU, P. E., PEVZNER, P. A. & TESLER, G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29**,** 987-991.

COX-FOSTER, D. L., CONLAN, S., HOLMES, E. C., PALACIOS, G., EVANS, J. D., MORAN, N. A., QUAN, P.-L., BRIESE, T., HORNIG, M. & GEISER, D. M. 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*, 318**,** 283-287.

CROOKS, A. J., LEE, J. M., EASTERBROOK, L. M., TIMOFEEV, A. V. & STEPHENSON, J. R. 1994. The NS1 protein of tick-borne encephalitis virus forms multimeric species upon secretion from the host cell. *The Journal of general virology*, 75 (Pt 12)**,** 3453-3460.

DAHL, F., STENBERG, J., FREDRIKSSON, S., WELCH, K., ZHANG, M., NILSSON, M., BICKNELL, D., BODMER, W. F., DAVIS, R. W. & JI, H. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proceedings of the National Academy of Sciences*, 104**,** 9387-9392.

DAVIS, B. S., CHANG, G.-J. J., CROPP, B., ROEHRIG, J. T., MARTIN, D. A., MITCHELL, C. J., BOWEN, R. & BUNNING, M. L. 2001. West Nile Virus Recombinant DNA Vaccine Protects Mouse and Horse from Virus Challenge and Expresses In Vitro a Noninfectious Recombinant Antigen That Can Be Used in Enzyme-Linked Immunosorbent Assays. *Journal of Virology,* 75**,** 4040-4047.

DE BRUIJN, N. G. & ERDOS, P. 1946. A combinatorial problem. *Koninklijke Netherlands: Academe Van Wetenschappen,* 49**,** 758-764.

DE LA TORRE, J. & HOLLAND, J. 1990. RNA virus quasispecies populations can suppress vastly superior mutant progeny. *Journal of Virology,* 64**,** 6278-6281.

DE MADRID, A. T. & PORTERFIELD, J. S. 1974. The Flaviviruses (Group B Arboviruses): a Cross-neutralization Study. *Journal of General Virology,* 23**,** 91-96.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research,* 36**,** e105-e105.

DOMINGO, E. 2000. Viruses at the edge of adaptation. *Virology,* 270**,** 251-253.

DOMINGO, E. 2002. Quasispecies Theory in Virology. *J Virol,* 76**,** 463 - 465.

DOMINGO, E. 2007. Virus evolution. *eLS*.

DOMINGO, E., BIEBRICHER, C., EIGEN, M. & HOLLAND, J. 2001. Georgetown, TX: Landes Bioscience.

DOMINGO, E., BRUN, A., NUÑEZ, J. I., CRISTINA, J., BRIONES, C. & ESCARMÍS, C. 2006. Genomics of viruses. *Pathogenomics: genome analysis of pathogenic microbes***,** 367-388.

DOMINGO, E., ESCARMÍS, C., SEVILLA, N. & BARANOWSKI, E. 1998. Population dynamics in the evolution of RNA viruses. *Advances in Experimental Medicine and Biology,* 440**,** 721-727.

DOMINGO, E. & HOLLAND, J. 1997. RNA virus mutations and fitness for survival. *Annu Rev Microbiol,* 51**,** 151 - 178.

DOMINGO, E. & HOLLAND, J. J. 2005. *The origin and evolution of viruses*, Wiley Online Library.

DOMINGO, E., SABO, D., TANIGUCHI, T. & WEISSMANN, C. 1978. Nucleotide sequence heterogeneity of an RNA phage population. *Cell,* 13**,** 735-744.

DOMINGO, E., SHELDON, J. & PERALES, C. 2012. Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews,* 76**,** 159-216.

DRAKE, J. & HOLLAND, J. 1999. Mutation rates among RNA viruses. *Proc Natl Acad Sci USA,* 96**,** 13910 - 13913.

DRESSMAN, D., YAN, H., TRAVERSO, G., KINZLER, K. W. & VOGELSTEIN, B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences,* 100**,** 8817-8822.

DUARTE, E. A., CLARKE, D., MOYA, A., ELENA, S., DOMINGO, E. & HOLLAND, J. 1993. Many-trillionfold amplification of single RNA virus particles fails to overcome the Muller's ratchet effect. *Journal of Virology,* 67**,** 3620-3623.

DUARTE, E. A., NOVELLA, I. S., LEDESMA, S., CLARKE, D. K., MOYA, A., ELENA, S. F., DOMINGO, E. & HOLLAND, J. J. 1994. Subclonal components of consensus fitness in an RNA virus clone. *Journal of Virology,* 68**,** 4295-4301.

ECKERT, K. A. & KUNKEL, T. A. 1991. DNA polymerase fidelity and the polymerase chain reaction. *Genome Research,* 1**,** 17-24.

EIDEN, M., VINA-RODRIGUEZ, A., HOFFMANN, B., ZIEGLER, U. & GROSCHUP, M. H. 2010. Two new real-time quantitative reverse transcription polymerase chain reaction assays with unique target sites for the specific and sensitive detection of lineages 1 and 2 West Nile virus strains. *Journal of Veterinary Diagnostic Investigation,* 22**,** 748-753.

EIGEN, M. 1971. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften,* 58**,** 456 - 523.

EIGEN, M., MCCASKILL, J. & SCHUSTER, P. 1988. Molecular Quasi-Species. *J Phys Chem,* 92**,** 6881 - 6891.

EIGEN, M. & SCHUSTER, P. 1979. *The hypercycle: a principle of natural self-organization,* Springer-Verlag Berlin.

ERIKSSON, N., PACHTER, L., MITSUYA, Y., RHEE, S.-Y., WANG, C., GHARIZADEH, B., RONAGHI, M., SHAFER, R. W. & BEERENWINKEL, N. 2008a. Viral population estimation using pyrosequencing. *PLoS Computational Biology,* 4**,** e1000074.

ERIKSSON, N., PACHTER, L., MITSUYA, Y., RHEE, S., WANG, C., GHARIZADEH, B., RONAGHI, M., SHAFER, R. & BEERENWINKEL, N. 2008b. Viral population estimation using pyrosequencing. *PLoS Comput Biol,* 4**,** e1000074.

EULER, L. 1741. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae,* 8**,** 128-140.

EWENS, W. J. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology,* 3**,** 87-112.

FALGOUT, B. & MARKOFF, L. 1995. Evidence that flavivirus NS1-NS2A cleavage is mediated by a membrane-bound host protease in the endoplasmic reticulum. *Journal of Virology,* 69**,** 7232-43.

FALGOUT, B., PETHEL, M., ZHANG, Y. M. & LAI, C. J. 1991. Both nonstructural proteins NS2B and NS3 are required for the proteolytic processing of dengue virus nonstructural proteins. *Journal of Virology,* 65**,** 2467-2475.

FAYZULIN, R., SCHOLLE, F., PETRAKOVA, O., FROLOV, I. & MASON, P. W. 2006. Evaluation of replicative capacity and genetic stability of West Nile virus replicons using highly efficient packaging cell lines. *Virology,* 351**,** 196-209.

FEDURCO, M., ROMIEU, A., WILLIAMS, S., LAWRENCE, I. & TURCATTI, G. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research,* 34**,** e22-e22.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution***,** 783-791.

FERGUSON, T. S. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics***,** 209-230.

FERRAGINA, P. & MANZINI, G. Year. Opportunistic data structures with applications. *In:* Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, 2000. IEEE, 390-398.

FITZPATRICK, K. A., DEARDORFF, E. R., PESKO, K., BRACKNEY, D. E., ZHANG, B., BEDRICK, E., SHI, P.-Y. & EBEL, G. D. 2010. Population variation of West Nile virus confers a host-specific fitness benefit in mosquitoes. *Virology,* 404**,** 89-95.

FLICEK, P. & BIRNEY, E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature Methods,* 6**,** S6-S12.

GEBAUER, F., DE LA TORRE, J., GOMES, I., MATEU, M., BARAHONA, H., TIRABOSCHI, B., BERGMANN, I., DE MELLO, P. A. & DOMINGO, E. 1988. Rapid selection of genetic and antigenic variants of foot-and-mouth disease virus during persistence in cattle. *Journal of Virology,* 62**,** 2041-2049.

GELDERBLOM, H. C., VATAKIS, D. N., BURKE, S. A., LAWRIE, S. D., BRISTOL, G. C. & LEVY, D. N. 2008. Viral complementation allows HIV-1 replication without integration. *Retrovirology,* 5**,** 60.

GOLLINS, S. & PORTERFIELD, J. 1985. Flavivirus infection enhancement in macrophages: an electron microscopic study of viral cellular entry. *Journal of General Virology,* 66**,** 1969-1982.

GOLLINS, S. W. & PORTERFIELD, J. S. 1986. The uncoating and infectivity of the flavivirus West Nile on interaction with cells: effects of pH and ammonium chloride. *Journal of General Virology,* 67**,** 1941-1950.

GORBALENYA, A. E., KOONIN, E. V., DONCHENKO, A. P. & BLINOV, V. M. 1989. Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Research,* 17**,** 4713-4730.

GOULD, L. H. & FIKRIG, E. 2004. West Nile virus: a growing concern? *The Journal of Clinical Investigation,* 113**,** 1102-1107.

GUBLER, D. J., KUNO, G. & MARKOFF, L. 2007. Flaviviruses. *Fields virology. 5th ed. Philadelphia: Lippincott Williams & Wilkins,* 1153-252.

GUINDON, S. & GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology,* 52**,** 696-704.

GUIRAKHOO, F., BOLIN, R. A. & ROEHRIG, J. T. 1992. The Murray Valley encephalitis virus prM protein confers acid resistance to virus particles and alters the expression of epitopes within the R2 domain of E glycoprotein. *Virology,* 191**,** 921-931.

HAHN, C. S., HAHN, Y. S., RICE, C. M., LEE, E., DALGARNO, L., STRAUSS, E. G. & STRAUSS, J. H. 1987. Conserved elements in the 3′ untranslated region of flavivirus RNAs and potential cyclization sequences. *Journal of Molecular Biology,* 198**,** 33-41.

HALL, T. A. Year. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *In:* Nucleic Acids Symposium Series, 1999. 95-98.

HAMILTON, W. R. Year. On quaternions. *In:* Proceedings of the Royal Irish Academy, 1847. Addison-Wesley, 1-16.

HAYES, C. G. 2001. West Nile Virus: Uganda, 1937, to New York City, 1999. *Annals of the New York Academy of Sciences,* 951**,** 25-37.

HAYES, E. B. & GUBLER, D. J. 2006. West Nile Virus: Epidemiology and Clinical Features of an Emerging Epidemic in the United States*. *Annual Review of Medicine,* 57**,** 181-194.

HODGES, E., XUAN, Z., BALIJA, V., KRAMER, M., MOLLA, M. N., SMITH, S. W., MIDDLE, C. M., RODESCH, M. J., ALBERT, T. J. & HANNON, G. J. 2007. Genome-wide in situ exon capture for selective resequencing. *Nature Genetics,* 39**,** 1522-1527.

HOLLAND, J., SPINDLER, K., HORODYSKI, F., GRABAU, E., NICHOL, S. & VANDEPOL, S. 1982a. Rapid evolution of RNA genomes. *Science,* 215**,** 1577-1585.

HOLLAND, J., SPINDLER, K., HORODYSKI, F., GRABAU, E., NICHOL, S. & VANDEPOL, S. 1982b. Rapid evolution of RNA genomes. *Science,* 215**,** 1577 - 1585.

HU, W. S., BOWMAN, E. H., DELVIKS, K. A. & PATHAK, V. K. 1997. Homologous recombination occurs in a distinct retroviral subpopulation and exhibits high negative interference. *Journal of Virology,* 71**,** 6028-6036.

HUBALEK, Z. 2000. European experience with the West Nile virus ecology and epidemiology: could it be relevant for the New World? *Viral Immunol,* 13**,** 415-26.

HUBÁLEK, Z. & HALOUZKA, J. 1999. West Nile fever--a reemerging mosquito-borne viral disease in Europe. Centers for Disease Control.

HUELSENBECK, J. P. & RONQUIST, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17**,** 754-755.

JERZAK, G. V. S., BERNARD, K., KRAMER, L. D., SHI, P.-Y. & EBEL, G. D. 2007. The West Nile virus mutant spectrum is host-dependant and a determinant of mortality in mice. *Virology,* 360**,** 469-476.

JERZAK, G. V. S., BROWN, I., SHI, P.-Y., KRAMER, L. D. & EBEL, G. D. 2008. Genetic diversity and purifying selection in West Nile virus populations are maintained during host switching. *Virology,* 374**,** 256-260.

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science Signalling,* 316**,** 1497.

JUPP, P. G., BLACKBURN, N. K., THOMPSON, D. L. & MEENEHAN, G. M. 1986. Sindbis and West Nile virus infections in the Witwatersrand-Pretoria region. *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde,* 70**,** 218-20.

KANAGAWA, T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering,* 96**,** 317-323.

KATOH, K. & TOH, H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics,* 9**,** 286-298.

KECECIOGLU, J. D. & MYERS, E. W. 1995. Combinatorial algorithms for DNA sequence assembly. *Algorithmica,* 13**,** 7-51.

KHROMYKH, A. A., SEDLAK, P. L., GUYATT, K. J., HALL, R. A. & WESTAWAY, E. G. 1999. Efficient trans-Complementation of the Flavivirus Kunjin NS5 Protein but Not of the NS1 Protein Requires Its Coexpression with Other Components of the Viral Replicase. *Journal of Virology,* 73**,** 10272-10280.

KHROMYKH, A. A., VARNAVSKI, A. N., SEDLAK, P. L. & WESTAWAY, E. G. 2001. Coupling between Replication and Packaging of Flavivirus RNA: Evidence Derived from the Use of DNA-Based Full-Length cDNA Clones of Kunjin Virus. *Journal of Virology,* 75**,** 4633-4640.

KIM, J. B., PORRECA, G. J., SONG, L., GREENWAY, S. C., GORHAM, J. M., CHURCH, G. M., SEIDMAN, C. E. & SEIDMAN, J. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science,* 316**,** 1481-1484.

KIRCHER, M., STENZEL, U. & KELSO, J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol,* 10**,** R83.

KOONIN, E. V. 1993. Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and lambda 2 protein of reovirus. *The Journal of general virology,* 74 (Pt 4)**,** 733-740.

KRZYWINSKI, M. I., SCHEIN, J. E., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J. & MARRA, M. A. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research.*

L'VOV, D. K., KOVTUNOV, A. I., IASHKULOV, K. B., GROMASHEVSKIĬ, V. L., DZHARKENOV, A. F., SHCHELKANOV, M., KULIKOVA, L. N., SAVAGE, H. M., CHIMIDOVA, N. M., MIKHALIAEVA, L. B., VASIL'EV, A. V., GALKINA, I. V., PRILIPOV, A. G., KINNEY, R. M., SAMOKHVALOV, E. I., BUSHKIEVA, B., GUBLER, D. J., AL'KHOVSKIĬ, S. K., ARISTOVA, V. A., DERIABIN, P. G., BUTENKO, A. M., MOSKVINA, T. M., L'VOV, D. N., ZLOBINA, L. V., LIAPINA, O. V., SADYKOVA, G. K., SHATALOV, A. G., USACHEV, V. E., VORONINA, A. G. & LUNEVA, L. I. 2004. Circulation of West Nile virus (Flaviviridae, Flavivirus) and some other arboviruses in the

ecosystems of Volga delta, Volga-Akhtuba flood-lands and adjoining arid regions (2000-2002). *Voprosy Virusologii,* 49**,** 45-51.

LAI, C. J., ZHAO, B. T., HORI, H. & BRAY, M. 1991. Infectious RNA transcribed from stably cloned full-length cDNA of dengue type 4 virus. *Proceedings of the National Academy of Sciences,* 88**,** 5139-5143.

LANCIOTTI, R. S., EBEL, G. D., DEUBEL, V., KERST, A. J., MURRI, S., MEYER, R., BOWEN, M., MCKINNEY, N., MORRILL, W. E., CRABTREE, M. B., KRAMER, L. D. & ROEHRIG, J. T. 2002. Complete Genome Sequences and Phylogenetic Analysis of West Nile Virus Strains Isolated from the United States, Europe, and the Middle East. *Virology,* 298**,** 96-105.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

LÁZARO, E., ESCARMÍS, C., PÉREZ-MERCADER, J., MANRUBIA, S. C. & DOMINGO, E. 2003. Resistance of virus to extinction on bottleneck passages: study of a decaying and fluctuating pattern of fitness loss. *Proceedings of the National Academy of Sciences,* 100**,** 10830-10835.

LEE, J. M., CROOKS, A. J. & STEPHENSON, J. R. 1989. The Synthesis and Maturation of a Non-structural Extracellular Antigen from Tick-borne Encephalitis Virus and Its Relationship to the Intracellular NS1 Protein. *Journal of General Virology,* 70**,** 335-343.

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.

LIBRADO, P. & ROZAS, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics,* 25**,** 1451-1452.

LIN, C., AMBERG, S. M., CHAMBERS, T. J. & RICE, C. M. 1993. Cleavage at a novel site in the NS4A region by the yellow fever virus NS2B-3 proteinase is a prerequisite for processing at the downstream 4A/4B signalase site. *Journal of Virology,* 67**,** 2327-2335.

LIN, Y.-J. & WU, S.-C. 2005. Histidine at Residue 99 and the Transmembrane Region of the Precursor Membrane prM Protein Are Important for the prM-E Heterodimeric Complex Formation of Japanese Encephalitis Virus. *Journal of Virology,* 79**,** 8535-8544.

LINDENBACH, B. D. & RICE, C. M. 1999. Genetic Interaction of Flavivirus Nonstructural Proteins NS1 and NS4A as a Determinant of Replicase Function. *Journal of Virology,* 73**,** 4611-4621.

LINDENBACH, B. D. & RICE, C. M. 2003. Molecular biology of flaviviruses. *Advances in Virus Research.* Academic Press.

LINDENBACH, B., THIEL, H.-J. & RICE, C. 2007. Flaviviridae: The 693 Viruses and Their Replication. *Fields,* 694, 1101-1152.

LISTER, R., O'MALLEY, R. C., TONTI-FILIPPINI, J., GREGORY, B. D., BERRY, C. C., MILLAR, A. H. & ECKER, J. R. 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in< i> Arabidopsis</i>. *Cell,* 133**,** 523-536.

LIU, W. J., WANG, X. J., CLARK, D. C., LOBIGS, M., HALL, R. A. & KHROMYKH, A. A. 2006. A Single Amino Acid Substitution in the West Nile Virus Nonstructural Protein NS2A Disables Its Ability To Inhibit Alpha/Beta Interferon Induction and Attenuates Virus Virulence in Mice. *Journal of Virology,* 80**,** 2396-2404.

LOBIGS, M. 1993. Flavivirus premembrane protein cleavage and spike heterodimer secretion require the function of the viral proteinase NS3. *Proceedings of the National Academy of Sciences,* 90**,** 6218-6222.

LORENZ, I. C., ALLISON, S. L., HEINZ, F. X. & HELENIUS, A. 2002. Folding and Dimerization of Tick-Borne Encephalitis Virus Envelope Proteins prM and E in the Endoplasmic Reticulum. *Journal of Virology,* 76**,** 5480-5491.

MA, L., JONES, C. T., GROESCH, T. D., KUHN, R. J. & POST, C. B. 2004. Solution structure of dengue virus capsid protein reveals another fold. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 3414-3419.

MACKENZIE, J. M., KHROMYKH, A. A., JONES, M. K. & WESTAWAY, E. G. 1998. Subcellular Localization and Some Biochemical Properties of the Flavivirus Kunjin Nonstructural Proteins NS2A and NS4A. *Virology,* 245**,** 203-215.

MACPHERSON, I. 1963. Characteristics of a hamster cell clone transformed by polyoma virus. *Journal of the National Cancer Institute,* 30**,** 795-815.

MANDL, C. W., GUIRAKHOO, F., HOLZMANN, H., HEINZ, F. X. & KUNZ, C. 1989. Antigenic structure of the flavivirus envelope protein E at the molecular level, using tick-borne encephalitis virus as a model. *Journal of Virology,* 63**,** 564-571.

MARBERG, K., GOLDBLTJM, N., STERK, V. V., JASINSKA-KLINGBEHG, W. & KLINGBERG, M. A. 1956. The Natural History of West Nile Fever. I. Clinical Observations during an Epidemic in Israel. *American Journal of Hygiene,* 64**,** 259-69.

MARDIS, E. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet,* 24**,** 133 - 41.

MATUSAN, A. E., PRYOR, M. J., DAVIDSON, A. D. & WRIGHT, P. J. 2001. Mutagenesis of the Dengue Virus Type 2 NS3 Protein within and outside Helicase Motifs: Effects on Enzyme Activity and Virus Replication. *Journal of Virology,* 75**,** 9633-9643.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics,* 21**,** 1087.

METZKER, M. L. 2010. Sequencing technologies [mdash] the next generation. *Nat Rev Genet,* 11**,** 31-46.

MEYER, M., STENZEL, U. & HOFREITER, M. 2008. Parallel tagged sequencing on the 454 platform. *Nature Protocols,* 3**,** 267-278.

MILLER, J. R., KOREN, S. & SUTTON, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* 95**,** 315.

MILLER, S., SPARACIO, S. & BARTENSCHLAGER, R. 2006. Subcellular Localization and Membrane Topology of the Dengue Virus Type 2 Non-structural Protein 4B. *Journal of Biological Chemistry,* 281**,** 8854-8863.

MORIN, R. D., O'CONNOR, M. D., GRIFFITH, M., KUCHENBAUER, F., DELANEY, A., PRABHU, A.-L., ZHAO, Y., MCDONALD, H., ZENG, T. & HIRST, M. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research,* 18**,** 610-621.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods,* 5**,** 621-628.

MULLER, H. 1964. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 1**,** 2-9.

MULLER, H. J. 1932. Some genetic aspects of sex. *The American naturalist,* 66**,** 118-138.

MUYLAERT, I. R., CHAMBERS, T. J., GALLER, R. & RICE, C. M. 1996. Mutagenesis of the N-Linked Glycosylation Sites of the Yellow Fever Virus NS1 Protein: Effects on Virus Replication and Mouse Neurovirulence. *Virology,* 222**,** 159-168.

MYERS, E. W. 1995. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology,* 2**,** 275-290.

NAKAMURA, K., OSHIMA, T., MORIMOTO, T., IKEDA, S., YOSHIKAWA, H., SHIWA, Y., ISHIKAWA, S., LINAK, M., HIRAI, A. & TAKAHASHI, H. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res,* 39**,** e90.

NICHOLAS, K. 2003. West Nile Virus: Epidemiology and Ecology in North America. *Advances in Virus Research.* Academic Press.

NOWAK, T., FÄRBER, P. M., WENGLER, G. & WENGLER, G. 1989. Analyses of the terminal sequences of west nile virus structural proteins and of the in vitro translation of these proteins allow the proposal of a complete scheme of the proteolytic cleavages involved in their synthesis. *Virology,* 169**,** 365-376.

NOWAK, T. & WENGLER, G. 1987. Analysis of disulfides present in the membrane proteins of the West Nile flavivirus. *Virology,* 156**,** 127-137.

OJOSNEGROS, S., GARCÍA-ARRIAZA, J., ESCARMÍS, C., MANRUBIA, S. C., PERALES, C., ARIAS, A., MATEU, M. G. & DOMINGO, E. 2011. Viral genome segmentation can result from a trade-off between genetic content and particle stability. *PLoS Genetics,* 7**,** e1001344.

OKOU, D. T., STEINBERG, K. M., MIDDLE, C., CUTLER, D. J., ALBERT, T. J. & ZWICK, M. E. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nature Methods,* 4**,** 907-909.

PIERSON, T. C. & DIAMOND, M. S. 2012. Degrees of maturity: the complex structure and biology of flaviviruses. *Current Opinion in Virology,* 2**,** 168-175.

PORRECA, G. J., ZHANG, K., LI, J. B., XIE, B., AUSTIN, D., VASSALLO, S. L., LEPROUST, E. M., PECK, B. J., EMIG, C. J. & DAHL, F. 2007. Multiplex amplification of large sets of human exons. *Nature Methods,* 4**,** 931-936.

POSADA, D. 2008. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution,* 25**,** 1253-1256.

PRILIPOV, A. G., KINNEY, R. M., SAMOKHVALOV, E. I., SAVAGE, H. M., AL'KHOVSKIĬ, S. V., TSUCHIYA, K. R., GROMASHEVSKIĬ, V. L., SADYKOVA, G. K., SHATALOV, A. G., VYSHEMIRSKIĬ, O. I., USACHEV, E. V., MOKHONOV, V. V., VORONINA, A. G., BUTENKO, A. M., LARICHEV, V. F., ZHUKOV, A. N., KOVTUNOV, A. I., GUBLER, D. J. & L'VOV, D. K. 2002. Analysis of new variants of West Nile fever virus. *Voprosy Virusologii,* 47**,** 36-41.

PUIG-BASAGOITI, F., DEAS, T. S., REN, P., TILGNER, M., FERGUSON, D. M. & SHI, P.-Y. 2005. High-Throughput Assays Using a Luciferase-Expressing Replicon, Virus-Like Particles, and Full-Length Virus for West Nile Virus Drug Discovery. *Antimicrobial Agents and Chemotherapy,* 49**,** 4980-4988.

QUINONES-MATEU, M. & ARTS, E. 2006. Virus fitness: concept, quantification, and application to HIV population dynamics. *Quasispecies: Concept and Implications for Virology***,** 83-140.

REY, F. A., HEINZ, F. X., MANDL, C., KUNZ, C. & HARRISON, S. C. 1995. The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. *Nature,* 375**,** 291-298.

RICE, C. M. 1985. *Science,* 229**,** 726-733.

ROBERTSON, G., HIRST, M., BAINBRIDGE, M., BILENKY, M., ZHAO, Y., ZENG, T., EUSKIRCHEN, G., BERNIER, B., VARHOL, R. & DELANEY, A. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods,* 4**,** 651-657.

ROSSI, S. L., ROSS, T. M. & EVANS, J. D. 2010. West Nile Virus. *Clinics in Laboratory Medicine,* 30**,** 47-65.

ROUX, L., SIMON, A. E. & HOLLAND, J. J. 1991. Effects of Defective Interfering Viruses on Virus Replication and Pathogenesis< i> In Vitro</i> and< i> In Vivo</i>. *Advances in Virus Research,* 40**,** 181-211.

RUGGLI, N. & RICE, C. M. 1999. Functional cDNA Clones of The< i> Flaviviridae</i>: Strategies and Applications. *Advances in Virus Research,* 53**,** 183-207.

RUIZ-JARABO, C. M., ARIAS, A., BARANOWSKI, E., ESCARMÍS, C. & DOMINGO, E. 2000. Memory in viral quasispecies. *Journal of Virology,* 74**,** 3543-3547.

RUíZ-JARABO, C. M., ARIAS, A., MOLINA-PARíS, C., BRIONES, C., BARANOWSKI, E., ESCARMíS, C. & DOMINGO, E. 2002. Duration and fitness dependence of quasispecies memory. *Journal of Molecular Biology,* 315**,** 285-296.

SCHLICK, P., TAUCHER, C., SCHITTL, B., TRAN, J. L., KOFLER, R. M., SCHUELER, W., VON GABAIN, A., MEINKE, A. & MANDL, C. W. 2009. Helices α2 and α3 of West Nile Virus Capsid Protein Are Dispensable for Assembly of Infectious Virions. *Journal of Virology,* 83**,** 5581-5591.

SCHNEIDER, W. L. & ROOSSINCK, M. J. 2001. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *Journal of Virology,* 75**,** 6566-6571.

SCHONES, D. E., CUI, K., CUDDAPAH, S., ROH, T.-Y., BARSKI, A., WANG, Z., WEI, G. & ZHAO, K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell,* 132**,** 887-898.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology,* 26**,** 1135-1145.

SHI, P.-Y., BRINTON, M. A., VEAL, J. M., ZHONG, Y. Y. & WILSON, W. D. 1996. Evidence for the Existence of a Pseudoknot Structure at the 3‘ Terminus of the Flavivirus Genomic RNA†. *Biochemistry,* 35**,** 4222-4230.

SHI, P.-Y., TILGNER, M. & LO, M. K. 2002. Construction and Characterization of Subgenomic Replicons of New York Strain of West Nile Virus. *Virology,* 296**,** 219-233.

SIMMS, D., CIZDZIEL, P. E. & CHOMCZYNSKI, P. 1993. TRIzol: A new reagent for optimal single-step isolation of RNA. *Focus,* 15**,** 99-102.

SMITH, D. R., ADAMS, A. P., KENNEY, J. L., WANG, E. & WEAVER, S. C. 2008. Venezuelan equine encephalitis virus in the mosquito vector< i> Aedes taeniorhynchus</i>: Infection initiated by a small number of susceptible epithelial cells and a population bottleneck. *Virology,* 372**,** 176-186.

SMITH, T. F. & WATERMAN, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology,* 147**,** 195-197.

STADLER, K., ALLISON, S. L., SCHALICH, J. & HEINZ, F. X. 1997. Proteolytic activation of tick-borne encephalitis virus by furin. *Journal of Virology,* 71**,** 8475-81.

SUGARBAKER, D. J., RICHARDS, W. G., GORDON, G. J., DONG, L., DE RIENZO, A., MAULIK, G., GLICKMAN, J. N., CHIRIEAC, L. R., HARTMAN, M.-L. & TAILLON, B. E. 2008. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proceedings of the National Academy of Sciences,* 105**,** 3521-3526.

SUTHAR, M. S., BRASSIL, M. M., BLAHNIK, G. & GALE, M. 2012. Infectious clones of novel lineage 1 and lineage 2 West Nile virus strains WNV-TX02 and WNV-Madagascar. *Journal of Virology*.

SWETINA, J. & SCHUSTER, P. 1982. Self-replication with errors: A model for polvnucleotide replication22This paper is considered as part II of Model Studies on RNA replication. Part I is the Gassner and Schuster [14]. *Biophysical Chemistry*, 16**,** 329-345.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28**,** 2731-2739.

TAUCHER, C., BERGER, A. & MANDL, C. W. 2010. A trans-Complementing Recombination Trap Demonstrates a Low Propensity of Flaviviruses for Intermolecular Recombination. *Journal of Virology*, 84**,** 599-611.

TAVARÉ, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci*, 17**,** 57-86.

VAN REGENMORTEL, M. H., INTERNATIONAL & FAUQUET, C. M. 2000. *Virus Taxonomy: Classification and Nomenclature of Viruses: Seventh Report of the International Committee on Taxonomy of Viruses*, Academic Press.

VAN TASSELL, C. P., SMITH, T. P., MATUKUMALLI, L. K., TAYLOR, J. F., SCHNABEL, R. D., LAWLEY, C. T., HAUDENSCHILD, C. D., MOORE, S. S., WARREN, W. C. & SONSTEGARD, T. S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, 5**,** 247-252.

VAN, V. 1973. A new evolutionary law. *Evolutionary theory*, 1**,** 1-30.

VENTER, M., HUMAN, S., ZAAYMAN, D., GERDES, G. H., WILLIAMS, J., STEYL, J., LEMAN, P. A., PAWESKA, J. T., SETZKORN, H., ROUS, G., MURRAY, S., PARKER, R., DONNELLAN, C. & SWANEPOEL, R. 2009. Lineage 2 west nile virus as cause of fatal neurologic disease in horses, South Africa. *Emerging Infectious Diseases*, 15**,** 877-84.

WANG, C., MITSUYA, Y., GHARIZADEH, B., RONAGHI, M. & SHAFER, R. 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Research*, 17**,** 1195 - 201.

WENGLER, G. & WENGLER, G. 1993. The NS 3 Nonstructural Protein of Flaviviruses Contains an RNA Triphosphatase Activity. *Virology*, 197**,** 265-273.

WERTHEIMER, A. M. 2012. West Nile Virus: an Update on Recent Developments. *Clinical Microbiology Newsletter*, 34, 67-71.

WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y.-J., MAKHIJANI, V. & ROTH, G. T. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452**,** 872-876.

WILHELM, B. T., MARGUERAT, S., WATT, S., SCHUBERT, F., WOOD, V., GOODHEAD, I., PENKETT, C. J., ROGERS, J. & BÄHLER, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature,* 453**,** 1239-1243.

WILKE, C. 2005. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology,* 5**,** 44.

WINKLER, G., MAXWELL, S. E., RUEMMLER, C. & STOLLAR, V. 1989. Newly synthesized dengue-2 virus nonstructural protein NS1 is a soluble protein but becomes partially hydrophobic and membrane-associated after dimerization. *Virology,* 171**,** 302-305.

WOLD, B. & MYERS, R. M. 2008. Sequence census methods for functional genomics. *Nature Methods,* 5**,** 19-21.

WOOLHOUSE, M. E. J. 2002. Population biology of emerging and re-emerging pathogens. *Trends in Microbiology,* 10**,** s3-s7.

WOOLHOUSE, M. E. J., HAYDON, D. T. & ANTIA, R. 2005. Emerging pathogens: the epidemiology and evolution of species jumps. *Trends in Ecology & Evolution,* 20**,** 238-244.

WOOLHOUSE, M. E. J., TAYLOR, L. H. & HAYDON, D. T. 2001. Population Biology of Multihost Pathogens. *Science,* 292**,** 1109-1112.

WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics,* 16**,** 97.

YAMSHCHIKOV, V. F., WENGLER, G., PERELYGIN, A. A., BRINTON, M. A. & COMPANS, R. W. 2001. An Infectious Clone of the West Nile Flavivirus. *Virology,* 281**,** 294-304.

YANAGI, M., PURCELL, R. H., EMERSON, S. U. & BUKH, J. 1997. Transcripts from a single full-length cDNA clone of hepatitis C virus are infectious when directly transfected into the liver of a chimpanzee. *Proceedings of the National Academy of Sciences,* 94**,** 8738-8743.

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**,** 119.

ZAGORDI, O., KLEIN, R., DAUMER, M. & BEERENWINKEL, N. 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res,* 38**,** 7400 - 7409.

Appendix A

Appendix B

Single Nucleotide Polymorphisms (SNPs)

The genome-wide distribution of Single Nucleotide Polymorphisms (SNPs) of WNV 1968 isolates are illustrated in Figure B1 as represented by the frequency of the major allele in each position. SNPs of isolate a (WNV 1968) and isolate b (WNV 1968) are listed in Table B1 and Table B2 respectively. Similarly, the genome-wide distribution of SNPs of WNV 349/77 isolates are illustrated in Figure B2. SNPs of isolate c (WNV 349/77), isolate d (WNV 349/77), isolate e (WNV 349/77), isolate f (WNV 349/77) and isolate g (WNV 349/77) are listed in Table B3, Table B4, Table B5, Table B6 and Table B7 respectively.

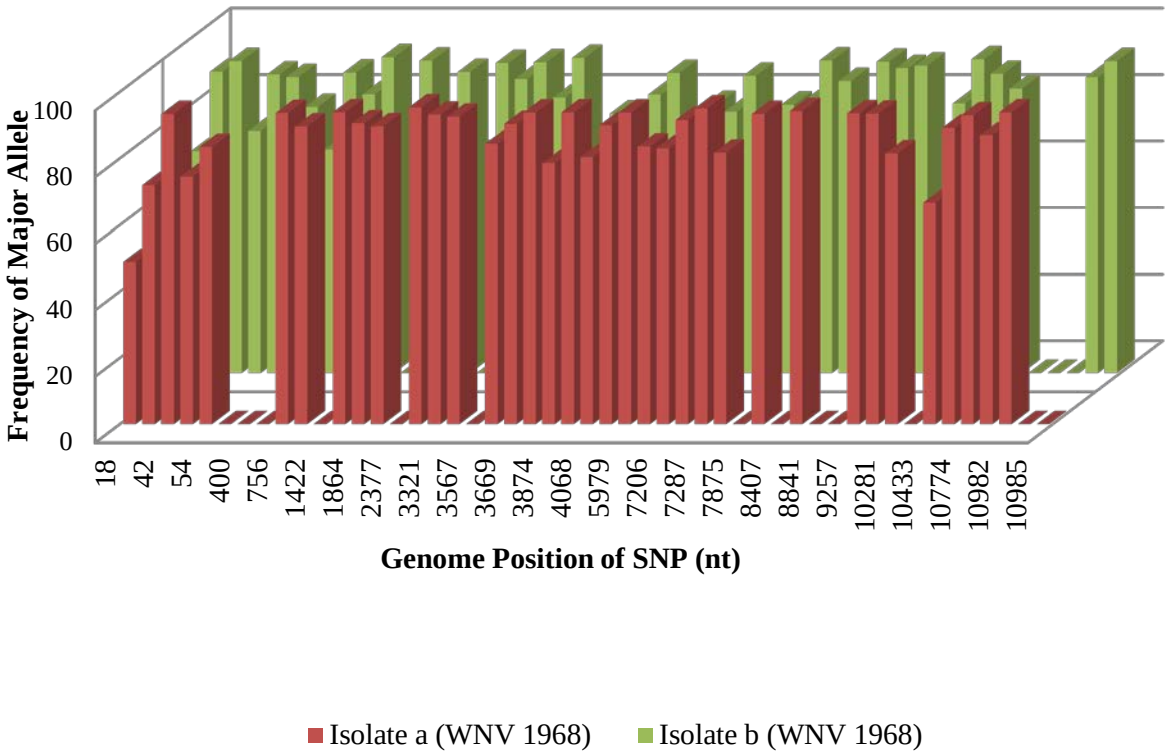

Figure B1. Genome-wide Single Nucleotide Polymorphisms (SNPs) of WNV 1968 isolates as represented by the frequency of the major allele. Isolate A3 is represented in red and isolate B1 is represented in green. The genome position where a SNP occurred is indicated on the horizontal axis. The frequency of the major allele in each position is indicated on the vertical axis.

Table B1. Single Nucleotide Polymorphisms (SNPs) of isolate A3 (WNV 1968)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
18	A/G/T	48.9/42.6/8.5	46/40/8	94	5'UTR
26	A/G	71.9/28.1	105/41	146	5'UTR
42	T/G	93.3/6.7	475/34	509	5'UTR
44	A/G	74.4/22.2	402/120	540	5'UTR
54	A/G	83.4/22.2	863/172	1035	5'UTR
756	G/A	93.6/5.8	931/58	995	Membrane
978	A/G	89.5/10.5	2509/293	2802	Membrane
1436	A/G	93.8/6.3	90/6	96	Envelope
1864	A/G	90.5/9.4	5639/585	6230	Envelope
2009	T/C	89.6/6.0	2332/155	2602	Membrane
2683	G/A	94.3/5.7	4038/243	4281	NS1
3321	A/G	93.1/6.9	892/66	958	NS1
3458	A/G	92.5/5.5	369/22	399	NS1
3573	C/G/T	84.3/7.8/7.8	582/54/54	690	NS2A
3669	T/C	90.3/9.6	2028/216	2245	NS2A
3814	A/G	93.7/6.3	4437/298	4736	NS2A
3874	T/C	78.5/21.4	2683/733	3419	NS2A
3917	T/C	93.8/6.2	2178/144	2323	NS2A
4068	A/T/G	80.3/11.1/8.5	720/100/76	897	NS2A
5057	A/G	89.9/5.7	5715/363	6356	NS3
5979	G/A	93.7/6.3	2030/136	2167	NS3
7126	T/C	83.5/13.0	167/26	200	NS4B
7206	G/A	83.0/17.0	376/77	453	NS4B
7221	C/T	91.5/6.0	721/47	788	NS4B
7287	C/T	95.0/5.0	3168/167	3335	NS4B
7305	A/G	81.6/18.4	2689/606	3296	NS4B
7974	G/A	93.3/6.7	6569/472	7041	NS5
8454	C/T	94.2/5.8	6218/382	6600	NS5
9257	C/T	93.5/6.4	3628/250	3879	NS5
9308	G/A	93.4/5.0	1354/73	1449	NS5
10281	T/C	81.4/18.6	2315/528	2845	NS5
10433	A/G/T	66.6/18.9/14.4	1009/287/218	1515	NS5
10439	C/T	89.1/10.3	2242/260	2516	NS5
10774	G/A	92.9/5.7	459/28	494	NS5

10953	C/G/T	86.9/13.1	53/8	61	3'UTR
10982	T/A	93.8/6.3	15/1X	16	3'UTR

Table B2. Single Nucleotide Polymorphisms (SNPs) of isolate B3 (WNV 1968)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
18	G/A	66.7/29.8	55/25	84	5'UTR
26	A/G	90.6/9.4	96/10	106	5'UTR
42	T/G	93.8/6.1	480/31	512	5'UTR
44	A/G	72.8/22.2	417/127	573	5'UTR
54	A/G	89.8/10.1	900/101	1002	5'UTR
222	C/T	89.0/9.4	20740/2192	23293	Nucleocapsid
400	C/A	80.0/20.0	4867/1216	6086	Nucleocapsid
405	G/A	67.1/32.8	3679/1798	5479	Nucleocapsid
756	G/A	90.3/8.9	607/60	672	Membrane
978	A/G	83.8/16.2	1316/254	1570	Membrane
1422	T/C	94.9/5.1	1803/96	1899	Envelope
1864	A/G	93.9/6.1	2971/193	3164	Envelope
2377	A/T	90.5/5.2	1630/95	1801	Envelope
3321	A/G	93.2/6.6	607/43	651	NS1
3458	A/G	88.4/8.9	336/34	380	NS1
3567	C/T	93.4/6.0	479/31	513	NS2A
3573	C/G/T	82.8/9.4/7.7	526/60/49	635	NS2A
3669	T/C	94.7/5.3	1923/107	2030	NS2A
3874	T/C	78.0/22.0	2963/836	3799	NS2A
4068	A/G/T	83.7/9.2/6.9	979/107/81	1169	NS2A
5057	A/G	90.2/5.6	5263/328	5838	NS3
7126	T/C/A	80.9/10.5/7.9	123/16/12	152	NS4B
7206	G/A	78.5/21.5	398/109	507	NS4B
7221	C/T	89.4/6.7	853/64	954	NS4B
7305	A/G	80.7/19.3	2990/714	3706	NS4B
7875	A/G/T	81.1/12.0/6.9	473/70/40	583	NS5
7974	G/A	94.0/6.0	4267/273	4540	NS5
8407	A/G	87.8/8.2	3034/282	3455	NS5
8841	A/G	93.6/5.0	836/45	893	NS5
9216	A/G	91.6/8.3	2129/194	2324	NS5
9257	C/T	92.4/7.6	2347/194	2541	NS5

10281	T/C	81.0/19.0	4687/1100	5788	NS5
10403	T/A	90/10	538/60	598	NS5
10429	T/C	94.4/5.2	473/26	501	NS5
10433	A/G/T	69.3/16.9/13.8	407/99/81	587	NS5
10439	C/T	85.5/14.2	728/121	851	NS5
10983	G/A	88.9/11.1	16/2X	18	3'UTR
10985	A/G	93.8/6.3	15.1	18	3'UTR

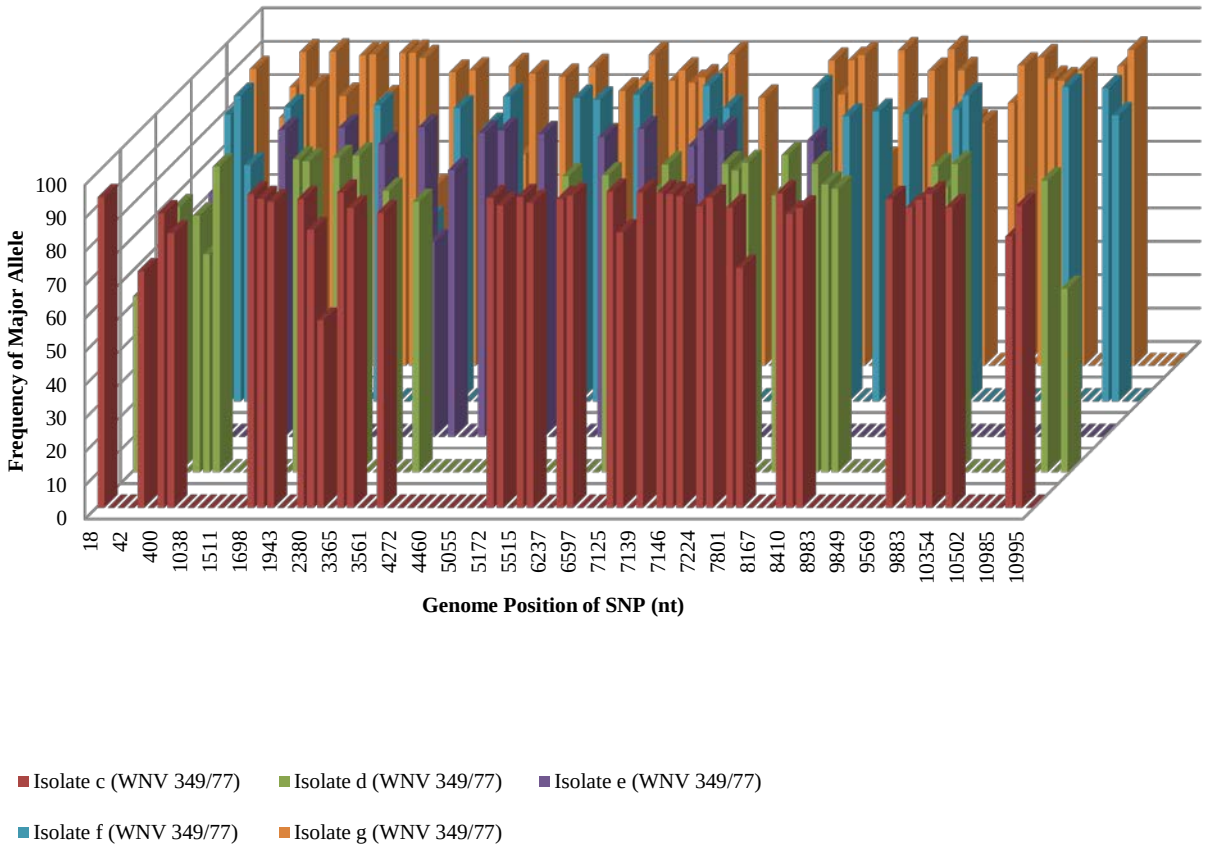

Figure B2. Genome-wide Single Nucleotide Polymorphisms (SNPs) of WNV 349/77 isolates as represented by the frequency of the major allele. Isolate C3 is represented in red, isolate D1 in green, isolate E1 in purple, isolate F1 in blue and isolate G2 in orange. The genome position where a SNP occurred is indicated on the horizontal axis. The frequency of the major allele in each position is indicated on the vertical axis.

Table B3. Single Nucleotide Polymorphisms (SNPs) of isolate C3 (WNV 249/77)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
42	T/G	93/7	53/4	57	5'UTR
44	A/G	70.8/24.6	46/16	65	5'UTR
400	C/A	88.2/11.8	1079/144	1223	Nucleocapsid
405	G/A	82.2/17.6	953/204	1160	Nucleocapsid
1698	G/A	93.9/5.0	1269/68	1351	Envelope
1862	G/A	92.5/6.1	457/30	494	Envelope
1867	A/G	91.6/8.4	362/33	395	Envelope
2316	T/C	92.5/7.5	210/17	227	Envelope
2380	A/T/G	83.2/20.6/6.2	440/56/33	529	Envelope
2491	G/A	56.1/43.9	991/775	1766	NS1
3365	G/A	94.4/5.6	203/12	215	NS1
3461	A/G	89.7/5.5	130/8	145	NS1
4071	A/T	88.3/8.0	143/13	162	NS2A
5172	A/G	92.8/7.2	3674/283	3958	NS3
5225	T/A	90.6/5.1	690/39	762	NS3
5515	G/A	93.1/5.0	149/8	160	NS3
5697	C/T	91.3/6.6	167/12	183	NS3
6308	A/G	92.5/5.9	628/40	679	NS3
6337	G/A	93.5/5.4	724/42	774	NS3
7125	G/A	94.7/5.3	89/5	94	NS4B
7129	T/C	82.5/15.9	52/10	63	NS4B
7139	T/C	94.3/5.7	33/2	35	NS4B
7144	G/A	94.3/5.7	33/2	35	NS4B
7146	C/T	93.9/6.1	31/2	33	NS4B
7182	T/G	93.3/6.7	42/3	45	NS4B
7224	C/T	90.3/5.6	371/23	411	NS4B
7230	C/T	92.9/5.9	533/34	574	NS4B
7801	G/A	90.0/10.0	108/12	120	NS5
7878	A/T/G	71.9/17.3/10.8	100/24/15	139	NS5
8404	C/T	94.2/5.0	452/24	480	NS5
8410	A/G	88.1/9.4	694/74	788	NS5
8879	A/G	89.7/8.7	269/26	300	NS5
9849	T/C	92.5/7.5	124/10	134	NS5
9883	C/T	90.0/7.6	189/16	210	NS5

9888	A/G	92.2/7.2	270/21	293	NS5
10259	G/A	93.9/5.5	833/49	887	NS5
10436	A/G	90/5.8	432/28	480	NS5
10986	G/A	81.3/17.2	52/11	64	3'UTR
10994	A/C	90.9/9.1	20/2X	22	3'UTR

Table B4. Single Nucleotide Polymorphisms (SNPs) of isolate D1 (WNV 249/77)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
18	A/G	52.7/43.6	29/24	55	5'UTR
44	A/G	79.4/15.9	370/74	466	5'UTR
400	C/A	76.9/22.9	5685/1695	7389	Nucleocapsid
405	G/A	65.3/34.6	4473/2374	6853	Nucleocapsid
593	T/C	91.6/5.8	1314/82	1435	Premembrane
1862	G/A	93.7/5.6	1401/83	1495	Envelope
1867	A/G	93.1/6.8	1239/91	1331	Envelope
2316	C/T	94.2/5.8	696/43	739	Envelope
2380	A/T/G	80.6/10.7/8.6	1696/224/181	2103	Envelope
2491	G/A	94.9/5.1	6858/369	7227	NS1
3461	A/G	84.3/13.3	418/66	496	NS1
4071	A/T/G	81.0/12.6/6.4	558/87/44	689	NS2A
5697	C/T	88.7/7.3	558/46	629	NS3
6337	G/A	89.0/8.8	2879/284	3235	NS3
7129	T/C	84.2/10.4	219/27	260	NS4B
7133	T/C	92.2/6.8	177/13	192	NS4B
7210	C/G/T	92.3/5.7	982/61	1064	NS4B
7224	C/T	90.4/6.3	1545/108	1710	NS4B
7230	C/T	92.7/5.7	2152/133	2322	NS4B
7878	A/G/T	82.9/9.1/8.0	436/48/42	526	NS5
7947	C/T	95.0/5.0	1250/66	1216	NS5
8404	C/T	92.3/5.2	2066/117	2239	NS5
8410	A/G	86.2/9.1	3198/338	3709	NS5
8879	A/G	84.9/11.6	959/131	1129	NS5
9878	C/T	91.6/7.2	557/44	608	NS5
9883	C/T	85.3/13.0	701/107	822	NS5

9888	A/G	92.3/7.0	1074/82	1164	NS5
10986	G/A	87.4/9.2	76/8	87	3'UTR
10995	G/A/C	55/25/20	11/05/2004	20	3'UTR

Table B5. Single Nucleotide Polymorphisms (SNPs) of isolate E1 (WNV 249/77)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
18	A/G	66.7/33.3	14/7X	21	5'UTR
44	A/G/T	73/21.4/5.6	246/72/19	337	5'UTR
1532	A/G	91.9/5.0	6025/328	6556	Envelope
1867	A/G	92.6/1.4	3054/243	3298	Envelope
2380	A/T/G	87.6/7.2/5.1	2854/235/166	3257	Envelope
3461	A/G	92.7/6.1	912/60	984	Envelope
3549	A/T	58.3/41.5	535/381	918	NS2A
4071	A/T/G	79.9/12.7/7.4	584/93/54	731	NS2A
4457	A/G	90.9/5.7	5451/340	5994	NS2B
4460	A/G	91.8/6.0	5500/365	5993	NS2B
5060	A/G	90.7/5.2	9088/519	10017	NS3
5697	C/T	89.7/6.7	1414/106	1576	NS3
6337	G/A	92.2/6.3	4117/282	4467	NS3
7129	T/C	87.0/9.8	685/77	787	NS4B
7133	T/C	91.7/5.5	585/35	638	NS4B
7198	T/C	92.0/7.9	712/61	774	NS4B
7878	A/T	88.8/6.7	1125/85	1267	NS5

Table B6. Single Nucleotide Polymorphisms (SNPs) of isolate F1 (WNV 249/77)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
18	A/G/T	47.4/41.0/11.5	37/32/9	78	5'UTR
26	A/G	85.9/14.1	110/18	128	5'UTR
42	T/G	91.5/8.5	377/35	412	5'UTR
44	A/G	70.7/24.5	311/108	440	5'UTR
593	T/C	88.1/7.9	1009/91	1145	Premembrane
1867	A/G	88.7/11.2	1923/242	2167	Envelope
2491	A/G	54.5/45.4	3435/2860	6297	NS1

3461	A/G	87.8/9.2	381/40	434	NS1
4071	A/T/G	82.5/9.6/7.9	430/50/41	531	NS2A
4272	C/T	91.3/6.9	1160/88	1271	NS2B
5060	A/G	90.9/5.1	5315/297	5844	NS3
5172	A/G	90.4/9.6	13377/1418	14796	NS3
5697	C/T	91.8/6.6	571/41	622	NS3
6955	A/G	94.5/5.5	3023/177	3200	NS4B
7129	T/C	87.9/8.4	305/29	347	NS4B
7230	C/T	93.9/5.1	1133/63	1207	NS4B
7878	A/T/G	85.4/7.7/6.9	633/57/51	741	NS5
8410	A/G	86.8/8.2	3408/322	3925	NS5
8879	A/G	86.1/10.3	549/66	638	NS5
9849	T/C	87.5/12.5	315/45	360	NS5
9883	C/T	91.6/5.9	468/30	511	NS5
10442	C/T	94.2/5.8	3396/210	3607	NS5
10985	T/A	93.8/6.3	15/1X	16	3'UTR
10986	G/A	85.7/14.3	12/2X	14	3'UTR

Table B7. Single Nucleotide Polymorphisms (SNPs) of isolate G2 (WNV 249/77)

Position	Allele Variations	Frequencies	Counts	Coverage	Gene
18	A/G/T	72.2/16/7/11.1	13/2/2X	18	5'UTR
20	C/G	88.9/11.1	16/2X	18	5'UTR
44	A/G	74.4/22.1	64/19	88	5'UTR
292	C/A	83.5/15.7	8970/1691	10740	Nucleocapsid
400	C/A	94.0/6.0	1798/115	1913	Nucleocapsid
405	G/A	83.6/15.9	1458/278	1744	Nucleocapsid
1038	A/G	94.1/5.2	1458/80	1550	Envelope
1378	G/A	80.9/19.1	1629/385	2014	Envelope
1511	A/T	93.0/6.0	493/32	530	Envelope
1514	A/T	93.3/5.8	500/31	536	Envelope
1520	A/G/T	80.2/10.8/9.0	594/80/67	741	Envelope
1862	G/A	93.7/5.4	1154/67	1232	Envelope
1867	A/G	93.8/6.2	935/62	997	Envelope
1943	A/G	92.2/5.6	1920/117	2082	Envelope

2134	G/A	53.7/46.2	697/600	1299	Envelope
2380	A/T/G	87.9/6.7/5.3	1598/122/96	1817	Envelope
3283	C/T	88.5/9.7	815/89	921	NS1
3561	A/G	89.8/8.3	334/31	372	NS2A
4071	A/T/G	63.4/20.3/16.4	147/47/38	232	NS2A
4245	T/C	87.6/10.1	641/74	732	NS2B
4372	C/T	86.6/9.9	681/78	786	NS2B
4704	C/T	75.6/24.4	1258/407	1665	NS3
5050	A/T	89.5/10.5	1053/123	1177	NS3
5055	A/G/T	68.8/21.4/9.8	1056/328/151	1535	NS3
5060	A/G/T	52.2/31.5/16.3	1011/610/316	1937	NS3
5065	A/G/T	82.4/10.8/5.9	2175/286/156	2640	NS3
5225	T/A/C	85.7/7.4/6.5	2384/206/181	2783	NS3
5297	C/G/T	93.2/5.2	1008/56	1081	NS3
5697	C/T/G	85.4/9.4/5.2	182/20/11	213	NS3
6100	G/A	88.5/9.3	780/82	881	NS3
6237	T/C	84.8/15.2	1195/214	1409	NS3
6308	A/G	86.4/10.0	542/63	627	NS3
6597	T/C	88.2/11.8	1425/191	1616	NS4A
6717	A/G	93.3/5.5	3176/186	3403	NS4A
7129	T/C	80.3/16.5	102/21	127	NS4B
7210	C/G	91.4/5.6	778/48	851	NS4B
7224	C/T	81.2/15.0	1178/217	1451	NS4B
7230	C/T	91.6/6.9	2116/227	2311	NS4B
7409	G/A	93.1/5.7	3716/227	3990	NS4B
7878	A/G/T	61.7/23.4/15/0	103/39/25	167	NS5
8167	G/A	94.5/5.5	1532/90	1622	NS5
8401	C/G/T	76.3/16.5/7.2	521/113/49	683	NS5
8404	C/T/G	74.8/16.4/8.8	535/117/63	715	NS5
8410	A/G	88.4/9.7	1281/140	1449	NS5
8904	T/C	94.9/5.1	130/7	137	NS5
8983	T/G	88.6/11.4	287/37	324	NS5
8987	A/G/T	73.0/17.1/9.9	443/104/60	607	NS5
8989	A/G/T	72.7/15.2/12.0	460/96/76	633	NS5
9563	A/T/G	78.7/12.0/9.3	170/26/20	216	NS5
9569	A/T	90.0/6.6	368/27	409	NS5

9878	C/T	92.4/5.7	306/19	331	NS5
9883	C/T	86.0/12.0	416/58	484	NS5
10259	G/A	85.5/13.2	1324/204	1549	NS5
10354	C/T	88.5/11.5	816/106	922	NS5
10502	T/C	77.5/22.3	597/172	770	NS5
10514	T/C	89.8/10.1	773/87	861	NS5
10956	C/G	94.7/5.3	36/2	38	3'UTR

Appendix C

Haplotypes

Viable haplotypes of WNV 349/77 isolates were grouped and renamed Haplotype 1 - Haplotype 30. The frequency of each haplotype as found for each isolate is indicated in Table C1. A hyphen replaces the frequency where any one haplotype was not found an isolate.

Table C1. Viable haplotypes of WNV 349/77 isolates

Haplotype	Frequency: isolate D1	Frequency: isolate E1	Frequency: isolate F1	Frequency: isolate G2	Gene Region Containing Variation
Haplotype 1	0.00036	0.00055	0.0007	-	NS5
Haplotype 2	0.00827	0.00873	0.00632	0.00631	NS5
Haplotype 3	0.00217	0.0042	0.00293	-	NS5
Haplotype 4	0.00661	-	0.00737	-	NS5
Haplotype 5	0.0046	0.00577	-	-	NS5
Haplotype 6	0.00208	-	0.00144	-	NS3
Haplotype 7	0.00268	0.00353	0.00372	0.00274	NS3
Haplotype 8	0.00094	-	-	0.00106	NS3
Haplotype 9	0.00286	0.00335	0.00269	0.00293	NS2A
Haplotype 10	0.00105	0.00201	-	-	NS1
Haplotype 11	0.00116	0.00212	-	-	NS1
Haplotype 12	0.00051	-	0.00072	0.00081	Envelope
Haplotype 13	0.00029	0.00018	0.00039	-	Envelope
Haplotype 14	0.0008	0.0027	0.00173	-	Envelope
Haplotype 15	0.00147	0.00365	0.00184	-	Envelope
Haplotype 16	0.00775	0.00309	-	-	Envelope
Haplotype 17	0.00237	-	0.00172	-	Envelope
Haplotype 18	0.00322	-	0.00175	0.00417	Envelope
Haplotype 19	0.00188	-	-	0.00078	Membrane
Haplotype 20	0.02395	-	0.01225	0.0074	Capsid
Haplotype 21	-	0.00928	-	0.00841	NS5
Haplotype 22	-	0.00912	0.00835	0.01083	NS2A
Haplotype 23	-	0.00025	0.00027	0.00014	NS1
Haplotype 24	-	0.00109	0.00061	-	Envelope
Haplotype 25	-	0.00148	0.00061	0.00221	Envelope

Haplotype 26	-	0.00074	0.00031	-	Envelope
Haplotype 27	-	0.00212	-	0.00194	Envelope
Haplotype 28	-	0.00192	-	0.001249	Envelope
Haplotype 29	-	0.0032	-	0.00426	NS1
Haplotype 30	-	0.00046	-	0.00104	Envelope

Appendix D

Table D1. Genome-wide nucleotide sequence p-distance values for all WNV strains in this study.

AF260967 (USA)																								
AY660002 (Mexico)	0.00																							
AY277251 (Russia)	0.03	0.03																						
AY701413 (Morocco)	0.02	0.02	0.03																					
AF404757 (Italy)	0.02	0.02	0.03	0.01																				
AF260967 (Romania)	0.02	0.03	0.03	0.01	0.02																			
AY490240 (China)	0.04	0.04	0.04	0.04	0.04	0.04																		
D00246 (Australia)	0.11	0.11	0.13	0.11	0.11	0.12	0.10																	
M12294 (Uganda)	0.19	0.20	0.19	0.19	0.19	0.19	0.19	0.20																
WNV 1968 Isolate A2 (SA)	0.19	0.19	0.19	0.18	0.19	0.19	0.18	0.20	0.00															
WNV 1968 Isolate B1 (SA)	0.19	0.19	0.19	0.18	0.19	0.19	0.18	0.20	0.00	0.00														
EF429198 (SA)	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.02	0.02	0.02													
WNV349/77 Isolate F1 (SA)	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.02	0.02	0.02	0.03												
WNV349/77 Isolate E1 (SA)	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.02	0.02	0.02	0.03	0.00											
WNV349/77 Isolate G2 (SA)	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.02	0.02	0.02	0.03	0.00	0.00										
WNV349/77 Isolate C3 (SA)	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.02	0.02	0.02	0.03	0.00	0.00	0.00									
WNV349/77 Isolate D1 (SA)	0.20	0.20	0.19	0.19	0.19	0.20	0.19	0.20	0.02	0.02	0.02	0.03	0.00	0.00	0.00	0.00								
WNV 1968 Isolate B1 (SA)	0.19	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03							
HS87/11 Isolate H2 (SA)	0.19	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02						
DQ176636 (Madagascar)	0.21	0.21	0.21	0.22	0.22	0.21	0.21	0.22	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.16	0.15	0.16					
AY765264 (Czech Republic)	0.24	0.24	0.25	0.24	0.24	0.24	0.24	0.25	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.26				
AY944247 (India)	0.22	0.22	0.23	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.25	0.25			
AY944242 (India)	0.23	0.23	0.23	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.25	0.25	0.01		
AY944241 (India)	0.22	0.22	0.23	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.23	0.23	0.25	0.25	0.01	0.00	
AY277251 (Russia)	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.27	0.27	0.26	0.26	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.25	0.28	0.26	0.27	0.26
AF075723 JEV	0.34	0.34	0.34	0.33	0.34	0.33	0.33	0.35	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.36	0.34	0.33	0.33	0.33	0.36

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Animal Ethics Committee

PROJECT TITLE	**Genomics of South African West Nile Viruses**
PROJECT NUMBER	**V006-13**
RESEARCHER/PRINCIPAL INVESTIGATOR	**Ms C Kortenhoeven**

STUDENT NUMBER (where applicable)	**25014553**
DISSERTATION/THESIS SUBMITTED FOR	**MSc**

ANIMAL SPECIES	**Mus musculus**	
NUMBER OF ANIMALS	**4**	
Approval period to use animals for research/testing purposes		**March – April 2013**
SUPERVISOR	**Prof. C Abolnik**	

KINDLY NOTE:

Should there be a change in the species or number of animal/s required, or the experimental procedure/s - please submit an amendment form to the UP Animal Ethics Committee for approval before commencing with the experiment

APPROVED Date **25 February 2013**

CHAIRMAN: UP Animal Ethics Committee Signature