THE MOLECULAR EVOLUTION AND EPIDEMIOLOGY OF RUBELLA VIRUS

By Leendert J. Cloete

Supervisor: Dr. Gordon W. Harkins

A thesis submitted in partial fulfilment of the requirements for the degree of Magister Scientiae (Bioinformatics), in the Department of the South African National Bioinformatics Institute, University of the Western Cape

November 2014

UNIVERSITY *of the*
WESTERN CAPE

KEYWORDS

Rubella Virus

Congenital Rubella Syndrome

Nucleic Acid Secondary Structures

Recombination

Positive Selection

Nucleotide Substitution Rate

Bayesian Evolutionary Analysis

Phylogeography

Generalized Linear Model

Phylodynamics

ORIGINAL PUBLICATIONS

Cloete, LJ, EP Tanov, BM Muhire, DP Martin, and GW Harkins. 2014. “The Influence of Secondary Structure, Selection and Recombination on Rubella Virus Nucleotide Substitution Rate Estimates.” Virol J 11 (1): 166.

ABBREVIATIONS

BEAST	bayesian evolutionary analysis by sampling trees
BEAUti	bayesian evolutionary analysis utility
BF	bayes factor
BSP	bayesian skygrid plot
CP	capsid protein
CRS	congenital rubella syndrome
CTMCs	continuous-time markov chains
DOOSS	data overlaid on secondary structures
E1	envelope glycoprotein 1
E2	envelope glycoprotein 2
ESS	effective sample sizes
FUBAR	fast, unconstrained bayesian approximation for inferring selection
GARD	genetic algorithm for recombination detection
GAVI	global vaccine alliance
GLM	generalized linear model
GNI	gross national income
HCSS	high confidence structure set
HPD	highest posterior density

kml	key markup language
MCC	maximum clade credibility
MCMC	markov chain monte carlo
MFE	minimum free energy
MR	measles-rubella-containing vaccine
MRCA	most recent common ancestor
MMR	measles-mumps-rubella-containing vaccine
mRNA	messenger RNA
NASP	nucleic acid structure predictor
N_e	effective population size
NCBI	national centre for biotechnology information
NSP	non-structural proteins
nt	nucleotides
ORF	open reading frame
PARRIS	partitioning approach for robust inference of selection
PCR	polymerase chain reaction
PR	paired
RCVs	rubella virus-containing vaccines
RDP	recombination detection program
RdRp	RNA-dependant RNA polymerase

RF	recombination-free
RI	recombination-included
RNA	ribonucleic acid
RV	rubella virus
SDT	species demarcation tool
SP	structural proteins
SPREAD	spatial phylogenetic reconstruction of evolutionary dynamics
TMRCA	time to the most recent common ancestor
UNICEF	united nations international children's emergency fund
UnPR	unpaired
UTR	untranslated region
WHO	world health organization

ABSTRACT

THE MOLECULAR EVOLUTION AND EPIDEMIOLOGY OF RUBELLA VIRUS

L.J. Cloete

M. Scientiae (Bioinformatics) thesis, South African National Bioinformatics Institute, University of the Western Cape

Despite widespread rubella virus (RV) vaccination programs, annually RV still causes severe congenital defects in an estimated 100,000 children globally. A concerted attempt to eradicate RV is currently underway and analytical tools to monitor the global decline of the last remaining RV lineages will be useful for assessing the effectiveness of this endeavour. Importantly, RV evolves rapidly enough that much of its epidemiological information might be inferable from RV genomic sequence data.

Using BEASTv1.8.0, I analysed publically available RV sequence data to estimate genome-wide and gene-specific nucleotide substitution rates, to test whether the current estimates of RV substitution rates are representative of the entire RV genome. During these investigations, I specifically accounted for possible confounders of nucleotide substitution rate estimates, such as temporally biased sampling, sporadic recombination, and natural selection favouring either increased or decreased genetic diversity (estimated by the PARRIS and FUBAR methods) at nucleotide sites within RV nucleic acid secondary structures (predicted by the NASP method).

I determined that RV nucleotide substitution rates range from 1.19×10^{-3} substitutions/site/year (in the E1 region) to 7.52×10^{-4} substitutions/site/year (in the P150 region). I found that these differences between nucleotide substitution rate estimates in various RV gene regions are largely attributable to temporal sampling biases, such that datasets containing a higher proportion of recently sampled sequences will tend to have inflated estimates of mean substitution rates. Although there exists little evidence of positive selection or natural genetic recombination in

RV, I revealed that RV genomes possess extensive biologically functional nucleic acid secondary structures and that purifying selection acting to maintain these structures contributes substantially to variations in estimated nucleotide substitution rates across RV genomes.

Although both temporal sampling biases and purifying selection favouring the conservation of RV nucleic acid secondary structures have an appreciable impact on substitution rate estimates, I find that these biases do not preclude the use of RV sequence data to date ancestral sequences and evaluate the associated RV phylodynamics. The combination of uniformly high substitution rates across the RV genome and strong temporal signal within the available sequence data enabled me to analyse the epidemiological and demographical dynamics of this virus during these attempts to eradicate it. By implementing a generalized linear model (GLM) and symmetrical model of discretized phylogeographic spread, I was able to identify several predictive variables of geographical RV spread and detect transmission linkages between distinct geographical regions. These results suggest that, in addition to strengthened vaccination strategies, there also needs to be an increased effort to educate people about the effects of vaccination and risks of RV infection.

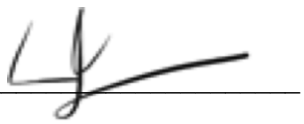
November 2014

DECLARATION

I declare that *The Molecular Evolution And Epidemiology Of Rubella Virus* is my own work, that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged as complete references.

Leendert J. Cloete

November 2014

Signed: 

ACKNOWLEDGEMENTS

Firstly, I would like to thank God for providing and entrusting me with this opportunity, and for giving me the strength and wisdom to see it through. I would also like to thank my supervisor, Dr. Gordon W. Harkins, for his advice, time, patience and guidance, both throughout this project and personally, and extent my gratitude towards Dr. Darren P. Martin, Emil P. Tanov, Brejnev M. Muhire and Dr. Jean-Baka Domelevo Entfellner.

Finally, I would like to acknowledge the South African National Research Foundation (NRF), The Poliomyelitis Research Foundation (PRF) and the Deutscher Akademischer Austausch Dienst (DAAD) for their joint funding, as well as the South African Center for High Performance Computing (CHPC), South African National Bioinformatics Institute (SANBI) and the University of Cape Town Institute of Infectious Diseases and Molecular Medicine (IIDMM).

"We are drowning in information, but are starving for knowledge."

- Rutherford D. Rogers

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

It is vital that gains made by various global rubella vaccination programs are not undone by resurgent *Rubella virus* (RV) outbreaks, such as that exemplified by the measles epidemic in South Africa (National Institute for Communicable Diseases (NICD), South Africa 2010) and the United Kingdom (Wise 2013). Comprehensive rubella and congenital rubella syndrome (CRS) surveillance systems to monitor immunity within vaccinated populations needs to be strengthened, to better anticipate changes in the epidemiological dynamics caused by vaccination programs, and to improve our understanding of the factors that influence the evolution of RV. In this respect, it is vital that we continue collecting and characterising circulating RV genome data as this could potentially be used to monitor the virus' evolutionary, demographic and epidemiological dynamics in the face of intensified control strategies.

Besides increased volumes of genomic sequence data, an important prerequisite for using RV sequences in such surveillance efforts is the demonstration that the rates at which RV genomes are evolving are high enough, that they can be reliably used to track both epidemiologically relevant fluctuations in virus population sizes, and viral movement events (such as transmission between individuals or migration between different countries or continents).

In this regard, it is very promising that RV structural E1 polyprotein gene region sequences display high degrees of clock-like evolution and mean nucleotide substitution rates ranging between 0.61×10^{-3} (Jenkins et al. 2002) and 1.65×10^{-3} substitutions/nucleotide/year (Zhu et al. 2011) - a rate of evolution that should be within the bounds required to extract meaningful phylogeographic and demographic information from RV genomic sequence data. It is noteworthy that Togavirus nucleotide substitution rates estimated by Jenkins et al. (2002), using the same strict-clock maximum likelihood-based methods employed on the RV structural E1 polyprotein region, are substantially slower than those estimated for RV, whereas a

study (Zhu et al. 2011) employing a more sophisticated Bayesian relaxed molecular clock–based inference method reported RV structural E1 polyprotein nucleotide substitution rates approximately equivalent to those of other Togaviruses (Cherian et al. 2009; Volk et al. 2010; Suwannakarn et al. 2011).

Using publically available RV full genome and gene-specific sequences sampled over the past 51 years I aimed to assess whether current RV nucleotide substitution rates estimates are representative of the entire RV genome. During these investigations I specifically accounted for possible confounders of nucleotide substitution rate estimates such as sporadic genetic recombination and natural selection favouring either increased genetic diversity in response to host immune pressures, or decreased genetic diversity at nucleotide sites involved in the formation of genomic secondary structures. In addition, I reconstructed a plausible history of RV's geographical spread and determined when in relation to the past rubella epidemics, the major globally circulating RV genotypes arose. Finally, I investigated how this virus persisted in the face of intensified vaccination efforts.

1. INTRODUCTION

1.1 Togaviridae

The family *Togaviridae* derived its name from the Latin word *toga* (meaning a Roman mantle or cloak), as members of the family were among the first well-characterized viruses known to contain a lipid envelope. Consequently, many enveloped viruses were incorrectly classified as Togaviruses. Several of the original viruses have since been reclassified into different families and as a result, the family *Togaviridae* currently comprises only two genera: *Alphavirus* and *Rubivirus*.

The primarily arthropod-borne genus *Alphavirus* contains approximately 30 species, has a wide range of hosts and is known to replicate in a variety of different cell types, whereas the monospecific genus *Rubivirus* is exclusively transmitted between humans by *Rubella virus* (RV). Members of both genera contain single-stranded positive sense ribonucleic acid (RNA) genomes enclosed within small, lipid-enveloped, icosahedral particles approximately 70 nm in diameter. The genomes range between 10,000–12,000 nucleotide (nt) in length (Wolinsky et al. 2001) and encode two open reading frames (ORF); the non-structural (NSP) and structural (SP) proteins. However, despite these similarities in genomic organization, structure and replication strategy, *Alphavirus* and *Rubivirus* share little genetic homology, and are in fact only distantly related (Frey 1994). A recent study even suggested that *Rubivirus* might be more closely related to the genus *Flavivirus* (family: *Flaviviridae*), based on analysis of the structural E1 polypotein (DuBois et al. 2013).

1.2 Background on Rubella virus

Rubella (*German: Rötheln*), or German measles as it is more commonly known, is caused by a RV infection. The disease was initially described in 1740, when Friedrich Hoffmann documented the first clinical description (Ackerknecht 1982). His findings were later independently confirmed by two German physicians (hence, the common English eponym), namely de Bergen (1752) and Orlow (1758; Wesselhoeft 1949). However, due to the similar nature of these diseases, both of these physicians incorrectly believed that rubella was considered a derivative of measles and it was only in 1814, that George de Maton first suggested that rubella be considered as a distinct disease.

During 1841, the British physician Henry Veale recorded a rubella outbreak in a boys' school in India. Prior to this outbreak, the disease was medically known as Rötheln, however, he documented his findings as *rubella* (a Latin diminutive meaning "little red") which he suggested was a more "soothing" term for the English ear (Veale 1866). Nonetheless, it was only in 1881 that rubella was officially recognised as a distinct disease (Forbes 1969).

In 1914, Alfred Fabian Hess first proposed that a virus was the cause of the disease known as rubella (Hess 1914) and in 1938, Hiro and Tosaka confirmed his results, when they successfully passed RV to children using filtered nasal washings (Atkinson et al. 2012). However, it was not until 1962, that the first RV was isolated (Parkman et al. 1962; Weller and Neva 1962).

Up until 1941, rubella was considered to be a relatively mild disease with few complications that occurs mostly during childhood. However, in the same year the Australian ophthalmologist Norman McAllister Gregg reported that infants with congenital cataracts and heart disease tended to have mothers with a history of RV infection during early pregnancy (Gregg 1941). Despite several years of scepticism against his findings, Gregg's observations were eventually confirmed by independent reports published in Australia (Pitt and Keir 1965), Sweden (Lundstorm 1962) and

the United States of America (Greenberg et al. 1957). These publications collectively established the role of RV in congenital cataracts, as well as the simultaneous association with heart disease and deafness in infants, and for the first time the associated effects of RV infection in infants were collectively termed congenital rubella syndrome (CRS). See Table 1.

Table 1: The main historical developments of Rubella virus.

1740

First clinical description of rubella by Friedrich Hoffmann

1881

Rubella officially recognised as a distinct disease at the International Congress on Medicine

1914

Alfred Fabian Hess proposed that rubella was caused by a virus

1938

Hiro and Tosaka successfully passed *Rubella virus* to children using filtered nasal washings, confirming Alfred Hess's findings

1941

Norman McAllister Gregg recognises the teratogenic effects of *Rubella virus*

1962

Rubella virus isolated in cell culture for the first time

Today, several viruses (Enterovirus, Adenovirus, Parvovirus B19 and Arbovirus) are known to cause rubella-like rashes and consequently, RV infections are often confused with the associated diseases, such as measles and dengue, if not examined using molecular diagnostics (Banatvala 2006). Due to the inability to distinguish rubella from these other infections, estimation of the prevalence of rubella and congenital rubella outbreaks prior to 1914 is not possible. However, major epidemics have been reported since the 1960s, both in developed and developing countries (Dudgeon 1975a; Cooper 1975; Donadio et al. 2003; Zheng et al. 2003; Wang et al. 2012). Since the first development of successful RV vaccines, indigenous RV

infection and CRS have been virtually eradicated in many developed countries around the world (Peltola et al. 2000; Song et al. 2012; Abernathy et al. 2013). However, this virus still continues to cause devastating epidemics throughout much of the world (Centers for Disease Control and Prevention 2013; Pham et al. 2013; Paradowska-Stankiewicz et al. 2013).

1.2.1 Structure and genome organization

RV is an enveloped virus with a ~9,762 nt positive-sense, single-stranded RNA genome which contains a 5'-methylated nucleotide cap and a 3'-polyadenylated tail and comprises two ORFs. The presence of a 5'-methylated nucleotide cap and a 3'-polyadenylated tail resembles cellular messenger RNA (mRNA) and allows RV genomes to be directly translated by the host enzymes. The 5' proximal ORF encodes the non-structural proteins (NSPs; P150 and P90) that function in RNA replication, whereas the 3' proximal ORF encodes the structural proteins (SPs; capsid protein, CP, and two envelope glycoproteins, E1 and E2) that together make up the virion (Figure 1). RV genomes also contain three untranslated regions (UTR's), which include 40 nt at the 5' end of the genome (5' UTR), ~118 nt between the SP and the NSP ORFs, and 59 nt at the 3' end of the genome (3' UTR). RV genomes also maintain the highest genomic GC content (~70%) of all known RNA viruses (Frey 1994).

The genomic RNA of RV serves as mRNA for the translation of the NSP, or as a template for anti-sense genomic RNA synthesis. The NSP in turn encode the viral proteins responsible for genome replication, by utilizing the cellular translational machinery. Embedded within the P150 gene is the methyl transferase- (involved in viral RNA capping) and cysteine protease domain (associated with the proteolytic cleavage of P200), a region of hyper-variability (2120–2440nt; contains higher than

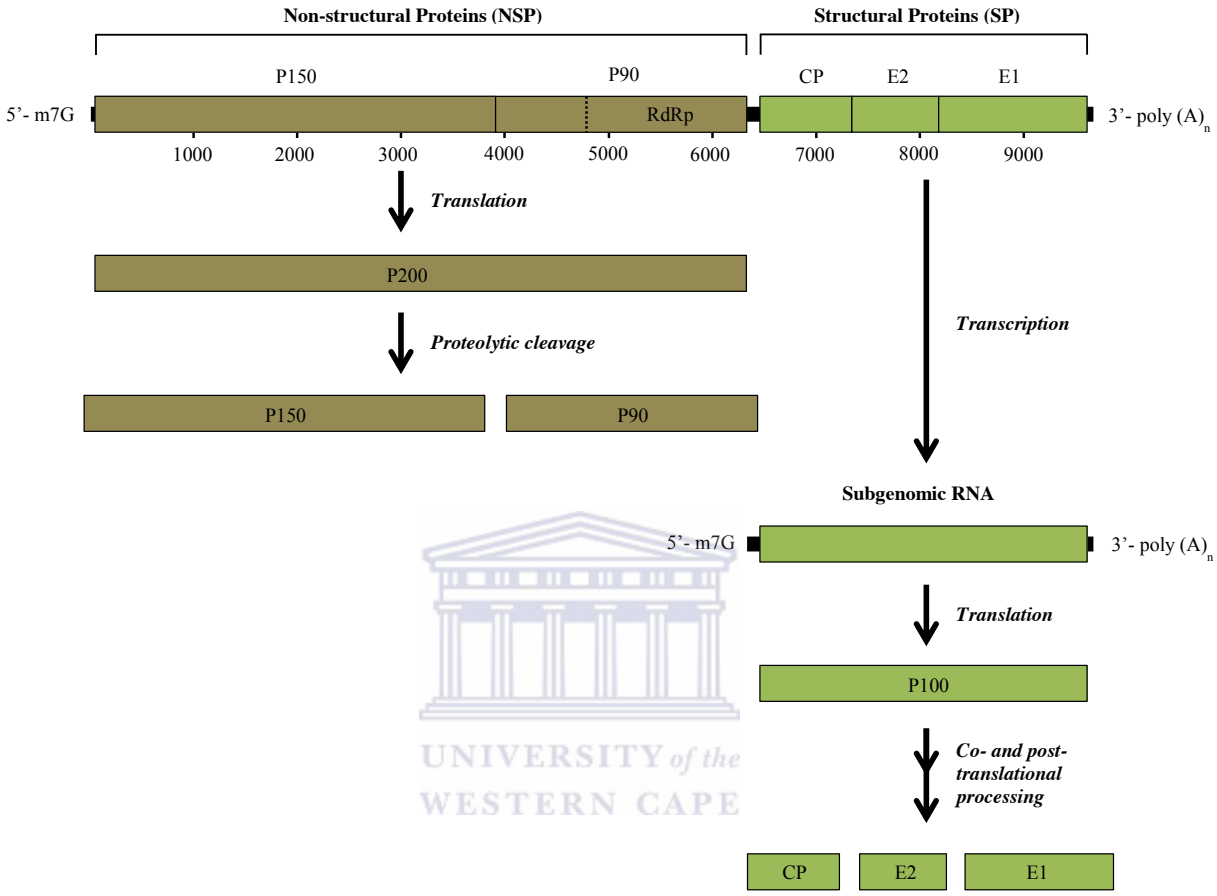


Figure 1. Genomic coding regions analysed. A schematic representation of the *Rubella virus* genome. The two ORFs, the non-structural (P150, P90; brown boxes) and structural polyproteins (CP, E2, E1; green boxes), are represented by 2 distinct boxes, and the UTRs as black horizontal lines. Boundaries of individual genes within the two coding regions are shown by solid vertical lines. The RdRp domain within the P90 gene is depicted by a vertical dotted-line. Specific coordinates of regions analysed: P150, 41–3943nt; RdRp, 4826–6388nt; CP, 6512–7411nt; E2, 7412–8257nt; E1, 8258–9703nt.

50% variability compared to other genes; Zhou et al. 2007) and the Q-domain (1491–2409nt; shares similar functions with the capsid protein, and is associated with RNA binding and interactions with cellular proteins; Tzeng and Frey 2009). Domains encoding the helicase and RNA-dependant RNA polymerase (RdRp) are located within the P90 gene (Frey 1994). It has been demonstrated that mutations within the NSPs result in an accumulation of anti-sense RNA, and a decrease in positive-sense RNA synthesis, indicating that translation of the NSPs are crucial for RNA replication (Liang and Gillam 2000).

The SP ORF is translated into a precursor polyprotein (P100) that is then cleaved into individual SPs. In contrast to *Alphavirus* capsid protein, which possess autoprotease activity, RV requires cellular signal peptidase for the capsid protein to be released from the P100 polyprotein. The principle antigenic components and neutralization domains are located on the E1 glycoprotein between 8900–9113nt (Katow and Sugiura 1985; Terry et al. 1988; Dominguez et al. 1990; Hobman et al. 1991; Chaye et al. 1992). The E1 glycoprotein also retains a putative neutralization domain (8882-8975nt) and various other antigenic sites involved in virus attachment and initiation of infection (Wolinsky et al. 1991).

RV virions are ~70 nm in diameter and typically consist of a lipid envelope containing the two viral glycoproteins, E1 and E2 and a nucleocapsid, containing the viral RNA and the capsid protein (Figure 2). The nucleocapsid core has a diameter of 30–35 nm with a T=4 icosahedral symmetry (Frey 1994; Liu et al. 1996) and comprises multiple copies of disulphide-linked homodimer capsid protein (M. Baron and Forsell 1991). The capsid protein is bound to the viral membrane by the C-termini and retains the putative signal peptide of the E2 glycoprotein. The N-termini, which is located within the viral envelope, contains a major RNA binding domain (6596–6680nt). This region is also involved in regulating subgenomic RNA

synthesis (Frey 1994; Liu et al. 1996). The viral lipids within the envelope are derived from the host-cell. The heavily glycosylated E1 and E2 glycoproteins are class 1 transmembrane proteins. Together they exist as heterodimers forming glycosylated spikes on the surface of the virion (Nakhasi et al. 2001). The crystal structure of the RV E1 glycoprotein is significantly different from homologous structures in *Alphavirus* and *Flavivirus* and it is thought that these differences likely originated as a result of several insertions within this RV gene region (DuBois et al. 2013). This is possibly due to stronger evolutionary constraints of viruses alternating between arthropod and vertebrate hosts, compared to RV which are exclusively transmitted between humans.

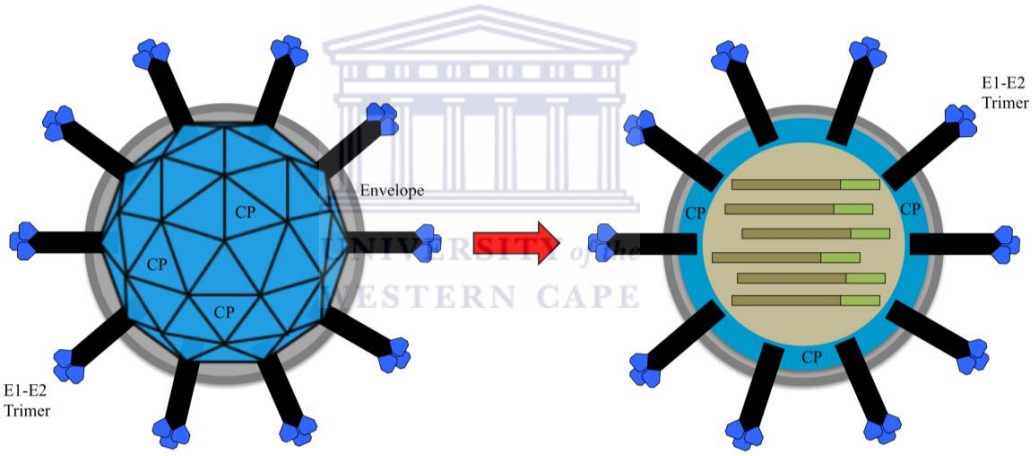


Figure 2. Schematic representation of the Rubella virus virion structure. The green and brown horizontal boxes within the *Rubella virus* particle represent the structural and non-structural polyproteins corresponding to Figure 1. The virion is ~70 nm in diameter, with the nucleocapsid core ~35 nm. It consists of a lipid envelope with two viral glycoproteins (E1 and E2), which exist as glycosylated transmembrane heterodimer spikes on the surface of the virion, and a T=4 icosahedral nucleocapsid contains the viral RNA and the capsid protein (CP).

1.2.2 Medical significance

RV is a disease causing agent that is associated with a nonthreatening self-limiting rash, low-grade fever and swelling of the lymph nodes (Ford et al. 1992). It is predominantly a childhood disease; however if contracted by a pregnant woman during early pregnancy, RV can be a powerful teratogenic agent, causing CRS (Wolinsky et al. 2001). Postnatally acquired RV symptoms among children are usually mild or absent, with most instances passing as subclinical or unrecognised occurrences. Adults however can develop malaise and fever associated with viremia (Figure 3), prior to the onset of a rash (Banatvala and Brown 2004).

Complications due to natural RV infection are rare and tend to occur more frequently in women than males and children (Atkinson et al. 2012). These include arthritis, encephalitis, hemorrhagic manifestations, orchitis, neuritis, and progressive panencephalitis.

While arthritis occurs in approximately 70% of adult women, the associated joint symptoms occur simultaneously with the onset of the rash and may persist for up to one month post infection. Encephalitis occurs in approximately 1 in 6000 cases, and haemorrhagic manifestations occur in around 1 in 3000 cases. These complications may last several days, but most individuals fully recover and chronic conditions are rare (Atkinson et al. 2012). Adverse effects to vaccination include symptoms such as acute- and chronic arthritis, swelling of the lymph nodes, neuropathies and thrombocytopenia and are usually mild and transient in nature (Ford et al. 1992).

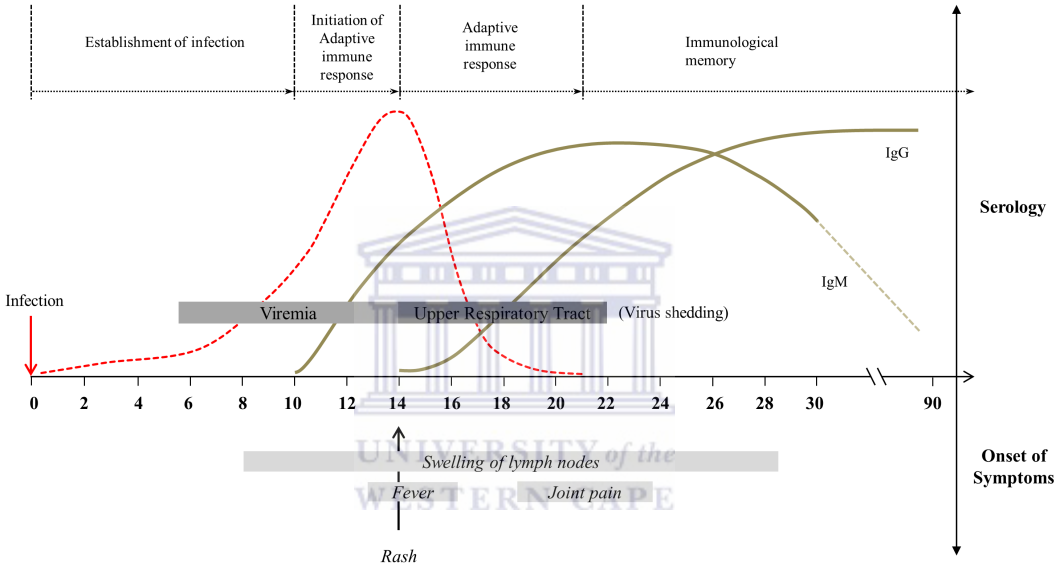


Figure 3. Immunological, virological and clinical feautures of a Rubella virus infection. A schematic representation of the immunological, virological and clinical feautures of *Rubella virus* infection. Primary infection and replication occurs in the nasopharynx, upper respiratory tract and the regional lymph nodes. Viraemia follows 5–7 days post-infection. The average time to the onset of symptoms is represented by light gray horisontal boxes below the x-axis. Serological anibody repsonses are represented by solid brown lines whereas virion concentration is represented by a dashed red line (adapted from Banatvala and Brown 2004).

1.2.3 Transmission and clinical diagnosis

RV is transmitted between hosts via large particle aerosols, which are secreted by the respiratory tract. Infection generally occurs in collectives of children, who may then infect their parents. Adult-to-adult infection is also more frequent among military recruits and on cruise ships (Ingalls et al. 1967) demonstrating that prolonged contact between hosts is necessary for RV to be successfully transmitted to susceptible individuals. Infants affected by CRS shed large quantities of RV from their body secretions for up to 1 year after birth, which could potentially result in RV transmission to susceptible adults caring for them (Atkinson et al. 2012). The average number of successful transmissions from a single case of rubella (basic reproductive number, R_0) in developed countries was estimated to be between 3–8 (Edmunds et al. 2000).

Primary implantation and replication occurs in the nasopharynx, upper respiratory tract and the regional lymph nodes and viraemia typically follows ~5–7 days after exposure, during which time transplacental infection of the fetus occurs (Heggie and Robbins 1969; Banatvala 2006; Atkinson et al. 2012). In the first week of exposure, rubella may be present atypically or with non-specific symptoms (Hemphill et al. 1988) and as a consequence, various other diseases can imitate RV infection, making accurate clinical diagnosis of rubella unreliable. In a study performed in the United Kingdom among children younger than five presenting a rash, only 3% of cases were positively confirmed as rubella (Ramsay et al. 2002) and in various tropical regions, *Alphavirus* and *Flavivirus* have been reported to cause rubella-like symptoms (Schmaljohn and McClain 1996).

Because of such difficulties distinguishing RV symptoms from various other diseases, only positive laboratory identification methods provide definitive means to achieve this. Reliable detection of an acute RV infection is achieved either by positive viral cultures, detection of RV by polymerase chain reaction (PCR), the presence of rubella-specific IgM antibodies, or a significant rise in IgG antibodies with paired acute- and convalescent-phase sera (Atkinson et al. 2012).

RV infection involves an initial latent period, with maximum virus production occurring 24–48 hours after infection (Hemphill et al. 1988), and is succeeded by an incubation period which persists for 14–21 days. During this time, swelling of the lymph nodes may occur, and in around two thirds of cases, individuals develop a rash which typically starts on the face and in the neck area. During the preceding 1–3 days, the rash continues to spread downward from the face and neck to the body and gradually begins to fade. Depending on the degree of skin pigmentation, the rash may be difficult to detect, but may be more prominent after hot showers or –baths. The rash is also only occasionally associated with an itch sensation, and the individual spots do not unite into larger bodies. Viraemia ends as humoral immune responses develop. However, RV may still be present in the pharynx and urine for up to 1–2 weeks (Reef and Plotkin 2013).

1.2.4 Congenital rubella syndrome and congenitally acquired Rubella virus infection

Internal organ development in a growing fetus occurs between 3–8 weeks (first trimester) past gestation. During this time, maternally acquired RV infection is likely to result in a generalized and persistent infection, leading to multiple defects that affect nearly all organs and as a result, may lead to foetal death, spontaneous abortion, or premature delivery (Atkinson et al. 2012). This is a result of the inability of the placental barrier to protect against vertical transmission, as well as the inability of the foetal defence mechanisms to launch an effective immune response (Banatvala 2006).

Although few animal models (Rayfield et al. 1986; Cusi et al. 1995) have been proposed to successfully study symptomatic RV infections, cell line studies suggests that a mechanism of RV-associated programmed cell death (Pugachev and Frey 1998a) and interaction between RV non-structural P90 and cellular proteins that regulate cell growth (cell-cycle regulatory *retinoblastoma protein*; cytokinesis

regulatory protein *citron-K kinase*) might be responsible for the teratogenicity (Atreya et al. 2004). It has however been shown that no CRS-specific mutations exist within RV genomes, and that samples from CRS patients do not form monophyletic clusters on phylogenetic trees (Katow 2004).

By the end of the first trimester, organogenesis is complete. During the second trimester, foetal humoral and cell-mediated immune responses gradually mature and passive transfer of maternal rubella-specific IgG occurs, consequently reducing the frequency and severity of congenital infection (Figure 4) and foetal damage (Miller et al. 1982; Banatvala 2006).

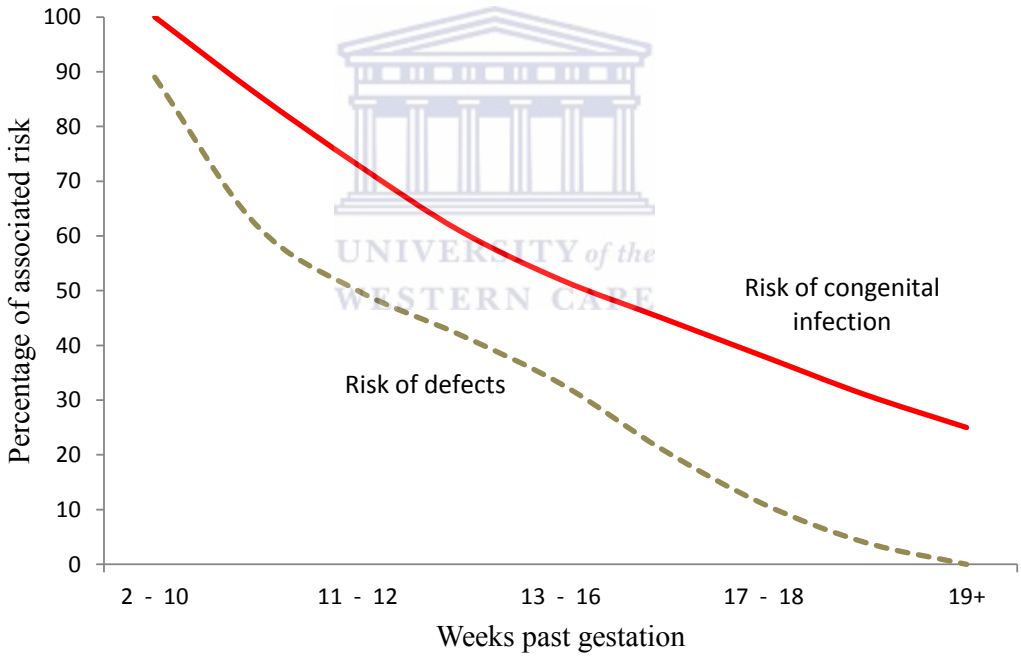


Figure 4. Associated risk of congenital rubella syndrome. The red solid line represents the risk of congenital *Rubella virus* infection in the weeks post gestation and the brown dashed line represents the risk of permanent foetal defects associated with congenital rubella syndrome (adapted from Miller et al. 1982).

The outcome and range of congenital infection abnormalities are largely associated with the gestational age at which the maternal rubella infection occurred. Deafness is the most common, and often the only, defect reported. It might however not become apparent for some time, but can persist indefinitely. Some clinical features (Table 2), including CRS anomalies include cataracts, glaucoma, retinopathy, cardiac defects, impaired foetal growth and mental retardation may show delayed onset until early adolescence or adulthood (Cutts et al. 1997, Atkinson et al. 2012).

Table 2. **Congenital rubella syndrome abnormalities, onset, and persistence of symptoms.** Adapted from (Dudgeon 1975b; Parkman 1996; Cutts et al. 1997).

Type of defect	Associated abnormalities	Time of symptom recognition	Transient features	Permanent features
General	Low birth weight	Neonatal	+	-
	Micrognathia	Neonatal	-	+
Ocular	Cataracts (unilateral/bilateral)	Infancy	-	+
	Microphtalmia	-	-	+
	Glaucoma	Infancy	-	+
	Pigmentary retinopathy	Infancy	-	+
Cardiovascular	Patent ductus arteriosis	Infancy	-	+
	Ventricular septal defect	Infancy	-	+
	Peripheral pulmonic artery stenosis	Infancy	-	+
	Myocarditis	-	+	-
Auditory	Sensorineural deafness	Infancy	-	+
	Deafness-associated speech defects	Infancy	-	+
Central nervous	Mental retardation	Infancy	-	+
system	Pyschomotor retardation	-	-	+
	Meningoencephalitis	Neonatal	+	-
	Progressive rubella panencephalitis	Neonatal	-	+
	Microcephaly	Neonatal	-	+

RV can be recovered from neonatal tear, nasopharynx, urine and stool samples, and continues to replicate in infant excretions, consequently infecting susceptible individuals. Children affected by CRS have also been shown to have a higher than expected incidence of autism (Atkinson et al. 2012). In addition to the apparent

abnormalities at birth or shortly thereafter, disease manifestations, such as diabetes and Dawsons disease, are generally delayed until early adolescence (Banatvala 2006).

1.2.5 Epidemiology

Phylogenetic studies have revealed that two major clades of RV exist with constituent members that differ from one another at between 8 and 10% of genomic sites. Whereas clade 1 consists of one provisional (1a) and nine recognised (1B, 1C, 1D, 1E, 1F, 1G, 1H, 1I, and 1J) RV genotypes, clade 2 contains three recognised (2A, 2B and 2C) genotypes (World Health Organization 2005; World Health Organization 2007; World Health Organization 2013a). Until the 2000s, clade 2 genotypes were restricted to Eurasia (Katow 2004; Zhou et al. 2007), however, genotype 2B viruses have subsequently become widely distributed geographically (Figure 5), and together with 1E and 1G, are the genotypes most frequently found among the more recently sampled isolates (Abernathy et al. 2011). Although globally there is only 1 serotype of RV (Zheng et al. 2003), it has been demonstrated that RV strains exist that differ in properties such as haemagglutination (Londesborough et al. 1995), plague morphology (Kouri et al. 1974), temperature sensitivity, virus yield and cell tropism (Chantler et al. 1993).

In 1941, widespread outbreaks of rubella were recorded amongst high concentrations of previously unexposed soldiers being mobilized to Sydney, Australia during the Second World War. Infection then spread to the general population when these troops returned home prior to serving overseas. This epidemic was particularly significant in that upon spreading to the general population, it yielded the first noted associations between RV and CRS (Gregg 1941).

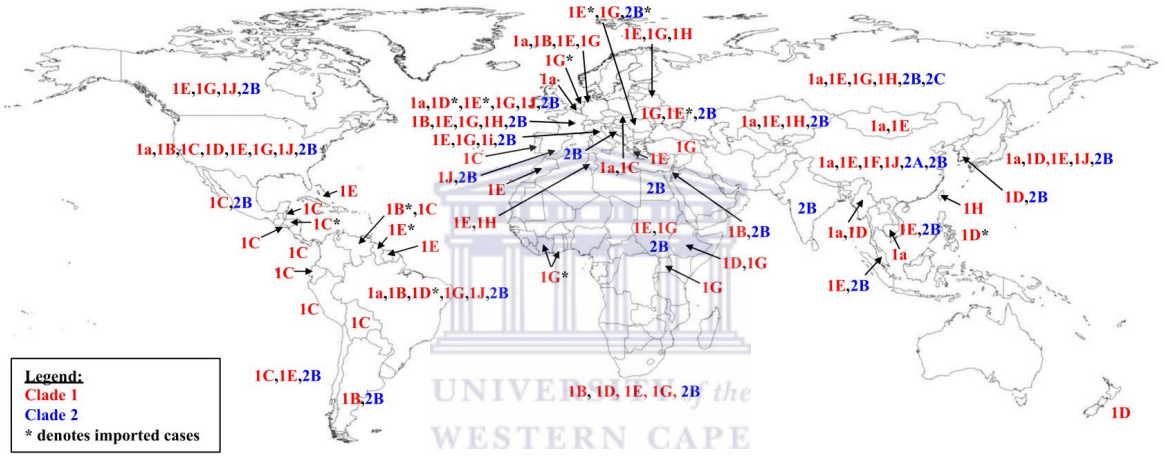


Figure 5. Global distribution of Rubella virus genotypes. Distribution of genotypes were mapped based on the documented sampling location (at the time of analysis) of publicly available sequences, as a means of supplementing the known genotypic geographical distribution (World Health Organization 2006; Abernathy et al. 2011). Names coloured in red and blue denote clade 1 and 2 genotypes, respectively.

In the spring of 1963, an outbreak of rubella began in Europe and later spread to North America (1964–1965) (Witte et al. 1969; Plotkin 2006) and Asia (1965–1969; Ueda 2009). During this global rubella pandemic (1962-1965), an estimated 12.5 million rubella cases were recorded in the United States alone, resulting in 11 250 foetal deaths, 2100 neonatal deaths and 20 000 infants born with CRS (Atkinson et al. 2012).

RV remains endemic throughout most of the world, even though comprehensive vaccine programs have been implemented in most developed countries. In these developed countries, vaccination programs have almost eradicated the disease, and more recent occurrences are mainly attributed to importations from non-vaccinated countries or countries with low-uptake rates. It is however of ongoing concern that in many developing countries, vaccination programmes are absent and consequently, the world's population is still being infected naturally (Wolinsky et al. 2001; Best et al. 2005; Centers for Disease Control and Prevention 2005a).

Epidemic cycles are highly variable between developed and developing countries, but tend to arise in the spring or early summer when the climate is temperate. In regions with low vaccination coverage, outbreaks typically recur with an average periodicity of between 5-9 years (World Health Organization 2011). The magnitude of outbreaks depends on a number of factors, including the number of susceptible individuals within a population, population densities, RV genotype distribution throughout the affected geographical area, socioeconomic factors and the quality of healthcare services available. Infection rates are also typically highest among individuals living in close proximity, such as student populations, cruise ship passengers and military establishments (Preblud and Alford 1983). Within non-vaccinated populations, CRS associated defects remain at least as high as for developed countries prior to the introduction of vaccination programs (Cutts et al. 1997). Because of the uneven adoption and coverage of rubella control programs among countries around the world, RV infections constitute a significant on-going global health threat.

Genotype diversity can be relatively heterogeneous in these populations (Donadio et al. 2003; Zheng et al. 2003), and presumably this pattern will be even more extreme if movements of people are unconstrained between neighbouring countries where rubella is endemic as the population may comprise many importations from independent epidemics. These relatively diverse RV populations are thought to exist in a source–sink relationship with other global RV populations with continuous low-level migration reseeding the regions of the world with high degrees of vaccine coverage where indigenous viruses no longer occur naturally (Tookey et al. 2000; Reef et al. 2002; Centers for Disease Control and Prevention 2005b).

In response to the ongoing global circulation of RV, the World Health Organization (WHO) has set a revised target for elimination of rubella by 2015, as well as the global eradication of rubella and CRS by 2020 (World Health Organization 2012).

1.2.6 Vaccines and vaccination strategies

During the rubella pandemic (1962–1965), it became increasingly apparent that the incidence of CRS was largely underestimated. This emphasized the need for an effective RV vaccine, and between 1965 and 1967, several RV vaccines were developed (Meyer et al. 1969; Prinzie et al. 1969; Plotkin et al. 1969). The first of these was licenced for commercial use during 1969-1970, and mass vaccination programs soon followed in several developed countries (Ueda 2009).

In the United States of America, three RV vaccines were licenced, including HPV-77 (attenuated in African green monkey kidney and later dog kidney cell cultures; Meyer et al. 1969), HPV-77 (attenuated in duck embryo cell cultures; Hilleman et al. 1969) and Cendehill (attenuated in rabbit kidney cell cultures; Prinzie et al. 1969) whereas the RA27/3 vaccine (attenuated in human diploid cells; Plotkin et al. 1969) was the only vaccine licenced in Europe. The initial vaccines licenced in Japan were the

Takahasi- (attenuated in rabbit kidney cell cultures) and Matsuura vaccines (attenuated in Japanese quail-embryo fibroblasts; Perkins 1985), however, five additional vaccines have been developed since then, including DCRB19, KRT, MEQ11, TO-336 and SK2. During 1980, the BRD-2 vaccine was developed in China (attenuated in human diploid cells; Zheng et al. 2003; Reef and Plotkin 2013).

In the United States of America, the HPV-77 (attenuated in duck embryo cell cultures) was widely distributed during 1969 and 1970 and it also formed part of the first measles-mumps-rubella-containing vaccine (MMR). This vaccine was shown to successfully protect around 65–94% of vaccinated individuals (Davis et al. 1971), however, comparative studies of HPV-77 and RA27/3 revealed that HPV-77 displayed lower antibody levels (Wallace and Isacson 1972), less persistent seropositivity (Balfour and Amren 1977), lower resistance to infection (Fogel et al. 1978), less herd immunity (Klock and Rachelefsky 1973) and a higher incidence of joint symptoms (Spruance and Smith 1971) compared to RA27/3. Additionally, RA27/3 can be administered intranasally (Ogra et al. 1971; Plotkin et al. 1973). In 1979, RA27/3 was licenced in the United States of America, and this resulted in the withdrawal of Cendehill from the American licensure and subsequently, HPV-77 was replaced by RA27/3 in a new MMR vaccine (MMR-II). In 2014, RA27/3 is the most widely used vaccine strain throughout the world, with the exception of countries such as China and Japan (Perkins 1985; Reef and Plotkin 2013).

Some countries initiated different vaccination strategies, most of which attained partial success. In the United Kingdom, Australia and Japan, adolescent girls were vaccinated (Dudgeon 1985; Cheffins et al. 1998; Ueda 2009) whereas the United States of America included RV-containing vaccines (RCVs) into their routine immunization schedule to vaccinate infants, in the hope of eventually depleting the reservoir of susceptible individuals (Preblud et al. 1980). In contrast, Iceland implemented serologic screening programs to identify and vaccinate only women detected as susceptible to RV (Farber and Finkelstein 1979).

Most of these strategies were eventually revised to include routine immunization of infants as well as targeted vaccination of women and adolescent girls (Watson et al. 1998; McLean et al. 2013). By maintaining high vaccination coverage, increasing monitoring systems, and introducing a second dose of MMR, many countries succeeded in eliminating indigenous RV (Peltola et al. 2000; Plotkin 2006; Song et al. 2012). By 2010, 131 of the 194 WHO Member States had adopted some form of prevention using either selective or universal mass immunization vaccination strategies, usually in response to national or regional rubella outbreaks (Banatvala and Brown 2004; World Health Organization 2012).

By 2013, most developed countries had included RCVs into their childhood immunization schedule, and as a result, countries in the WHO Pan American region have eliminated rubella and CRS (Castillo-Solórzano et al. 2011), whereas the European region has registered a 98% reduction between 2000 and 2009 (number of reporting member states increased from 41 to 46). In contrast, only a few developing countries have included RCVs in their schedule (Figure 6) and consequently the WHO eastern Mediterranean region, comprising no developed countries, only registered a 35% reduction in rubella and CRS cases between 2000 and 2009 (number of reporting member states increased from 11 to 15). The WHO African and southeast Asian regions registered a 20-fold (number of reporting member states increased from 7 to 38) and 14-fold (number of reporting member states increased from 3 to 9) increase in rubella cases over the same period (Reef et al. 2011).

Unsurprisingly, the incidence of CRS in regions that had not included RCVs into their childhood immunization schedule is much higher compared to regions that had introduced some form of mass vaccination (Cutts and Vynnycky 1999). In 1996, approximately 22,000 children were born with CRS in Africa, 46,000 children in South-East Asia and 12,634 children in the Western Pacific regions. Very few countries in these regions have introduced immunization vaccination strategies since then, and therefore, there is no reason to believe that the current burden of CRS

is likely to be different to the estimates for 1996, and could in fact, have increased since then (Cutts and Vynnycky 1999). Unfortunately the incidence of CRS in countries with either, limited vaccination programs, or poor vaccine-uptake, remains at least as high as that for developed countries prior to the introduction of vaccine programs (Cutts et al. 1997).

Figure 6. Countries presently including Rubella virus vaccination into their routine immunization schedules. Most developed countries have included *Rubella virus*-containing vacccines into their childhood immunization schedule. In contrast, only a few developing countries have included *Rubella virus*-containing vacccines in their schedule.

Source:*www.childinfo.org/files/immunization_summary_2012_en.pdf*

Table 3: The main epidemiological and vaccination developments in the history of Rubella virus.

1963–1965

Extensive European and North American epidemics, resulting in an estimated 12.5 million cases in USA alone

1965–1969

Extensive Asian epidemics

1969–1970

Attenuated *Rubella virus* vaccines licensed in the United States and the United Kingdom (USA universal childhood vaccination program; UK selective vaccination of prepubertal school girls)

1971

Measles-mumps-rubella-containing vaccine (MMR) vaccine was licensed in the USA

1976

First live attenuated *Rubella virus* vaccine was developed in Japan

1978–1983

Severe UK epidemics

1980

BRD-II vaccine was developed in China

1988

UK adapts vaccination policy to offer MMR to preschool children of both sexes

1989-1991

Resurgence of rubella in the USA

2010

131 (68%) of the World Health Organization states include *Rubella virus* vaccination in their national immunization schedules

2012

World Health Organization revised target for global eradication of *Rubella virus* to 2020

1.2.7 Economic impact

The cost associated with the rubella pandemic (1962–1965) in the United States of America was estimated at $840 million (Atkinson et al. 2012). If current estimates are accurate, then with 100,000 cases of rubella and CRS still being reported worldwide annually (World Health Organization 2013b), it is clear that the associated human and socio-economic costs remain extremely high and extract a heavy toll, particularly in developing countries where rubella remains endemic. Cost-benefit analysis studies (World Health Organization, Department of Vaccines and Biologicals 2000; Hinman et al. 2002; Bennett et al. 2002) that have been conducted, both in developed and

developing countries, to investigate the benefit of rubella elimination against the cost of vaccination, found a 13.3 benefit-to-cost ratio (Irons et al. 2000). All of these studies, except a study performed in Finland in 1979 (Elo 1979), have confirmed that the benefits of routine immunization and mass vaccination outweigh the costs, suggesting that rubella vaccination is medically and economically justifiable, especially when combined with the measles vaccine.

Different immunization delivery strategies have also been studied with sometimes contrasting results. For example, studies in Denmark (Bjerregaard 1991) and Israel (Golden and Shapiro 1984; Berger et al. 1990) estimated it to be more cost-beneficial to vaccinate infants and adolescent girls, whereas research performed in the United States of America reported that it was more cost-effective to immunize 12-year old girls (Stray-Pedersen 1982).

During 2012, CRS was estimated to cost between 4,200 and 57,000 US dollar per case annually in middle-income countries, and up to $140,000 over a lifetime in high-income countries (Babigumira et al. 2013). Nonetheless, even with the relatively low price of RV vaccines (especially when supplied by UNICEF or subsidised by GAVI) developing countries are still slow to add RV vaccines to their national immunization schedules. A recent study (Babigumira et al. 2013) suggested that this could be due to the fact that public health interventions in many of these developing countries might be more cost-effective. Additionally, when vaccine coverage (less then 80%) sufficiently decreases viral circulation in a population, there could be a shift in susceptibility from children to young mothers, and countries could be at risk of increasing the incidence of CRS (Schoub et al. 2009). By 2011, measles vaccination already formed part of the routine immunization schedule of all developing countries and substituting monovalent measles vaccines for measles-rubella-containing vaccine (MR) or MMR, could serve as a means of eliminating RV (World Health Organization 2011).

2. METHODS

2.1 Assembly of Rubella virus datasets

Sequences analysed in this thesis (sampled between 1961-2013) were retrieved from the NCBI GenBank (see Appendix 1 for Python script used to retrieve publically available sequences). Consequently, all sequence accession numbers used refer to those of the NCBI GenBank. Alignments of the RV datasets described below were performed using MUSCLE (Edgar 2004) and subsequently manually edited using MEGA v5.05 (Tamura et al. 2011).

Fourteen separate RV multiple sequence alignment datasets were analysed (see Table 4): (dataset i) a full genome dataset, containing a representative sample of RV genotypes, was generated to predict the presence of genome-wide nucleic acid secondary structural elements. At the time of the analysis, only 34 full genome sequences were available in GenBank, excluding vaccine strains and multiple sequences generated from particular isolates. The reason that only ten of the 34 available full genome sequences were selected for the prediction of genome-wide nucleic secondary structural elements was to reduce the computational burden imposed by the Nucleic Acid Structure Prediction (NASP) software (Semegni et al. 2011). These ten sequences were selected from distinct clades within a neighbour joining phylogenetic tree (calculated using MEGA v5.05; Tamura et al. 2011) after which the most divergent sequences within each of the selected clades were identified using pairwise genetic distances (calculated using SDT v1.0; Muhire et al. 2013; see Appendix 2).

Since recombination can have a pronounced undesirable effect on the accurate inference of phylogenetic trees (Schierup and Hein 2000; Posada and Crandall 2002), the estimation of precise nucleotide substitution rates (Martin et al. 2011) and the inference of positive selection (Anisimova et al. 2003), I opted to test the effect of recombination on my RV genome-wide nucleotide substitution rate estimates, by creating both (dataset ii) a *full genome recombination-included (RI)* dataset

containing 34 full genome sequences and (dataset iii) a *full genome recombination-free (RF)* dataset containing 32 full genome sequences from which the two sequences identified by the computer program RDP v4.17 (Martin et al. 2010) as having been derived through recombination were excluded.

Table 4. Summary description of the various datasets used in the thesis (also see Appendix 3).

Dataset	Description	Acronym	Number of sequences	Temporal range	Alignment length
i	Full genome, representative sample containing 10 *Rubella virus* genotypes (extracted from dataset ii)	-	10	1961-2008	9762 nt
ii	Full genome (not tested for recombination)	*Full Genome RI*	34	1961-2009	9762 nt
iii	Full genome (without 2 detected recombinant isolates)	*Full Genome RF*	32	1961-2009	9762 nt
iv	Capsid structural protein	*CP*	52	1961-2009	900 nt
v	RNA-dependent RNA polymerase	*RdRp*	56	1961-2009	672 nt
vi	Envelope glycoprotein 2	*E2*	54	1961-2009	846 nt
vii	P150 non-structural protein	*P150*	34	1961-2009	3943 nt
viii	Envelope glycoprotein 1	*E1*	640	1961-2012	739 nt
ix	*Unbiased* envelope glycoprotein 1, extracted from dataset ii	*Unbiased E1*	34	1961-2009	739 nt
x	Temporally balanced envelope glycoprotein 1	*Temporally Balanced E1*	45	1961-2012	739 nt
xi	Envelope glycoprotein 1, without 2 detected recombinant isolates and 437nt NASP predicted base-paired nucleotide sites	*E1 RF UnPR*	638	1961-2012	302 nt
xii	Envelope glycoprotein 1, without 2 detected recombinant isolates, containing only 437nt NASP predicted base-paired nucleotide sites	*E1 RF PR*	638	1961-2012	437 nt
xiii	Full genome, without 2 detected recombinant isolates and 1960nt NASP predicted base-paired nucleotide sites.	*Full Genome RF UnPR*	32	1961-2009	7802 nt
xiv	Full genome, without 2 detected recombinant isolates, containing only 1960nt NASP predicted base-paired nucleotide sites	*Full Genome RF PR*	32	1961-2009	1960 nt

For the NSP and SP datasets, the various gene regions were excised from the available 34 full genome sequences, and supplemented by additional publically available sequences from GenBank for the specific gene region of interest. The result being (dataset iv) a *Capsid* gene region dataset (CP) containing 52 sequences (dataset v) a *RNA-dependent RNA polymerase (RdRp)* gene region dataset containing 56 sequences. Only 672nt of the full 2445nt RdRp gene region was used for analyses, as some of the supplementary sequences did not contain the entire gene region. (dataset vi) an *E2* gene region dataset (E2) containing 54 sequences (vii) a *P150* gene region dataset containing 34 sequences, and (dataset viii) an *E1* gene region dataset (E1) containing 640 sequences. The 739nt used during analysis of the E1 gene region correspond to the RV genotyping window, and is therefore the most sampled dataset.

However, since only 5% of the sequences within the E1 gene region dataset were collected prior to 1990, it is likely that the estimated nucleotide substitution rates (the rate at which persistent mutations become fixed in a population) still comprise mutations which are yet to be purged from the sampled population by neutral genetic drift, consequently resulting in inflated nucleotide substitution rate estimates (Duffy et al. 2008). Similarly, the maintenance of nucleic acid secondary structures within single-stranded RNA molecules potentially also impose significant constraints on the evolutionary dynamics of the underlying nucleotide sequences, as the structures exist as meta-stable conformations. Thus, to test the effect of temporal bias within my E1 gene region dataset and investigate the constraints imposed on nucleotide substitution rate estimates by nucleic acid secondary structures, I created (dataset ix) an *unbiased E1* dataset containing only the E1 gene region, extracted from the 34 full genome recombination-included (RI) dataset sequences (dataset x) and *temporally balanced E1* datasets (see Appendix 4 for the Python script written to generate the temporally balanced datasets) containing 53 sequences. To generate the temporally balanced E1 datasets, I sorted the E1 gene region dataset sequences into their respective decades and a maximum of 13 sequences from each decade were randomly selected for analysis, as this was the actual number of sequences available from the 1960s. As

only two and three sequences were available from the 1970s and 1980s, respectively, I decided to include all of these sequences. This random selection process was repeated to generate 10 replicate datasets, each of which was analysed independently. Furthermore, I created (dataset xi) an E1 recombination-free dataset of 638 sequences with all nucleotide sites removed that were predicted to be base-paired within nucleic acid secondary structures identified by the computer program NASP (*E1 RF UnPR*), (dataset xii) an E1 recombination-free dataset of 638 sequences containing only sites that were predicted by NASP to be base-paired (*E1 RF PR*), (dataset xiii) a full genome recombination-free dataset of 32 sequences with all sites removed that were predicted to be base-paired within nucleic acid secondary structures (*Full Genome RF UnPR*) and (dataset xiv) a full genome recombination-free dataset of 32 sequences containing only sites that were predicted by NASP to be base-paired (*Full Genome RF PR*). See Figure 7 for the relationship between these datasets, as well as an analysis pipeline of the software and methods used during this thesis.

2.2 Evolutionary model selection

The best-fit nucleotide substitution model was estimated using model test as implemented in MEGA v5.05, and the degree of clock-like evolution evident within the analysed sequence datasets was evaluated using root-to-tip genetic distance vs. sampling date regression analyses as implemented in the computer program, Path-O-Gen v1.4 (available from http://tree.bio.ed.ac.uk/software/pathogen/; Drummond et al. 2003). Identification of the best-fit combined molecular clock and demographic model was determined using Bayes factor (BF) tests, calculated as the ratio of the marginal likelihoods of the alternative models as determined using the computer program Tracer v1.5 (Rambaut et al. 2009).

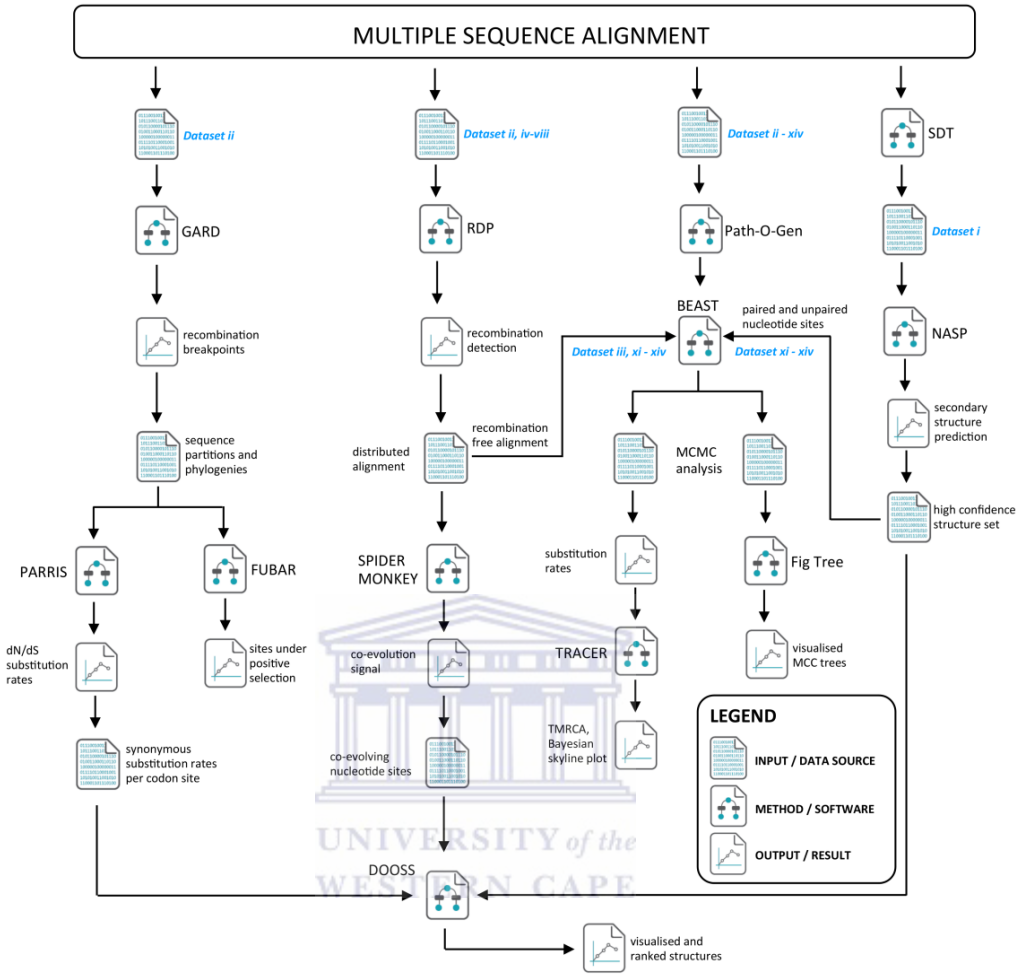


Figure 7. Graphical representation of the analysis pipeline. Sequence alignments and datasets were prepared using MEGA v5.05. SDT v1.0 was used to calculate pairwise genetic distances between sequences and the NASP method implemented to predicted potentially paired sites across the *Rubella virus* genome. GARD and RDP v4.17 was utilised to detected potential recombination breakpoints, which served as input for the selection analysis methods (PARRIS, FUBAR, and SPIDERMONKEY). Both PARRIS and FUBAR were used to determine synonymous substitution rates across the coding regions, whereas SPIDERMONKEY was used to detect sites which may be coevolving while still mainting complementary base-pairings. DOOSS v1.0 was used to rank and annotate the NASP predicted nucleic acid secondary structures from the FUBAR and SPIDERMONKEY results. BEAST v1.8 was implemented to estimate nucleotide substitution rates and the time to the most recent common ancestors (TMRCA) and the resulting trace files analysed in TRACER. Finally, the BEAST generated maximum clade credibility (MCC) tree files were summarised and annotated using FigTree.

2.3 Identification of nucleic acid secondary structures within Rubella virus genomes

Nucleic acid secondary structures are created through the formation of hydrogen bonds between complementary bases of a nucleotide sequence. Extensive nucleic acid secondary structures exists within the genomes of many mammalian and plant single-stranded RNA viruses (Simmonds et al. 2004) with the most biologically relevant structural elements displaying high degrees of conservation. Within viral RNA genomes such as RV, the collection of pairings between complementary bases (A-U, C-G and G-U; the latter being referred to as a weak 'wobble' pair) allows single-stranded RNA molecules to assume a meta-stable structural conformation.

RV genomes contain a number of known biologically functional genomic secondary structures (Dominguez et al. 1990; Nakhasi et al. 1994; Pogue et al. 1996; Chen and Frey 1999; Zheng et al. 2003). However, these genomes have an extremely high GC content and it is therefore likely that they may contain additional currently uncharacterised evolutionarily conserved structures that might in turn, constrain nucleotide substitution rates. Thus, to identify these evolutionarily conserved RV nucleic acid secondary structures, I implemented the computer program NASP (Semegni et al. 2011) using the default settings.

NASP utilised the hybrid-ss software (Markham and Zuker 2008) to predict groups of plausible secondary structural elements present within the ten RV full genome sequences that were previously identified from a neighbour joining phylogenetic tree (calculated using MEGA v5.05; Tamura et al. 2011) and pairwise genetic distances (calculated using SDT v1.0; Muhire et al. 2013) as reflecting the most representative sample of RV genotype diversity (dataset i; see Appendix 2 and 3). If it can be assumed that RV genomes have evolved to form meta-stable nucleic acid secondary structures, then randomly shuffling nucleotides within these genomes would influence their base pairing potential, resulting in higher minimum free energies (MFE) estimates. Thus, to establish support for the predicted structures, NASP implemented a series of randomised nucleotide-shuffling permutation tests to

determine which of the structures represent predicted folds associated with lower minimum free energies (MFE) estimates than could be accounted for by chance. These are collectively referred to as the high confidence structure set (HCSS).

To assess whether individual nucleotides predicted to be base-paired (within the high confidence structure set) were coevolving in a way consistent with selection favouring the maintenance of complementary base-pairing, I used a modification (Muhire et al. 2014) of the SPIDERMONKEY (Poon et al. 2008) method. Synonymous nucleotide substitution rates at the third codon position within coding regions were subsequently estimated using the maximum likelihood phylogenetic-based selection characterization methods PARRIS (Scheffler et al. 2006) and FUBAR (Murrell et al. 2013). Third codon positions were utilised by PARRIS and FUBAR as only 30% of nucleotide changes at this position result in non-synonymous substitutions, whereas nucleotide changes at the first and second codon position will result in non-synonymous substitutions 96% and 100% of the time, respectively.

The NASP predicted structural elements were visualised using DOOSS v1.0 (Golden and Martin 2013) and ranked in order of their likely biological functionality according to the: (i) associated degrees of conservation (determined by NASP); (ii) degrees of synonymous substitution rate reduction at codon sites containing paired nucleotides (determined by PARRIS; Scheffler et al. 2006); (iii) degree of complementary coevolution between nucleotides predicted to be base-paired, as determined by a SPIDERMONKEY-based method (Poon et al. 2008) described in Muhire et al. (2014).

I also tested for evidence of genome-wide associations between (i) base-paired nucleotides (within the high confidence structure set) and decreased synonymous substitution rates, and (ii) base-paired nucleotides (within the high confidence structure set) and nucleotide sites coevolving to maintain complementary base-pairings. The former was tested using a Mann Whitney U-test to compare median synonymous nucleotide substitution rate estimates (determined by PARRIS) at third codon positions between paired and unpaired sites whereas the latter employed a

Fishers exact test to find associations between nucleotide sites coevolving to maintain complementary base-pairings (determined by the SPIDERMONKEY-based method) and base-pairing between nucleotide site pairs (determined by NASP).

2.4 Detection of sporadic Rubella virus recombination

To account for the potentially confounding effects of recombination on RV nucleotide substitution rate estimates, I analysed the 34 sequence full-genome and 640 sequence E1 datasets (dataset ii and viii, respectively) for evidence of inter and intra-strain recombination using RDP v4.17. Using this program I was able to characterise probable recombination events, identify recombinants and likely parental sequences, and localize possible recombination breakpoints. Only potential recombination events detected by three or more out of the seven independent recombination detection methods implemented in RDP v4.17 were considered as genuine recombination events. The Genetic Algorithm for Recombination Detection (GARD; Kosakovsky Pond et al. 2006) was also used to detect recombination breakpoints during the previously mentioned PARRIS and FUBAR analyses (dataset ii). Similarly, the SPIDERMONKEY-based analyses was performed on the recombination-free datasets produced by RDP v4.17.

2.5 Positive selection analysis

Because positive selection results in the fixation of advantageous mutations at a faster rate than neutral mutations, it can have a pronounced undesirable effect on the accurate estimation of precise long-term nucleotide substitution rates. To test whether there is evidence for positive selection acting at codon positions within the RV genome, I analysed the full genome recombination-included (RI) dataset (dataset ii) using the fixed effects likelihood-based parametric selection inference method (FUBAR) implemented on the DATAMONKEY website (available from http://www.datamonkey.org/; Delport et al. 2010).

2.6 Bayesian phylogenetic analysis

A Bayesian Markov chain Monte Carlo (MCMC) method implemented in BEAST v.1.8.0 (Drummond and Rambaut 2007; Drummond et al. 2012) was used to estimate the overall nucleotide substitution rates (in contrast to the previously estimated synonymous nucleotide substitution rates exclusively at all third codon position) and the times to the most recent common ancestors (TMRCAs) for datasets ii-xiv (Table 4, Appendix 3). BEAST is a flexible probabilistic method for testing hypotheses and estimating evolutionary parameters, such as nucleotide substitution models, demographic models and molecular clocks, from an inferred posterior distribution of phylogenetic trees measured through time.

Four different evolutionary models were investigated, including either a non-parametric (Bayesian skygrid plot; BSP; Gill et al. 2013) or parametric (constant population size; Kingman 1982) demographic model together with a strict (a molecular clock model adopts uniform nucleotide substitution rates across all ancestral branches of a phylogenetic tree) or uncorrelated lognormal relaxed molecular clock model (a molecular clock model that permits rate variation among all branches of a phylogenetic tree, uncorrelated to the rate of the ancestral branch). For each dataset, between three and ten independent replicate runs of a Markov chain were performed using BEAST, ranging between 2.0×10^6 and 4.0×10^8 steps in length. All analyses were continued until the effective sample sizes (ESS) of all relevant model parameters were above 200: a criterion encompassing the number of uncorrelated parameter samples and the speed at which sampling occurred prior to convergence of the Markov chain Monte Carlo (MCMC) method to stationarity. When similar results were obtained from independent runs of the Markov chain, these were combined using LogCombiner v1.8.0, which is available in the BEAST package (Drummond and Rambaut 2007; Drummond et al. 2012).

Bayes factor tests, which compares the ratio of the marginal likelihoods between two independent models, were implemented to identify the best-fit clock and demographic model. Unlike other methods of testing (such as likelihood ratio tests and Akaike

Information Criterion), Bayes factor tests allow the comparison of non-nested models (non-parametric Bayesian skygrid plot vs. parametric constant population size demographic models).

2.7 Phylodynamics of Rubella virus

As nucleotide sequences contain a "molecular footprint" of historical adaptations and geographical spread (Holmes 2004), I decided to reconstruct the RV spatiotemporal history of the 640 sequence E1 gene region dataset using a symmetrical diffusion model (Lemey et al. 2009) and Bayesian MCMC method implemented in BEAST v1.8.0. The diffusion model considers the geographical spread among a finite number of discrete sampling locations and accounts for uncertainty both in the evolutionary relationships of the analysed sequences, and in the geographical locations of ancestral sequences. This is achieved by modelling the locations for taxa as continuous-time Markov chains (CTMCs). Using this method, it is possible to predict when, where and with what degree of certainty ancestral RVs most likely existed. Bayes factor (BF) tests with a cut-off of BF = 5.0 were performed, using the computer program, SPREAD v1.0.6 (available from http://www.phylogeography.org/SPREAD.html; Bielejec et al. 2011), to evaluate the relative degree of statistical support for inferred epidemiological linkages between sampling locations (BF > 100 represents decisive support; BF > 5.0 represents substantial support; BF < 5.0 represents negligible support). To further quantify the spatial spread and assess source-sink dynamics, "Markov jump counts" were implemented (Minin and Suchard 2008), which permits a measure of both the number of transitions among sampling locations (Markov jumps) and the waiting times between these transitions (Markov rewards).

To improve computational performance on the large E1 gene region dataset, I capitalised on the BEAGLE v2.1 (Ayres et al. 2012) high-performance phylogenetic library in conjunction with BEAST. In addition, by subdividing sequences into clusters of geographically proximate sampling locations, I was able to include all

sequence data while keeping the number of samples per location as balanced as possible. To optimally define groups of sequences displaying definite geographical clustering, I used the sampling geocoordinates and a hierarchical clustering method (called *hclust;* see Appendix 5) implemented in R (R Development Core Team 2008). The geocoordinates at the centroids of 11 discrete geographical clusters identified by this approach were used as the sampling locations for phylogeographic analyses performed in this thesis.

Tools available in SPREAD v1.0.6 were used to produce a graphical animation in .kml (key markup language) file format of the spatio-temporal movement dynamics of ancestral RV sequences. These .kml files contain information on statistically supported routes and times of virus movements as identified using Bayes factor tests (BF > 5), and can be viewed using Google Earth (available from http://earth.google.com).

By adopting a probabilistic model-based approach developed by Lemey *et al*. (2014), I also tested a range of pre-defined predictive variables and identified the key drivers of geographical RV spread. To achieve this, I employed a generalized linear model (GLM), which parameterizes rates of pairwise location movements as a log linear function (Lemey et al. 2014). This method uses nucleotide sequence data to determine the predictive variables (among those investigated) with the most explanatory power and estimate their effective contributions in explaining the inferred patterns of phylogeographic spread. In this regard, several predictive variables were considered, including log-transformed measures of geographical distance, demographic and economic data, the average number of years of education completed and location sample sizes. To account for temporal changes in predictor measurements between 1961-2013 (which currently might only be fully assessed using an extremely parameter-rich discrete epoch approach, Bielejec et al. 2014; not implemented in this thesis), I opted to average values for each sampling location across their entire range.

This could be done, since the relative differences remain similar over the time period, and all predictive variables are standardized (with a mean of 0 and a variance of 1) after log-transformation. Subsequently, these values were aggregated to obtain a single discrete value per geographical cluster.

The predictive variables of geographical RV spread implemented in the GLM include:

(i) **Geographical distance**. To assess whether geographical proximity could predict RV spread, I calculated great circle distances (as the crow flies) between the centroids of all pairs of geographical clusters, using a R-script (see Appendix 6).

(ii) **Level of education.** Education is a major component of social and economical well-being, irrespective of whether a country is classified as developed or developing. Consequently, the World Bank Education Index, calculated from the mean and expected years of schooling, was used as empirical data for including this predictor into the GLM (http://knoema.com/WBKEI2013/knowledge-economy-index-world-bank-2012?tsId=1017860).

(iii) **Vaccination coverage.** Data reported from the WHO Immunization Summary (http://www.childinfo.org/files/immunization_summary_2012_en.pdf) was used to determine whether RV-containing vaccines were included in the national childhood immunization schedule for each sampling location considered. For countries utilizing the combination MR or MMR, estimates for RV vaccination coverage were based on the WHO-UNICEF estimates (http://www.who.int/gho/immunization/measles/en/) of first dose measles-containing vaccine.

(iv) **Population size and density.** Estimates on population sizes and densities per sampling location, measured as "thousands" and "number of people per square metre", were obtained from the United Nations World Population Prospects.

(v) **Annual gross national income (GNI) per capita.** Estimates for the historical income level of each geographical cluster were obtained from the *World Bank*.

(vi) **Sample sizes.** To test whether sample sizes differences had any potential to bias my estimations, I considered sampling sizes at both the origin and destination of each geographical cluster separately.

3. RESULTS AND DISCUSSION

3.1 Biologically relevant nucleic acid secondary structures within Rubella virus genomes

NASP identified 661 potentially conserved nucleic acid secondary structural elements; 121 of which, account for >95% difference in the estimated minimum free energy between the actual sequences and the randomised versions of the sequences. Collectively, these formed the high confidence structure set (HCSS) upon which I focused further analyses. Furthermore, approximately 21% of the nucleotides within the 121 conserved structural elements of the HCSS were predicted to be base-paired (Figure 8 and Table 5).

Well-supported nucleic acid secondary structural elements within the HCSS were identified in both the NSP and SP coding gene regions, with the majority inferred to have occurred in the SP ORF (Figure 8; Table 5). All four of the previously characterised RV genomic structural elements (within RV coding regions) were within the top 20 of those highlighted in the DOOSS consensus ranking. In this ranking, structures (Appendix 7) are ordered according to their associated degrees of conservation, synonymous substitution rate reduction at codon sites containing paired nucleotides and the amount of evidence for complementary coevolution between nucleotides predicted to be base-paired (see Methods section). In the SP coding gene region, two well-characterized structural elements known to be involved in calreticulin binding (Chen and Frey 1999) were ranked first and seventh (Figure 9) and a structural element serving as a template for the sub-genomic RNA promoter on the negative-sense strand (Tzeng and Frey 2002) was ranked fourth. In the NSP gene coding region, a structural element promoting genomic positive strand synthesis (Pugachev and Frey 1998b), was ranked eighteenth (Table 5). Notably, whereas four

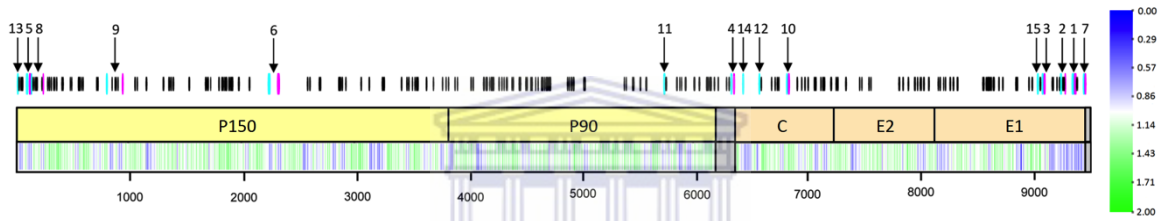


Figure 8. Genome-wide predicted high confident structure set and synonymous substitution rates. Pairs of vertical lines above the genome represent the base-paired nucleotide regions within the high confidence structure set (HCSS; Table 5) whereas the positions of the fifteen highest ranked structures are indicated by arrows (see Methods). Genome coordinates are displayed on the x-axis. The vertical lines below the gene map indicate site-to-site variation in synonymous nucleotide substitution rate estimates (see colour key on right). Blue and green coloured lines represent codon sites displaying elevated or reduced synonymous nucleotide substitution rates, respectively, relative to the mean.

of the top 10 ranked structures were situated within the E1 gene region (including the three highest ranked structures), none of the top 20 ranked structures were located in the E2 NSP region (Figure 8).

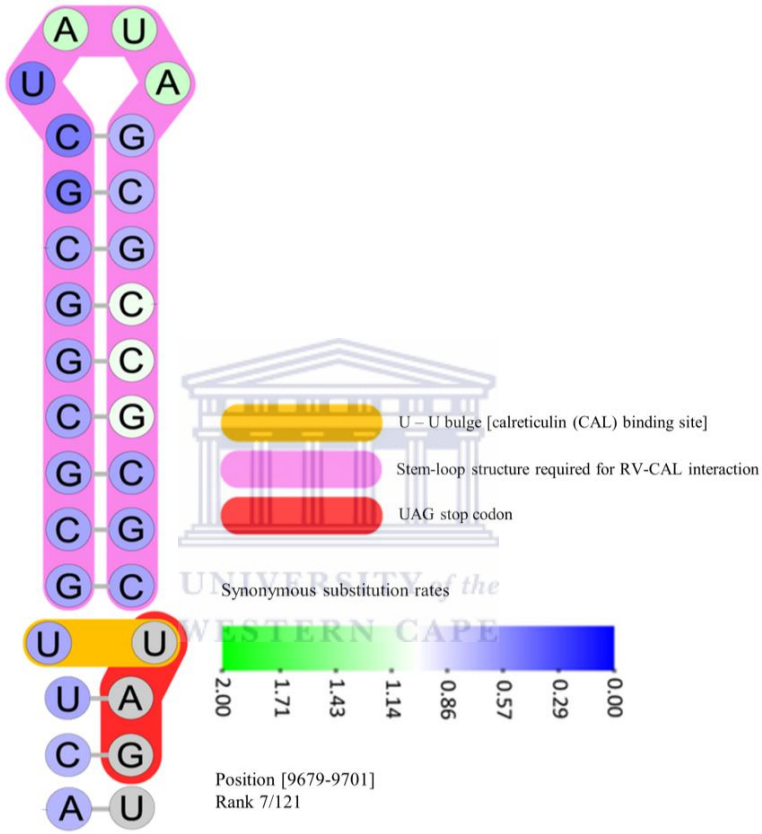


Figure 9. Example of NASP predicted nucleotide secondary structures of Rubella virus. This structure (labelled SL2) has been previously proposed (Chen and Frey 1999) to interact with human calreticulin (CAL). The rank refers to the DOOSS consensus rank of the specific predicted nucleotide secondary structure as it forms part of the high confidence structure set (HCSS; see Figure 8 and Table 5). Site-to-site variations in synonymous nucleotide substitution rates are indicated by colours ranging from blue to green (see colour key). Nucleotides falling outside the coding region are shaded in grey. The proposed CAL binding site (U-U bulge), is highlighted in orange, while the region critical for RV-CAL interaction and the stop codon are highlighted in purple and red, respectively.

Table 5. Consensus ranking of NASP predicted high confidence structure set (HCSS) structural elements across the entire *Rubella virus* genome.

Consensus Rank	Coordinates [1]	Structure length (nt)	Proposed biological function
1	9576-9614	38	Involved in CAL binding (E1)[2]
2	9469-9518	49	-
3	9304-9333	29	-
4	6483-6517	34	Subgenomic promoter (CP)[3]
5	96-136	40	-
6	2289-2390	101	-
7	9679-9701	22	Involved in CAL binding (E1)[2]
8	140-253	113	-
9	2321-2347	26	-
10	822-974	152	-
11	853-942	89	-
12	6845-6871	26	-
13	847-949	102	-
14	5875-5903	28	-
15	6596-6621	25	-
16	204-230	26	-
17	148-247	99	-
18	15-66	51	(5' UTR - P150)[4]
19	9477-9510	33	-
20	9261-9294	33	-
21	6329-6374	45	-
22	9614-9629	15	-
23	4999-5022	23	-
24	1929-1955	26	-
25	4261-4317	56	-
26	8229-8283	54	-
27	9064-9221	157	-
28	8004-8545	541	-
29	4378-4411	33	-
30	8146-8385	239	-
31	8173-8203	30	-
32	6260-6309	49	-

Table 5. continued.

Consensus Rank	Coordinates [1]	Structure length (nt)	Proposed biological function
33	153-190	37	-
34	8207-8366	159	-
35	286-309	23	-
36	1385-1412	27	-
37	7321-7337	16	-
38	7122-7159	37	-
39	7296-7320	24	-
40	3320-3345	25	-
41	565-2948	2383	-
42	8839-8868	29	-
43	8040-8097	57	-
44	6274-6297	23	-
45	9076-9191	115	-
46	7237-7259	22	-
47	6441-6472	31	-
48	7001-7022	21	-
49	4239-4330	91	-
50	4792-4815	23	-
51	8773-8822	49	-
52	1866-1884	18	-
53	416-433	17	-
54	25-56	31	-
55	6707-6746	39	-
56	165-182	17	-
57	4561-5165	604	-
58	194-239	45	-
59	3594-3622	28	-
60	4622-4641	19	-
61	1903-1970	67	-
62	3494-3637	143	-
63	898-914	16	-
64	731-749	18	-
65	4124-4151	27	-

Table 5. continued.

Consensus Rank	Coordinates [1]	Structure length (nt)	Proposed biological function
66	1078-1106	28	-
67	492-511	19	-
68	8409-8433	24	-
69	6757-6779	22	-
70	4496-4514	18	-
71	1839-1899	60	-
72	7732-7757	25	-
73	7387-7533	146	-
74	4445-4459	14	-
75	5590-5607	17	-
76	341-357	16	-
77	9437-9466	29	-
78	9489-9502	13	-
79	1423-1578	155	-
80	8830-8874	44	-
81	7081-7188	107	-
82	4822-4838	16	-
83	30-47	17	-
84	5989-6203	214	-
85	3642-3728	86	-
86	2642-2672	30	-
87	7434-7457	23	-
88	1718-1737	19	-
89	870-931	61	-
90	4711-4739	28	-
91	319-375	56	-
92	5664-5716	52	-
93	2748-2767	19	-
94	5658-5722	64	-
95	4573-5154	581	-
96	4667-4685	18	-
97	4771-4845	74	-
98	6067-6159	92	-

Table 5. continued.

Consensus Rank	Coordinates [1]	Structure length (nt)	Proposed biological function
99	3979-4008	29	-
100	4764-4852	88	-
101	8910-8953	43	-
102	3898-3927	29	-
103	1743-1770	27	-
104	5515-5538	23	-
105	5032-5063	31	-
106	3542-3567	25	-
107	3127-3210	83	-
108	8765-8828	63	-
109	3783-3874	91	-
110	1180-1344	164	-
111	595-614	19	-
112	8467-8506	39	-
113	8795-8813	18	-
114	9394-9410	16	-
115	2956-2997	41	-
116	4331-4431	100	-
117	4778-4820	42	-
118	7648-7674	26	-
119	2017-2124	107	-
120	6018-6039	21	-
121	4274-4298	24	-

[1] Coordinates in reference sequence [GenBank: JN635281].

[2] Chen, M, Frey, T. 1999. "Mutagenic analysis of the 3′ cis-acting elements of the rubella virus genome." *J Virol* 73:3386–403.

[3] Tzeng, W, Frey, T. 2002. "Mapping the rubella virus subgenomic promoter." *J Virol* 77:3189-201

[4] Pugachev, K, Frey, T. 1998. "Effects of defined mutations in the 5′ nontranslated region of rubella virus genomic RNA on virus viability and macromolecule synthesis." *J Virol*, 72:641–50.

3.2 Coevolution selection tests and synonymous nucleotide substitution rate estimates at base-paired vs. unpaired sites

Given the very high GC contents of RV genomes, it is expected that they will have a reasonably high degree of nucleic acid secondary structure irrespective of any potential roles on the biology of this virus. If most of the detected structural elements have no biological function, then there should be little evidence of natural selection operating to maintain these structures. If, however, base-paired nucleotides within structural elements are either evolving under stronger negative selection than unpaired sites (selection against change), or are co-evolving with their base-paired partners (i.e. they are evolving non-independently), this could plausibly have an effect on nucleotide substitution rate estimates.

To test this hypothesis I used the FUBAR (Murrell et al. 2013) and PARRIS (Scheffler et al. 2006) methods to estimate RV synonymous nucleotide substitution rates within the NSP and SP coding gene regions (see Figure 8). I specifically tested for evidence of selection against synonymous substitutions at codons containing paired nucleotides at their third positions (referred to as "paired codon sites"). Using a Mann-Whitney U-test, I compared median estimated nucleotide substitution rates at paired and unpaired codon sites. These tests revealed that both the NSP and SP coding gene regions displayed significantly lower synonymous nucleotide substitution rates at paired codon sites than at unpaired codon sites (PARRIS p-value = 2.288×10^{-2} and FUBAR p-value = 4.068×10^{-5} for the SP coding gene region; PARRIS p-value = 5.205×10^{-3} and FUBAR p-value = 1.118×10^{-6} for the NSP coding gene region).

To further test whether base-paired sites were co-evolving so as to maintain complementary nucleotide base-pairings, I used a SPIDERMONKEY-based method (Poon et al. 2008; Muhire et al. 2014). Consequently, I found a significant association between the NASP predicted base-paired nucleotide sites (within the high confidence structure set; HCSS) and genomic sites predicted to be coevolving with one another in a complementary fashion (using the SPIDERMONKEY-based method; p-value =

2.2×10^{-16}). Although this finding suggests that a large proportion of nucleotides within RV genomes are not independently evolving, it is not possible to quantify the ratio of sites co-evolving against those that are not, using this method.

These results then suggest that the 117 previously unreported structures predicted by NASP are likely biologically relevant as their constituent nucleotides are not evolving in a strictly neutral fashion. It is however not possible to determine, based on these analyses, which individual structural elements are biologically functional, as this would require further validation (using SHAPE analysis to confirm the existence of the NASP predicted nucleotide secondary structures; Wilkinson et al. 2006) and extensive molecular testing (using mutagenesis; Shepherd et al. 2005) which is beyond the scope of this thesis.

3.3 Recombination within Rubella virus genome sequences

Since recombination undermines the accuracy of phylogenetic inference (Schierup and Hein 2000; Posada and Crandall 2002), and some evidence of recombination has previously been reported in the RV sequences deposited in sequence databases (Zheng et al. 2003; Zhou et al. 2007; Abernathy et al. 2013), I opted to scan my datasets for evidence of recombinant sequences. Collectively, evidence of only two recombinant sequences were identified; [GenBank:JN635285; isolated in USA in 1988] and [GenBank:AF435866; isolated in Slovakia in 1974].

I detected a previously unreported intra-genotype (1a) recombination event within the P150 NSP gene region of isolate [GenBank:AF435866] (Figure 10A), which is currently provisionally classified as genotype 1a, and has not previously been investigated for evidence of recombination using a full genome sequence. It is noteworthy that the sequence for isolate [GenBank:AF435865, isolated in Germany in 1984], which was identified by RDP4.17 as the isolate that contributed a large nucleotide segment to the detected recombinant isolate sequence

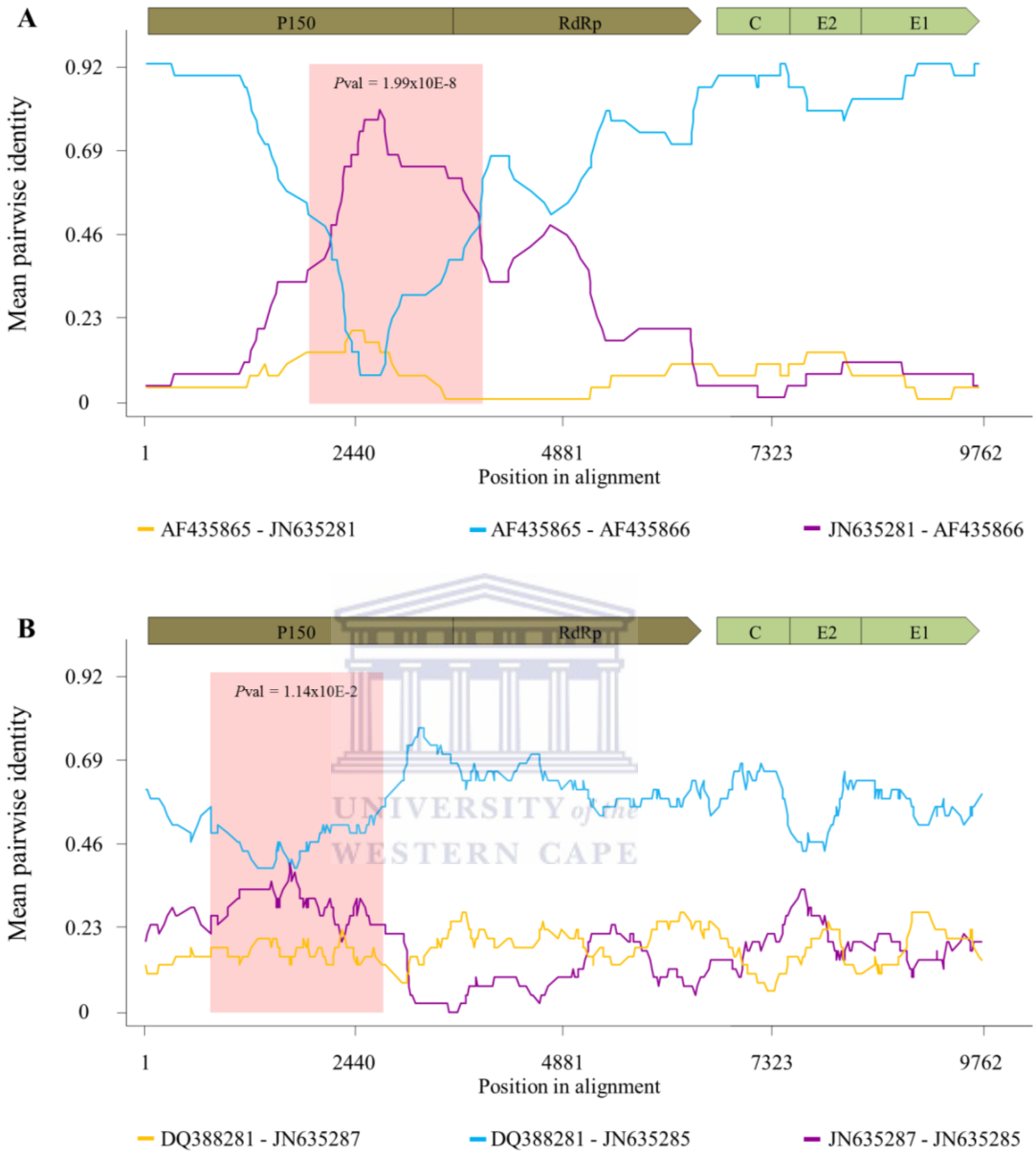


Figure 10. Pairwise identity plots of the potential recombination events detected in the 34 sequence full genome Rubella virus dataset. The non-structural and structural coding gene regions are shown above each plot in brown and green boxes, resprectively. The y-axis represents the mean pairwise identity between the compared sequences within a 30-nucleotide window, moved one nucleotide at a time along the length of the genome. Pairwise comparisons between the major (isolate contributing a larger segment of nucleotide sequence) and minor (isolate contributing a smaller segment of nucleotide sequence) parental sequences are shown in orange, between the major parental sequence and the detected recombinant sequences in blue, and between the minor parental sequence and the detected recombinant sequences, in purple. The area outlined in pink demarcates the boundries of the potential recombinant regions (*P* value < 0.05).

[GenBank:AF435866], was determined in the same laboratory (Hofmann et al. 2003) as [GenBank:AF435866] – a fact which suggests that [GenBank:AF435866] may be a laboratory artefact rather than a genuine natural recombinant (Han and Worobey 2011). I also detected significant evidence for an inter-genotype recombination event within the NSP P150 gene region of sequence [GenBank:JN635285] (Figure 10B), which is consistent with the results of Abernathy et al. (2013).

3.4 Positive selection within Rubella virus coding regions

In contrast to the results of a previous study (Hofmann et al. 2003), my analysis of selection pressures acting on individual codon sites using the FUBAR method found no significant evidence (highest posterior probability = 0.77 that dN/dS >1) of sites within the RV coding regions that were detectably evolving under positive selection pressure. Instead, around 91% of the NSP and 81% of the SP gene coding region sites, respectively, were inferred to be evolving under negative selection pressure with posterior probability estimates of greater then/equal to 0.9. This finding is consistent with results obtained by Zhou et al. (2007).

3.5 Temporal structure of Rubella virus genome sequences

The degree of clock-like evolution evident within the various sequence datasets was analysed using root-to-tip genetic distance versus sampling date regression analyses with the computer program, Path-O-Gen v1.4 (Drummond et al. 2003; Rambaut 2013). These analyses revealed high degrees of temporal structure in all datasets as evidenced by correlation coefficients ranging between 0.9 (for the full genome recombination-included (RI) dataset) and 0.67 (for the E1 dataset) [datasets ii and viii, respectively, see Methods section]. In the absence of pervasive recombination and positive selection, this indicated that all of the assembled datasets could be productively used to produce unbiased estimates of nucleotide substitution rates and times to the most recent common ancestors (TMRCAs).

3.6 Nucleotide substitution rates across the Rubella virus genome

Also consistent with previous studies (Zhou et al. 2007; Abernathy et al. 2013), the best fit nucleotide substitution models for the different RV datasets were TN93 with either a calculated proportion of invariant sites (I) or gamma distributed rate variation (G). For all analysed datasets (see Appendix 3) the uncorrelated lognormal relaxed-clock models had significantly higher likelihoods than the strict-clock models under both demographic models tested (constant population size, Bayesian skygrid plot). However, both demographic models fitted the data equally well.

Of the genomic regions analysed, the E1 SP-coding gene region displayed the highest estimated nucleotide substitution rate (1.19×10^{-3} substitutions/site/year; 95% highest posterior density (HPD) Bayesian confidence intervals = 1.04×10^{-3} – 1.35×10^{-3}), and the P150 NSP region the lowest (7.52×10^{-4} substitutions/site/year; 95% HPD = 5.85×10^{-4} – 9.26×10^{-4}; Figure 11A). All of these estimates, with the exception of the E1 gene region (dataset viii), had substantially overlapping 95% HPD's with the rates reported previously for RV by Jenkins et al. (2002). The E1 gene region nucleotide substitution rate estimate was however roughly twice as high as that previously estimated using a dataset of 50 sequences sampled between 1961 and 2001 (Jenkins et al. 2002). Nevertheless, all of my estimates were substantially lower than the rates reported for the E1 gene region within RV genotype 1E isolates sampled in China between 2001 and 2009 (Zhu et al. 2011).

Similar genome-wide nucleotide substitution rate estimates to those reported here have also been reported for *Chikungunya virus* (Cherian et al. 2009; Volk et al. 2010; Suwannakarn et al. 2011), another Togavirus in the genus *Alphavirus*, using the same

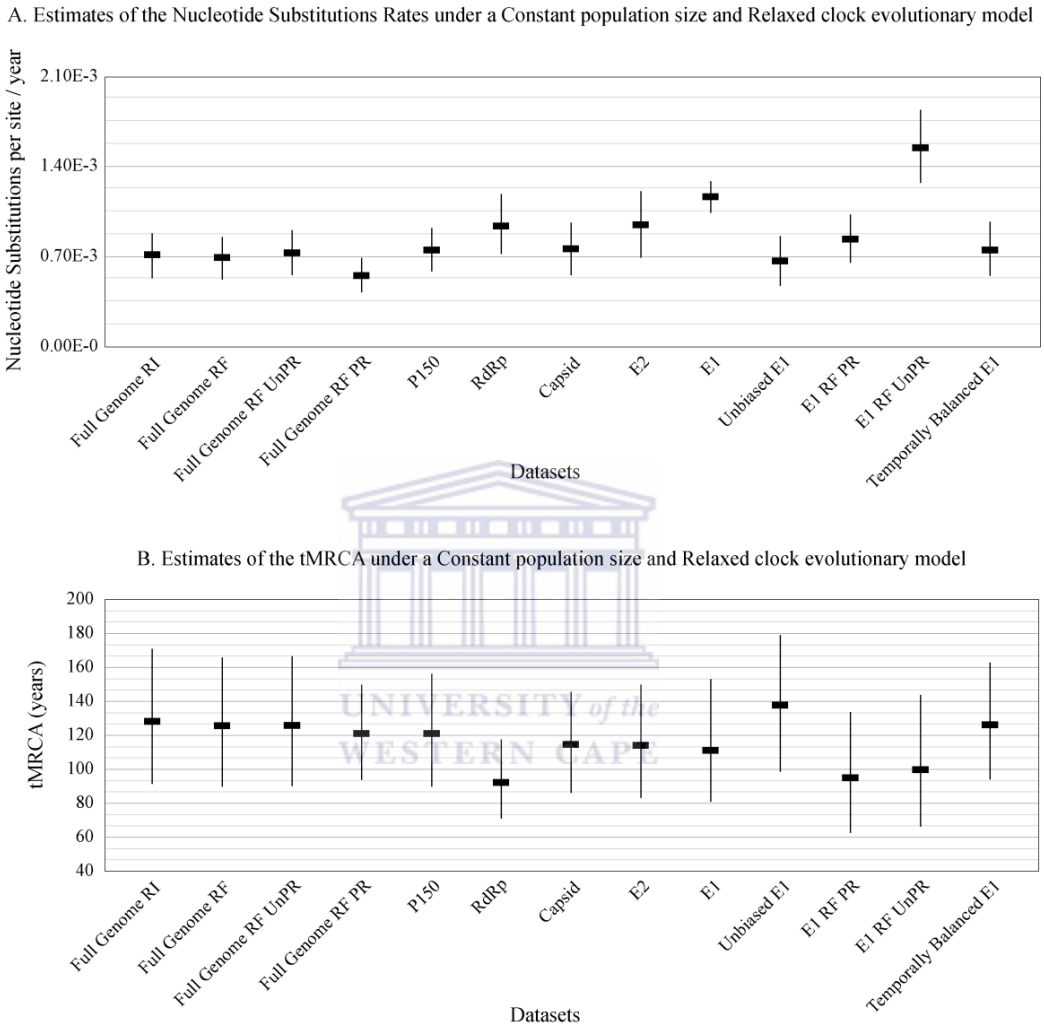


Figure 11. Nucleotide substitution rate and mean TMRCA estimates for the different Rubella virus sequence datasets. A. Nucleotide substitution rate estimates for the different *Rubella virus* datasets under the appropriate nucleotide substitution model run under a constant population size and relaxed-clock evolutionary model. **B**. Estimates of the time to the most recent common ancestor (TMRCA) for the different *Rubella virus* sequence datasets under a constant population size and an uncorrelated lognormal relaxed molecular clock model.

methods of estimation as those employed here. However, it is impossible to enumerate the proportion of the nucleotide changes present in my datasets that represent transient mutations that will ultimately be purged from the population by genetic drift (or weak purifying selection).

It is likely that, due to the inclusion of larger numbers of recently sampled E1 gene region sequences compared to those used to obtain the Jenkins et al. (2002) estimates (only 5% of the 640 samples considered in this thesis were collected prior to 1990), my nucleotide substitution rate estimates for this gene region are inflated and reflect a composite of the RV basal mutation rate (i.e. the rate at which all mutations occur) and its substitution rate (i.e. the rate at which persistent mutations become fixed in a population; Duffy et al. 2008).

To test whether my nucleotide substitution rate estimates for the E1 gene region were inflated due to the inclusion of larger numbers of recently sampled sequences, I analysed a dataset including only the E1 gene region extracted from the available 34 full genome sequences (dataset ix, see Methods section). I found that the estimated nucleotide substitution rate was similar to the rates inferred for the other RV genomic regions (see "Unbiased E1" in Figure 11A). Similarly, lower nucleotide substitution rates were also inferred from the analysis of the "temporally balanced" E1 dataset (dataset x) containing only a random subset of 53 E1 gene region sequences sampled between 1961 and 2012 (see "Temporally Balanced E1" in Figure 11A). These results therefore suggest that the nucleotide substitution rate estimates for the E1 gene region are strongly affected by the temporal imbalance within this dataset, and when this bias is taken into account, the nucleotide substitution rate estimates for the E1 gene region are similar to the remainder of the genome.

3.7 Estimated dates of the time to the most recent common ancestor of Rubella virus

Regardless of differences between the datasets with respect to estimated nucleotide substitution rates, the associated estimates of the mean TMRCAs for the different RV lineages analysed here all ranged between 1884 (95% HPD = 1841–1921) for the full genome recombination-included (RI) dataset and 1926 (95% HPD = 1904–1947) for the RdRp dataset (see Figure 11B and Figure 12). The mean TMRCA estimates for the E1 dataset with the various evolutionary models tested here were well within this range (between 1901; 95% HPD = 1858-1932) implying that sampling biases such as those evident in the E1 dataset have not had a particularly large impact on TMRCA estimates.

Regardless of the fact that the clade 2 genotypes were rendered paraphyletic by the genotype 2A and 2C isolates from China and Russia, respectively, the estimated ages for both the ancestral nodes of the clade 2 genotypes were older than that of the clade 1 genotypes, irrespective of the evolutionary model and dataset used, and were also positioned basally to the clade 1 genotypes, however with negligible posterior support (Figure 12). This result indicates that the RV isolates sample in this thesis likely originated in China/Russia, which is consistent with previous reports on the origin of RV (Katow 2004). Finally, it is important to stress that these estimates do not indicate the date when RV first emerged. They simply indicate when the most recent common ancestor (MRCA) of the RV genotypes analysed in this thesis likely existed.

3.8 The effects of recombination, selection and nucleic acid secondary structure on Rubella virus substitution rate estimates

To evaluate the potentially confounding effects of recombination and nucleic acid secondary structure on the estimation of nucleotide substitution rates, all the previously performed Bayesian phylogenetic analyses were repeated on the

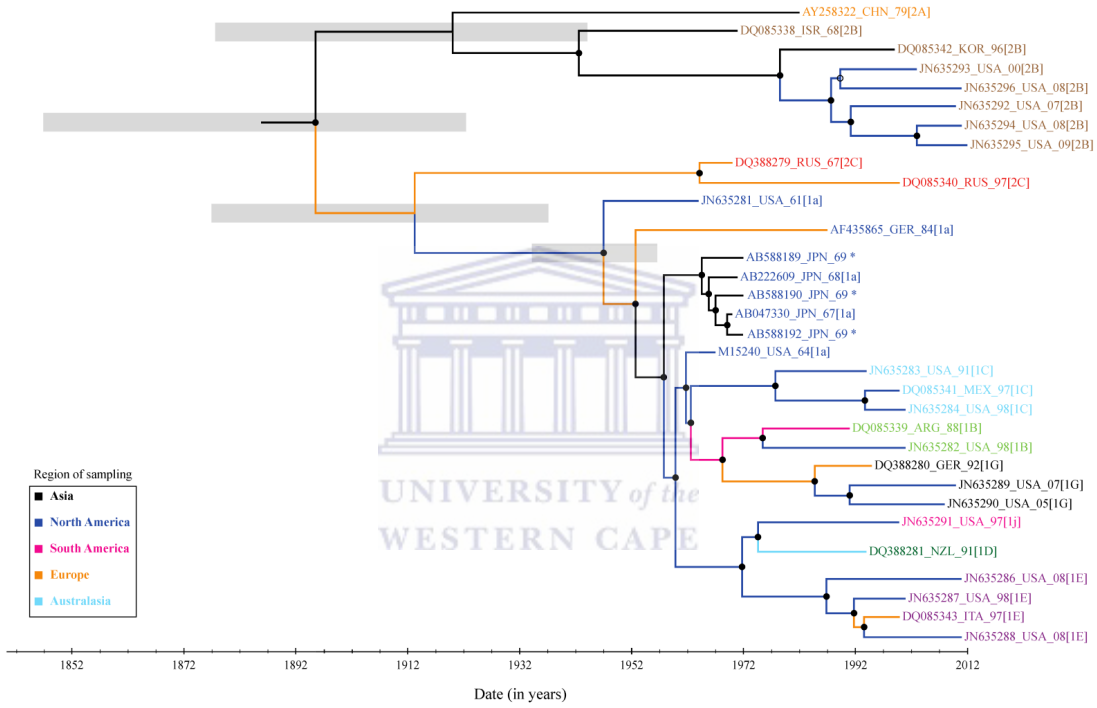


Figure 12. Maximum clade credibility tree for the full genome recombination-free dataset. Maximum clade credibility tree constructed from the 32 full genome recombination-free dataset under the TN93+G+I nucleotide substitution model and the Bayesian skyline plot, relaxed-clock evolutionary model. Branches are coloured according to the region of sampling and the taxon labels according to the genotype. Internal nodes with posterior support greater than 90% are indicated by a filled circle and greater than 80% by an open circle. Thick grey boxes at the root and most basal nodes of clade 1 and 2 genotypes, respectively, represent the range of 95% highest posterior density (HPD) confidence intervals of the time to the most recent common ancestor.

34 sequence full genome recombination-included (RI) dataset and 640 sequence E1 gene region datasets (dataset ii and viii, respectively), subsequent to removing the two detected recombinants and excising all NASP predicted base-paired nucleotide sites within nucleic acid secondary structures of the high confidence structure set (HCSS).The mean nucleotide substitution rate estimate for the 32 sequence full genome recombination-free (RF) dataset was similar to the rate inferred from the 34 sequence full genome recombination-included (RI) dataset (Figure 11A). Likewise, nucleotide substitution rate estimates remained unchanged after all NASP predicted base-paired nucleotide sites were excised from the full genome recombination-free (RF) dataset (compare "Full Genome RI", "Full Genome RF" and "Full Genome RF UnPR", respectively). However, when only the NASP predicted base-paired nucleotide sites within the full genome recombination-free (RF) dataset were considered, a substantially lower nucleotide substitution rate was inferred (compare "Full Genome RF UnPR" and "Full Genome RF PR").

Similar to results from the "Full Genome RF PR" dataset, reduced nucleotide substitution rates were inferred for the E1 gene region dataset after the two detected recombinant sequences were removed and only the NASP predicted base-paired nucleotide sites were considered ("E1 RF PR"). This suggests that the constraints imposed by the combined effects of recombination and nucleic acid secondary structures act to significantly reduce nucleotide substitution rate estimates.

3.9 A global view of Rubella virus geographical spread

In order to determine how RV attained its current global distribution, I analysed the 640 sequence E1 gene region dataset under a symmetric continuous-time Markov chain (CTMC) model. This method constructs a reversible diffusion rate matrix between the geographical sampling clusters (Figure 13; determined using hclust, see Methods section) and implements Euclidian distance between these clusters as a prior for inferring geographical spread.

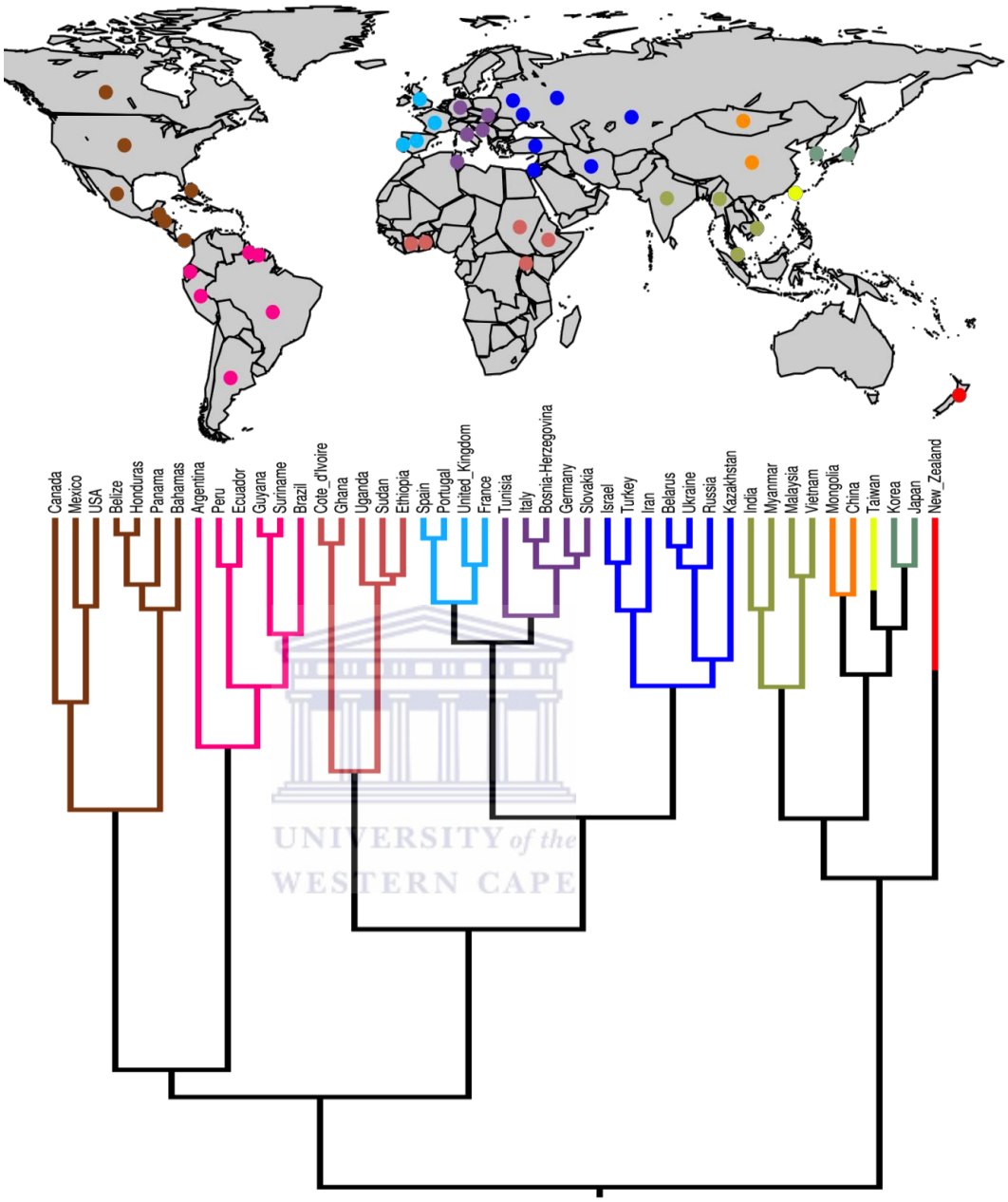


Figure 13. Hierachical clustering of 11 sampling regions. Hierachical clustering analysis was done using the R function *hclust* (Appendix 5, publicly available from various online repositotires) by implementing the *Complete-linkage clustering method*. Sampling locations were grouped based on the clustering analysis and geographical proximity. The 11 clustered regions were used for phylogeographic analysis by assigning each region as a single discrete trait. Clustering of sampling locations into 11 geographical regions, which resulted in the highest posterior support, are shown above. Each grouping is represented as a separate colour.

Forty-four statistically supported (Bayes factor > 5.0) geographical linkages were reported from this analysis (Figure 14), denoting that every inferred movement of RV between two statistically supported geographically clusters could be considered as reliable. Consequently, more than one instance of geographical spread could be inferred between two geographical clusters (Figure 15). Collectively, 62% of the inferred RV movements between the supported geographical linkages were intercontinental events. Unsurprisingly, the genotypes with the most widespread geographical distribution (World Health Organization 2007), namely genotype 1E (present in 22 countries over 5 continents) and 2B (present in 19 countries over 5 continents), were involved in the majority of these. Contrary to previous reports (Icenogle et al. 2011) that RV genotype 2B only started circulating endemically in the Americas during 2006-2007, my analysis inferred that genotype 2B was already introduced into North America from Western Europe sometime between 1981-1991.

My analysis also identified a statistically supported link relating to reports that the 1962–1965 global rubella pandemic originated in Europe and later spread to the USA (Witte et al. 1969). However, since RV genotype 1a isolates (GenBank:JN635218 and GenBank:L16233) have been sampled in the USA prior to the pandemic, it is apparent that genotype 1a (the genotype proposed to be largely responsible for the spread of RV from Europe into USA during the 1962–1965 pandemic) was already present in at least 2 continents prior to the pandemic (Frey et al. 1998). However, further elucidation of the epidemiology prior to, and during, the 1962–1965 global rubella pandemic is constrained by a lack of both sequence data from this period, and detailed epidemiological information relating to known historical RV geographical spread.

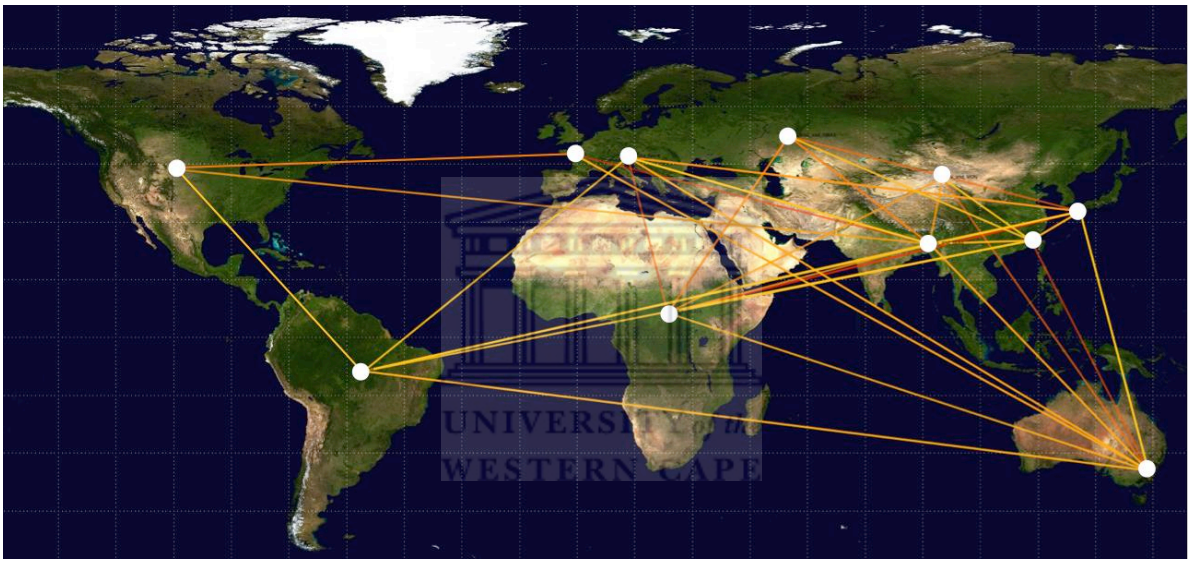


Figure 14. Bayes factor supported Rubella virus movement pathways. Phylogeographical analysis of the 11 geographical clusters (white dots, see Figure 13) revealed 44 statistically supported (Bayes factor > 5.0) viral movement pathways, indicated the connecting lines between white dots. Movement pathways are coloured with a gradient ranging from yellow (lowest support) to red (highest support).

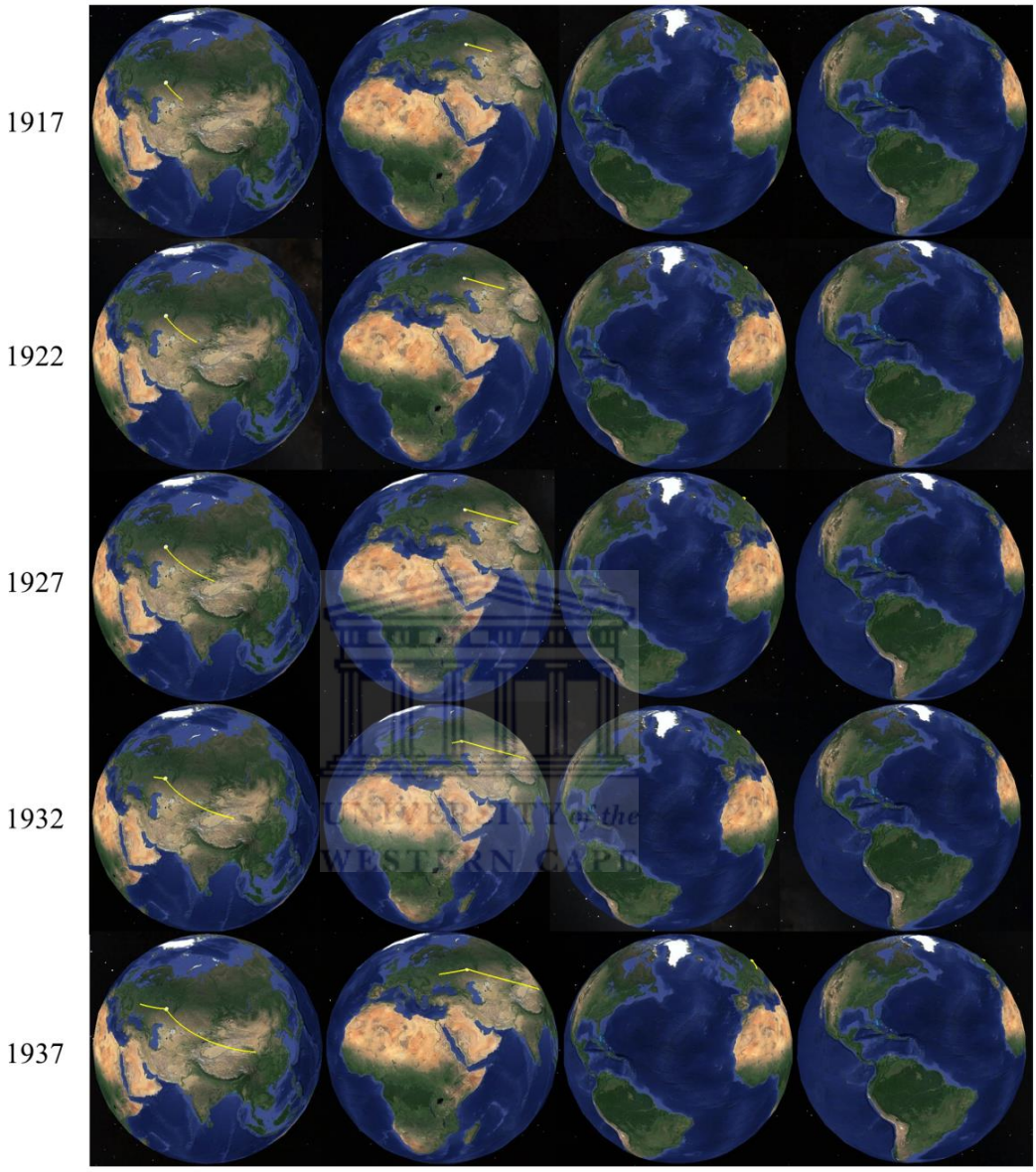


Figure 15. Geographical spread of Rubella virus. The spread of *Rubella virus* between geographically distinct regions are represented by yellow lines on the globes. The y-axis corresponds to the inferred date of geographical spread (in years), represented in 5-year intervals. For each inferred date, four separate images are shown, representing rotation of the earth at the specific time period. From left, the geographical regions shown are Asia, Africa & Europe, the Atlantic Ocean and the Americas.

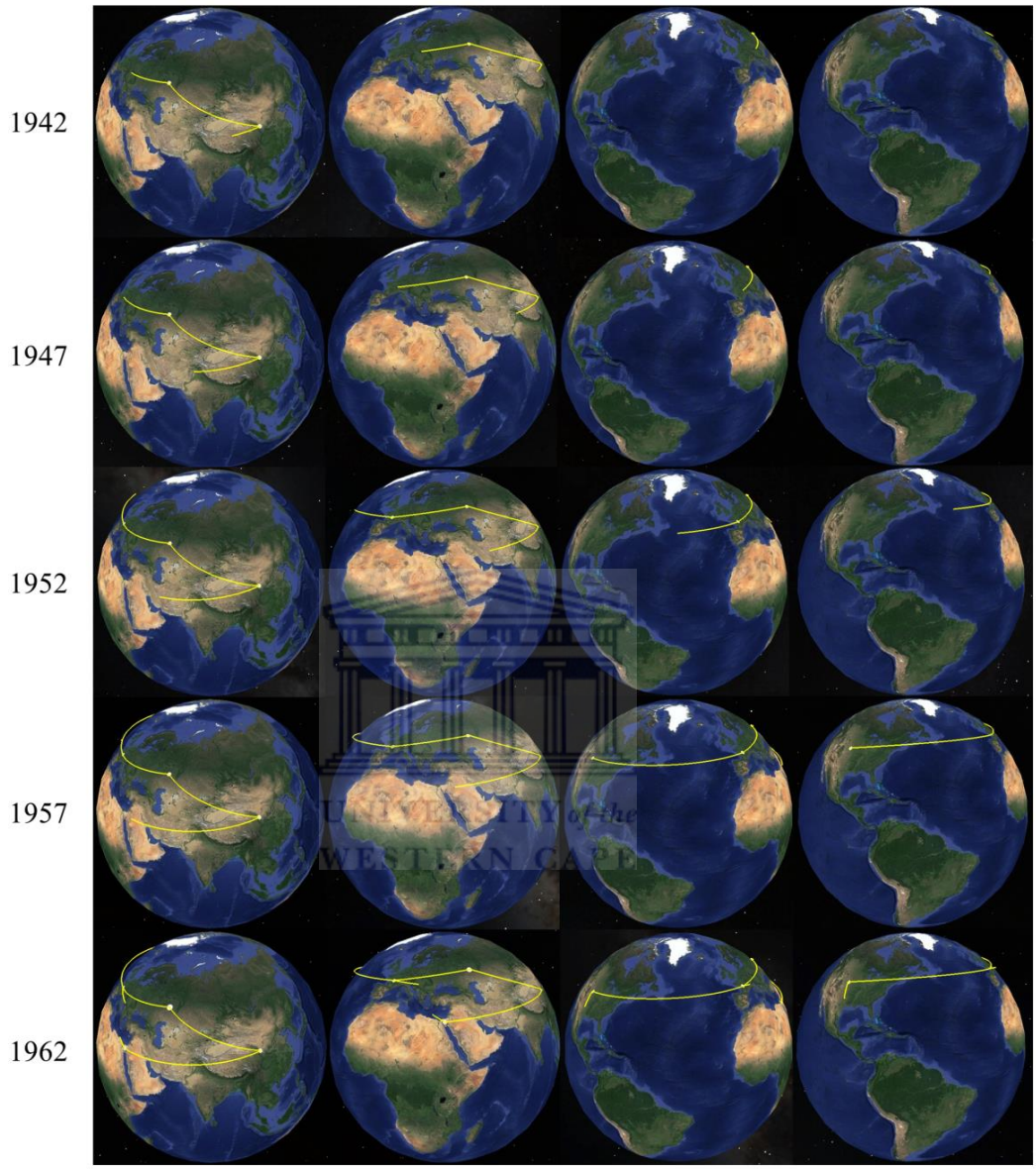


Figure 15. Geographical spread of Rubella virus. The spread of *Rubella virus* between geographically distinct regions are represented by yellow lines on the globes. The y-axis corresponds to the inferred date of geographical spread (in years), represented in 5-year intervals. For each inferred date, four separate images are shown, representing rotation of the earth at the specific time period. From left, the geographical regions shown are Asia, Africa & Europe, the Atlantic Ocean and the Americas.

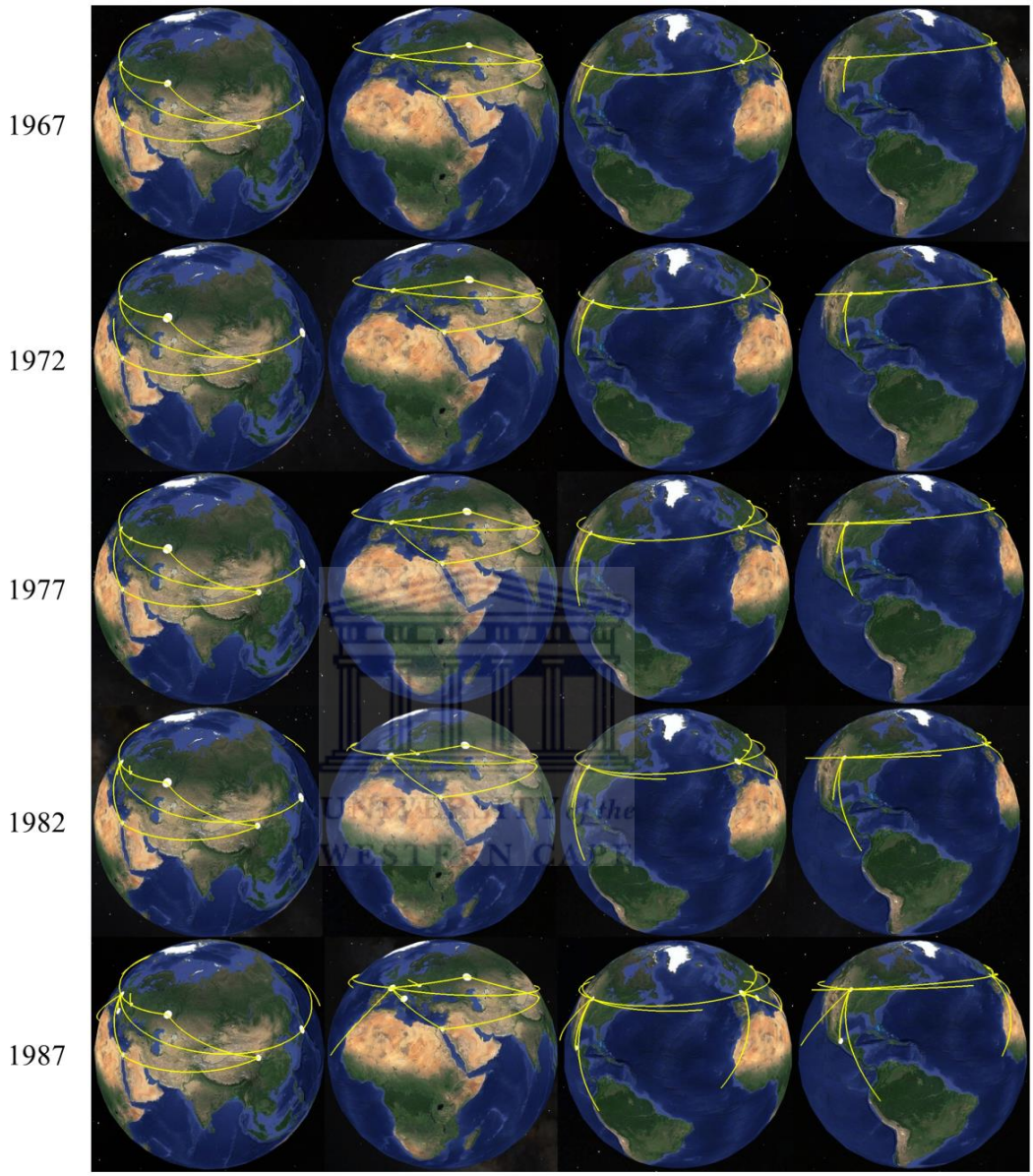


Figure 15. Geographical spread of Rubella virus. The spread of *Rubella virus* between geographically distinct regions are represented by yellow lines on the globes. The y-axis corresponds to the inferred date of geographical spread (in years), represented in 5-year intervals. For each inferred date, four separate images are shown, representing rotation of the earth at the specific time period. From left, the geographical regions shown are Asia, Africa & Europe, the Atlantic Ocean and the Americas.

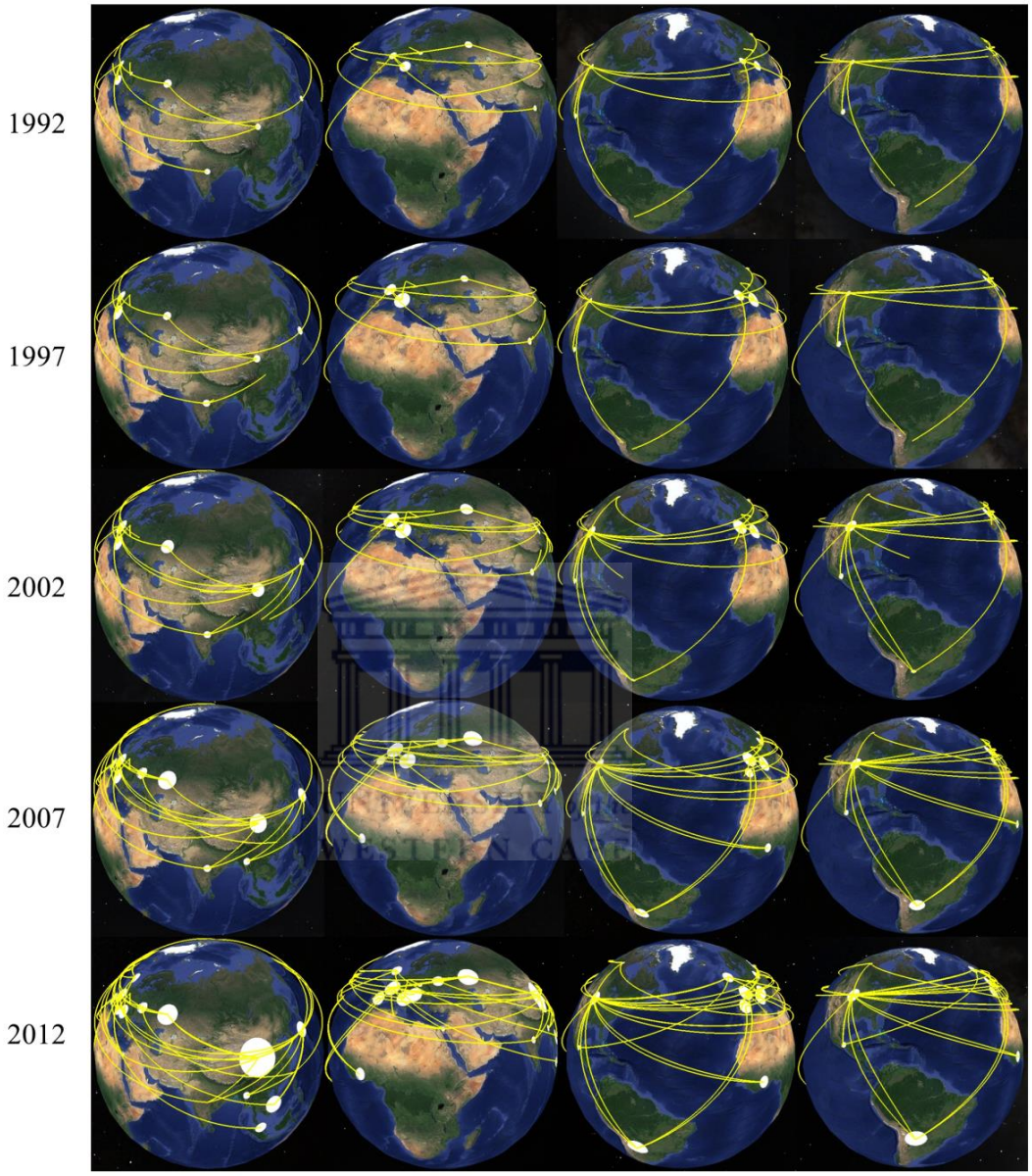


Figure 15. Geographical spread of Rubella virus. The spread of *Rubella virus* between geographically distinct regions are represented by yellow lines on the globes. The y-axis corresponds to the inferred date of geographical spread (in years), represented in 5-year intervals. For each inferred date, four separate images are shown, representing rotation of the earth at the specific time period. From left, the geographical regions shown are Asia, Africa & Europe, the Atlantic Ocean and the Americas.

3.10 The geographical origin of the most recent common Rubella virus ancestor

The most probable location of the MRCA of contemporary RV isolates was inferred to be the *Eastern Europe, Western & Northwest Asia* region (with 26.85% posterior probability support, Figure 16A). Notably, an Asian origin of the modern RV genotypes corresponds with previous reports of this virus (Katow 2004). It should however be stressed, that both the regions indicated for ancestral viruses in my analysis and movements inferred from these regions, are simply the most plausible given the regions from which RV E1 sequences were sampled. Such unavoidable sampling biases in my analysis mean that it is possible that the actual location of the MRCA might, for example, be in another region from which no samples are available.

It is also noteworthy that the time in the late 19th century (1898; 95% HPD = 1858–1932) when the most recent common ancestor of the sampled RV isolates is inferred to have most likely existed, is consistent with historical accounts of RV infections (Forbes 1969), and with the most recent common ancestor date estimates of the various other RV genomic regions analysed in this thesis (Figure 12B). It is however important to stress that these estimates do not specify the actual date when RV first emerged. It simply indicates when and with which degree of uncertainty the most recent common ancestor of the RV isolates analysed in this thesis likely existed.

To evaluate the degree to which sample size differences from the different geographical regions considered might have biased my estimates of the most probable ancestral location of the MRCA, I ran the same data under a tip-swap null model where the sampling locations of all the analysed sequences were randomized across the tips of the posterior distribution of trees. Consequently, by randomly selecting regions as the most probable ancestral location of the most recent common ancestor, you would expect approximately equal support for each region.

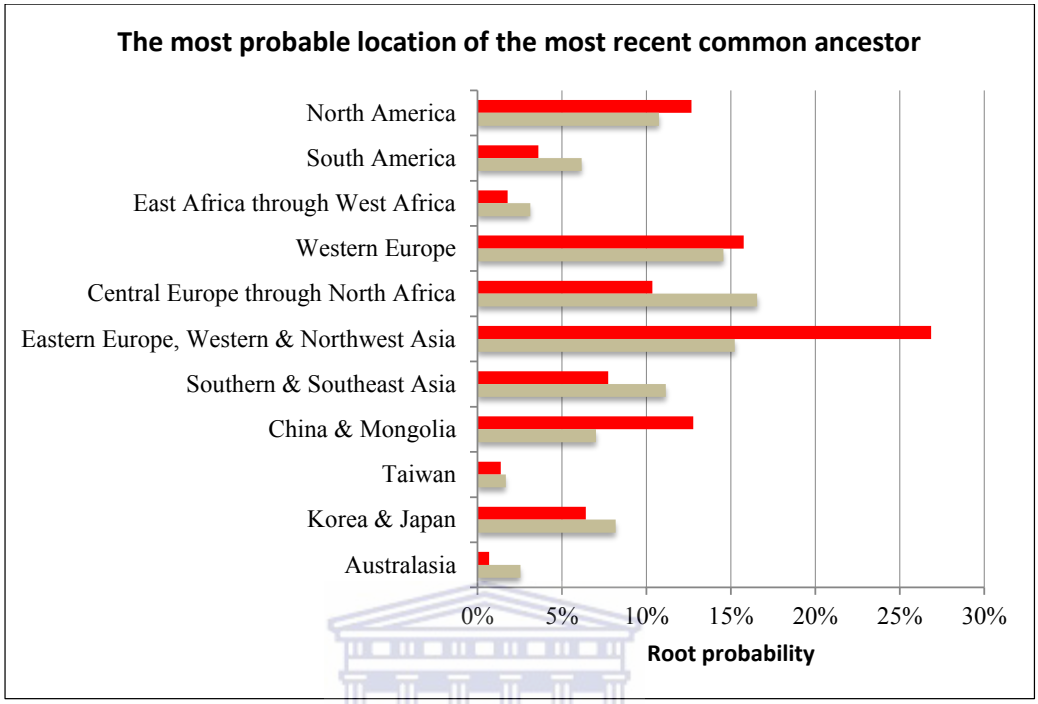


Figure 16. Most probable location of the most recent common ancestor of Rubella virus and tip-swap null model. The analysis indicated that Eastern Europe, Western & Northwest Asia was the most probable location of the most recent common ancestor of the sampled *Rubella virus* isolates (red lines). To evaluate the degree to which sample size variations biased my estimates, I ran the same data under a tip-swap null model (brown lines), where sampling locations of all the analysed sequences were randomized across the tips of the posterior distribution of trees. These results indicated that the phylogenetic signal present in the data is sufficiently strong to overcome any sample size biases that might exist.

However, if a sample size bias exists within the data, then by chance the regions with the highest sample sizes would be selected more often. However, as expected from this model, several regions (with almost equal probabilities) were inferred as the most probable location of the MRCA (Figure 16B), indicating that the phylogenetic signal in the data is sufficiently strong to overcome the systematic bias affecting these estimates, towards the regions with the largest sample size (caused by the unevenness in the sampling scheme with respect to geographical sampling locations).

3.11 A global view of Rubella virus spatial diffusion and phylodynamics

To test a range of variables that could potentially predict the geographical spread of RV, I employed a probabilistic generalized linear model (GLM) approach. This method was used to quantify the frequency that each predefined variable was utilized to predict geographical spread (*inclusion probability*) and estimate Bayes factor support for each variable as the ratio between the prior probability (defined before analysis as 0.052, an independent Bernoulli probability distributions reflecting a 50% probability that no predictive variable will be utilized) and the inclusion probabilities (BF > 10 represents decisive support; BF > 3.0 represents substantial support; BF < 3.0 represents negligible support). Additionally, the GLM analysis provided regression coefficients for each predictive variable (*predictor coefficient*), which quantified changes in the rate of RV transmission as a function of changes in the frequency that each variable was utilized to predict geographical spread.

This analysis revealed several predictive variables that were strongly associated with the geographical spread of RV (Figure 17). Of the predictive variables considered, education was estimated as the predictor best explaining geographical spread. This is reflected in the frequency (Bayes factor support = 20.69) at which this predictor was included in the GLM, and the contribution of this variable (predictor coefficient = -0.579) when included. This indicated that transmission rates were higher out of geographical regions with a lower education index (calculated from the mean years of schooling completed and the expected years of schooling in the respective region) compared to regions with a higher education index. Unsurprisingly, the rate of RV transmission was also higher amongst regions with low vaccination coverage compared to regions that maintained a high vaccination coverage, as is apparent by the statistical support (Bayes factor > 3) for the inclusion of this predictive variable, and the negative predictor coefficient.

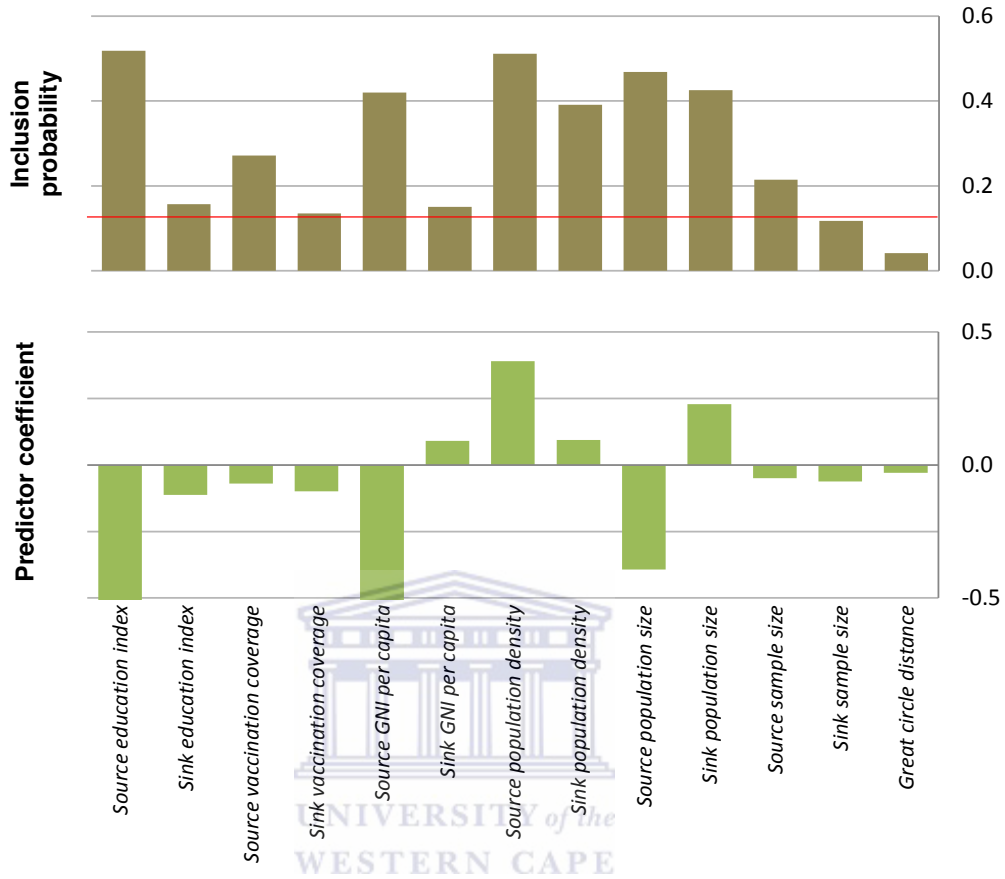


Figure 17. Generalized linear model results from predictive variables of Rubella virus spread. The inclusion probabilities indicate the frequency at which this variable was included in the generalized linear model, and the predictor coefficient the estimated size of the effect when included. The red line represents Bayes factor support equal to three.

The high inclusion probabilities of income (Bayes factor = 13.93 and 3.40) and education (Bayes factor = 20.69 and 3.58) provided additional support for vaccination coverage as a strong predictive variable, since it has been proposed that people with lower levels of education (defined by the education index) are less likely to vaccinate their children (Wright and Polack 2006), and generally also have a lower level of income (Ladd 2012). Altogether, the GLM analysis indicated that the geographical spread of RV tended to occur more frequently from regions with a low level of education and income into regions that maintain high levels of both education and

income. If these historical patterns of geographical spread inferred by the GLM remain unchanged, then these results suggests that, in addition to strengthened vaccination strategies, there also needs to be an increased effort to educate people about the effects of vaccination and the risks of RV infection, in the attempt to eradicate RV globally.

Interestingly, the analysis did not reveal any support for geographical proximity as a predictive variable of RV geographical spread (Bayes factor support < 1), instead transmission rates between the sampled geographical regions were inferred to be higher when considering regions with both high population sizes (Bayes factor support > 15) and densities (Bayes factor support > 12). This finding is not completely unexpected, since prolonged close proximity with an infected individual is often needed for transmission of RV (Ingalls et al. 1967).

To further evaluate the geographical spread of RV, I employed Markov Jump counts and rewards, which quantified the observed number of transitions between the sampled regions along the phylogenetic tree, and estimates the contribution of each region to the persistence of RV through time. This analysis inferred that Taiwan was the geographical region that donated and received the highest number of RVs, followed by North America, China & Mongolia, and Southern & Southeast Asia. However, since China & Mongolia and Taiwan comprise the highest number of samples within my 640 sequence E1 gene region dataset, respectively, it is likely that in relation to the other geographical regions, the inferred number of transitions to and from these regions would be higher. To test this hypothesis, I subsampled the 640 sequence E1 gene region dataset to contain an equal number of isolates per sampling region. Additionally, I also subsampled the 640 sequence E1 gene region dataset so that each region contained a sample size proportionate to the population size of that specific region. The results inferred from both of these subsampled datasets were congruent with those inferred before subsampling. This confirmed that estimates from both the GLM predictor and Markov Jump analysis are likely reliable, , and that sampling biases have not influenced these results.

4. CONCLUSION

My analysis identified 117 previously unreported nucleotide secondary structures, most of which were inferred to likely be biologically functional. Consistent with the results of previous studies, I have shown that evidence for recombination (Zheng et al. 2003; Zhou et al. 2007; Abernathy et al. 2013) and positive selection (Hofmann et al. 2003) is sparse. However, my results indicated that nucleotides in RV genomes are likely not evolving in a strictly neutral fashion, as base-paired nucleotides involved in the formation of nucleotide secondary structures displayed significantly lower nucleotides substitution rate estimates compare to nucleotides that were unpaired. Similarly, I demonstrated that temporally biased sampling in RV gene regions, such as the E1 structural glycoprotein, resulted in higher mean nucleotide substitution rate estimates. Fortunately, such biases had a negligibly negative impact on the utility of E1 gene region sequences for dating ancestral RV sequences under uncorrelated lognormal relaxed-clock evolutionary models. The inferred nucleotide substitution rate estimates were also sufficiently high, indicating that RV E1 gene region sequences (the most frequently sampled RV genome region) contained sufficient phylogenetic signal to be appropriate for sequence-based inferences of RV demographic and movement dynamics.

As a result, I was able to identify several geographical regions that acted as transmission hotspots for the geographical spread of RV. Furthermore, my analyses indicated that a lower level of education, rather then vaccination coverage, largely drove the spread of RV on an intra- and intercontinental scale. Finally, this thesis hopes to inform future policy makers about the past demographic patterns of RV, enabling them to make well-informed decisions in the effort to eradicate RV by 2020.

5. REFERENCES

Abernathy, ES, MH Chen, J Bera, S Shrivastava, E Kirkness, Q Zheng, W Bellini, and J Icenogle. 2013. "Analysis of Whole Genome Sequences of 16 Strains of Rubella Virus from the United States, 1961–2009." *Virol J* 10 (1): 1–9. doi:10.1186/1743-422X-10-32.

Abernathy, ES, JM Hübschen, CP Muller, L Jin, D Brown, K Komase, Y Mori, et al. 2011. "Status of Global Virologic Surveillance for Rubella Viruses." *J Infect Dis* 204 (Suppl 1): S524–S532. doi:10.1093/infdis/jir099.

Ackerknecht, EH. 1982. "Medicine in the Eighteenth Century." In *A Short History of Medicine*, 128–44. Baltimore: Johns Hopkins University Press.

Anisimova, M, R Nielsen, and Z Yang. 2003. "Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites." *Genetics* 164 (3): 1229–36.

Atkinson, W, S Wolfe, and J Hamborsky. 2012. "Chapter 19: Rubella." In *Epidemiology and Prevention of Vaccine-Preventable Diseases*, 12th Edition. Vol. Second Printing. Washington DC: Public Health Foundation: Centers for Disease Control and Prevention.

Atreya, CD, KV Mohan, and S Kulkarni. 2004. "Rubella Virus and Birth Defects: Molecular Insights into the Viral Teratogenesis at the Cellular Level." *Birth Defects Res A Clin Mol Teratol* 70 (7): 431–37.

Ayres, DL, A Darling, DJ Zwickl, P Beerli, MT Holder, PO Lewis, JP Huelsenbeck, et al. 2012. "BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics." *Syst Biol* 61 (1): 170–73. doi:10.1093/sysbio/syr100.

Babigumira, JB, I Morgan, and A Levin. 2013. "Health Economics of Rubella: A Systematic Review to Assess the Value of Rubella Vaccination." *BMC Public Health* 13 (1): 406. doi:10.1186/1471-2458-13-406.

Balfour, HH, Jr, and DP Amren. 1977. "Rubella Vaccine (HPV-77 DE5 Strain) Fails to Sustain Antibody Titres." *Lancet* 2 (8048): 1130–31.

Banatvala, JE. 2006. "Clinical Feautures: Post-Natally Acquired Rubella." *Perspect Med Virol* 15: 19–37. doi:10.1016/S0168-7069(06)15002-8.

Banatvala, JE, and DWG Brown. 2004. "Rubella." *Lancet* 363 (9415): 1127–37.

Baron, MD, and K Forsell. 1991. "Oligomerization of the Structural Proteins of Rubella Virus." *Virology* 185 (2): 811–19.

Bennett, JV, J Fernandez de Castro, JL Valdespino-Gomez, ML Garcia-Garcia, R Islas-Romero, G Echaniz-Aviles, A Jimenez-Corona, and J Sepulveda-Amor. 2002. "Aerosolized Measles and Measles-Rubella Vaccines Induce Better Measles Antibody Booster Responses than Injected Vaccines: Randomized Trials in Mexican Schoolchildren." *Bull World Health Organ* 80 (10): 806–12.

Berger, SA, GM Ginsberg, and PE Slater. 1990. "Cost-Benefit Analysis of Routine Mumps and Rubella Vaccination for Israeli Infants." *Isr J Med Sci* 26 (2): 74–80.

Best, JM, C Castillo-Solorzano, JS Spika, JP Icenogle, JW Glasser, NJ Gay, J Andrus, and AM Arvin. 2005. "Reducing the Global Burden of Congenital Rubella Syndrome: Report of the World Health Organization Steering Committee On Research Related To Measles and Rubella Vaccines and Vaccination, June 2004." *J Infect Dis* 192 (11): 1890–97.

Bielejec, F, P Lemey, G Baele, A Rambaut, and MA Suchard. 2014. “Inferring Heterogeneous Evolutionary Processes through Time: From Sequence Substitution to Phylogeography.” *Syst Biol* 63 (4): 493–504. doi:10.1093/sysbio/syu015.

Bielejec, F, A Rambaut, MA Suchard, and P Lemey. 2011. “SPREAD: Spatial Phylogenetic Reconstruction of Evolutionary Dynamics.” *Bioinformatics* 27 (20): 2910–12. doi:10.1093/bioinformatics/btr481.

Bjerregaard, P. 1991. “Economic Analysis of Immunization Programmes.” *Scand J Soc Med Suppl* 46: 115–19.

Castillo-Solórzano, C, C Marsigli, P Bravo-Alcántara, B Flannery, C Ruiz Matus, G Tambini, S Gross-Galiano, and JK Andrus. 2011. “Elimination of Rubella and Congenital Rubella Syndrome in the Americas.” *J Infect Dis* 204 (Suppl 2): S571–S578. doi:10.1093/infdis/jir472.

Centers for Disease Control and Prevention. 2005a. “Brief Report: Imported Case of Congenital Rubella Syndrome -- New Hampshire, 2005.” *MMWR Morb Mortal Wkly Rep* 54 (45): 1149–76.

Centers for Disease Control and Prevention. 2005b. “Elimination of Rubella and Congenital Rubella Syndrome - United States, 1969 - 2004.” *MMWR Morb Mortal Wkly Rep* 54 (11): 279–82.

Centers for Disease Control and Prevention. 2013. “Nationwide Rubella epidemic—Japan, 2013.” *MMWR Morb Mortal Wkly Rep* 62 (23): 457–62.

Chantler, JK, KD Lund, NP Miki, CA Berkowitz, and G Tai. 1993. “Characterization of Rubella Virus Strain Differences Associated with Attenuation.” *Intervirology* 36 (4): 225–36.

Chaye, H, P Chong, B Tripet, B Brush, and S Gillam. 1992. “Localization of the Virus Neutralizing and Hemagglutinin Epitopes of E1 Glycoprotein of Rubella Virus.” *Virology* 189 (2): 483–92.

Cheffins, T, A Chan, RJ Keane, EA Haan, and R Hall. 1998. “The Impact of Rubella Immunisation on the Incidence of Rubella, Congenital Rubella Syndrome and Rubella-Related Terminations of Pregnancy in South Australia.” *Br J Obstet Gynaecol* 105 (9): 998–1004.

Chen, MH, and TK Frey. 1999. “Mutagenic Analysis of the 3’ Cis-Acting Elements of the Rubella Virus Genome.” *J Virol* 73 (4): 3386–3403.

Cherian, S, A Walimbe, S Jadhav, S Gandhe, S Hundekar, A Mishra, and V Arankalle. 2009. “Evolutionary Rates and Timescale Comparison of Chikungunya Viruses Inferred from the Whole genome/E1 Gene with Special Reference to the 2005–07 Outbreak in the Indian Subcontinent.” *Infect Genet Evol* 9 (1): 16–23. doi:10.1016/j.meegid.2008.09.004.

Cooper, LZ. 1975. “Congenital Rubella in the United States.” In *Infections of the Fetus and the Newborn Infant: Proceedings of a Symposium Held in New York City, March 1975. Presented by New Yrsity Medical Center. Sponsored by the National Founork Univedation - March of Dimes*, 3:1–22. Progress in Clinical and Biological Research. New York: A.R. Liss.

Cusi, M G, M Valassina, S Bianchi, W Wunner, and P E Valensin. 1995. “Evaluation of Rubella Virus E2 and C Proteins in Protection against Rubella Virus in a Mouse Model.” *Virus Research* 37 (3): 199–208.

Cutts, FT, SE Robertson, JL Diaz-Ortega, and R Samuel. 1997. “Control of Rubella and Congenital Rubella Syndrome (CRS) in Developing Countries, Part 1: Burden of Disease from CRS.” *Bull World Health Organ* 75: 55–68.

Cutts, FT, and E Vynnycky. 1999. “Modelling the Incidence of Congenital Rubella Syndrome in Developing Countries.” *Int J Epidemiol* 28: 1176–84.

Davis, WJ, HE Larson, JP Simsarian, PD Parkman, and HM Meyer Jr. 1971. “A Study of Rubella Immunity and Resistance to Infection.” *JAMA* 215 (4): 600–608.

Delport, W, AF Poon, SD Frost, and SL Kosakovsky Pond. 2010. “Datamonkey 2010: A Suite of Phylogenetic Analysis Tools for Evolutionary Biology.” *Bioinformatics* 26 (19): 2455–57. doi:10.1093/bioinformatics/btq429.

Dominguez, G, CY Wang, and TK Frey. 1990. “Sequences of the Genome RNA of Rubella Virus: Evidence of Genetic Rearrangement during Togavirus Evolution.” *Virology* 177: 225.

Donadio, FF, MM Siqueira, A Vyse, L Jin, and SA Oliveira. 2003. “The Genomic Analysis of Rubella Virus Detected from Outbreak and Sporadic Cases in Rio de Janeiro State, Brazil.” *J Clin Virol* 27 (2): 205–9.

Drummond, AJ, OG Pybus, and A Rambaut. 2003. “Inference of Viral Evolutionary Rates from Molecular Sequences.” *Adv Parasitol* 54: 331–58.

Drummond, AJ, and A Rambaut. 2007. “BEAST: Bayesian Evolutionary Analysis by Sampling Trees.” *BMC Evol Biol* 7 (1): 214.

Drummond, AJ, MA Suchard, D Xie, and A Rambaut. 2012. “Bayesian Phylogenetics with BEAUti and the BEAST 1.7.” *Mol Biol Evol* 29 (8): 1969–73. doi:10.1093/molbev/mss075.

DuBois, RM, MC Vaney, MA Tortorici, RA Kurdi, G Barba-Spaeth, T Krey, and FA Rey. 2013. “Functional and Evolutionary Insight from the Crystal Structure of Rubella Virus Protein E1.” *Nature* 493 (7433): 552–56. doi:10.1038/nature11741.

Dudgeon, JA. 1975a. “Congenital Rubella in the United Kingdom of Great Britain.” In *Progress in Clinical and Biological Research - Infections of the Fetus and the Newborn Infant*, 3:23–34. Progress in Clinical and Biological Research. New York: A.R. Liss.

Dudgeon, JA. 1975b. “Congenital Rubella.” *J Pediatr* 87 (6): 1078–86.

Dudgeon, JA. 1985. “Selective Immunization: Protection of the Individual.” *Rev Infect Dis* 7 (Suppl 1): S185–S190.

Duffy, S, LA Shackelton, and EC Holmes. 2008. “Rates of Evolutionary Change in Viruses: Patterns and Determinants.” *Nat Rev Genet* 9 (4): 267–76. doi:10.1038/nrg2323.

Edgar, RC. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucl Acids Res* 32 (5): 1792–97.

Edmunds, WJ, NJ Gay, M Kretzschmar, RG Pebody, H Wachmann, and ESEN Project. European Sero-epidemiology Network. 2000. “The Pre-Vaccination Epidemiology of Measles, Mumps and Rubella in Europe: Implications for Modelling Studies.” *Epidemiol Infect* 125 (3): 635–50.

Elo, O. 1979. “Cost-Benefit Studies of Vaccinations in Finland.” *Dev Biol Stand* 43: 419–28.

Farber, ME, and SN Finkelstein. 1979. “A Cost-Benefit Analysis of a Mandatory Premarital Rubella-Antibody Screening Program.” *N Engl J Med* 300 (15): 856–59. doi:10.1056/NEJM197904123001512.

Fogel, A, CB Gerichter, B Barnea, R Handsher, and E Heeger. 1978. “Response to Experimental Challenge in Persons Immunized with Different Rubella Vaccines.” *J Pediatr* 92 (1): 26–29.
Forbes, JA. 1969. “Rubella: Historical Aspects.” *Am J Dis Child* 118: 5–11.
Ford, DK, GD Reid, AJ Tingle, LA Mitchell, and M Schulzer. 1992. “Sequential Follow up Observations of a Patient with Rubella Associated Persistent Arthritis.” *Ann Rheum Dis* 51 (3): 407–10.
Frey, TK. 1994. “Molecular Biology of Rubella Virus.” In *Advances in Virus Research*, 44:69–160. Elsevier.
Frey, TK, ES Abernathy, TJ Bosma, WG Starkey, KM Corbett, JM Best, S Katow, and SC Weaver. 1998. “Molecular Analysis of Rubella Virus Epidemiology across Three Continents, North America, Europe, and Asia, 1961–1997.” *J Infect Dis* 178 (3): 642–50.
Gill, MS, P Lemey, NR Faria, A Rambaut, B Shapiro, and MA Suchard. 2013. “Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci.” *Mol Biol Evol* 30 (3): 713–24. doi:10.1093/molbev/mss265.
Golden, M, and DP Martin. 2013. “DOOSS: A Tool for Visual Analysis of Data Overlaid on Secondary Structures.” *Bioinformatics* 29 (2): 271–72. doi:10.1093/bioinformatics/bts667.
Golden, M, and GL Shapiro. 1984. “Cost-Benefit Analysis of Alternative Programs of Vaccination against Rubella in Israel.” *Public Health* 98 (3): 179–90.
Greenberg, M, O Pellitteri, and J Barton. 1957. “Frequency of Defects in Infants Whose Mothers Had Rubella during Pregnancy.” *JAMA* 165: 675–78.
Gregg, NM. 1941. “Congenital Cataract Following German Measles in the Mother.” *Trans Ophthalmol Soc Aust* 3: 35–46.
Han, GZ, and M Worobey. 2011. “Homologous Recombination in Negative Sense RNA Viruses.” *Viruses* 3 (12): 1358–73. doi:10.3390/v3081358.
Heggie, A D, and F C Robbins. 1969. “Natural Rubella Acquired after Birth. Clinical Features and Complications.” *American Journal of Diseases of Children (1960)* 118 (1): 12–17.
Hemphill, ML, RY Forng, ES Abernathy, and TK Frey. 1988. “Time Course of Virus-Specific Macromolecular Synthesis during Rubella Virus Infection in Vero Cells.” *Virology* 162 (1): 65–75.
Hess, AF. 1914. “German Measles (rubella): An Experimental Study.” *Arch Intern Med* 13 (6): 913–16.
Hilleman, MR, EB Buynak, JE Whitman Jr, RW Weibel, and J Stokes Jr. 1969. “Live Attenuated Rubella Virus Vaccines: Experiences with Duck Embryo Cell Preparations.” *Am J Dis Child* 118: 166–71.
Hinman, AR, B Irons, M Lewis, and K Kandola. 2002. “Economic Analyses of Rubella and Rubella Vaccines: A Global Review.” *Bull World Health Organ* 80 (4): 264–70.
Hobman, TC, ZY Qiu, H Chaye, and S Gillam. 1991. “Analysis of Rubella Virus E1 Glycosylation Mutants Expressed in COS Cells.” *Virology* 181 (2): 768–72.
Hofmann, J, M Renz, S Meyer, A von Haeseler, and UG Liebert. 2003. “Phylogenetic Analysis of Rubella Virus Including New Genotype I Isolates.” *Virus Res* 96 (1-2): 123–28.
Holmes, EC. 2004. “The Phylogeography of Human Viruses.” *Mol Ecol* 13: 745–56.

Icenogle, JP, MM Siqueira, ES Abernathy, XR Lemos, RA Fasce, G Torres, and SE Reef. 2011. "Virologic Surveillance for Wild-Type Rubella Viruses in the Americas." *J Infect Dis* 204 (Suppl 2): S647–S651. doi:10.1093/infdis/jir431.

Ingalls, TH, SA Plotkin, HM Meyer Jr, and PD Parkman. 1967. "Rubella: Epidemiology, Virology, and Immunology." *Am J Med Sci* 253 (3): 349–73.

Irons, B, MJ Lewis, M Dahl-Regis, C Castillo-Solórzano, PA Carrasco, and CA de Quadros. 2000. "Strategies to Eradicate Rubella in the English-Speaking Caribbean." *Am J Public Health* 90 (10): 1545–49.

Jenkins, GM, A Rambaut, OG Pybus, and EC Holmes. 2002. "Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis." *J Mol Evol* 54 (2): 156–65.

Katow, S. 2004. "Molecular Epidemiology of Rubella Virus in Asia: Utility for Reduction in the Burden of Diseases due to Congenital Rubella Syndrome." *Pediatr Int* 46 (2): 207–13.

Katow, S, and A Sugiura. 1985. "Antibody Response to Individual Rubella Virus Proteins in Congenital and Other Rubella Virus Infections." *J Clin Microbiol* 21 (3): 449–51.

Kingman, JFC. 1982. "The Coalescent." *Stoch Proc Appl* 13 (3): 235–48.

Klock, LE, and GS Rachelefsky. 1973. "Failure of Rubella Herd Immunity during an Epidemic." *N Engl J Med* 288 (2): 69–72.

Kosakovsky Pond, SL, D Posada, MB Gravenor, CH Woelk, and SD Frost. 2006. "GARD: A Genetic Algorithm for Recombination Detection." *Bioinformatics* 22 (24): 3096–98. doi:10.1093/bioinformatics/btl474.

Kouri, G, A Aguilera, P Rodriguez, and M Korolev. 1974. "A Study of Microfoci and Inclusion Bodies Produced by Rubella Virus in the RK-13 Cell Line." *J Gen Virol* 22 (1): 73–80.

Ladd, HF. 2012. "Education and Poverty: Confronting the Evidence." *J Pol Anal Manag* 31 (2): 203–27.

Lemey, P, A Rambaut, T Bedford, N Faria, F Bielejec, G Baele, CA Russell, et al. 2014. "Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2." *PLoS Pathogens* 10 (2): e1003932. doi:10.1371/journal.ppat.1003932.

Lemey, P, A Rambaut, AJ Drummond, and MA Suchard. 2009. "Bayesian Phylogeography Finds Its Roots." *PLoS Comput Biol* 5 (9): e1000520. doi:10.1371/journal.pcbi.1000520.

Liang, Y, and S Gillam. 2000. "Mutational Analysis of the Rubella Virus Nonstructural Polyprotein and Its Cleavage Products in Virus Replication and RNA Synthesis." *J Virol* 74 (11): 5133–41.

Liu, Z, D Yang, Z Qiu, KT Lim, P Chong, and S Gillam. 1996. "Identification of Domains in Rubella Virus Genomic RNA and Capsid Protein Necessary for Specific Interaction." *J Virol* 70 (4): 2184–90.

Londesborough, P, L Ho-Terry, and G Terry. 1995. "Sequence Variation and Biological Activity of Rubella Virus Isolates." *Archives of Virology* 140 (3): 563–70.

Lundstorm, R. 1962. "Rubella during Pregnancy: A Follow-up Study of Children Born after an Epidemic of Rubella in Sweden, 1951, with Additional Investigations on Proplylaxis and Treatment of Maternal Rubella." *Acta Paediatr Scand* 133 (Suppl): 1–110.

Markham, NR, and M Zuker. 2008. "UNAFold: Software for Nucleic Acid Folding and Hybridization." *Methods Mol Biol* 453: 3–31. doi:10.1007/978-1-60327-429-6_1.

Martin, DP, P Lemey, M Lott, V Moulton, D Posada, and P Lefeuvre. 2010. “RDP3: A Flexible and Fast Computer Program for Analyzing Recombination.” *Bioinformatics* 26 (19): 2462–63. doi:10.1093/bioinformatics/btq467.

Martin, DP, P Lemey, and D Posada. 2011. “Analysing Recombination in Nucleotide Sequences.” *Mol Ecol Res* 11 (6): 943–55. doi:10.1111/j.1755-0998.2011.03026.x.

McLean, HQ, AP Fiebelkorn, JL Temte, and GS Wallace. 2013. “Prevention of Measles, Rubella, Congenital Rubella Syndrome, and Mumps, 2013: Summary Recommendations of the Advisory Committee on Immunization Practices (ACIP).” *MMWR Recomm Rep* 62 (RR-04): 1–34.

Meyer, HM, Jr, PD Parkman, TE Hobbins, HE Larson, WJ Davis, JP Simsarian, and HE Hopps. 1969. “Attenuated Rubella Viruses: Laboratory and Clinical Characteristics.” *Am J Dis Child* 118 (2): 155–69.

Miller, E, JE Cradock-Watson, and TM Pollock. 1982. “Consequences of Confirmed Maternal Rubella at Successive Stages of Pregnancy.” *Lancet* 2 (8302): 781–84.

Minin, VN, and MA Suchard. 2008. “Counting Labeled Transitions in Continuous-Time Markov Models of Evolution.” *J Math Biol* 56 (3): 391–412. doi:10.1007/s00285-007-0120-8.

Muhire, BM, M Golden, B Murrell, P Lefeuvre, JM Lett, A Gray, AYF Poon, et al. 2014. “Evidence of Pervasive Biologically Functional Secondary Structures within the Genomes of Eukaryotic Single-Stranded DNA Viruses.” *J Virol* 88 (4): 1972–89. doi:10.1128/JVI.03031-13.

Muhire, BM, DP Martin, JK Brown, J Navas-Castillo, E Moriones, FM Zerbini, R Rivera-Bustamante, VG Malathi, RW Briddon, and A Varsani. 2013. “A Genome-Wide Pairwise-Identity-Based Proposal for the Classification of Viruses in the Genus Mastrevirus (family Geminiviridae).” *Arch Virol* 158 (6): 1411–24. doi:10.1007/s00705-012-1601-7.

Murrell, B, S Moola, A Mabona, T Weighill, D Sheward, SL Kosakovsky Pond, and K Scheffler. 2013. “FUBAR: A Fast, Unconstrained Bayesian Approximation for Inferring Selection.” *Mol Biol Evol* 30 (5): 1196–1205. doi:10.1093/molbev/mst030.

Nakhasi, HL, M Ramanujam, CD Atreya, TC Hobman, N Lee, A Esmaili, and RC Duncan. 2001. “Rubella Virus Glycoprotein Interaction with the Endoplasmic Reticulum Calreticulin and Calnexin.” *Arch Virol* 146 (1): 1–14.

Nakhasi, HL, NK Singh, GP Pogue, XQ Cao, and TA Rouault. 1994. “Identification and Characterization of Host Factor Interactions with Cis-Acting Elements of Rubella Virus RNA.” *Arch Virol Suppl* 9: 255–67.

National Institute for Communicable Diseases (NICD), South Africa. 2010. “Measles Outbreak.” *Communicable Diseases Communiqué* 9 (5): 2–3.

Ogra, PL, D Kerr-Grant, G Umana, J Dzierba, and D Weintraub. 1971. “Antibody Response in Serum and Nasopharynx after Naturally Acquired and Vaccine-Induced Infection with Rubella Virus.” *N Engl J Med* 285 (24): 1333–39.

Paradowska-Stankiewicz, I, MP Czarkowski, T Derrough, and P Stefanoff. 2013. “Ongoing Outbreak of Rubella among Young Male Adults in Poland: Increased Risk of Congenital Rubella Infections.” *Euro Surveill* 18 (21).

Parkman, PD. 1996. “Togaviruses: Rubella Virus.” In *Medical Microbiology*, edited by S Baron, 4th Edition. Galveston, Texas: University of Texas Medical Branch at Galveston.

Parkman, PD, EL Buescher, and MS Artenstein. 1962. “Recovery of Rubella Virus from Army Recruits.” *Proc Soc Exp Biol Med* 111: 225–30.

Peltola, H, I Davidkin, M Paunio, M Valle, P Leinikki, and OP Heinonen. 2000. “Mumps and Rubella Eliminated from Finland.” *JAMA* 284 (20): 2643–47.

Perkins, FT. 1985. “Licensed Vaccines.” *Rev Infect Dis* 7 (Suppl 1): S73–S76.

Pham, VH, TV Nguyen, TT Nguyen, LD Dang, NH Hoang, TV Nguyen, and K Abe. 2013. “Rubella Epidemic in Vietnam: Characteristic of Rubella Virus Genes from Pregnant Women and Their Fetuses/newborns with Congenital Rubella Syndrome.” *J Clin Virol* 57 (2): 152–56. doi:10.1016/j.jcv.2013.02.008.

Pitt, D, and EH Keir. 1965. “Results of Rubella in Pregnancy.” *Med J Aust* 2: 647–51.

Plotkin, S.A., J. Farguhar, M. Katz, and F. Buser. 1969. “Attenuation of RA27/3 Rubella Virus in WI-38 Human Diploid Cells.” *Am J Dis Child.* 118: 178–85.

Plotkin, SA. 2006. “The History of Rubella and Rubella Vaccination Leading to Elimination.” *Clin Infect Dis* 43 (Suppl 3): S164–S168. doi:10.1086/505950.

Plotkin, SA, JD Farquhar, and PL Ogra. 1973. “Immunologic Properties of RA27-3 Rubella Virus Vaccine. A Comparison with Strains Presently Licensed in the United States.” *JAMA* 225 (6): 585–90.

Pogue, GP, J Hofmann, R Duncan, JM Best, J Etherington, RD Sontheimer, and HL Nakhasi. 1996. “Autoantigens Interact with Cis-Acting Elements of Rubella Virus RNA.” *J Virol* 70 (9): 6269–77.

Poon, AF, FI Lewis, SD Frost, and SL Kosakovsky Pond. 2008. “Spidermonkey: Rapid Detection of Co-Evolving Sites Using Bayesian Graphical Models.” *Bioinformatics* 24 (17): 1949–50. doi:10.1093/bioinformatics/btn313.

Posada, D, and KA Crandall. 2002. “The Effect of Recombination on the Accuracy of Phylogeny Estimation.” *J Mol Evol* 54 (3): 396–402. doi:10.1007/s00239-001-0034-9.

Preblud, SR, and CA Alford. 1983. “Rubella.” In *Infectious Diseases of the Fetus and Newborn Infant*, 3rd Edition, 196–240. Philadelphia: W.B. Saunders.

Preblud, SR, MK Serdula, JA Frank Jr, AD Brandling-Bennett, and AR Hinman. 1980. “Rubella Vaccination in the United States: A Ten-Year Review.” *Epidemiol Rev* 2: 171–94.

Prinzie, A, C Huygelen, J Gold, J Farguhar, and J McKee. 1969. “Experimental Live Attenuated Rubella Virus Vaccine. Clinical Evaluation of Cendehill Strain.” *Am J Dis Child* 118 (2): 172–77.

Pugachev, KV, and TK Frey. 1998a. “Rubella Virus Induces Apoptosis in Culture Cells.” *Virology* 250 (2): 359–70. doi:10.1006/viro.1998.9395.

Pugachev, KV, and TK Frey. 1998b. “Effects of Defined Mutations in the 5’ Nontranslated Region of Rubella Virus Genomic RNA on Virus Viability and Macromolecule Synthesis.” *J Virol* 72 (1): 641–50.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rambaut, A. 2013. *Path-O-Gen* (version v1.4). http://tree.bio.ed.ac.uk/software/pathogen/.

Rambaut, A, MA Suchard, and AJ Drummond. 2009. *Tracer* (version v1.6). http://beast.bio.ed.ac.uk/Tracer.

Ramsay, M, M Reacher, C O’Flynn, R Buttery, F Hadden, B Cohen, W Knowles, T Wreghitt, and D Brown. 2002. “Causes of Morbilliform Rash in a Highly Immunised English Population.” *Arch Dis Child* 87 (3): 202–6.

Rayfield, EJ, KJ Kelly, and JW Yoon. 1986. “Rubella Virus-Induced Diabetes in the Hamster.” *Diabetes* 35 (11): 1278–81.

Reef, S E, P Strebel, A Dabbagh, M Gacic-Dobo, and S Cochi. 2011. “Progress toward Control of Rubella and Prevention of Congenital Rubella Syndrome--Worldwide, 2009.” *The Journal of Infectious Diseases* 204 Suppl 1 (July): S24–27. doi:10.1093/infdis/jir155.

Reef, SE, TK Frey, K Theall, ES Abernathy, CL Burnett, JP Icenogle, MM McCauley, and M Wharton. 2002. “The Changing Epidemiology of Rubella in the 1990s.” *JAMA* 287 (4): 464–72.

Reef, SE, and SA Plotkin. 2013. “31 - Rubella Vaccine.” In *Vaccines*, Sixth Edition, 688–717. London: W.B. Saunders.

Scheffler, K, DP Martin, and C Seoighe. 2006. “Robust Inference of Positive Selection from Recombining Coding Sequences.” *Bioinformatics* 22 (20): 2493–99.

Schierup, MH, and J Hein. 2000. “Recombination and the Molecular Clock.” *Mol Biol Evol* 17 (10): 1578–79.

Schmaljohn, AL, and D McClain. 1996. “Alphaviruses (Togaviridae) and Flaviviruses (Flaviviridae).” In *Medical Microbiology*, edited by S Baron, 4th Edition. Galveston, Texas: University of Texas Medical Branch at Galveston.

Schoub, BD, BN Harris, J McAnerney, and L Blumberg. 2009. “Rubella in South Africa: An Impending Greek Tragedy.” *S Afr Med J* 99 (7).

Semegni, J Y, M Wamalwa, R Gaujoux, G W Harkins, A Gray, and D P Martin. 2011. “NASP: A Parallel Program for Identifying Evolutionarily Conserved Nucleic Acid Secondary Structures from Nucleotide Sequence Alignments.” *Bioinformatics (Oxford, England)* 27 (17): 2443–45. doi:10.1093/bioinformatics/btr417.

Shepherd, DN, DP Martin, DR McGivern, MI Boulton, JA Thomson, and EP Rybicki. 2005. “A Three-Nucleotide Mutation Altering the Maize Streak Virus Rep pRBR-Interaction Motif Reduces Symptom Severity in Maize and Partially Reverts at High Frequency without Restoring pRBR-Rep Binding.” *J Gen Virol* 86 (3): 803–13. doi:10.1099/vir.0.80694-0.

Simmonds, P, A Tuplin, and DJ Evans. 2004. “Detection of Genome-Scale Ordered RNA Structure (GORS) in Genomes of Positive-Stranded RNA Viruses: Implications for Virus Evolution and Host Persistence.” *RNA* 10 (9): 1337–51. doi:10.1261/rna.7640104.

Song, N, Z Gao, JG Wood, L Hueston, GL Gilbert, CR MacIntyre, HE Quinn, R Menzies, and P McIntyre. 2012. “Current Epidemiology of Rubella and Congenital Rubella Syndrome in Australia: Progress towards Elimination.” *Vaccine* 30 (27): 4073–78. doi:10.1016/j.vaccine.2012.04.025.

Spruance, SL, and CB Smith. 1971. “Joint Complications Associated with Derivatives of HPV-77 Rubella Virus Vaccine.” *Am J Dis Child* 122 (2): 105–11.

Stray-Pedersen, B. 1982. “Economic Evaluation of Different Vaccination Programmes to Prevent Congenital Rubella.” *NIPH Ann* 5 (2): 69–83.

Suwannakarn, K, A Theamboonlers, and Y Poovorawan. 2011. “Molecular Genome Tracking of East, Central and South African Genotype of Chikungunya Virus in South–east Asia between 2006 and 2009.” *Asian Pac J Trop Med* 4 (7): 535–40.

Tamura, K, D Peterson, N Peterson, G Stecher, M Nei, and S Kumar. 2011. “MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.” *Mol Biol Evol* 28 (10): 2731–39. doi:10.1093/molbev/msr121.

Terry, GM, L Ho-Terry, P Londesborough, and KR Rees. 1988. “Localization of the Rubella E1 Epitopes.” *Arch Virol* 98 (3-4): 189–97.

Tookey, P, P Molyneaux, and P Helms. 2000. “UK Case of Congenital Rubella Can Be Linked to Greek Cases.” *BMJ* 321 (7263): 766.

Tzeng, WP, and TK Frey. 2002. “Mapping the Rubella Virus Subgenomic Promoter.” *J Virol* 76 (7): 3189–3201. doi:10.1128/JVI.76.7.3189-3201.2002.

Tzeng, WP, and TK Frey. 2009. “Functional Replacement of a Domain in the Rubella Virus p150 Replicase Protein by the Virus Capsid Protein.” *J Virol* 83 (8): 3549–55. doi:10.1128/JVI.02411-08.

Ueda, K. 2009. “Development of the Rubella Vaccine and Vaccination Strategy in Japan.” *Vaccine* 27 (24): 3232–33. doi:10.1016/j.vaccine.2009.02.076.

Veale, H. 1866. “History of an Epidemic of Rotheln, with Observations on Its Pathology.” *Edinburgh Med J* 12: 404–14.

Volk, SM, R Chen, KA Tsetsarkin, AP Adams, TI Garcia, AA Sall, F Nasar, et al. 2010. “Genome-Scale Phylogenetic Analyses of Chikungunya Virus Reveal Independent Emergences of Recent Epidemics and Various Evolutionary Rates.” *J Virol* 84 (13): 6497–6504.

Wallace, RB, and P Isacson. 1972. “Comparative Trial of HPV-77, DE-5 and RA 27-3 Live-Attenuated Rubella Vaccines.” *Am J Dis Child* 124 (4): 536–38.

Wang, C, Z Zhu, Q Xu, A Xu, X Fang, L Song, W Li, P Xiong, and W Xu. 2012. “Rubella Epidemics and Genotypic Distribution of the Rubella Virus in Shandong Province, China, in 1999-2010.” *PLoS ONE* 7 (7): e42013. doi:10.1371/journal.pone.0042013.

Watson, JC, SC Hadler, CA Dykewicz, SE Reef, and L Phillips. 1998. “Measles, Mumps, and Rubella--Vaccine Use and Strategies for Elimination of Measles, Rubella, and Congenital Rubella Syndrome and Control of Mumps: Recommendations of the Advisory Committee on Immunization Practices (ACIP).” *MMWR Recomm Rep* 47 (RR-8): 1–57.

Weller, TH, and FA Neva. 1962. “Propagation in Tissue Culture of Cytopathic Agents from Patients with Rubella-like Illness.” *Proc Soc Exp Biol Med* 111 (1): 215–25. doi:10.3181/00379727-111-27749.

Wesselhoeft, C. 1949. “Rubella (german Measles) and Congenital Deformities.” *N Engl J Med* 240 (7): 258–61.

Wilkinson, KA, EJ Merino, and KM Weeks. 2006. “Selective 2’-Hydroxyl Acylation Analyzed by Primer Extension (SHAPE): Quantitative RNA Structure Analysis at Single Nucleotide Resolution.” *Nat Protoc* 1 (3): 1610–16. doi:10.1038/nprot.2006.249.

Wise, J. 2013. “Measles Outbreak Hits Northeast England.” *BMJ* 346 (1): 662. doi:10.1136/bmj.f662.

Witte, JJ, AW Karchmer, G Case, KL Herrmann, E Abrutyn, I Kassanoff, and JS Neill. 1969. “Epidemiology of Rubella.” *Am J Dis Child* 118 (1): 107–11.

Wolinsky, JS, JK Chantler, and AJ Tingle. 2001. “Rubella Virus.” In *Fields Virology*, 4th Edition, 963–90. Philadelphia: Lippincott Willams & Wilkins.

Wolinsky, JS, M McCarthy, O Allen-Cannady, WT Moore, R Jin, SN Cao, A Lovett, and D Simmons. 1991. “Monoclonal Antibody-Defined Epitope Map of Expressed Rubella Virus Protein Domains.” *J Virol* 65 (8): 3986–94.

World Health Organization. 2005. “Standardization of the Nomenclature for Genetic Characteristics of Wild-Type Rubella Viruses.” *Wkly Epidemiol Rec* 80 (14): 126–32.

World Health Organization. 2007. “Update of Standard Nomenclature for Wild-Type Rubella Viruses.” *Wkly Epidemiol Rec* 82: 209–24.

World Health Organization. 2011. “Rubella Vaccines: WHO Position Paper.” *Wkly Epidemiol Rec* 86 (29): 301–16.

World Health Organization. 2012. “Global Measles and Rubella Strategic Plan: Strategic Plan 2012- 2020.”

World Health Organization. 2013a. “Rubella Virus Nomenclature Update: 2013.” *Wkly Epidemiol Rec* 88 (32): 337–48.

World Health Organization. 2013b. “World Health Organization Immunization Surveillance, Assessment and Monitoring.” *World Health Organization: Rubella Reported Cases*. http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tsincidencerubella.html.

World Health Organization, Department of Vaccines and Biologicals. 2000. *Report of a Meeting on Preventing Congenital Rubella Syndrome: Immunization Strategies, Surveillance Needs*. Geneva: World Health Orginization.

Wright, JA, and C Polack. 2006. “Understanding Variation in Measles-Mumps-Rubella Immunization Coverage -- A Population-Based Study.” *Eur J Public Health* 16 (2): 137–42. doi:10.1093/eurpub/cki194.

Zheng, DP, TK Frey, JP Icenogle, S Katow, ES Abernathy, KJ Song, WB Xu, V Yarulin, RG Desjatskova, and Y Aboudy. 2003. “Global Distribution of Rubella Virus Genotypes.” *Emerg Infect Dis* 9 (12): 1523.

Zheng, DP, YM Zhou, K Zhao, YR Han, and TK Frey. 2003. “Characterization of Genotype II Rubella Virus Strains.” *Arch Virol* 148 (9): 1835–50.

Zheng, DP, H Zhu, MG Revello, G Gerna, and TK Frey. 2003. “Phylogenetic Analysis of Rubella Virus Isolated during a Period of Epidemic Transmission in Italy, 1991-1997.” *J Infect Dis* 187 (10): 1587–97.

Zhou, Y, H Ushijima, and TK Frey. 2007. “Genomic Analysis of Diverse Rubella Virus Genotypes.” *J Gen Virol* 88 (3): 932–41.

Zhu, Z, A Cui, H Wang, Y Zhang, C Liu, C Wang, S Zhou, et al. 2011. “Emergence and Continuous Evolution of Genotype 1E Rubella Viruses in China.” *J Clin Microbiol* 50 (2): 353–63. doi:10.1128/JCM.01264-11.

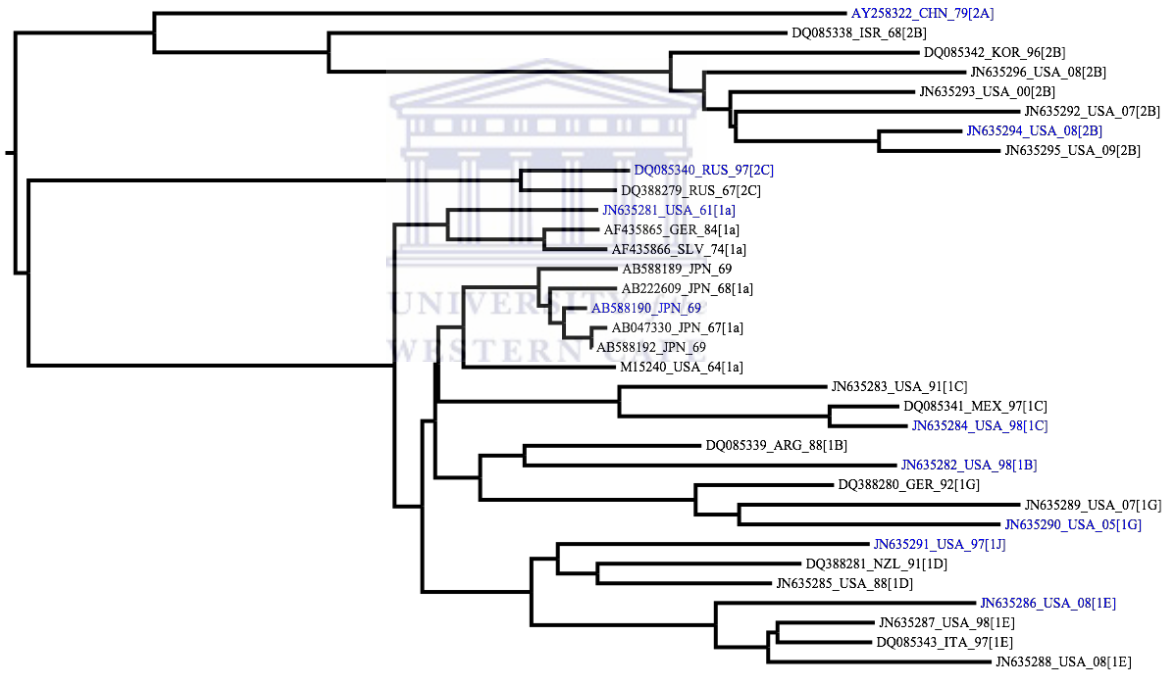
6. APPENDICES

Appendix 1. Python script used to retrieve publically available sequences, with user specific search query, from the NCBI GenBank. At the time of analysis, 1254 sequences were available for the search term "*Rubella virus*". Using pairwise identities, as determined by SDT v1.0, known resequenced and duplicate samples were identified and removed. Sequences were subsequently subdivided into various datasets for further analysis (see Table 4).

Usage: python script_name.py <GenBank_search_term>

```
from Bio import Entrez

search_term = sys.argv[1]

search_handle = Entrez.esearch(db="nucleotide", term=search_term, usehistory="y")

search_results = Entrez.read(search_handle)

search_handle.close()


gi_list = search_results["IdList"]

count = int(search_results["Count"])

webenv = search_results["WebEnv"]

query_key = search_results["QueryKey"]

batch_size = 5

out_handle = open("output_file_name.fasta", "w")

for start in range(0, count, batch_size):

    end = min(count, start+batch_size)

    print "Going to download record %i to %i" % (start+1, end)

    fetch_handle = Entrez.efetch(db="nucleotide", rettype="fasta", retmode="text",retstart=start,
    retmax=batch_size, webenv=webenv, query_key=query_key)

    data = fetch_handle.read()

    fetch_handle.close()

    out_handle.write(data)

out_handle.close()

print ("Script Done")
```

Appendix 2. Neighbour joining phylogenetic tree generated for Nucleic Acid Structure Prediction (NASP) of genome-wide nucleic acid secondary structural elements. Neighbour joining phylogenetic tree, containing a representative sample of RV genotypes (dataset i), was constructed from 34 full genome sequences available in GenBank at the time of analysis. Of these 34 full genome sequences, only ten were used for genome-wide nucleic acid secondary structure prediction, to reduce the computational burden imposed by the NASP software. These ten sequences (taxa labelled in blue) were selected from distinct clades within the neighbour joining phylogenetic tree, after which the most divergent sequences within each of the selected clades were identified using pairwise genetic distances. Isolate genotypes are depicted in square brackets next to taxon labels, respectively.



Appendix 3. Full description of datasets used in thesis. A full description of the *Rubella virus* sequences and datasets used in this study, including the accession number, genotype assignment, collection date, country of origin and dataset assignment (also see Table 4 in text).

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
AB047330	Japan	1967	1a		x	x	x	x	x	x	x	x	x	x	x	x	x
AB071280	Japan	1993	-								x			x	x		
AB222609	Japan	1968	1a		x	x	x	x	x	x	x	x	x	x	x	x	x
AB233430	Japan	2002	-						x								
AB285128	Japan	2003	-				x		x		x			x	x		
AB285129	Japan	2004	-				x		x		x			x	x		
AB285130	Japan	2001	-				x		x		x			x	x		
AB285131	Japan	2004	-				x		x		x			x	x		
AB285132	Japan	2002	-				x		x		x			x	x		
AB285133	Japan	2002	-				x		x		x			x	x		
AB285134	Japan	2002	-				x				x			x	x		
AB285135	Japan	2002	-				x		x		x			x	x		
AB285136	Japan	2002	-				x		x		x			x	x		
AB285137	Japan	1994	-				x		x		x			x	x		
AB285138	Japan	2004	-				x		x		x			x	x		
AB285139	Japan	2004	-				x		x		x			x	x		
AB285140	Japan	2004	-				x		x		x		x	x	x		
AB285141	Japan	2004	-				x		x		x			x	x		
AB285142	Japan	2004	-				x		x		x			x	x		
AB285143	Japan	2004	-				x		x		x			x	x		
AB285144	Japan	2004	-				x		x		x			x	x		
AB285145	Japan	2003	-				x		x		x			x	x		
AB546233	Vietnam	2009	2B								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
AB588189	Japan	1969	-		x	x	x	x	x	x	x	x	x	x	x	x	x
AB588190	Japan	1969	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x
AB588192	Japan	1969	-		x	x	x	x	x	x	x	x	x	x	x	x	x
AB621553	Japan	2011	1J								x			x	x		
AB632389	Japan	2011	2B								x			x	x		
AB640794	Japan	2011	1E								x			x	x		
AB646368	Japan	2010	1E								x			x	x		
AB665169	Japan	2011	2B								x			x	x		
AB674470	Japan	2011	2B								x			x	x		
AB674471	Japan	2011	1E								x			x	x		
AB683468	Japan	2011	1E								x			x	x		
AB683469	Japan	2011	1E								x			x	x		
AB702680	Japan	2011	2B								x			x	x		
AB702681	Japan	2011	2B								x			x	x		
AB702682	Japan	2011	2B								x			x	x		
AB702683	Japan	2011	2B								x			x	x		
AB702684	Japan	2011	2B								x			x	x		
AB702685	Japan	2011	2B								x			x	x		
AB702686	Japan	2012	2B								x			x	x		
AB735186	Japan	2011	1E								x			x	x		
AB735187	Japan	2012	2B								x			x	x		
AB735188	Japan	2012	1E								x			x	x		
AB735189	Japan	2012	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset														
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
AB739704	Japan	2012	1E								x			x	x		
AB745027	Vietnam	2011	2B								x			x	x		
AB745028	Vietnam	2011	2B								x			x	x		
AB745029	Vietnam	2011	2B								x			x	x		
AB745030	Vietnam	2011	2B								x			x	x		
AB745031	Vietnam	2011	2B								x			x	x		
AB745032	Vietnam	2011	2B								x			x	x		
AB745033	Vietnam	2011	2B								x			x	x		
AB745034	Vietnam	2011	2B								x			x	x		
AB745035	Vietnam	2011	2B								x			x	x		
AB745036	Vietnam	2011	2B								x			x	x		
AB745037	Vietnam	2011	2B								x			x	x		
AB745038	Vietnam	2012	2B								x			x	x		
AB745039	Vietnam	2011	2B								x			x	x		
AB753257	Japan	2012	1E								x			x	x		
AB753258	Japan	2012	2B								x			x	x		
AB753259	Japan	2012	2B								x			x	x		
AB753260	Japan	2012	2B								x			x	x		
AB753261	Japan	2012	2B								x			x	x		
AB753262	Japan	2012	2B								x			x	x		
AF039107	USA	1961	-								x			x	x		
AF435865	Germany	1984	1a		x	x	x	x	x	x	x	x	x	x	x	x	x
AF435866	Slovakia	1974	1a		x		x	x	x	x	x	x	x				

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
AF551761	Germany	1999	-								x		x	x	x		
AM258944	Belarus	2005	-								x			x	x		
AM258945	Belarus	2004	1G						x								
AM258946	Belarus	2005	-								x			x	x		
AM258947	Belarus	2005	-								x			x	x		
AM258948	Belarus	2005	-								x			x	x		
AM258949	Belarus	2005	-								x			x	x		
AM258950	Belarus	2004	-								x			x	x		
AM258951	Belarus	2004	-								x			x	x		
AM258952	Belarus	2004	-								x			x	x		
AM258953	Belarus	2005	1H						x								
AM258954	Belarus	2004	1E								x			x	x		
AM258955	Belarus	2005	1E								x			x	x		
AM258956	Belarus	2005	1E								x			x	x		
AM258957	Belarus	2005	1E								x			x	x		
AY161349	Italy	1991	-								x		x	x	x		
AY161350	Italy	1991	-								x			x	x		
AY161351	Italy	1991	-								x		x	x	x		
AY161352	Italy	1991	1I								x			x	x		
AY161353	Italy	1991	-								x			x	x		
AY161354	Italy	1991	-								x			x	x		
AY161355	Italy	1991	-								x			x	x		
AY161356	Italy	1991	-								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
AY161357	Italy	1991	-								x		x	x	x		
AY161358	Italy	1991	-								x			x	x		
AY161359	Italy	1992	-								x			x	x		
AY161360	Italy	1992	1I								x			x	x		
AY161361	Italy	1993	-								x			x	x		
AY161362	Italy	1993	-								x			x	x		
AY161363	Italy	1993	-								x			x	x		
AY161364	Italy	1993	-								x			x	x		
AY161365	Italy	1994	-								x			x	x		
AY161366	Italy	1994	-								x			x	x		
AY161367	Italy	1994	-								x			x	x		
AY161368	Italy	1994	-								x		x	x	x		
AY161369	Italy	1994	-								x			x	x		
AY161370	Italy	1994	-								x			x	x		
AY161371	Italy	1995	-								x			x	x		
AY161372	Italy	1995	-								x			x	x		
AY161373	Italy	1995	-								x			x	x		
AY161375	Italy	1997	-								x			x	x		
AY161376	Italy	1997	-								x			x	x		
AY161377	Italy	1997	-								x			x	x		
AY161378	Italy	1997	-								x			x	x		
AY161379	Italy	1997	-								x			x	x		
AY247016	Russia	1968	-								x		x	x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset														
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
AY247017	Russia	1969	-								x		x	x	x		
AY247018	Russia	1973	-								x		x	x	x		
AY247019	Russia	1997	-								x		x	x	x		
AY253148	Russia	1969	-								x		x	x	x		
AY258322	China	1979	2A	x	x	x	x	x	x	x	x	x	x	x	x	x	x
AY280704	Myanmar	2001	-								x			x	x		
AY280705	Myanmar	2001	-								x			x	x		
AY280706	Myanmar	2001	-								x			x	x		
AY280707	Myanmar	2002	-								x			x	x		
AY397695	Japan	1997	-								x			x	x		
AY397696	Japan	1997	-								x		x	x	x		
DQ085331	Italy	1997	1E								x			x	x		
DQ085332	USA	1998	-								x		x	x	x		
DQ085333	USA	1999	-								x			x	x		
DQ085334	USA	1998	-								x			x	x		
DQ085335	USA	1998	-								x			x	x		
DQ085336	USA	1998	-								x			x	x		
DQ085337	Canada	1997	-								x			x	x		
DQ085338	Israel	1968	2B		x	x	x	x	x	x	x	x	x	x	x	x	x
DQ085339	Argentina	1988	1B		x	x	x	x	x	x	x	x	x	x	x	x	x
DQ085340	Russia	1997	2C	x	x	x	x	x	x	x	x	x	x	x	x	x	x
DQ085341	Mexico	1997	1C		x	x	x	x	x	x	x	x		x	x	x	x
DQ085342	Korea	1996	2B		x	x	x	x	x	x	x	x		x	x	x	x

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
DQ085343	Italy	1997	1E		x	x	x	x	x	x	x	x	x	x	x	x	x
DQ388279	Russia	1967	2C		x	x	x	x	x	x	x	x	x	x	x	x	x
DQ388280	Germany	1992	1G		x	x	x	x	x	x	x	x		x	x	x	x
DQ388281	New Zealand	1991	1D		x	x	x	x	x	x	x	x		x	x	x	x
DQ388282	USA	1998	1C					x									
DQ388283	USA	1999	1C					x									
DQ388284	USA	1998	1E					x									
DQ388285	USA	1998	1E					x									
DQ388286	United Kingdom	1986	1a					x									
DQ388287	USA	1997	1D					x									
DQ388288	USA	1998	1C					x									
DQ388289	USA	1999	1C					x									
DQ388290	USA	1998	1E					x									
DQ388291	Italy	1991	1B					x									
DQ388292	Germany	1992	1B					x									
DQ388293	Italy	1993	2B					x									
DQ388294	Italy	1994	2B					x									
DQ388295	Italy	1997	1E					x									
DQ388296	Korea	1996	1D					x									
DQ388297	Russia	1973	1a					x									
DQ388298	United Kingdom	1978	1B					x									
DQ388299	USA	1997	1C					x									
DQ388301	Israel	1992	1B					x									

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
DQ388302	USA	1998	1C					x									
DQ388303	Canada	1997	1E					x									
DQ388305	United Kingdom	1986	1B					x									
DQ454161	Russia	2005	1H								x			x	x		
DQ454162	Russia	2004	1G								x			x	x		
DQ458965	Brazil	1999	1a								x			x	x		
DQ897934	Russia	2006	-								x			x	x		
DQ897935	Russia	2006	-								x			x	x		
EF182759	Russia	2005	1G								x			x	x		
EF182760	Russia	2005	1G								x			x	x		
EF182761	Russia	2005	1G								x		x	x	x		
EF182762	Russia	2005	1G								x			x	x		
EF182763	Russia	2006	1G								x			x	x		
EF182764	Russia	2006	1G								x			x	x		
EF182765	Russia	2006	1G								x			x	x		
EF199889	Russia	2006	1G								x			x	x		
EF199893	Russia	2004	1G								x			x	x		
EF649760	Russia	2000	2C								x			x	x		
EF649761	Russia	2000	2C								x			x	x		
EF649762	Russia	2000	2C								x			x	x		
EF649763	Russia	2000	2C								x			x	x		
EF649764	Russia	2000	2C								x			x	x		
EF649765	Russia	2000	2C								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
EF649766	Russia	1999	2C								x			x	x		
EF649767	Russia	2004	2C								x			x	x		
EF649768	Russia	2005	2C								x			x	x		
EF649769	Russia	2002	2C								x			x	x		
EF649770	Russia	2002	2C								x			x	x		
EF649771	Russia	2004	2C								x			x	x		
EF672032	Taiwan	2005	1H								x			x	x		
EU240899	Great Britain	2007	2B								x			x	x		
EU240900	Great Britain	2007	2B								x			x	x		
EU518606	Spain	2005	1J								x			x	x		
EU518607	Spain	2004	1J								x			x	x		
EU518608	Spain	2005	1J								x			x	x		
EU518609	Spain	2005	1J								x			x	x		
EU518610	Spain	2005	1J								x			x	x		
EU518611	Spain	2005	1J								x			x	x		
EU518612	Spain	2005	1J								x			x	x		
EU518613	Spain	2005	1J								x			x	x		
EU518614	Spain	2005	1J								x			x	x		
EU518615	Spain	2005	1J								x			x	x		
EU518616	Spain	2005	1J								x			x	x		
EU518617	Spain	2005	1J								x			x	x		
EU518618	Spain	2008	1J								x			x	x		
EU622498	Peru	2004	-								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
EU622499	Peru	2004	-								x			x	x		
EU622500	Peru	2004	-								x			x	x		
EU622501	Peru	2004	-								x			x	x		
EU622502	Peru	2005	-								x			x	x		
EU622503	Peru	2005	-								x			x	x		
EU622504	Peru	2005	-								x			x	x		
EU622505	Peru	2005	-								x			x	x		
EU622506	Peru	2005	-								x			x	x		
FJ436377	China	2008	1E								x			x	x		
FJ436378	China	2008	1E								x			x	x		
FJ656218	China	2008	2B								x			x	x		
FJ656219	China	2008	2B								x			x	x		
FJ711660	Russia	2006	1G								x			x	x		
FJ711661	Russia	2008	1G								x			x	x		
FJ711662	Russia	2008	1G								x			x	x		
FJ711663	Russia	2008	1G								x			x	x		
FJ711664	Russia	2008	1G								x			x	x		
FJ711665	Russia	2008	1G								x			x	x		
FJ711666	Russia	2008	1G								x		x	x	x		
FJ711667	Kazakhstan	2006	-								x		x	x	x		
FJ711668	Russia	2008	1E								x		x	x	x		
FJ711669	Russia	2008	1E								x			x	x		
FJ711670	Russia	2008	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FJ711671	Russia	2008	1E								x			x	x		
FJ711672	Russia	2008	1E								x			x	x		
FJ711673	Russia	2005	1E								x			x	x		
FJ711674	Russia	2007	1E								x			x	x		
FJ711675	Russia	2008	1E								x			x	x		
FJ711676	Russia	2007	1E								x			x	x		
FJ711677	Russia	2008	1E								x			x	x		
FJ711678	Russia	2008	1E								x			x	x		
FJ711679	Russia	2008	1E								x			x	x		
FJ711680	Russia	2008	1E								x			x	x		
FJ711681	Russia	2008	1E								x			x	x		
FJ711682	Russia	2006	1E								x			x	x		
FJ711683	Ukraine	2007	1E								x			x	x		
FJ711684	Kazakhstan	2006	1E								x			x	x		
FJ711685	Kazakhstan	2008	2B								x			x	x		
FJ711686	Russia	2008	1H								x			x	x		
FJ711687	Russia	2007	1H								x			x	x		
FJ711688	Russia	2007	1H								x			x	x		
FJ711689	Russia	2004	1H								x			x	x		
FJ711690	Russia	2008	1H								x			x	x		
FJ711691	Russia	2007	1H								x			x	x		
FJ711692	Russia	2008	1H								x			x	x		
FJ711693	Russia	2008	1H								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FJ711694	Russia	2006	1H								x			x	x		
FJ711695	Russia	2008	1H								x			x	x		
FJ711696	Russia	2008	1H								x			x	x		
FJ711697	Russia	2004	1H								x			x	x		
FJ711698	Russia	2004	1H								x			x	x		
FJ711699	Kazakhstan	2008	1H								x			x	x		
FJ774999	Sudan	2005	1G								x			x	x		
FJ775000	Sudan	2005	1E								x			x	x		
FJ875029	China	2000	1E								x			x	x		
FJ875030	China	2001	2A								x			x	x		
FJ875031	China	2001	2A								x			x	x		
FJ875032	China	2000	2A								x			x	x		
FJ875033	China	2006	1E								x			x	x		
FJ875034	China	1999	1F								x			x	x		
FJ875035	China	1999	1F								x			x	x		
FJ875036	China	2001	1E								x			x	x		
FJ875037	China	2001	1E								x			x	x		
FJ875038	China	2001	1E								x			x	x		
FJ875039	China	2007	1E								x			x	x		
FJ875040	China	2007	1E								x			x	x		
FJ875041	China	2007	1E								x			x	x		
FJ875042	China	2007	1E								x			x	x		
FJ875043	China	2007	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FJ875044	China	2001	1E								x			x	x		
FJ875045	China	2002	1F								x			x	x		
FJ875046	China	2002	1F								x			x	x		
FJ875047	China	2006	1E								x			x	x		
FJ875048	China	2007	1E								x			x	x		
FJ875049	China	2003	1E								x			x	x		
FJ875050	China	2003	1E								x			x	x		
FJ875051	China	2005	1E								x			x	x		
FJ875052	China	2007	1E								x			x	x		
FJ875053	China	2007	1E								x			x	x		
FJ875054	China	2004	1E								x			x	x		
FJ875055	China	2005	1E								x			x	x		
FJ875056	China	2006	1E								x			x	x		
FJ875057	China	2006	2B								x			x	x		
FJ875058	China	2007	1E								x			x	x		
FJ875059	China	2007	1E								x			x	x		
FJ875060	China	2007	1E								x			x	x		
FJ875061	China	2006	1E								x			x	x		
FJ875062	China	2006	1E								x			x	x		
FJ875063	China	2007	1E								x			x	x		
FJ875064	China	2007	1E								x			x	x		
FJ875065	China	2007	1E								x			x	x		
FJ875066	China	2007	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FJ875067	China	2007	1E								x			x	x		
FJ875068	China	2007	1E								x			x	x		
FJ875069	China	2007	1E								x			x	x		
FJ875070	China	2007	1E								x			x	x		
FJ875071	China	2007	1E								x			x	x		
FJ971761	Argentina	2008	2B								x			x	x		
FJ971762	Argentina	2008	2B								x			x	x		
FJ971763	Argentina	2008	2B								x			x	x		
FJ971764	Argentina	2008	2B								x			x	x		
FJ971765	Argentina	2008	2B								x			x	x		
FJ971766	Argentina	2008	2B								x			x	x		
FJ971767	Argentina	2008	2B								x			x	x		
FJ971768	Argentina	2008	2B								x			x	x		
FJ971769	Argentina	2008	2B								x			x	x		
FJ971770	Argentina	2008	2B								x			x	x		
FJ971771	Argentina	2008	2B								x			x	x		
FJ971772	Argentina	2008	2B								x			x	x		
FJ971773	Argentina	2008	2B								x			x	x		
FJ971774	Argentina	2008	2B								x		x	x	x		
FJ971775	Argentina	2008	2B								x			x	x		
FJ971776	Argentina	2008	2B								x			x	x		
FJ971777	Argentina	2008	2B								x			x	x		
FJ971778	Argentina	2008	2B								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FJ971779	Argentina	2008	2B								x			x	x		
FJ971780	Argentina	2008	2B								x			x	x		
FJ971781	Argentina	2008	2B								x			x	x		
FJ971782	Argentina	2008	2B								x			x	x		
FJ971783	Argentina	2008	2B								x			x	x		
FN546966	France	1995	-								x			x	x		
FN546967	France	1995	1E								x			x	x		
FN546968	France	1995	1H								x			x	x		
FN546969	France	1996	1G								x			x	x		
FN546970	France	1997	1E								x			x	x		
FN546971	France	1997	1E								x		x	x	x		
FN546972	France	1997	1E								x			x	x		
FN546973	France	1997	1E								x			x	x		
FN546974	France	1997	1E								x			x	x		
FN546975	France	1997	1E								x			x	x		
FN546976	France	1997	1E								x			x	x		
FN546977	France	1997	1E								x			x	x		
FN546978	France	1997	1E								x			x	x		
FN546979	France	1997	1E								x			x	x		
FN546980	France	1997	1E								x			x	x		
FN546981	France	1997	1E								x			x	x		
FN546982	France	1997	1E								x		x	x	x		
FN546983	France	1998	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FN546984	Portugal	1998	1E								x			x	x		
FN546985	France	1998	1G								x			x	x		
FN546986	France	1999	1E								x			x	x		
FN546987	France	1999	1E								x			x	x		
FN546988	France	1999	1E								x			x	x		
FN546989	France	1999	1E								x			x	x		
FN546990	France	1999	1E								x			x	x		
FN546991	France	1999	1E								x		x	x	x		
FN546992	France	1999	1E								x			x	x		
FN546993	France	1999	1E								x			x	x		
FN546994	France	2000	1E								x			x	x		
FN546995	France	2000	1E								x			x	x		
FN546996	France	2000	1E								x			x	x		
FN546997	France	2000	1E								x			x	x		
FN546998	France	2000	1E								x			x	x		
FN546999	France	2000	1E								x			x	x		
FN547000	France	2000	1E								x			x	x		
FN547002	France	2001	1E								x			x	x		
FN547003	France	2001	1E								x			x	x		
FN547004	France	2001	1E								x			x	x		
FN547005	France	2001	1B								x			x	x		
FN547006	France	2002	1E								x			x	x		
FN547007	France	2002	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FN547008	France	2002	1E								x			x	x		
FN547009	France	2002	1E								x			x	x		
FN547010	France	2002	1E								x			x	x		
FN547011	France	2002	1E								x			x	x		
FN547012	France	2002	1E								x			x	x		
FN547013	France	2002	1E								x			x	x		
FN547014	Tunisia	2003	1E								x			x	x		
FN547015	France	2003	1E								x			x	x		
FN547016	France	2003	1E								x			x	x		
FN547017	France	2004	2B								x			x	x		
FN547018	France	2004	1E								x			x	x		
FN547019	France	2005	1E								x			x	x		
FN547020	France	2005	1E								x			x	x		
FN547021	France	2009	2B								x			x	x		
FR717206	Bosnia-Herzegovina	2009	2B								x			x	x		
FR717207	Bosnia-Herzegovina	2009	2B								x			x	x		
FR717208	Bosnia-Herzegovina	2009	2B								x			x	x		
FR717209	Bosnia-Herzegovina	2009	2B								x			x	x		
FR717210	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717211	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717212	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717213	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717214	Bosnia-Herzegovina	2010	2B								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
FR717215	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717216	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717217	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717218	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717219	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717220	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717221	Bosnia-Herzegovina	2010	2B								x			x	x		
FR717222	Bosnia-Herzegovina	2010	2B								x			x	x		
GQ329848	Brazil	2005	1J								x			x	x		
GQ329849	Brazil	2005	1J								x			x	x		
GQ329850	Brazil	2005	1J								x			x	x		
GU174756	Canada	2009	2B								x			x	x		
GU254251	Brazil	2006	2B								x			x	x		
GU254252	Brazil	2007	2B								x			x	x		
GU254253	Brazil	2007	2B								x			x	x		
GU254254	Brazil	2008	2B								x			x	x		
GU254255	Brazil	2008	2B								x			x	x		
GU289729	China	2009	2B								x			x	x		
GU289730	China	2009	1E								x			x	x		
GU289731	China	2009	2B								x			x	x		
GU353072	USA	2008	2B								x			x	x		
GU353076	USA	2005	1E								x			x	x		
GU968187	Brazil	2007	1a								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
HM211177	China	2010	1E								x			x	x		
HM212630	Brazil	2008	2B								x			x	x		
HM212631	Brazil	2008	2B								x			x	x		
HM212632	Brazil	2009	2B								x			x	x		
HM212633	Brazil	2008	2B								x			x	x		
HM212634	Brazil	2009	2B								x			x	x		
HM461998	China	2010	1J								x			x	x		
HQ199838	China	2010	1E								x			x	x		
HQ893749	Vietnam	2010	2B								x			x	x		
HQ893750	Vietnam	2010	2B								x			x	x		
HQ893751	Vietnam	2010	2B								x			x	x		
HQ893752	Vietnam	2010	2B								x			x	x		
HQ893753	Vietnam	2010	2B								x			x	x		
HQ893754	Vietnam	2010	2B								x			x	x		
HQ893755	Vietnam	2010	2B								x		x	x	x		
HQ893756	Vietnam	2010	2B								x			x	x		
HQ893757	Vietnam	2010	2B								x			x	x		
HQ893758	Vietnam	2010	2B								x		x	x	x		
JF702819	China	2009	1E								x			x	x		
JF702820	China	2009	1E								x			x	x		
JF702821	China	2009	1E								x			x	x		
JF702822	China	2009	1E								x			x	x		
JF702823	China	2009	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JF702824	China	2009	1E								x			x	x		
JF702825	China	2009	1E								x			x	x		
JF702826	China	2009	1E								x			x	x		
JF702827	China	2009	1E								x			x	x		
JF702828	China	2009	1E								x			x	x		
JF702829	China	2009	1E								x			x	x		
JF702830	China	2009	1E								x			x	x		
JF702831	China	2009	1E								x			x	x		
JF702832	China	2009	1E								x			x	x		
JF702833	China	2009	1E								x			x	x		
JF702834	China	2009	1E								x			x	x		
JF702835	China	2009	1E								x			x	x		
JF702836	China	2009	1E								x			x	x		
JF702837	China	2009	1E								x			x	x		
JF702838	China	2009	1E								x			x	x		
JF702839	China	2009	1E								x			x	x		
JF702840	China	2009	1E								x		x	x	x		
JF702841	China	2009	1E								x			x	x		
JF702842	China	2009	1E								x			x	x		
JF702843	China	2009	1E								x			x	x		
JF702844	China	2009	1E								x			x	x		
JF702845	China	2008	1E								x			x	x		
JF702846	China	2008	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JF702847	China	2008	1E								x			x	x		
JF702848	China	2008	1E								x			x	x		
JF702849	China	2008	1E								x			x	x		
JF702850	China	2008	1E								x			x	x		
JF702851	China	2008	1E								x			x	x		
JF702852	China	2008	1E								x			x	x		
JF702853	China	2008	1E								x			x	x		
JF702854	China	2008	1E								x			x	x		
JF702855	China	2008	1E								x			x	x		
JF702856	China	2008	1E								x			x	x		
JF702858	China	2008	1E								x			x	x		
JF702859	China	2008	1E								x			x	x		
JF702860	China	2008	1E								x			x	x		
JF702861	China	2008	1E								x			x	x		
JF702862	China	2008	1E								x			x	x		
JF702863	China	2008	1E								x			x	x		
JF702864	China	2008	1E								x			x	x		
JF702865	China	2008	1E								x			x	x		
JF702866	China	2008	1E								x		x	x	x		
JF702867	China	2008	1E								x			x	x		
JF702868	China	2008	1E								x			x	x		
JF702869	China	2008	1E								x			x	x		
JF702870	China	2008	2B								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JF702871	China	2008	2B								x			x	x		
JF911797	Canada	2011	2B								x			x	x		
JN036398	China	2011	1E								x			x	x		
JN036399	China	2011	1E								x			x	x		
JN544447	China	2011	1E								x			x	x		
JN544448	China	2011	1E								x			x	x		
JN575762	Canada	2011	1J								x			x	x		
JN582035	Argentina	2009	2B								x			x	x		
JN582036	Argentina	2009	2B								x			x	x		
JN582037	Argentina	2009	2B								x			x	x		
JN635281	USA	1961	1a	x	x	x	x	x	x	x	x	x	x	x	x	x	x
JN635282	USA	1998	1B	x	x	x	x	x	x	x	x	x		x	x	x	x
JN635283	USA	1991	1C		x	x	x	x	x	x	x	x		x	x	x	x
JN635284	USA	1998	1C	x	x	x	x	x	x	x	x	x		x	x	x	x
JN635285	USA	1988	1D		x		x	x	x	x	x	x	x				
JN635286	USA	2008	1E	x	x	x	x	x	x	x	x	x		x	x	x	x
JN635287	USA	1998	1E		x	x	x	x	x	x	x	x	x	x	x	x	x
JN635288	USA	2008	1E		x	x	x	x	x	x	x	x		x	x	x	x
JN635289	USA	2007	1G		x	x	x	x	x	x	x	x		x	x	x	x
JN635290	USA	2005	1G	x	x	x	x	x	x	x	x	x		x	x	x	x
JN635291	USA	1997	1J	x	x	x	x	x	x	x	x	x		x	x	x	x
JN635292	USA	2007	2B		x	x	x	x	x	x	x	x		x	x	x	x
JN635293	USA	2000	2B		x	x	x	x	x	x	x	x		x	x	x	x

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JN635294	USA	2008	2B	x	x	x	x	x	x	x	x	x		x	x	x	x
JN635295	USA	2009	2B		x	x	x	x	x	x	x	x		x	x	x	x
JN635296	USA	2008	2B		x	x	x	x	x	x	x	x		x	x	x	x
JN661163	Malaysia	2011	2B								x			x	x		
JN661164	Malaysia	2011	2B								x			x	x		
JN661165	Malaysia	2011	2B								x			x	x		
JN661166	Malaysia	2011	2B								x			x	x		
JN661167	Malaysia	2011	2B								x			x	x		
JN661168	Malaysia	2011	2B								x			x	x		
JN661169	Malaysia	2011	2B								x			x	x		
JN661170	Malaysia	2011	2B								x			x	x		
JN661171	Malaysia	2011	2B								x			x	x		
JN661172	Malaysia	2011	2B								x			x	x		
JN827384	China	2011	2B								x			x	x		
JQ031213	China	2011	2B								x			x	x		
JQ283993	India	2011	2B								x			x	x		
JQ283994	India	2007	2B								x			x	x		
JQ283995	India	2008	2B								x			x	x		
JQ413980	India	2009	2B								x		x	x	x		
JQ639404	China	2006	1E								x			x	x		
JQ639405	China	2006	1E								x			x	x		
JQ639406	China	2006	1E								x			x	x		
JQ639407	China	2007	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JQ639408	China	2007	1E								x			x	x		
JQ639409	China	2007	1E								x			x	x		
JQ639410	China	2007	1E								x			x	x		
JQ639411	China	2008	1E								x			x	x		
JQ639412	China	2009	1E								x			x	x		
JQ979489	China	2007	1E								x			x	x		
JQ979490	China	2007	1E								x			x	x		
JQ979491	China	2007	1E								x			x	x		
JQ979492	China	2007	1E								x			x	x		
JQ979493	China	2007	1E								x			x	x		
JQ979494	China	2007	1E								x			x	x		
JQ979495	China	2007	1E								x			x	x		
JQ979496	China	2007	1E								x			x	x		
JQ979497	China	2007	1E								x			x	x		
JQ979498	China	2007	1E								x			x	x		
JQ979499	China	2007	1E								x			x	x		
JQ979500	China	2007	1E								x			x	x		
JQ979501	China	2008	1E								x			x	x		
JQ979502	China	2008	1E								x			x	x		
JQ979503	China	2008	1E								x			x	x		
JQ979504	China	2008	1E								x			x	x		
JQ979505	China	2008	1E								x			x	x		
JQ979506	China	2008	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JQ979507	China	2008	1E								x			x	x		
JQ979508	China	2008	1E								x			x	x		
JQ979509	China	2008	1E								x			x	x		
JQ979510	China	2008	1E								x			x	x		
JQ979511	China	2008	1E								x			x	x		
JQ979512	China	2008	1E								x			x	x		
JQ979513	China	2008	1E								x			x	x		
JQ979514	China	2008	1E								x			x	x		
JQ979515	China	2008	1E								x			x	x		
JQ979516	China	2008	1E								x			x	x		
JQ979517	China	2008	1E								x			x	x		
JQ979518	China	2008	1E								x			x	x		
JQ979519	China	2008	1E								x			x	x		
JQ979520	China	2009	1E								x			x	x		
JQ979521	China	2009	1E								x			x	x		
JQ979522	China	2009	1E								x			x	x		
JQ979523	China	2009	1E								x			x	x		
JQ979524	China	2010	1E								x			x	x		
JQ979525	China	2010	1E								x			x	x		
JQ979526	China	2010	1E								x			x	x		
JQ979527	China	2010	1E								x			x	x		
JQ979528	China	2010	1E								x			x	x		
JQ979529	China	2010	1E								x		x	x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JQ979530	China	2010	1E								x			x	x		
JQ979531	China	2010	1E								x			x	x		
JQ979532	China	2010	1E								x			x	x		
JQ979533	China	2010	1E								x			x	x		
JQ979534	China	2010	1E								x			x	x		
JQ979535	China	2010	1E								x			x	x		
JQ979536	China	2010	1E								x			x	x		
JQ979537	China	2010	1E								x			x	x		
JQ979538	China	2010	1E								x			x	x		
JQ979539	China	2011	1E								x			x	x		
JQ979540	China	2011	1E								x			x	x		
JQ979541	China	2011	1E								x			x	x		
JQ979542	China	2011	1E								x			x	x		
JQ979543	China	2011	1E								x			x	x		
JQ979544	China	2011	1E								x			x	x		
JQ979545	China	2011	1E								x			x	x		
JQ979546	China	2011	1E								x			x	x		
JQ979547	China	2011	1E								x			x	x		
JQ979548	China	2011	1E								x			x	x		
JQ979549	China	2011	1E								x			x	x		
JQ979550	China	2011	1E								x			x	x		
JQ979551	China	2011	1E								x			x	x		
JQ979552	China	2011	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JX036507	China	2012	2B								x			x	x		
JX036508	China	2012	2B								x			x	x		
JX036509	China	2012	1E								x			x	x		
JX036510	China	2012	1E								x			x	x		
JX112763	China	2012	2B								x			x	x		
JX112764	China	2012	1E								x			x	x		
JX112765	China	2012	1E								x			x	x		
JX171315	Russia	2011	1E								x			x	x		
JX398300	Great Britain	2010	2B								x			x	x		
JX398301	Great Britain	2011	2B								x			x	x		
JX398302	Great Britain	2012	2B								x			x	x		
JX398303	Great Britain	2012	2B								x			x	x		
JX398304	Great Britain	2012	2B								x			x	x		
JX398305	Great Britain	2012	2B								x			x	x		
JX398306	Great Britain	2012	2B								x			x	x		
JX398307	Great Britain	2012	2B								x			x	x		
JX398308	Great Britain	2012	2B								x			x	x		
JX398309	Great Britain	2012	2B								x			x	x		
JX398310	Great Britain	2012	1G								x			x	x		
JX398311	Great Britain	2012	1G								x			x	x		
JX398312	Great Britain	2012	1G								x			x	x		
JX398313	Great Britain	2012	2B								x			x	x		
JX477651	USA	2011	1E								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset													
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JX477652	USA	2012	1E								x			x	x		
JX477653	USA	2010	1J								x			x	x		
JX477654	USA	2010	1G								x			x	x		
JX477655	USA	2012	1G								x			x	x		
JX477656	USA	2012	1G								x			x	x		
JX477657	USA	2010	2B								x			x	x		
JX477658	USA	2010	2B								x			x	x		
JX477659	USA	2011	2B								x			x	x		
JX477660	USA	2011	2B								x			x	x		
JX477661	USA	2012	2B								x			x	x		
JX477662	USA	2012	2B								x		x	x	x		
JX531652	China	2012	1E								x			x	x		
JX646676	Mexico	2012	2B								x			x	x		
JX679257	Peru	2005	1C								x			x	x		
JX679258	Peru	2004	1C								x			x	x		
JX679259	Peru	2004	1C								x			x	x		
JX679260	Peru	2004	1C								x			x	x		
JX679261	Peru	2004	1C								x			x	x		
JX679262	China	2008	1E								x			x	x		
JX679263	Guyana	1997	1E								x			x	x		
JX679264	Guyana	1997	1E								x			x	x		
JX679265	Cote d'Ivoire	2008	1G								x			x	x		
JX679266	Cote d'Ivoire	2008	1G								x			x	x		

Appendix 3. Full description of datasets used in thesis. Continued…

GenBank Accession No.	Country of sampling	Collection Date	Genotype	Dataset														
				i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv
JX679267	Cote d'Ivoire	2008	1G								x			x	x		
JX679268	Cote d'Ivoire	2008	1G								x			x	x		
JX679269	Cote d'Ivoire	2008	1G								x			x	x		
JX679270	Cote d'Ivoire	2008	1G								x			x	x		
JX679271	Cote d'Ivoire	2008	1G								x			x	x		
JX679272	Cote d'Ivoire	2008	1G								x			x	x		
JX679273	Ghana	2008	1G								x			x	x		
JX679274	Ghana	2005	1G								x			x	x		
JX679275	Ghana	2004	1G								x			x	x		
JX679276	Ghana	2005	1G								x			x	x		
JX679277	Belize	1994	1C								x			x	x		
JX679278	Honduras	2004	1C								x			x	x		
JX679279	Peru	2005	1C								x			x	x		
JX679280	Peru	2004	1C								x			x	x		
JX679281	Peru	2004	1C								x			x	x		
JX679282	Peru	2004	1C								x			x	x		
JX913763	China	2012	1E								x			x	x		
KC138719	China	2012	1E								x			x	x		
KC288128	India	1992	2B								x			x	x		
KC288129	India	1992	2B								x			x	x		
M15240	USA	1964	1a		x	x	x	x	x	x	x	x	x	x	x	x	x

Appendix 4. Python script written to generate temporally balanced random subsamples. All isolates are sorted into their respective decades. The user is then queried about the number of isolates that should be selected from each decade. The user-specific number of isolates from each decade are then randomly selected, and written to file as a new alignment. The user can specify how many times this process should be repeated.

Usage: python script_name.py <input.fasta> <date_of_oldest_isolate> <date_of_most_recent_isolate> <number of_random_output_replicates>

```
import sys
from Bio import SeqIO
from random import sample

seq_fasta_file = open(sys.argv[1])
start_decade = sys.argv[2]
end_decade = sys.argv[3]

start_decade = str(start_decade[:-1]) + "0"
end_decade = str(end_decade[:-1]) + "0"
end_decade = int(end_decade)

decades_dict = {}
decade = int(start_decade)
while decade <= end_decade:
        decades_dict[decade] = []
        decade += 10

count = int(start_decade[2:3])
decade = int(start_decade)
number_per_decade_dict = {}
while decade <= end_decade:
        for record in SeqIO.parse(seq_fasta_file, "fasta"):
```

```
                    if count > 9:
                            count = 0
                    name_split = record.id.split("_")
                    if len(name_split[2]) == 4:
                            name_split = int(name_split[2][2:3])
                    else:
                            name_split = int(name_split[2][0])
                    if name_split == count:
                            decades_dict[decade].append(">" + str(record.id) + "\n" + str(record.seq)
                            + "\n")
            count += 1
            number_per_decade = raw_input("There are "+str(len(decades_dict[decade])) + " samples
            from " + str(decade) + ". Amount to use? ")
            number_per_decade_dict[decade] = int(number_per_decade)
            decade += 10
            seq_fasta_file.seek(0)
seq_fasta_file.close()


number_output_files = sys.argv[4]
output_name_count = 1
while output_name_count <= int(number_output_files):
            output_file = open("output_filename_" + str(output_name_count) + ".fasta", "w")
            decade = int(start_decade)
            while decade <= end_decade:
                    random_sample = sample(decades_dict[decade],
                    int(number_per_decade_dict[decade]))
                    output_file.writelines(random_sample)
                    decade += 10
            output_name_count += 1
            output_file.close()
print ('Script Done")
```

Appendix 5. H-clust R script used to group sequences into geographically proximate regions. This hierarchical clustering method defines optimal geographical groupings, using the centroid geocoordinates of each sampling location.

Usage: R-package

```
library(SoDA)
library(maps)
library(mapdata)


distGPS <- function(input)
{
 dMat <- matrix(0,ncol=nrow(input),nrow=nrow(input))
 colnames(dMat) <- input[,1]
 rownames(dMat) <- input[,1]
 for(i in 1:(nrow(input)-1))  {
  for(j in (i+1):nrow(input))
  {
   a <- geoDist(input[i,2],input[i,3],input[j,2],input[j,3])
   dMat[i,j] <- a
   dMat[j,i] <- a
  }
 }
dMat
centroid <- function(clustout,input)
{
 seqGrp <- lapply(clustout,names)
 out <- lapply(seqGrp,centerGrp,input)
 sequenceN <- unlist(seqGrp)
 Lon <- NULL
 Lat <- NULL
 Nom <- NULL
for(i in 1:length(sequenceN))
 {
```

```
Lon <- c(Lon,input[,3][input[,1]==sequenceN[i]])
   Lat <- c(Lat,input[,2][input[,1]==sequenceN[i]])
   Nom <- c(Nom,sequenceN[i])
 }
 Grp <- NULL
 X <- NULL
 Y <- NULL
 for(i in 1:length(out))
 {
   coord <- out[[i]]
   seqName <- seqGrp[[i]]
   Nom <- c(Nom,seqName)
   Grp <- c(Grp,rep(i,length(seqName)))
   X <- c(X,rep(out[[i]][1],length(seqName)))
   Y <- c(Y,rep(out[[i]][2],length(seqName)))
 }
 result <- data.frame(nom=Nom,Grp=Grp,CentroidLon=X,CentroidLat=Y,Lon=Lon,Lat=Lat)
}


centerGrp <- function(X,input)
{
 dataGrp <- input[match(X,input[,1]),]
 Xm <- mean(as.numeric(as.character(dataGrp[,3])))
 Ym <- mean(as.numeric(as.character(dataGrp[,2])))
 out <- c(Xm,Ym)
 out
}


setwd("/output/working/directory/")
geocoordinates <- read.csv("geocoordinates.csv", dec=".", sep=",", header=TRUE)
```

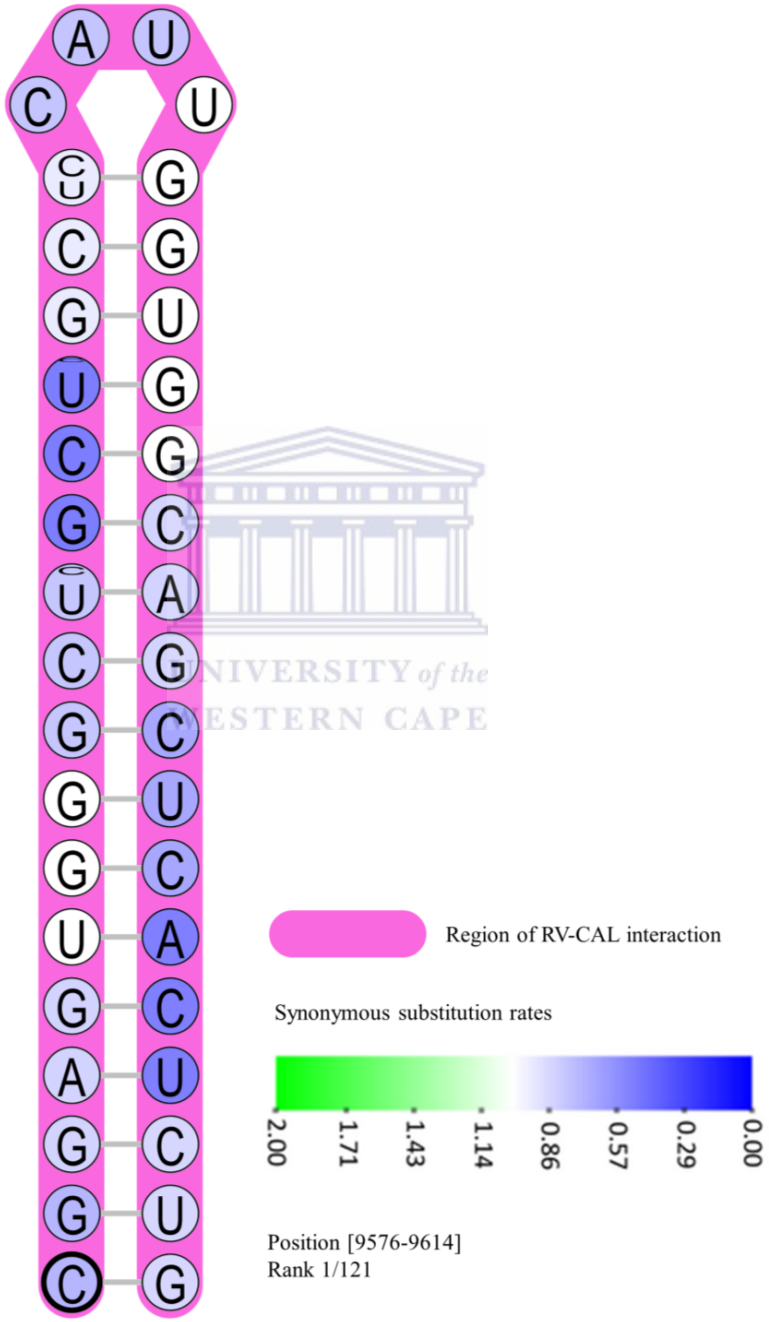
```
m <- c("complete","ward", "single", "centroid", "average", "mcquitty" , "median")

pdf("clustering.pdf")

for(i in 1:length(m))
plot(hclust(as.dist(distance_matrix_GPS),method=m[i]),ylim=c(0,1),cex=0.5,main=m[i],xlab="",ylab="")

dev.off()



hierachical_clustering_GPS <- hclust(as.dist(distance_matrix_GPS),method="complete")

plot(hierachical_clustering_GPS,ylim=c(0,1),cex=0.5)

clustering_out <- identify(hierachical_clustering_GPS)

grouping <- centroid(clustering_out,geocoordinates)
```

Appendix 6. R script written to estimate great circle distances (as-the-crow-flies). This script implements the Haversine formula to determine the pairwise great circle distances between the centroid geocoordinates of all sampling locations.

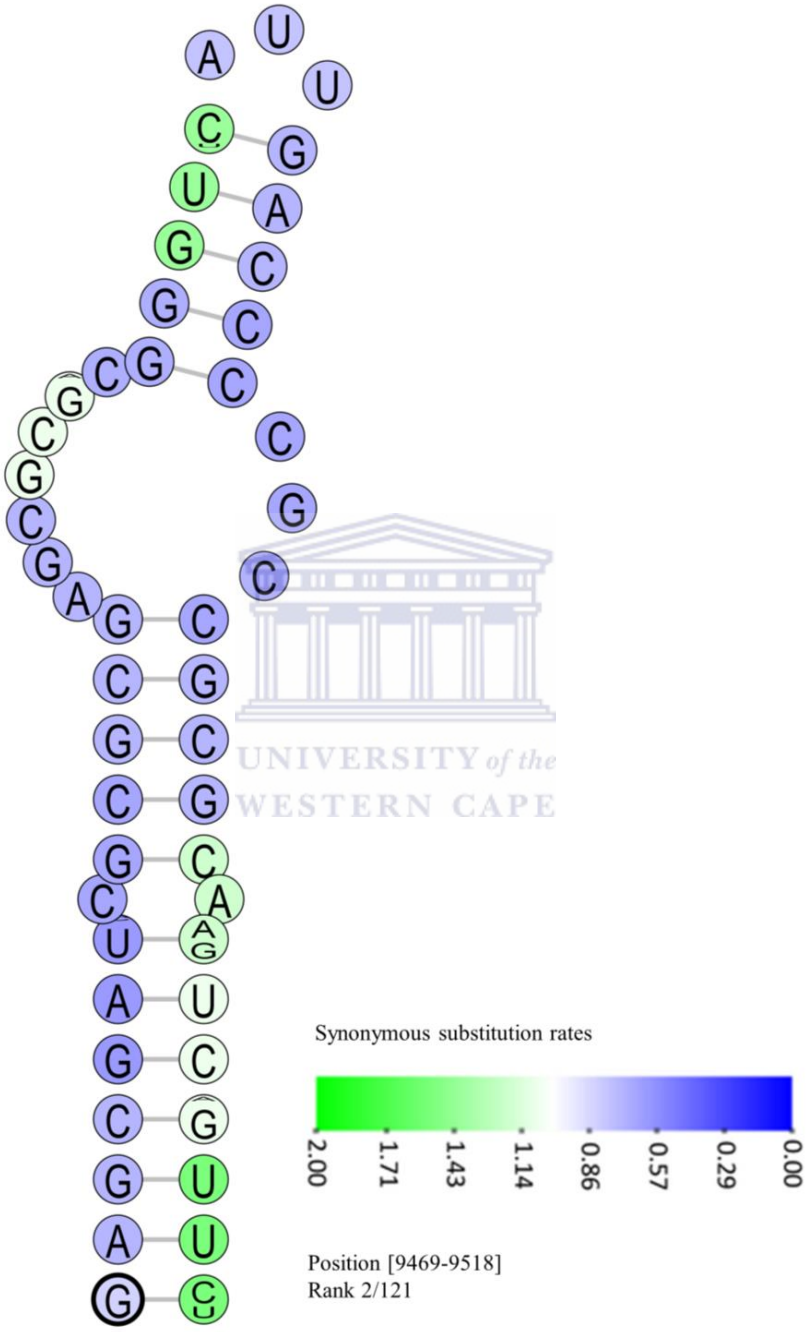
Usage: R-package

```
library (fields)


setwd("/output/working/directory/")

latlong_mat <- read.csv("locations_geoccordinates.csv",sep=",",na.strings="",row.names=1)


great_dist_mat <- rdist.earth(matrix(c(latlong_mat$Longitude,latlong_mat$Latitude),
ncol=2),matrix(c(latlong_mat$Longitude,latlong_mat$Latitude), ncol=2), miles=FALSE, R=6371)


rownames (great_dist_mat) <- rownames(latlong_mat)

colnames (great_dist_mat) <- rownames(latlong_mat)

diag(great_dist_mat) <- 0


write.table((great_dist_mat), file = "great_circ_dist.csv", sep = ';', col.names=NA)
```

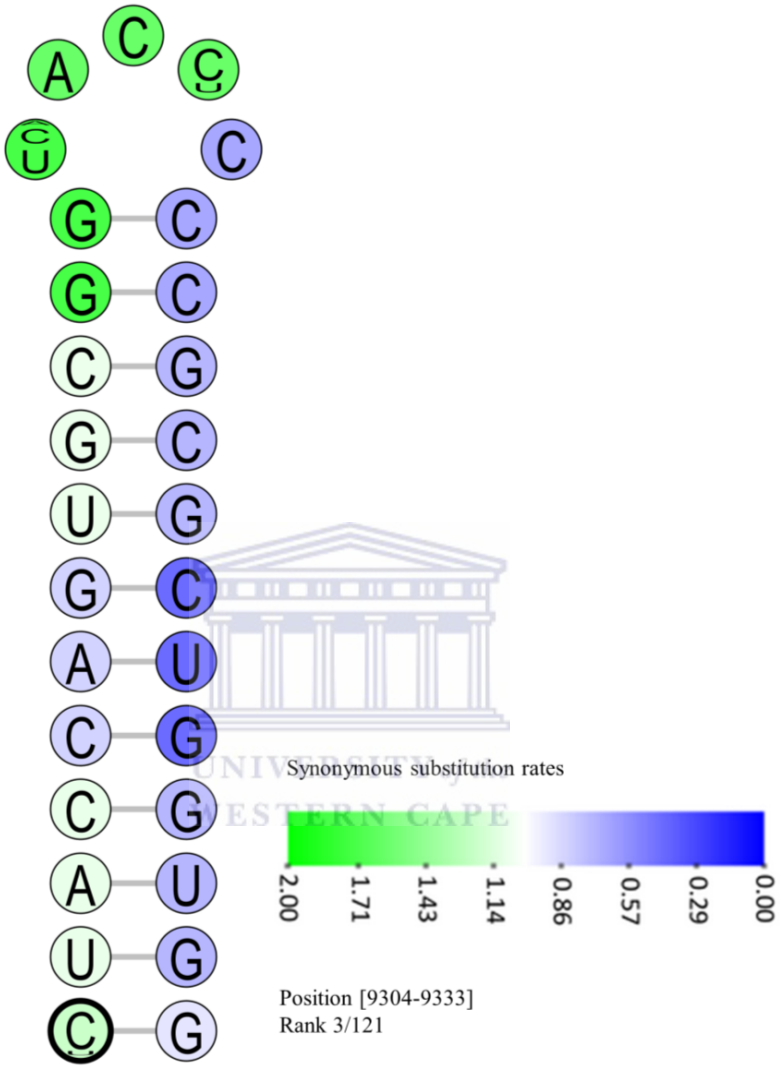
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. The rank refers to the DOOSS consensus rank of the specific predicted nucleotide secondary structure as it forms part of the high confidence structure set (HCSS; see Figure 9 and Table 5). Site-to-site variations in synonymous nucleotide substitution rates are indicated by colours ranging from blue to green (see colour key). Nucleotides falling outside the coding region are shaded in grey.



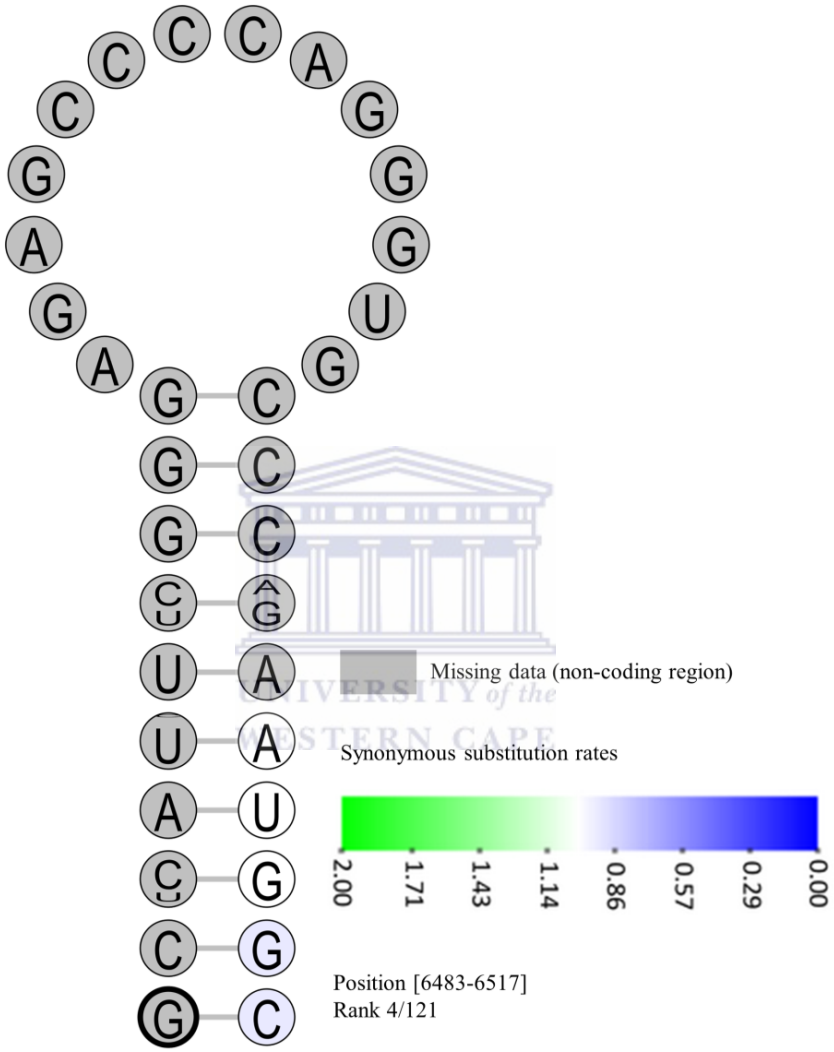
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



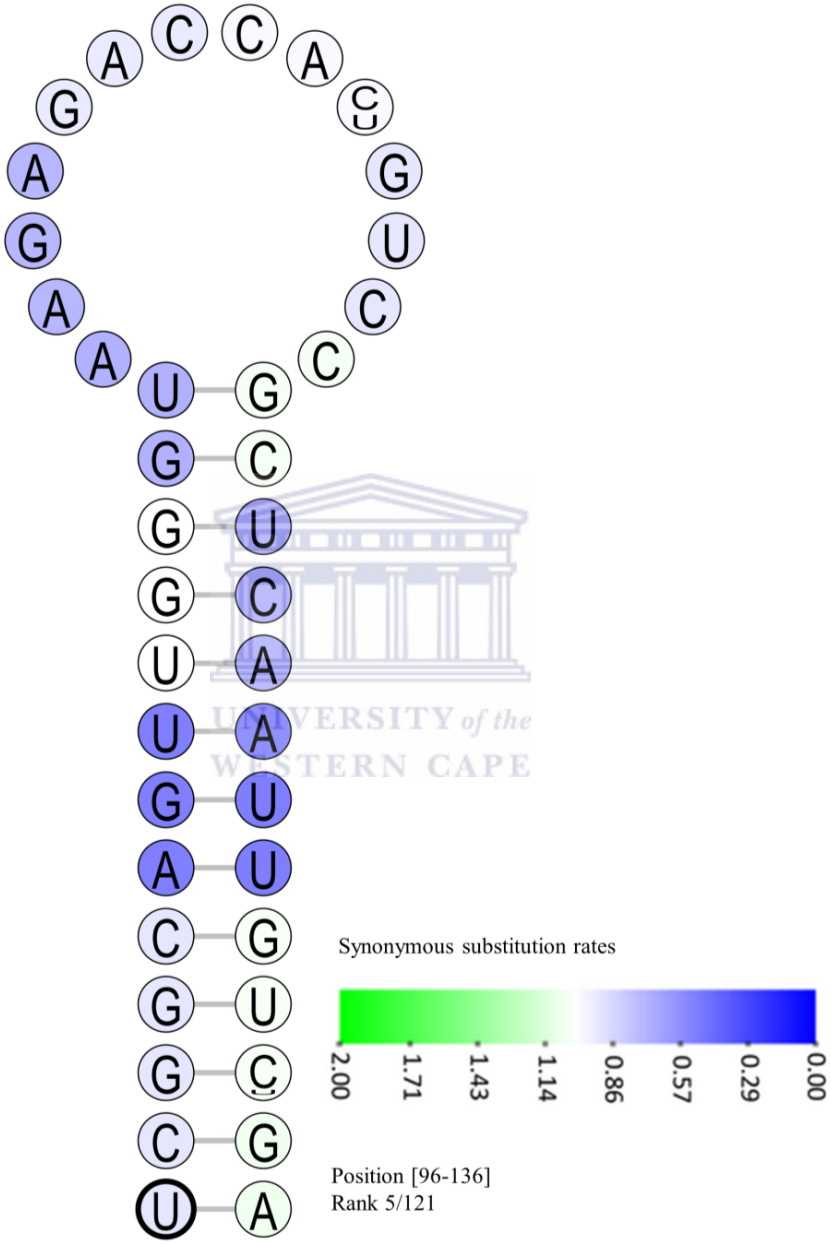
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



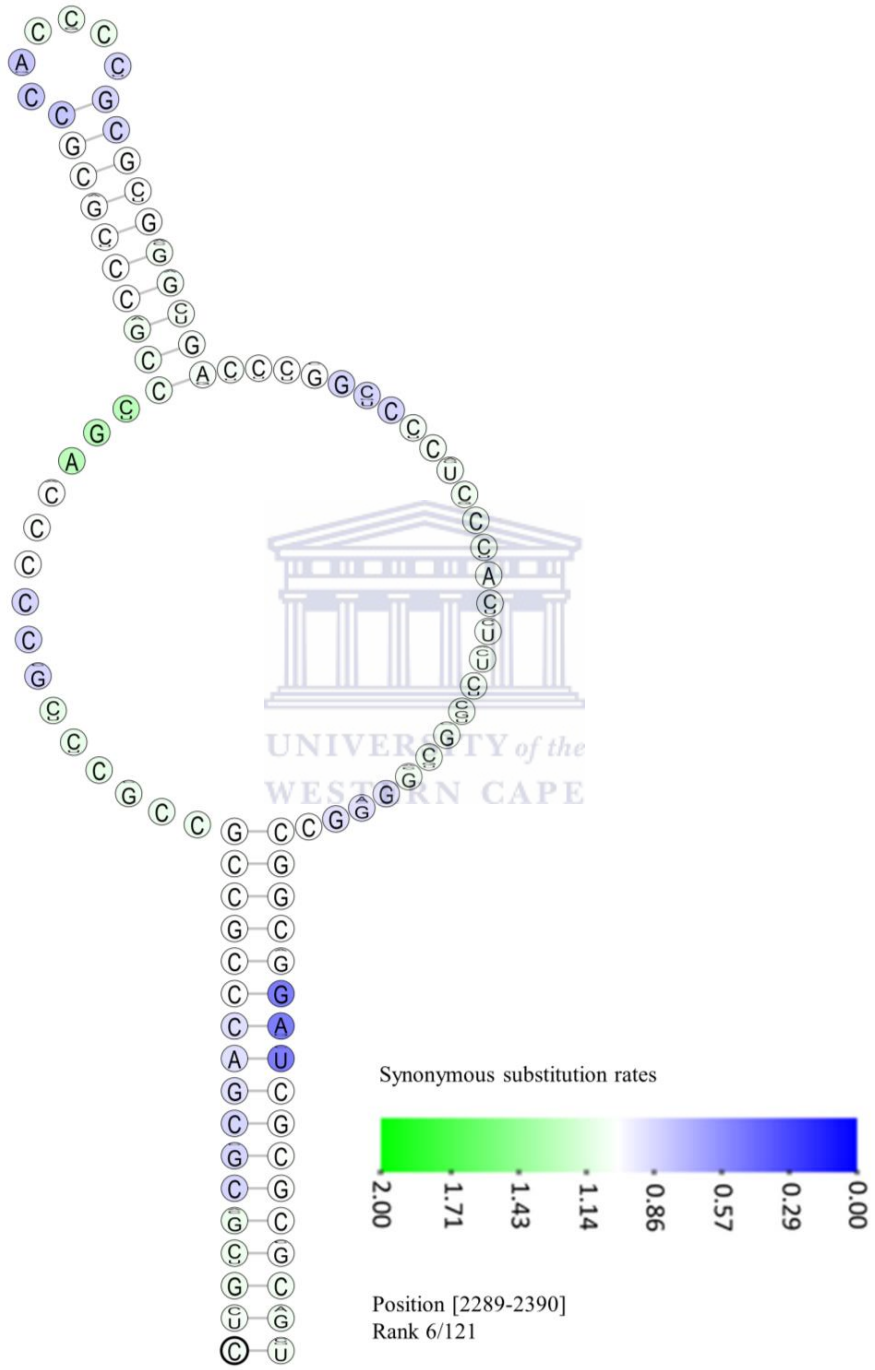
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



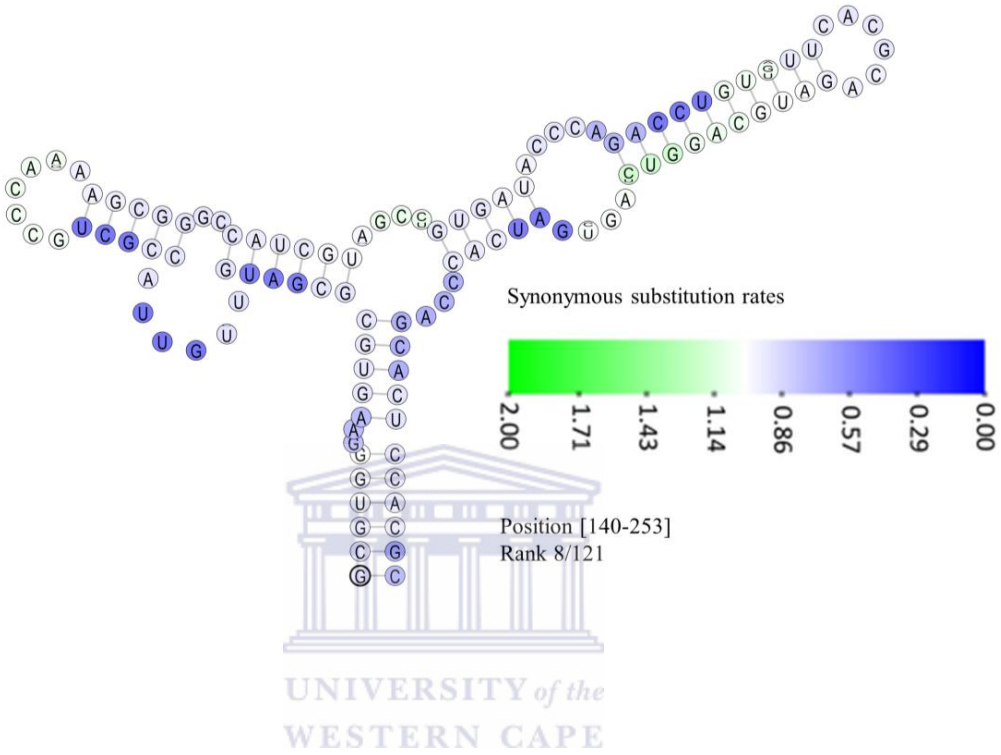
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



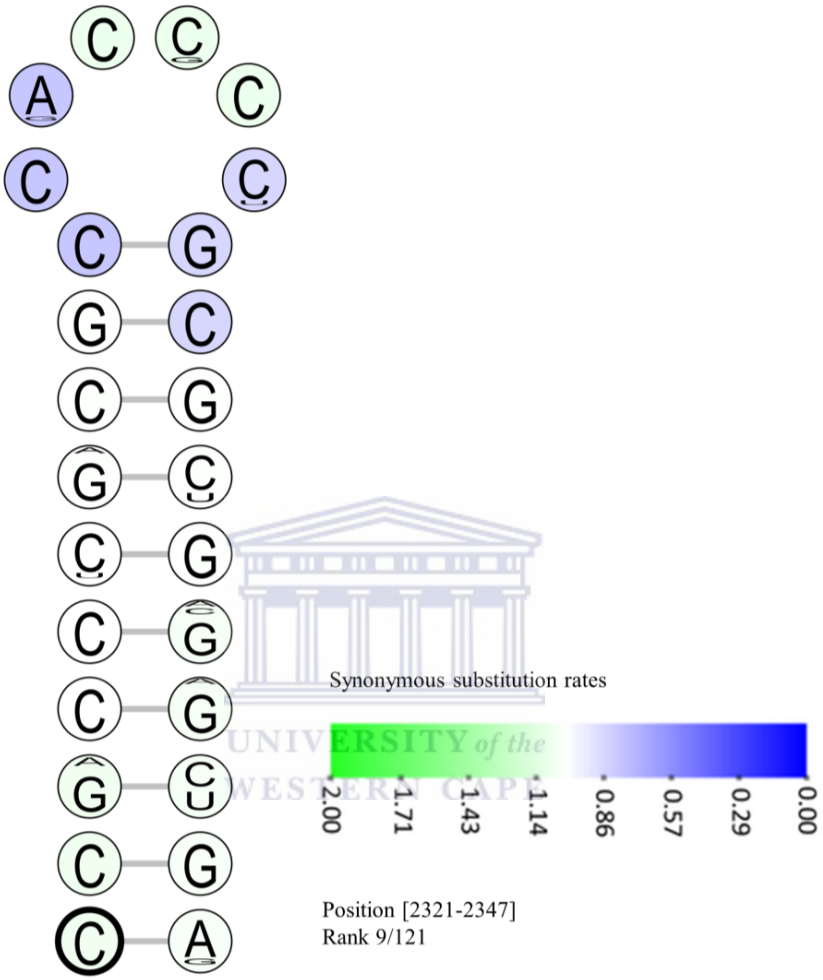
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



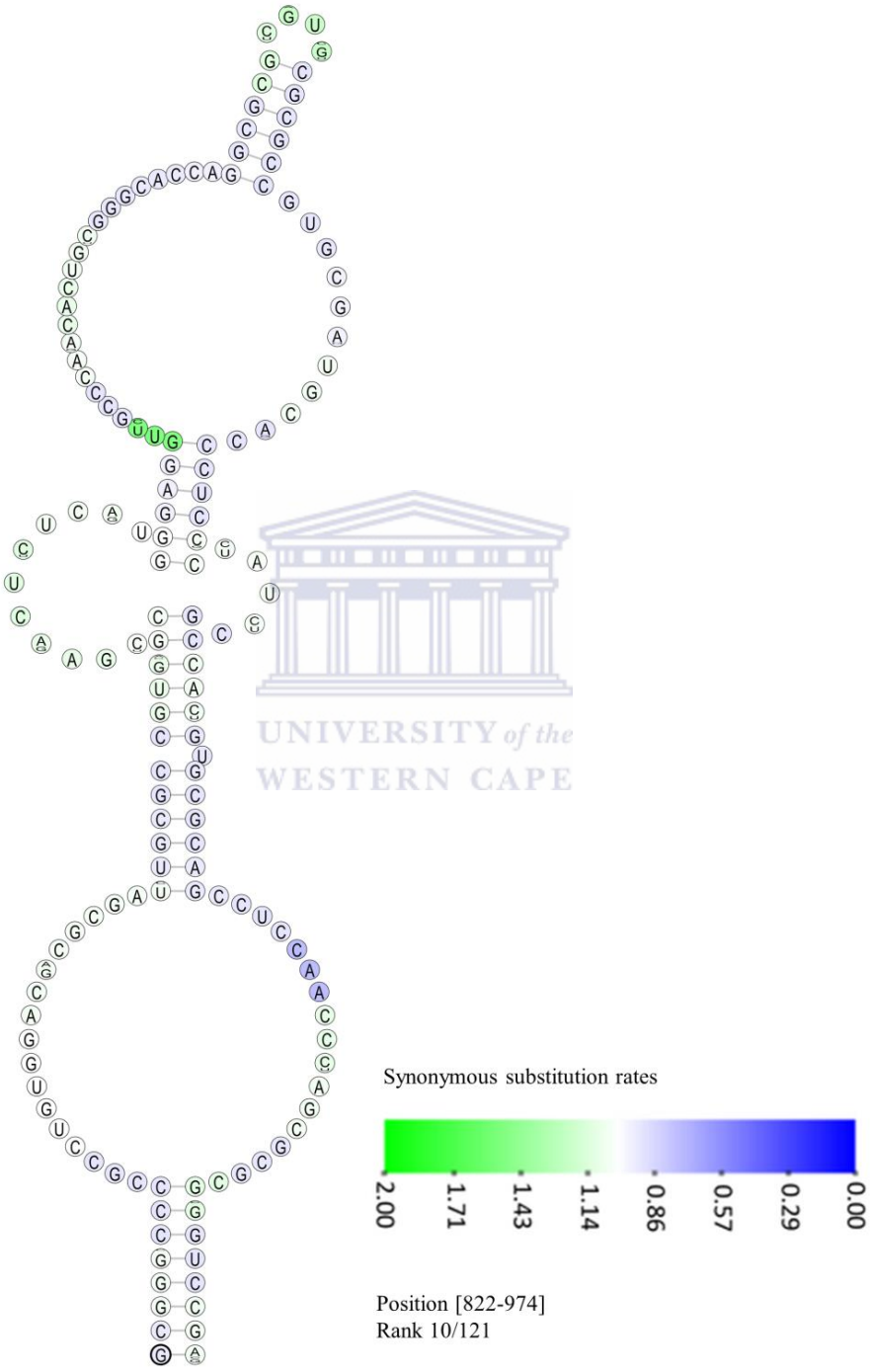
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



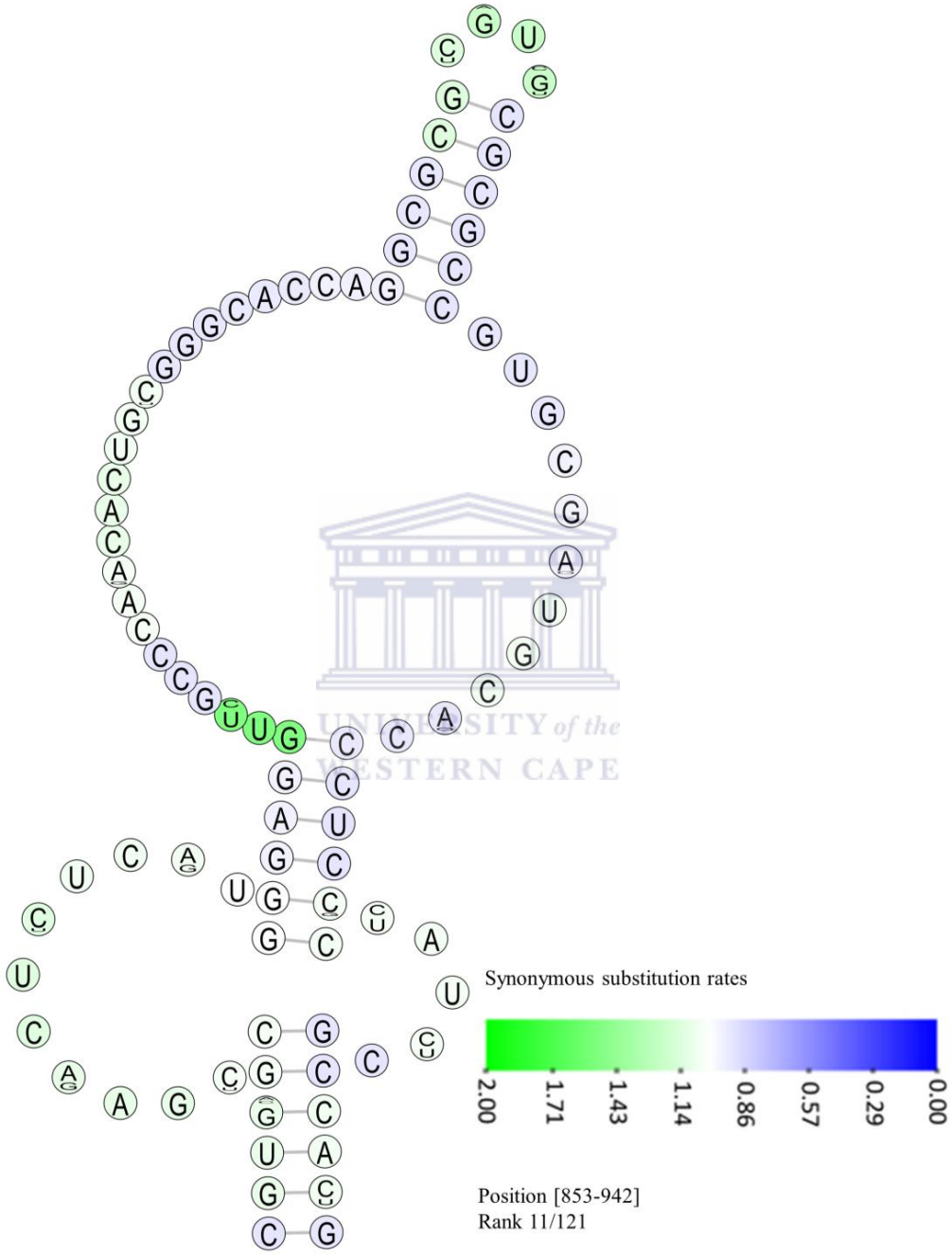
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



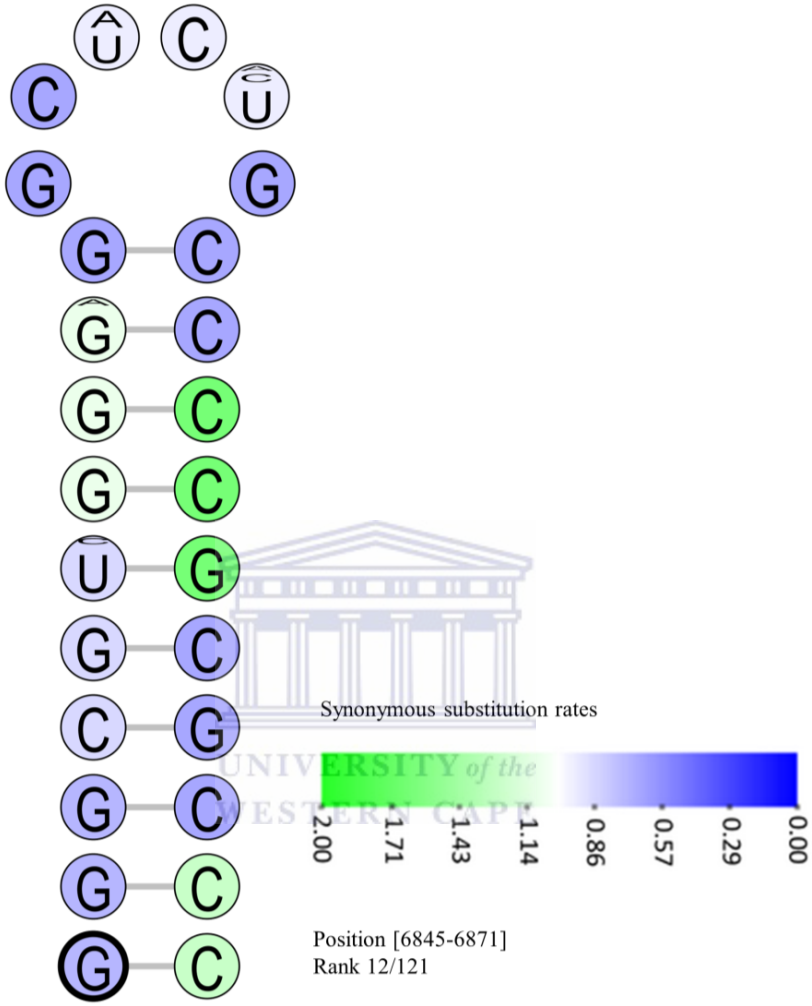
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



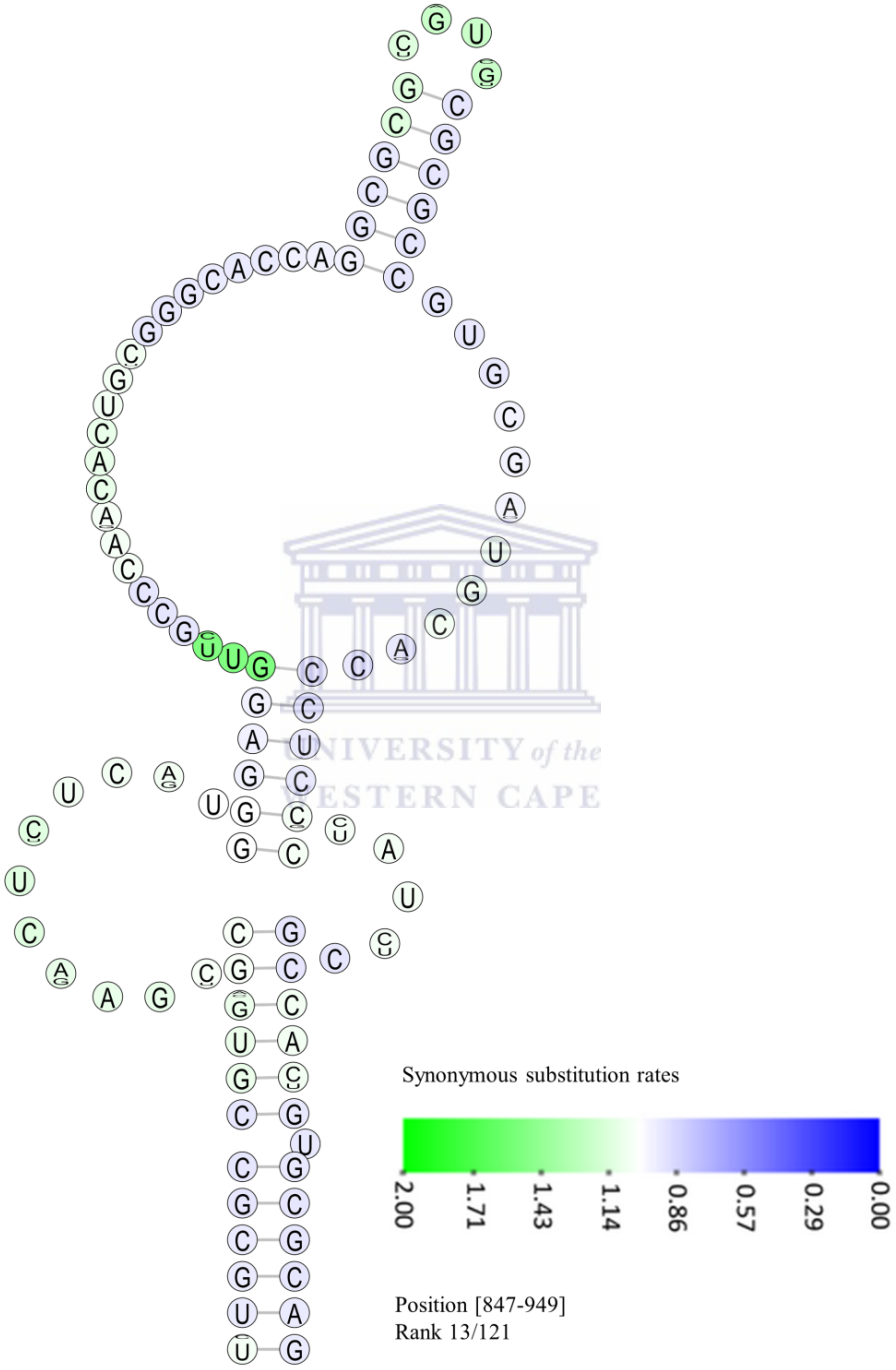
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



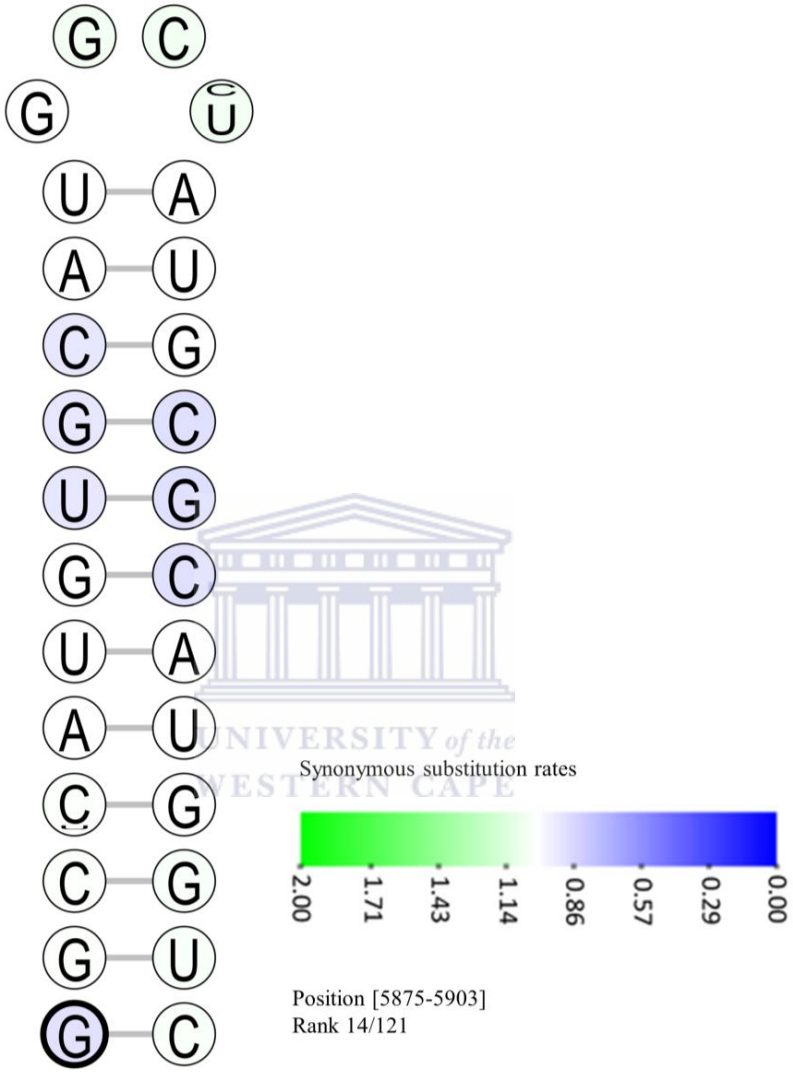
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



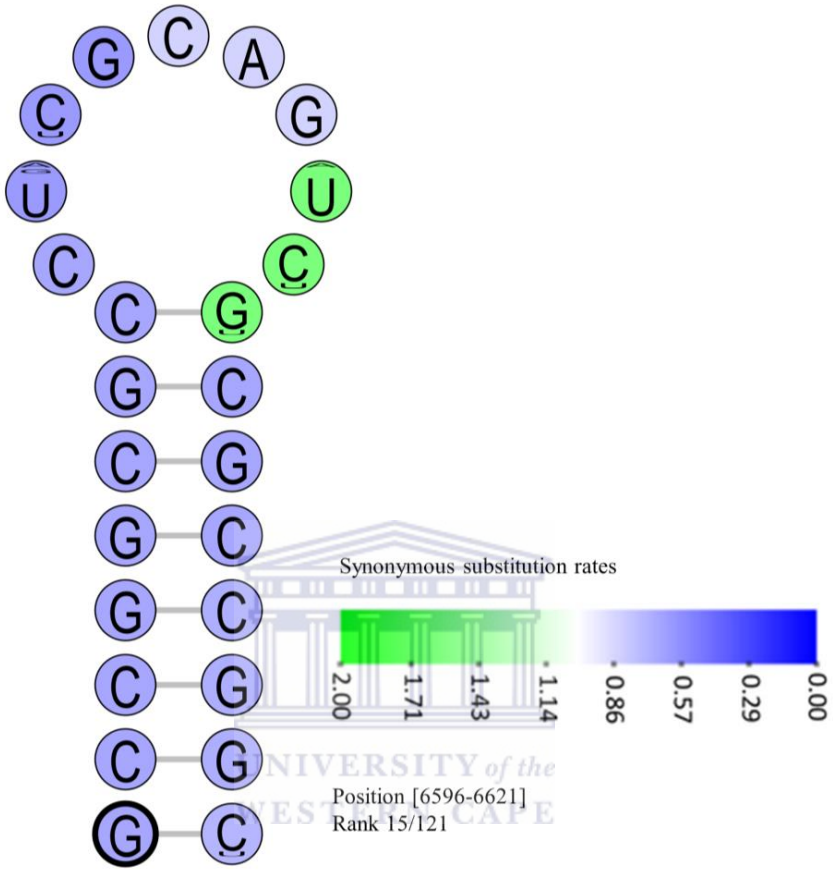
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



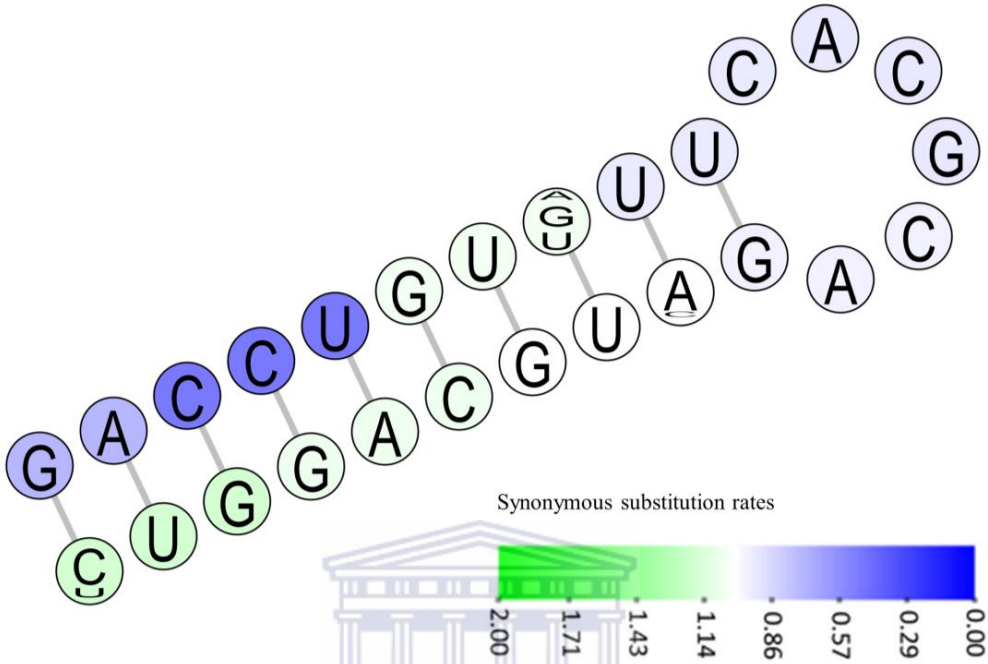
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



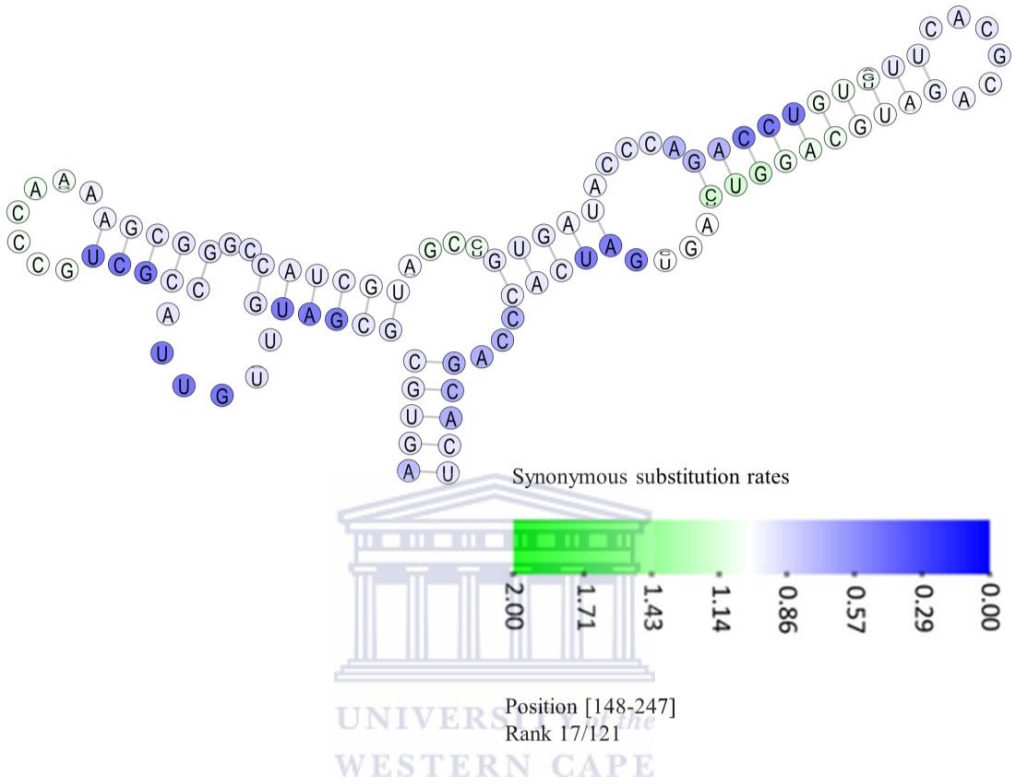
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



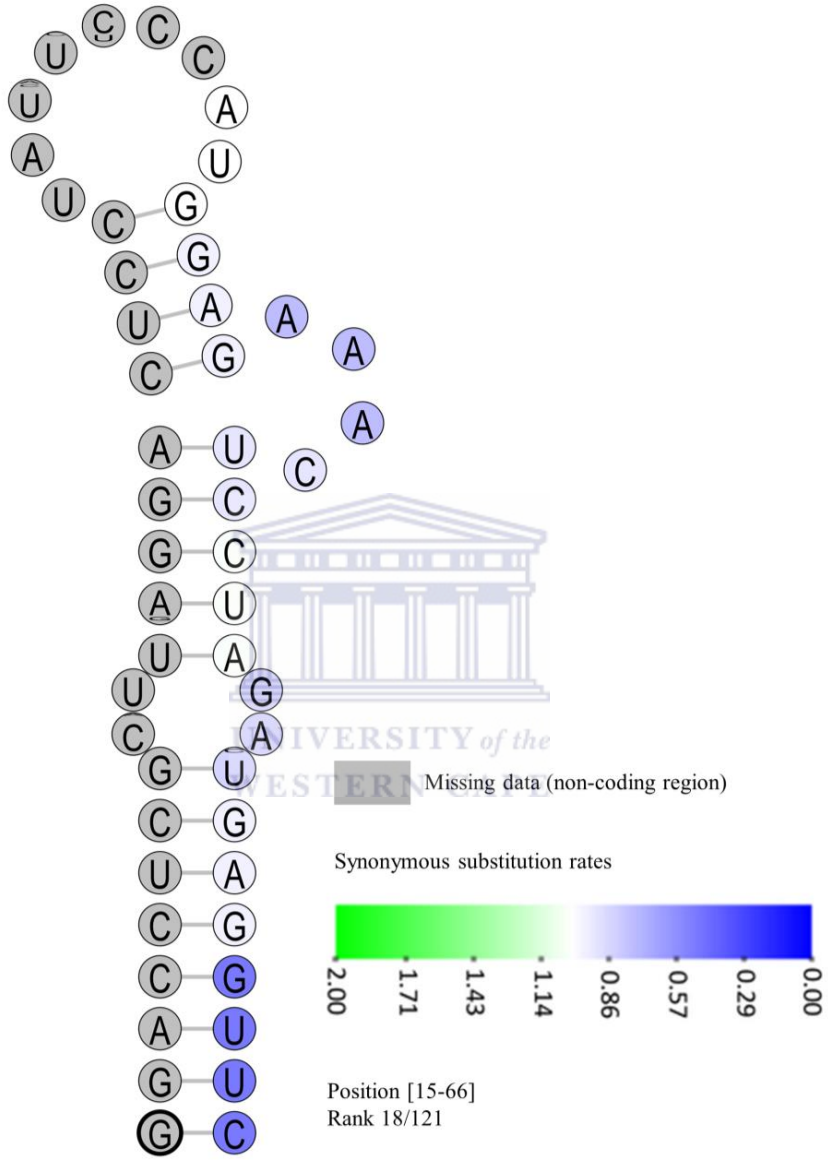
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



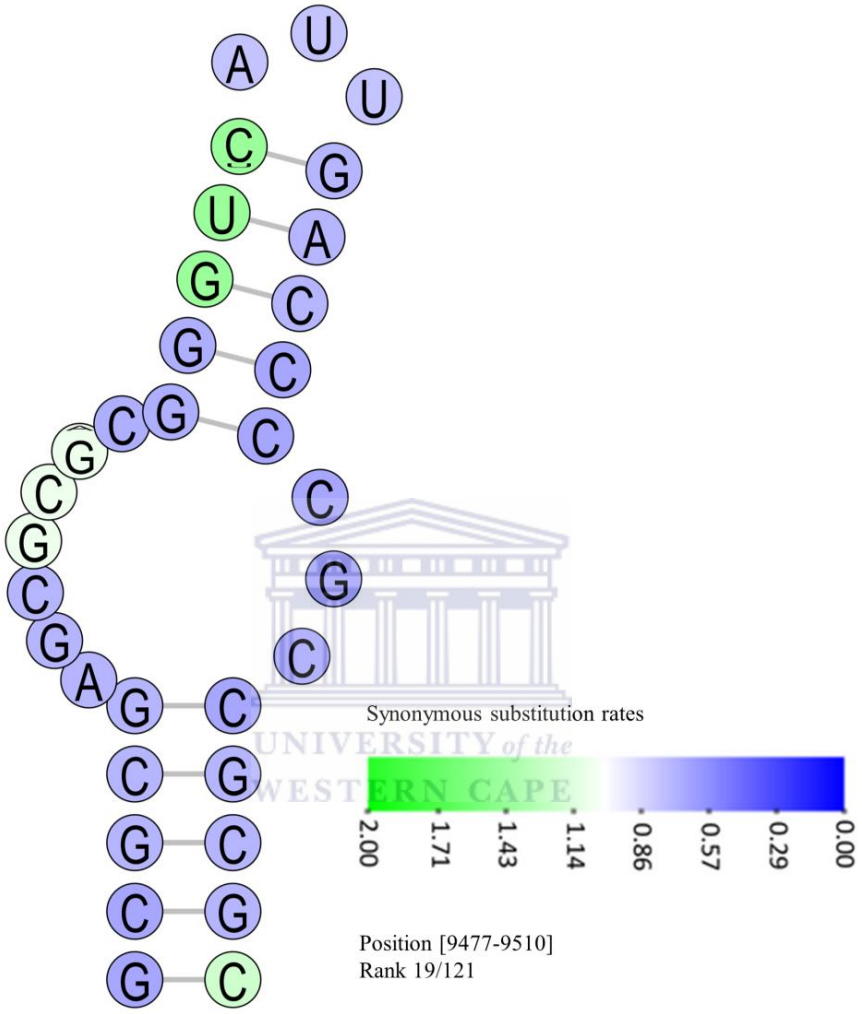
Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…



Appendix 7. Top-20 NASP predicted nucleotide secondary structures of Rubella virus. Continued…