PHYLODYNAMIC ANALYSES REVEAL RAPID EVOLUTION OF RUBELLA VIRUS

By

CODY ANDREW WAIN

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Microbial Biology

Written under the direction of

Dr. Siobain Duffy

And approved by

_____

_____

_____

New Brunswick, New Jersey

May, 2019

ABSTRACT OF THE THESIS

Phylodynamic Analyses Reveal Rapid Evolution of Rubella Virus

By CODY ANDREW WAIN

Thesis Director:

Siobain Duffy

Rubella is an infection caused by rubella virus. Rubella virus belongs to the *Togaviridae* family and is a single stranded positive sense RNA virus of about 9,762 nucleotides. Rubella, once known as German Measles, causes an iconic red rash all over the body and the teratogenic congenital rubella syndrome in pregnant women. Due to Measles-Mumps-Rubella vaccination rates falling throughout the developed world and measles and mumps becoming resurgent, the evolution of rubella virus is important to study prior to its potential resurgence. The E1 gene of the rubella virus is responsible for interaction with the human immune system, and it is the antigen to which antibodies are formed. The evolutionary rate of E1 along with the full rubella genome was determined using phylodynamic analysis. Both the whole genome and the E1 gene were evolving in a clocklike manner, and the evolution of both were successfully analyzed with BEAST2 software. A difference between the best-fitting priors between the two datasets was the kind of molecular clock preferred: the whole genome was best fit by a relaxed molecular clock, while the E1 gene preferred the strict molecular clock. This difference had some impact on the results, with the estimated evolutionary rate for the E1 gene from the strict clock being

lower than the whole genome, but still within the 95% highest posterior density range at 1.08 x 10^{-3} substitutions per site per year (ssy) while the whole genome had an evolutionary rate of 1.60 x 10^{-3} ssy with a 95% Highest Posterior Density (HPD) of 1.06 x 10^{-3} to 2.18 x 10^{-3} ssy. Reconducting the E1 analysis with a relaxed molecular clock resulted in a similar evolutionary rate as the whole genome of 1.51 x 10^{-3} ssy with a 95% HPD of 1.23 x 10^{-3} to 1.80 x 10^{-3} ssy. This is one of the first cases where there was a statistically significant difference in substitution rate (non-overlapping HPDs) between analyses of the same dataset calculated with different clock priors. The relaxed clock estimates of nucleotide substitution rate are higher than has been estimated for rubella virus in the past and agrees with the more rapid rate of evolution seen in a single decade in China. These results suggest that rubella evolves faster than expected, though it is not undergoing substantial positive selection, and that choice of clock model is a more significant determinant of substitution rate than previously considered.

ACKNOWLEDGEMENTS:

I would like to thank my thesis advisor Dr. Siobain Duffy for her guidance and patience during the course of this research. Her informative support and willingness to give much of her time to ensure the completion of this research was deeply appreciated. Special thanks are also given to Natasia Jacko and the rest of the Duffy lab for being extremely informative throughout my work and providing much appreciated assistance whenever I was stuck. I would also like to thank Chris Njagi who provided clear methodology which was of great assistance for path sampling analysis.

I would also like to acknowledge the Microbial Biology Program Director Dr. Gerben Zylstra for his incredible help and support throughout the course of the microbial biology program and providing answers to any questions I had.

Finally, I would like to thank my family, specifically my mom, dad, brother, and grandparents for their continued moral support. Thank you for always being there for me and believing that I could do this. I wouldn't have been able to do it without you all.

Table of Contents

LIST OF TABLES:

LIST OF FIGURES:

Section 1: Introduction

1.1 History of Rubella:

The infection known as rubella, also sometimes known as German Measles or three-day measles, is caused by the rubella virus. In the mid 1700s and early 1800s, the infection was commonly mistaken to either be scarlet fever or some derivative of measles, due to their commonality in sharing the iconic red rash of these infections (Cooper, 1985). It wasn't until George Maton, a German physician working in England, suggested that rubella could potentially be a distinct illness due to not sharing several key characteristics of scarlet fever or measles and thus gave it a name, Rötheln (Wesselhoeft, 1947). While this is the first instance in which the disease was named, the more common name of the disease that we use today was not developed until later. In 1866 an English surgeon by the name of Henry Veale was in India where he witnessed an outbreak of the virus within schoolchildren, and gave it the distinct name rubella, as he believed that the German word Rötheln was a harsh word (Veale, 1866). With clinical observations of the disease showing differences from measles and scarlet fever, the International Congress of Medicine met in 1881 and officially recognized rubella as a distinct disease (IMC, 1881).

With the discovery and study of viruses underway in the 1890s, infection by rubella was first proposed to be caused from a virus in 1914 by Alfred Hess, who inoculated monkeys with the blood of children infected with rubella (Hess, 1914). This was later confirmed to be true in 1937 when the disease was passed successfully to children from people with severe cases of the infection (CDC, 2015). While rubella was initially thought to be a mostly harmless rash that quickly passed (hence the name, three-day measles), later clinical observations proved this to be false. A potential link between the rubella infection and serious birth defects happened in 1940 when it was discovered that babies who were born to mothers who contracted rubella had high

rates of cataracts, which led to the belief that the infection played a role (Gregg, 1941). These initial findings were later shown to be supported in several studies where the rate of newborn deformities were much higher in women who contracted rubella during pregnancy compared to women who did not contract rubella (Fox and Bortin, 1946, Ober et al, 1947, Mackenzie et al, 1948).This outbreak and subsequent observation of birth defects in children was instrumental as it illuminated the study of viruses to be included as possible teratogenic agents, or things which can cause birth defects (O'Connell, 2013). Finally, the rubella virus, which was already believed to be the causative agent of rubella, was finally isolated in tissue culture in 1962 which allowed for it to be studied in lab (Cooper, 1985).

Since the discoveries of fetal abnormalities associated with rubella infections during pregnancy, congenital rubella syndrome (CRS) has become the most troubling effect of rubella infections. CRS occurs when a pregnant woman is infected with the rubella virus and transfers the infection to her developing fetus, resulting in either a stillborn, miscarriage, or a baby with several defects. These defects include cataracts, mental retardation, hearing loss, congenital heart disease, bone disease, and more (Lanzieri et al, 2018). CRS typically has the highest chance of occurring when a pregnant woman is infected in her first trimester, with the chance going down significantly as gestation period increases (Lee and Bowden, 2000). The current prevalence of congenital rubella syndrome is not globally reported; however, it was estimated that in 2001, over 100,000 cases of congenital rubella syndrome occurred with a global prevalence that year of 836,321 cases (Robertson et al, 2003). This makes the spread of rubella an important global concern to prevent the transmission of rubella to pregnant women.

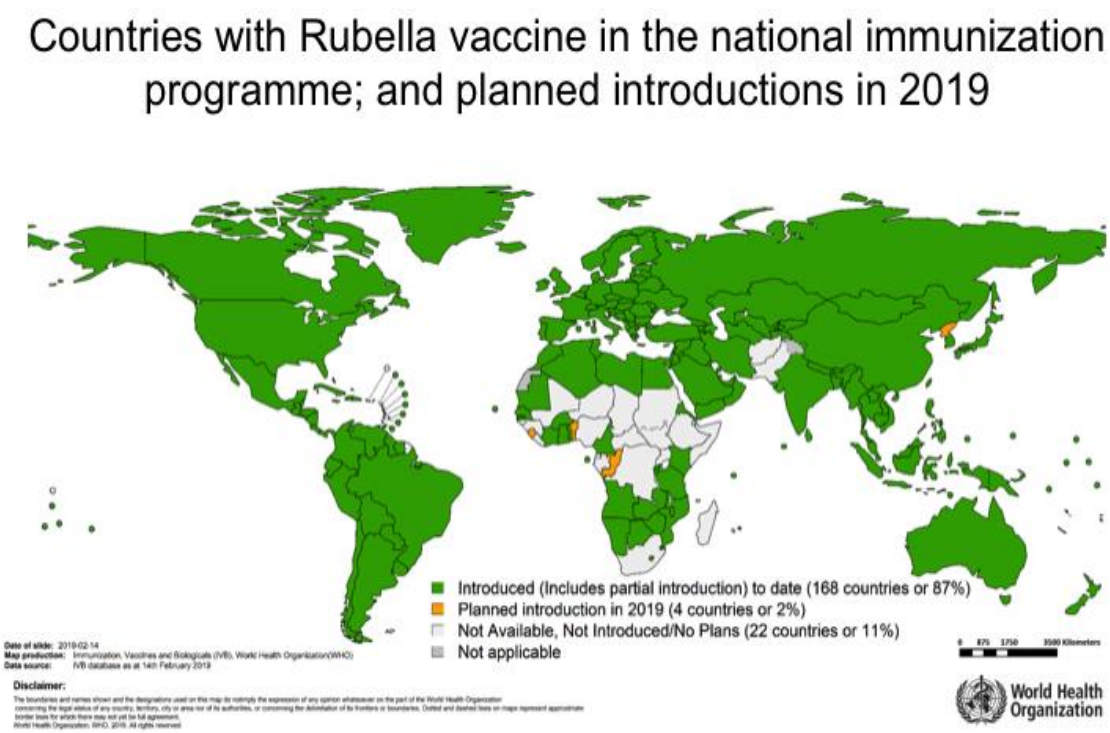

Figure 1**.** Current global range of countries immunizing for rubella or planning to immunize for rubella as of 2019. Adapted from WHO immunization schedule.

1.2 Modern Significance and Reach:

Increased availability of vaccines against rubella along with vaccination programs and control strategies carried out throughout an increasing number of countries has led to a global decrease of the number of rubella cases in the 21st century. The number of global reported cases of rubella has fallen from 670,894 cases in 2000 and 836,356 cases in 2001 to 22,361 cases in 2016 (Grant et al, 2016, Robertson et al, 2003). Although still endemic to many regions of the world, particularly those with no implemented vaccination strategies against rubella (gray countries in Figure 1), many regions are making massive gains into the elimination of the disease. As of 2015, rubella has been declared eliminated from the region of the Americas (CDC, 2015). In the United States, children are routinely vaccinated with a trivalent vaccine against measles, mumps and rubella (MMR), which contains a live attenuated vaccine for rubella (CDC, 2019). Worldwide, a variety of live attenuated vaccines are given for rubella, and vaccination campaigns were

adopted more slowly than for measles or mumps. For instance, China only started offering rubella vaccinations beginning in 1993 and it entered the nationwide immunization schedule in 2008 (Su et al, 2018). Further expansion is continuing throughout the world, with 4 different countries planning to add rubella vaccination to their immunization schedule in 2019 (yellow countries in Figure 1). Despite this, further expansion into Africa is sorely needed as they are the last continent and major region within the world still with a significant number of countries and population at risk for rubella outbreaks and epidemics (WHO, 2016).

With the rise of the internet making information widely available, the spread of misinformation has infected the discussion regarding vaccinations. Misinformation, misleading claims, misinterpretation of data, and fraudulent data are easily spread amongst the populace and in certain cases become engrained within public discourse despite being demonstrably false. In 1998, former doctor Andrew Wakefield along with a number of other researchers published a now redacted paper in which it was concluded that the MMR vaccine was the probable cause of developmental delays in children, which was described as autism (Wakefield et al, 1998). Despite unethical practices, scientific misconduct, and fraud that was later revealed which caused the paper to be retracted and Andrew Wakefield's medical license to be revoked, the lasting damage and mistrust of vaccines spawned from this paper are still felt to this day. The findings and conclusions within the paper were immediately proven false (Taylor et al, 2000) and no follow up studies have corroborated Wakefield's claims. Very recently, another study has corroborated the findings that the MMR vaccine has no links to development of autism within children, showing that 21 years later there is still no evidence for Wakefield's claims (Hviid et al, 2019). However, as vaccine compliance has fallen in many developed countries, measles has become resurgent (Phadke et al, 2016), with sporadic outbreaks of mumps (CDC, 2018). Rubella is less common than these other diseases included in the MMR vaccine with measles and

mumps being more common, especially in the US (WHO, 2016) but it would not be unexpected to see a rubella outbreak in the USA in upcoming years. The CDC currently recommends a targeted vaccination rate of 95% for the MMR vaccine to prevent widespread outbreaks of measles, mumps, and rubella. As of 2017, only 6 states have higher than 95% of children aged 19-35 months inoculated with the MMR vaccine, with 16 states falling below 90% vaccination rate for MMR (CDC, 2017). With measles and mumps outbreaks being widespread with low MMR vaccination rates, rubella may be expected to make a comeback in the US, as is already speculated in places such as France (Beraud, 2018). What was once thought of a disease that would eventually be eradicated due to vaccination now has a high chance of making a resurgence and due to this it is imperative that the evolution of the rubella virus can be understood to help control and prevent future outbreaks.

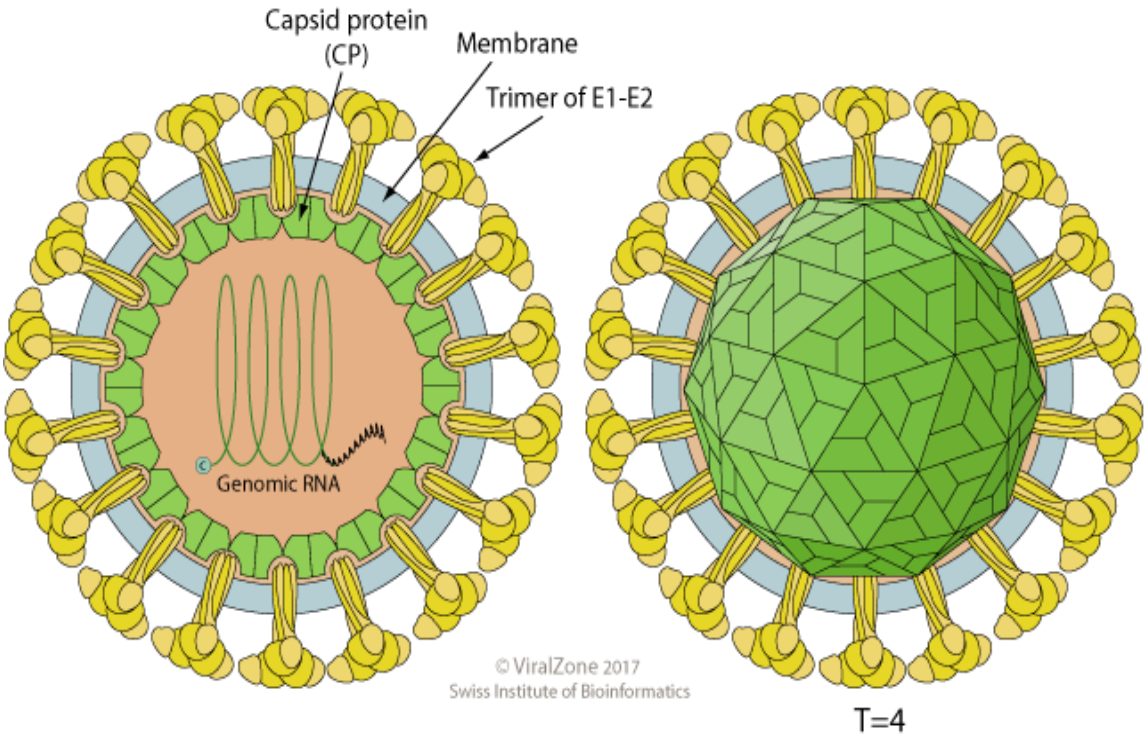

Figure 2. Graphical representation of the rubella virus along with structural proteins (CP,E1, and E2) being shown in their correct positions. The trimer formed by the E1 and E2 proteins are shown in yellow on the surface of the virus. Image from Viralzone: www.expasy.org/viralzone, SIB Swiss Institute of Bioinformatics, 2017. Image CC4.0 BY-NC-ND

1.3 Rubella Virus:

The etiological agent of rubella, the rubella virus or rubivirus, is a 9,762-nucleotide single stranded RNA mammalian virus. It belongs to the viral family *Togaviridae* which are defined by being spherical viroids, single stranded, linear, positive sense RNA viruses with a 5' methyl cap and 3' polyadenylated tail. This methyl cap and polyadenylated tail allow it to be easily translated when within a host as this mimics mRNA (Jose et al, 2009). Like other viruses within the *Togaviridae* family, the rubella virus has non-structural proteins encoded within the 5' end of the virus while the envelope and capsid proteins are encoded at the 3' end (ViralZone, 2019) The rubella virus genome consists of genes that encode for two RNA replication proteins, p150 and p90, along with three structural polyproteins, E1, E2, and the capsid protein. The E1, E2, and capsid protein are constructed as a polyprotein with the proteins being cut by signal peptidase, and the p150 and p90 proteins are also encoded as a polyprotein with them being cut by p150 (Frey, 1994). The E1-E2 trimer that is formed on the surface of the virus is especially important in the infection with rubella virus. The complex that is formed from the E1 and E2 proteins coming together has been shown to be necessary in transfer out of the endoplasmic reticulum, through the Golgi apparatus, and to the cell membrane (Yang et al, 1998). The E1 protein also specifically has been shown to be the protein directly responsible for attachment and fusion into human cells (Yang et al, 1998). In addition to this, the E1 protein is especially important as the human immune system forms antibodies targeting antigens against the E1 protein, both from natural infection and through vaccination (Chaye et al, 1992).

1.4 Rubella virus evolution:

Recent literature regarding the evolutionary rate of the rubella virus has mainly focused on country specific outbreaks and genotypes of the rubella virus (Zhu et al, 2012, Zhu et al, 2015,

Yalcinkaya, 2015) Additionally, the studies mostly focus on the 739-nucleotide window within the E1 gene recommended by the WHO for rubella genotype classification (WHO, 2015). The studies that have been done generally see the rate of evolution of the E1 gene increasing as time progresses, with significantly higher rates being reported depending on year of the study along with very little to no recombination being present within rubella (Cloete et al, 2014, Zhu et al, 2012, Jenkins et al, 2002). In China, it has also been found that the E1 gene has been evolving at even a significantly higher rate than E1 isolates globally (Zhu et al, 2012, Zhu et al, 2015).

The genotypes of rubella are typically divided into two distinct clades. Clade 1 contains 10 genotypes, these being 1a, 1B, 1C, 1D, 1E, 1F, 1G, 1H, 1I, and 1J with 1a being a provisional genotype while clade 2 contains 3 genotypes, these being 2A, 2B, and 2C (WHO, 2005). The most common current circulating rubella genotypes globally are 1E, 1G, 1J and 2B with 1E and 2B being global while 1G is more restricted to Africa and 1J more restricted to Asia (Abernathy et al, 2011, Zheng et al, 2003). Identification of the genotype is important for characterization by the WHO into these genotypes, however it is argued whether more precise genotypes should be established between the most common genotypes (Rivailler, 2017).

1.5 Molecular Modeling:

In order to effectively and accurately gauge a given nucleotide substitution rate over time, several models are able to be selected from regarding rate, time, and effective population sizes. Among these are the site model, which determines rates of change within specific sites of a genome, the molecular clock model, which infers a given model of divergence at a certain time and then bases the evolution rate over time from that, and the prior models, which are associated with the population of a genomic dataset.

There are several different nucleotide substitution models available for use, although the one used within the context of this study was the Generalized Time Reversible Model (GTR). The GTR model assumes that the frequency of each base is different, and that all pairings of nucleotides have independent substitution rates (Tavare, 1986). It does not allow each potential substitution to have an independent rate because that would require *a priori* knowledge of the root of the tree – and would make the model not time reversible. Additionally, modeling included within the different site models themselves are used to show different rates of variation within different sites in the sequence. These two additional models are the gamma distribution models, typically denoted by the letter G, and the proportion of invariable sites, typically denoted by the letter I. The gamma distribution, from its name gives a changing probability distribution within the nucleotide substitution rates. The proportion of invariable sites examines the number of sites within a genome that have very low rates of evolution or are highly conserved and takes into account their impact on analysis (Huelsenback, 2001).

Of the molecular clock models this study included the strict clock model, the relaxed logarithmic clock model, and relaxed exponential clock model, each corresponding to how the rubella virus potentially evolved over time. The strict clock model assumes that evolutionary rate is constant throughout time. The relaxed clock models both allow for each branch in the phylogeny to have its own evolutionary rate regardless of placement within a phylogenetic tree. The differences between the log normal and exponential models are in distribution of the rates of the nucleotide substitutions.

Different tree priors are also tested for their use in demographic reconstruction of the rubella virus over time. The constant coalescent tree prior assumes a constant effective population size over the time period of samples whereas the exponential coalescent model assumes an exponentially increasing population size. Bayesian skyline utilizes sampling at posterior trees at

different time points to accurately assess effective population size. Extended Bayesian skyline is quite similar but is better fit for bottleneck events, which one might expect if there exist events that would rapidly decrease transmission.

Section 2: Methodology

2.1 Acquisition of Data:

A total of 63 whole genome isolates of rubella virus originating between 1961-2013 from 12 different countries were available in GenBank in February 2018, downloaded and saved. Separately, in a Microsoft Excel worksheet, details of isolation and passaging of the virus were noted. Of the 63 whole genomes, 17 were excluded from further phylogenetic analysis due to potential issues clouding the accuracy in any further analysis. Substitution rate analyses require sequences with rigorous dates of isolation from its global gene pool (Rioux and Belloux, 2016) so genome sequences without sufficient details of the year of their isolation, extensive passaging in the lab prior to sequencing which would allow for changes in genome to occur within lab, and sequences derived from vaccine strains cannot be used for accurate analyses of the natural rate of rubella virus evolution. Further, known recombinant sequences are not appropriate for phylogenetic analyses in general, and were excluded from this study (Cloete et al, 2014). Sequences were named to reflect their GenBank accession number, country of isolation (ISO two letter country code), and year of isolation.

Following similar methodology as described above, a total of 309 E1 gene sequences from 1961-2013 in 15 different countries were downloaded and saved from GenBank along with similar details about isolation. Of the 309 E1 gene sequences, 67 were excluded from further analysis due to similar reasons mentioned above. Sequence names were changed to reflect GenBank accession number, country of isolate, and year of isolation. The 242 E1 isolates used included 46

E1 genes from the whole genome dataset. It should be noted that there is a primer-amplified 739 base region of E1 that covers just over half (51.2%) of the full E1 gene that is frequently sequenced and deposited into GenBank (WHO, 2005). More than a thousand such sequences, most with rigorous dates of isolation, were available in February 2018, which likely would have exceeded the time limits of analyses allowed on the CIPRES Gateway server that was used.

2.2 Alignment:

Multiple sequence alignment was done using Clustal Omega for both the whole genomes and the E1 isolates (Chojnacki et al, 2017). Alignment outputs were saved in both the clustal and nexus formats for further use. E1 isolates were imported into Geneious Prime v.2019.1.1 (Geneious, 2019) and the E1 gene regions from the whole genomes were extracted and cut using the sites of start and stop codons from reference sequences in GenBank. Using Geneious Prime, neighbor joining trees were built from the whole genome dataset and the E1 dataset for further use. Of the 1443 nucleotide E1 gene region, the first 27 nucleotides were excised from most sequences to match the shortest E1 sequences in the dataset, allowing each sequence to have equal information content (final alignment length = 1416 nt). A third dataset (WGE1), of simply the extracted E1 gene region (minus the first 27 nt to match the size of the larger E1 dataset) from the whole genome alignment was made as a control for the larger E1 analysis.

2.3 Recombination Detection:

Recombination is a means of increasing genetic variance that is important to test for prior to phylodynamic analysis because recombination can blur the phylogenetic history of the isolates used. Recombination within datasets has been shown to diminish both phylogenetic accuracy and cause overestimation of nucleotide substitution rates (Posada and Crandall, 2002, Schierup and Hein, 2000, Lanier and Knowles, 2012). Recombination was tested for using RDP4 which

uses seven different independent detection methods to find recombination, if present, within a dataset (Martin et al, 2015). The seven detection methods that were used were RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al, 1999), Bootscan/Rescan (Martin et al, 2005), MaxChi (Smith, 1992), Chimaera (Posada and Crandall, 2001), SiScan (Gibbs et al, 2000), and 3Seq (Lam et al, 2018). Of the seven detection algorithms used, at least three of the algorithms needed to detect an event of recombination in order to be considered significant. The general settings of the RDP4 detection methods that were implemented were analysis of linear sequences with a p value threshold of 0.05 and use of Bonferroni correction, which is necessary when making many multiple comparisons for recombinant detection. Additional settings changes included using the Kimura model (instead of the Jukes-Cantor model) for the Bootscan detection method to help be used in nucleotide substitution rates regarding transitions and transversions. (Kimura et al, 1980). Recombinant sequences were removed from the dataset to prevent inaccurate phylogenetic analysis.

2.4 Temporal Signal Detection:

In order to determine whether or not further phylodynamic analysis would be appropriate with these datasets, we determined whether these rubella sequences were evolving in a clock-like manner. This is done by using different sequences isolated at different points in time in order to see whether they have a measurable amount of genetic divergence that correlates positively with time. TempEst v.1.5.1 was used to measure this correlation, using a tip-dated (year of sequence collection) neighbor joining tree. Genetic differences between taxa create a root to tip linear regression graph that shows correlation of genetic divergence over years since divergence (Rambaut et al, 2016). The neighbor joining trees previously made using Geneious Prime of the whole genome and E1 datasets were used as input files for TempEst and dates were manually inputted to run the program.

2.5 Phylodynamic models to assess rate of evolution:

To use BEAST2 (Bouckaert et al, 2014) to estimate rates of evolution and other population parameters for rubella virus, first the most appropriate priors had to be selected for each dataset. The best fitting nucleotide substitution model was chosen by JModelTest v2.1.6 (Darriba et al, 2012). JModelTest utilizes five different model selection strategies with each allowing for rate variation to make a likely estimate of the best fitting nucleotide substitution model given the dataset. Path sampling in BEAST2 first determined (by marginal likelihood estimates) the best fitting molecular clock model (assuming a constant viral population size) and then using the chosen clock, the best fitting demographic model was chosen. There were three different clock models used: a strict molecular clock model that assumes that every branch of a phylogenetic tree has the same rate of evolution, a relaxed lognormal clock model that assumes that evolution occurs as a lognormal function among branches, and a relaxed exponential clock model that assumes that substitutions vary following an exponential function amongst branches. After the best fitting clock was selected (assuming a constant viral population size), four different tree priors were tested for each of the datasets. These included constant population size, exponentially growing population size (as would match many emerging viruses), allowing the BEAST2 analysis to determine population sizes at different time periods during the analysis (Bayesian skyline), and the extended Bayesian skyline, which accommodates smaller numbers of sequences using the original Bayesian skyline model. BEAUTi v2.5.0 was used for the generation of the XML files for path sampling and subsequent analyses (Bouckaert et al, 2014). The XML files were modified for path sampling by replacing the run command with:

<run spec='beast.inference.PathSampler' chainLength="2000000" alpha='0.3' rootdir='/tmp/' burnInPercentage='50' preBurnin="0" deleteOldLogs='true' nrOfSteps='100'>

cd $(dir)

java -cp $(java.class.path) beast.app.beastapp.BeastMain $(resume/overwrite) -java -seed $(seed) beast.xml

(Adapted from github.com/BEAST2-Dev/BEASTLabs/examples/testPathSampler.xml)

The downgraded v.2.5.0 version of BEAST2 and BEAUTi were used for XML file generation and analysis due to several infinity errors within the marginal likelihood estimates which terminated several runs when attempted with the most updated version of BEAST2. Additionally, chain lengths and logs were standardized at 2,000,000 and 10000, respectively. All analyses were run through the CIPRES Science Gateway (Miller et al, 2010).

2.6 Bayesian Phylogenetic Analysis

Using the best fit clock model-prior pairing as identified by previously done path sampling, BEAST2 was run for the whole genomes, E1 genes, and E1 genes solely for the whole genomes. When a dataset did not select the Bayesian skyline demographic prior as the best fit, it was run in parallel with the selected analysis to visualize the effective viral population over the timespan of the sampled sequences. The xml files for the final runs were generated in BEAUTi v2.5.0 using the parameters defined by previously done path sampling. A chain length of 1 billion with logs being filed every 10,000 steps was done for each of the final runs as these settings have been previously shown to produce sufficiently resolved results (Njagi, 2018). Two independent runs of each dataset were done to ensure that the analysis converged on the same outputs. All analyses were run through the CIPRES Science Gateway (Miller et al, 2010). Log and tree output files of the BEAST runs for each datatset were imported into Tracer which allows for visualization and creation of skyline plots from the completed BEAST2 runs (Rambaut et al, 2018). Nucleotide substitution rates were recorded from the finalized clockRate value (strick clock prior) or ucedMean value (relaxed

clock prior). Lastly, maximum clade credibility (MCC) trees were constructed using TreeAnnotator v1.8.4 from the BEAST2 package software. Trees were then viewed and colored according to country of isolation using FigTree v1.4.4 (Rambaut, 2018)

2.7 Selection Pressure Analysis:

Positive and negative selection of the E1 dataset was conducted using three separate analyses using the Datamonkey server (Weaver et al, 2018, Delport et al, 2010, Pond et al, 2005). The three analyses used were Mixed Effects Models of Evolution (MEME), Fast Unconstrained Bayesian Approximation (FUBAR), and Single Likelihood Ancestor Counting (SLAC). The FUBAR analysis measures the nonsynonymous (dN) and synonymous (dS) substitution rates of a dataset and takes into account large numbers of varying site classes which allows the evolutionary rate to be different although this also assumes a constant diversifying pressure (Murrell et al, 2013). The MEME analysis observes both pervasive and episodic positive selection throughout a dataset at an individual site level (Murrell et al, 2012). The SLAC analysis also measures dN and dS substitution rates and assumes constant selection pressure through maximum likelihoods and counting approaches (Pond and Frost, 2005). The p value thresholds for each analysis was set at less than 0.1.

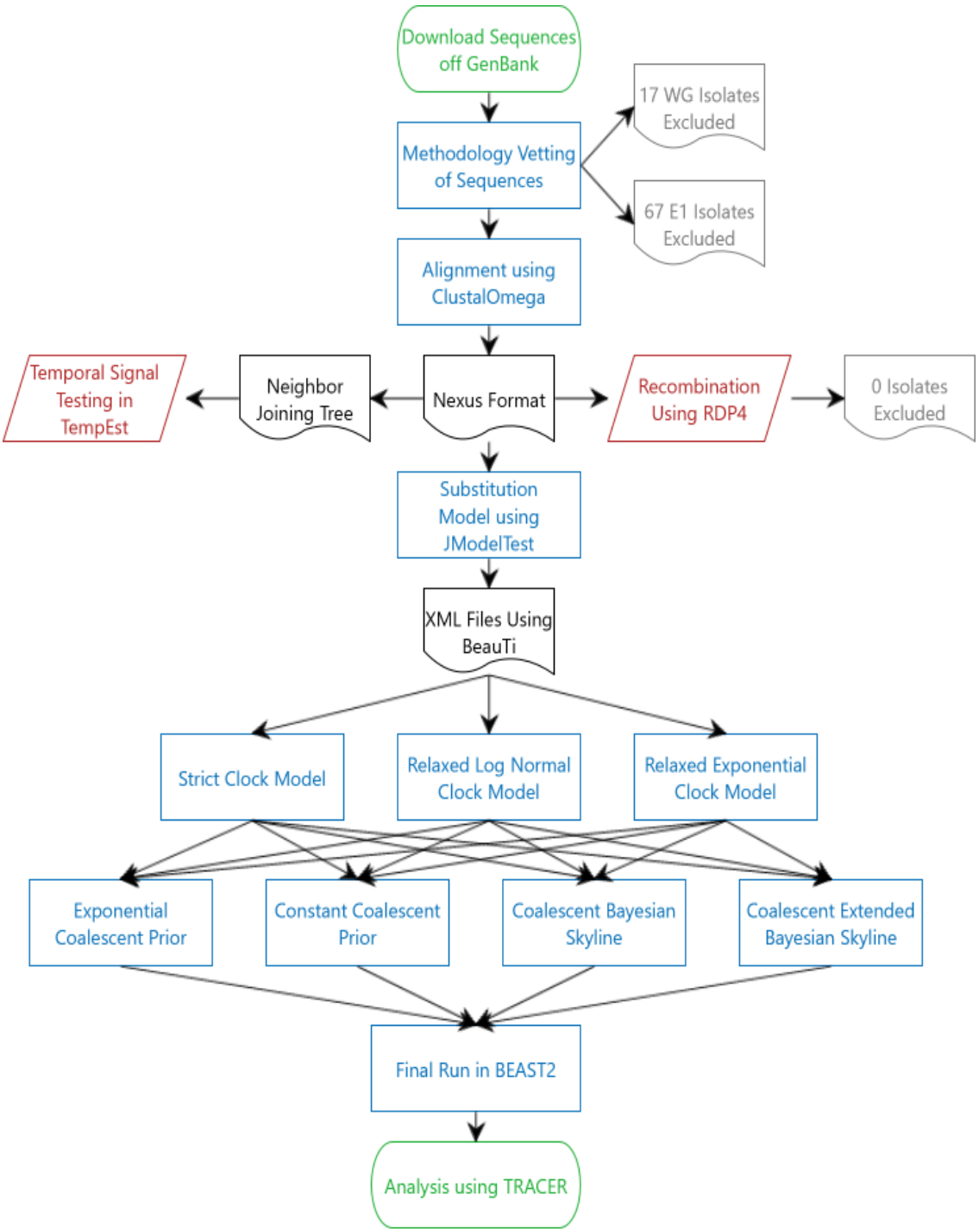

Figure 3. Flowchart of the methods used along with file formatting for the analysis of the whole genome and E1 rubella datasets. Greyed out symbols indicate exclusion of isolates, black symbols indicate correct file formatting, red parallelograms indicate necessary side steps before continuation, and blue rectangles represent main sequence of events.

Section 3: Results

3.1 Alignment:

After vigorous background vetting of several whole genome and E1 isolates, 46 whole genome isolates were used for analysis. Almost half of the 242 E1 gene isolates used for alignment were slightly smaller than the full gene, as they were missing the 27 nt at the 5' end of the gene. We excluded these 27nt from all isolates. Clustal Omega produced good alignments that did not need to be corrected by hand; there were few insertions or deletions in the whole genome alignment (9778 nt) and none in the trimmed E1 alignment (1416 nt). The whole genome had a 94% pairwise identity score and the E1 dataset had a 95% pairwise identity score. The G/C content of the whole genome was 69.7% and 66.2% for the E1 gene dataset, which is very high for an RNA virus but quite normal for rubella virus (Takkinen et al, 1988). A full list of whole genome sequences used is given Appendix 1. In addition to the E1 portion of all sequences in Appendix 1, the E1 gene sequences used are given in Appendix 2. As one whole genome isolate had already been excluded due to previous evidence of recombination (Cloete et al, 2014), no recombination was detected within the whole genome or E1 datasets.

3.2 Temporal Signal Detection:

Correlation coefficients for both the whole genome and E1 datasets were extremely high. The whole genome dataset had a correlation coefficient of the best fitting root to tip of 0.9246, indicating that the whole genome dataset is evolving in a strongly clocklike manner (Figure 4). For the E1 gene dataset, the correlation coefficient was 0.8461, also providing good evidence that rubella's E1 gene is evolving in a predictable, clocklike manner (Figure 5). These datasets were appropriate for further phylodynamic analysis. The WGE1 dataset, which was extracted

from the whole genome alignment, was assumed to have a positive correlation between genetic divergence and time.

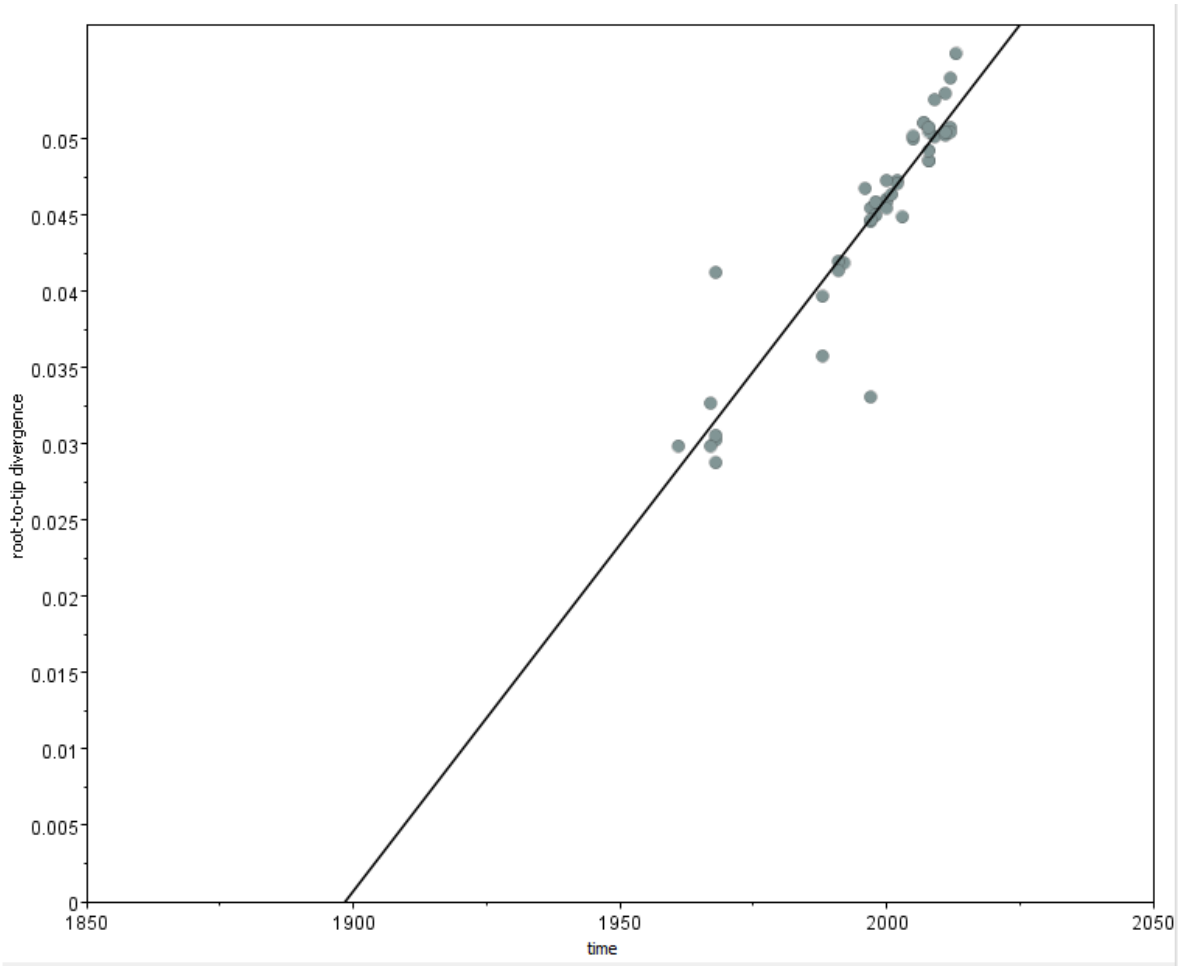

Figure 4. Temporal signal detection of the 46 whole genome isolates using best fitting root-to-tip divergence. The correlation coefficient for the best fitting root to tip was 0.9246 with an r squared value of 0.8549.

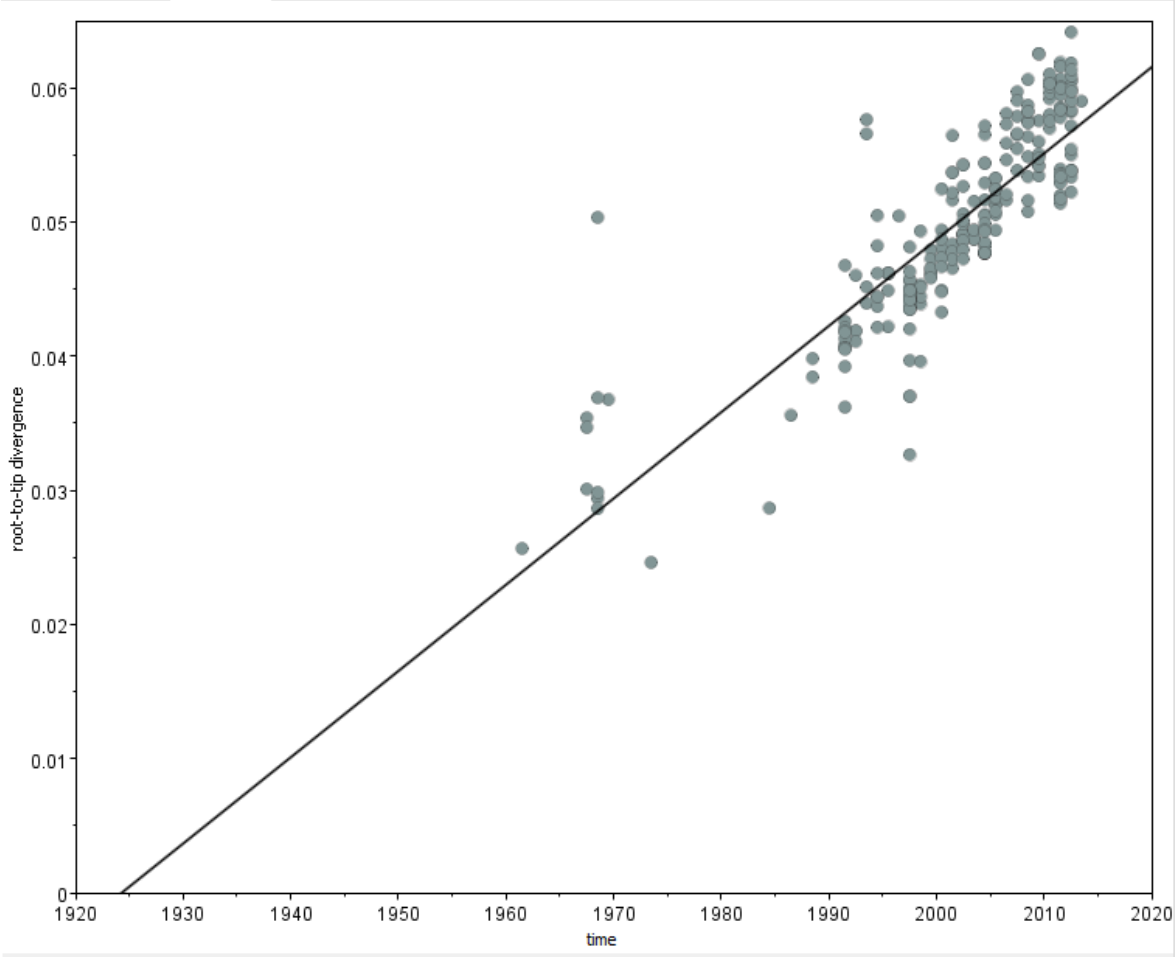

Figure 5. Temporal signal detection of the 242 E1 gene isolates using best fitting root-to-tip divergence. The correlation coefficient for the best fitting root to tip was 0.8461 with the r squared value being 0.7158.

3.3 Phylodynamic modeling selection:

The best fitting site model for both the whole genome and the E1 datasets based on the marginal likelihood estimate was the GTR I + G model (chosen by the corrected Akaike Information Criterion), indicating that gamma distribution and proportion invariant additions are necessary in accurately determining nucleotide substitution rate.

For the clock models tested, the relaxed exponential clock model was the best fitting clock model for the whole genome dataset, which accommodates the substitution rate varying over branches of the phylogenetic tree. The best fitting clock model for the E1 dataset was the strict clock model, which requires each branch on the phylogenetic tree of the E1 dataset to evolve at the same rate. As the E1 dataset did not have as good of a fit to a linear correlation as the whole genomes in TempEst (Figures 4 and 5), this was surprising.

With the constant population size being used to test which clock model is best to then go on and test the other tree priors to see which the best population dynamic is, further tree prior testing for each clock model is not needed. However, for the whole genome dataset, each and every clock model and tree prior pairing was tested in order to confirm that the best fitting clock model selection did not differ based on tree priors being used. This factorial approach was not taken for the E1 and WGE1 datasets, thus the exponentially increasing, Bayesian, and extended Bayesian tree prior models were not tested for the relaxed lognormal clock model in either the E1 dataset or the whole genome E1 dataset. Additionally, the exponentially increasing and extended Bayesian priors were not tested for the relaxed exponential clock model. Regardless of best model choice, the Bayesian skyline model was run as subsequent BEAST2 analysis with the selected clock model because that produces estimates of rubella virus effective population size over time. Tables 1-3 show the clock models and tree priors tested with their marginal likelihood

estimates. The differences in marginal likelihoods were not always large, meaning that potentially multiple priors could be appropriate. The whole genome dataset showed a very clear best fitting clock model and tree prior pairing which was the relaxed exponential clock model with a constant population size. The full E1 dataset preferred the strict clock model with the Bayesian skyline tree prior. The WGE1 dataset, despite coming from the whole genome dataset which favored a relaxed molecular clock preferred the same strict clock and Bayesian skyline priors as the larger E1 dataset.

Table 1. Path sampling analysis of the **whole genome** dataset. Marginal likelihood estimates are listed for each tree prior and clock model pairing with best selected model bolded.

	Molecular Clock Models Tested		
Tree Priors	**Strict**	**Relaxed LogNormal**	**Relaxed Exponential**
Constant Population	-48634.7	-46786.2	**-45677.6**
Exponentially Increasing Population	-45708.0	-45856.2	-45696.6
Bayesian Skyline	-45707.6	-45998.8	-45710.1
Extended Bayesian	-45707.4	-46028.0	-45753.4

Table 2. Path sampling analysis of the **E1 gene** dataset. Marginal likelihood estimates are listed for each tree prior and clock model pairing with best selected model bolded. Further tree priors were not tested for the relaxed lognormal and two priors were not tested for the relaxed exponential molecular clock.

	Molecular Clock Models Tested		
Tree Priors	**Strict**	**Relaxed LogNormal**	**Relaxed Exponential**
Constant Population	-13663.3	-13665.6	-13665.6
Exponentially Increasing Population	-13655.4	-	-
Bayesian Skyline	**-13650.1**	-	-13664.5
Extended Bayesian	-13651.9	-	-

Table 3. Path sampling analysis of the **Whole Genome E1** dataset. Marginal likelihood estimates are listed for each tree prior and clock model pairing with the best selected model bolded. Further tree priors were not tested for the relaxed lognormal and two priors were not tested for the relaxed exponential molecular clock.

	Molecular Clock Models Tested		
Tree Priors	**Strict**	**Relaxed LogNormal**	**Relaxed Exponential**
Constant Population	-6722.8	-6733.6	-6733.6
Exponentially Increasing Population	-6721.8	-	-
Bayesian Skyline	**-6720.7**	-	-6730.9
Extended Bayesian	-6733.6	-	-

3.4 Nucleotide Substitution and Analysis:

The mean nucleotide substitution rate per year of the whole genomes using the relaxed exponential clock model with constant population tree prior resulted in a rate of 1.60×10^{-3} substitutions per site per year (ssy) with the 95% Highest Posterior Density (HPD) interval range being from 1.06×10^{-3} to 2.18×10^{-3} ssy (Figure 6). The effective population size visualized using the Bayesian skyline plot (Figure 7), shows us that over time the effective population size for the whole genome dataset has remained fairly constant, giving further credence to the constant population size being the best tree prior to use.

For the E1 dataset, the strict clock model with the Bayesian prior resulted in a mean nucleotide substitution rate of 1.08×10^{-3} ssy with a 95% HPD interval of 9.50×10^{-4} to 1.21×10^{-3} ssy. This rate is lower than the mean substitution rate of the whole genome, but due to the overlap between the two HPDs should be considered a similar rate. Genes that interact with mammalian immune systems tend to not evolve slowly and often are under positive selection due to a co-evolutionary arms race (Stern and Sorek, 2010), so as a control BEAST2 analyses were run on the E1 gene from the whole genome alignment only (WGE1). Surprisingly, WGE1 had a substantially lower rate of evolution than both the whole genome dataset (8.72×10^{-4} ssy, 95% HPD 7.26×10^{-4}-1.02×10^{-3} ssy, Figure 6) and the full E1 dataset. The WGE1 HPD does not overlap with the HPD of the whole genome at all, showing that the E1 of the whole genomes is evolving significantly more slowly than the rest of the genome.

As this is an atypical result for antigenic proteins in mammalian viruses (Hicks and Duffy, 2014), we employed relaxed molecular clocks on these two E1 datasets as well. Previous substitution rate analyses of rubella virus E1 had used relaxed molecular clocks along with the Bayesian

skyline prior (Padhi and Ma, 2014, Cloete et al, 2014, Zhu et al, 2012, Jenkins 2002), and the E1 dataset had showed less of a fit to a strict linear relationship between genetic divergence and time in TempEst (5). The two E1 datasets were reanalyzed with relaxed molecular clocks and the Bayesian skyline demographic priors.

These parameters resulted in a rate of 1.507×10^{-3} substitutions per sites per year with a 95% HPD interval of 1.23×10^{-3} to 1.80×10^{-3} ssy which is quite similar to the substitution rate of the whole genomes. Incredibly, the rate estimated with the relaxed molecular clock has non-overlapping HPDs with the rate estimated with the strict molecular clock meaning that it is statistically significantly higher. The reanalyzed WGE1 dataset estimated a higher substitution rate of 1.643×10^{-3} ssy with the 95% HPD interval range being from 1.04×10^{-3} to 2.287×10^{-3} ssy, which is again statistically significantly higher than the estimate from the same dataset using a strick clock prior. Importantly, both E1 analyses with relaxed molecular clocks showed evidence that the relaxed clock was a necessary assumption for analyzing the datasets. Both analyses' distribution of results for the coefficient of variation (CoV) around the molecular clock excluded zero, which is what the CoV would frequently be for a dataset evolving in a strict clocklike manner. The CoV results indicate that the relaxed clock model was more appropriate to use than a strict molecular clock.

The time of most recent common ancestor was also reported in BEAST2, with very similar findings for the E1 and whole genome datasets. For the whole genome, the most recent common ancestor was estimated at about 111 years ago (105 years before 2013, the most recent tip date in the analysis), which would be around 1908 with a 95% HPD intervals between 1846-1951. The E1 dataset had a time of most recent common ancestor estimated of about 92 years ago (86 years before 2013, the most recent tip date in the analysis), which would be around 1927 with 95% HPD intervals between 1881-1957.

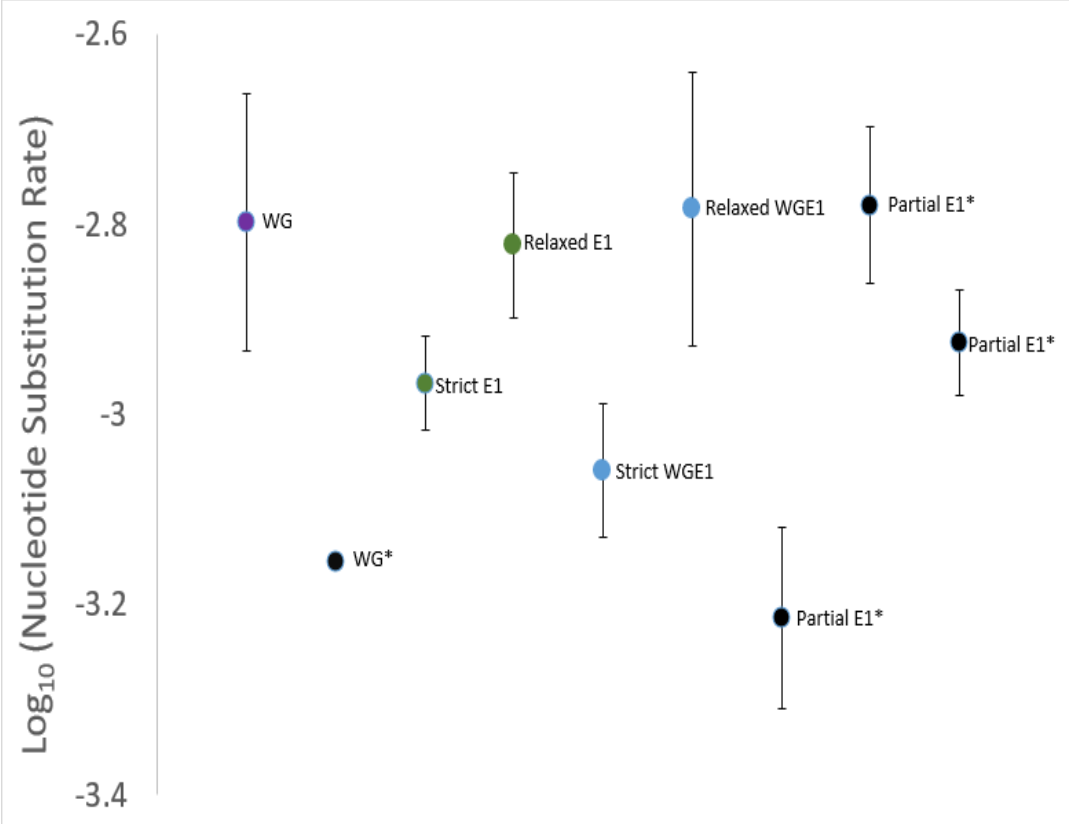

Figure 6. Log10 graph of nucleotide substitution rates of whole genome (WG), E1, and whole genome E1 (WGE1) with 95% HPD intervals. Substitution rates indicated by a (*) were taken from Cloete et al, 2014, Zhu et al, 2012, and Jenkins et al, 2002.

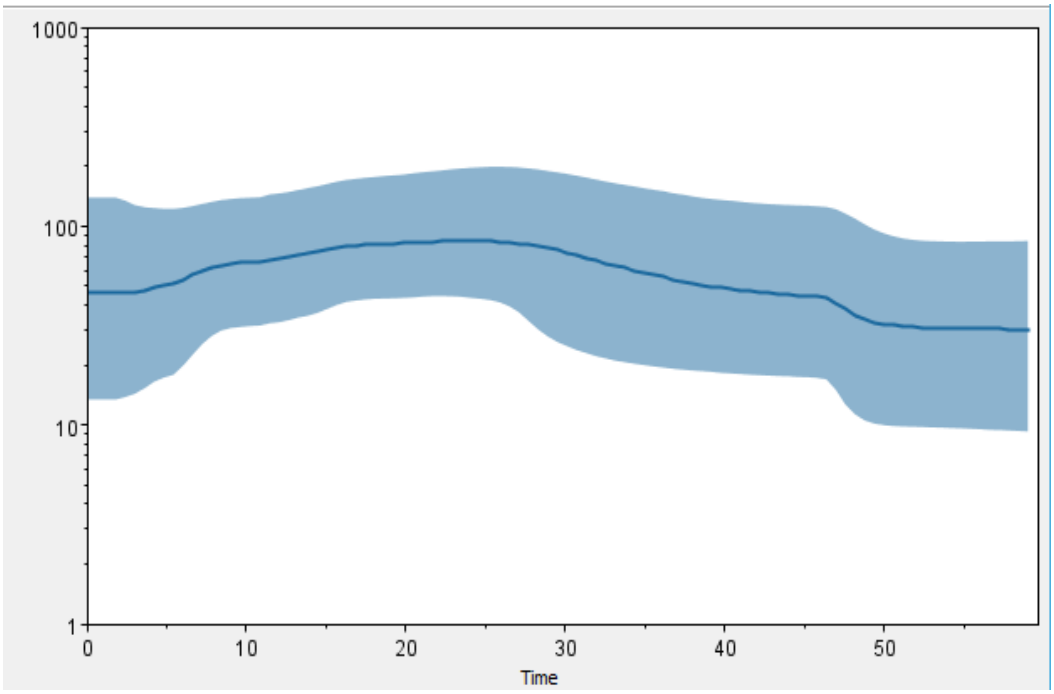

Figure 7. Bayesian skyline plot of the whole genome relaxed exponential clock model. The Y axis is showing the effective population size with the x axis showing the timeline with the numbers representing number of years from 2013. Mean nucleotide substitution rate for the whole genomes was 1.595 x 10^{-3} sites/year.

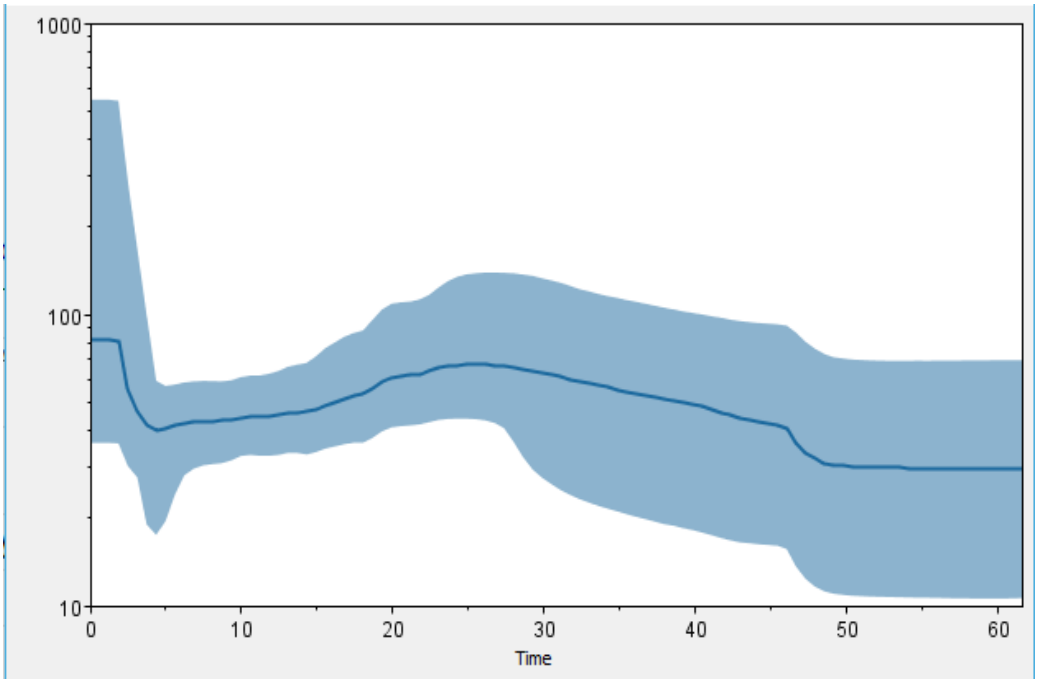

Figure 8. Bayesian skyline plot of the E1 gene relaxed exponential clock model. The Y axis is showing the effective population size with the x axis showing the timeline with the numbers representing number of years from 2013. The mean nucleotide substitution rate for the E1 dataset was 1.507 x 10^{-3} sites/year.

3.5 Phylogenetic resolution:

Maximum Clade Credibility (MCC) Trees constructed (Figures 9 and 10) show grouping primarily based on genotype and country of origin. While not all genotypes were able to be identified for all of the isolates, a significant majority were still able to be found. For the E1 gene, 10 of the genotypes present within the E1 clade were present within the dataset and 2 of the 3 genotypes present within clade 2 were present. The notable exception was the absence of the 2A genotype in the entire sample. There was very strong support for the branching and clades in the E1 and whole genome MCC trees. This is supported by the isolates coming together highly based on genotype. The only genotypes observed that were not monophyletic were 1a, 1B, and 1D, although these genotypes are not common. The two clades of genotypes for rubella cluster quite distinctly for both the E1 and WG datasets.

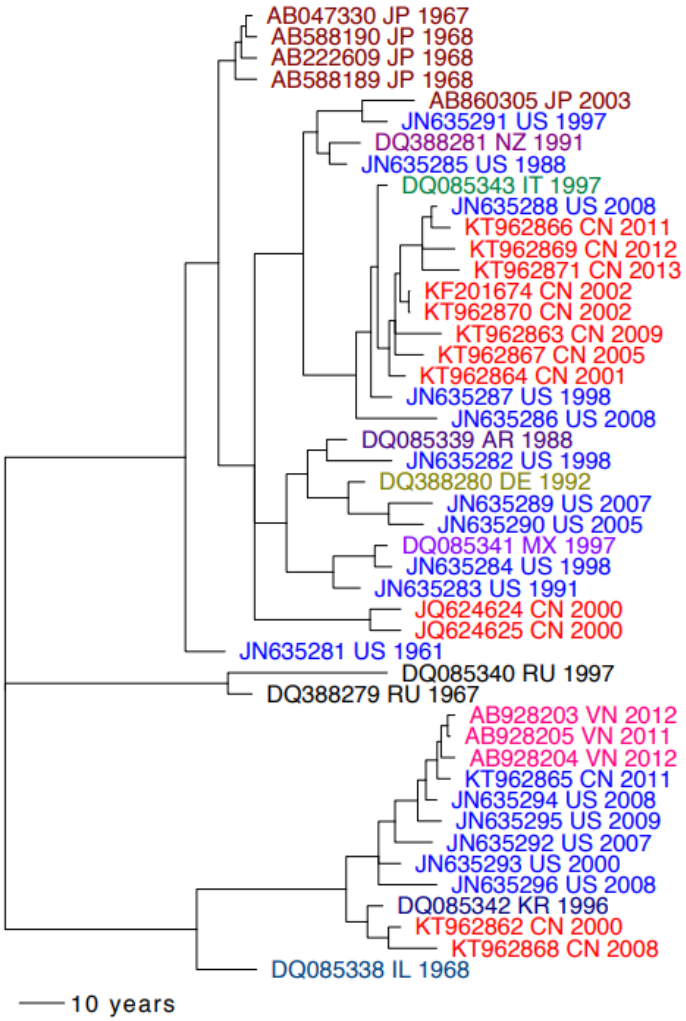

Figure 9. Maximum Clade Credibility (MCC) Tree of the Whole genome dataset. Taxa are scaled to time of isolation; branch lengths are in years. Branches showing less than 0.9 probability were collapsed and were color coded by country (also given by two letter ISO country code): Argentina China Germany Israel Italy Japan Korea Mexico New Zealand Russia United States Vietnam.

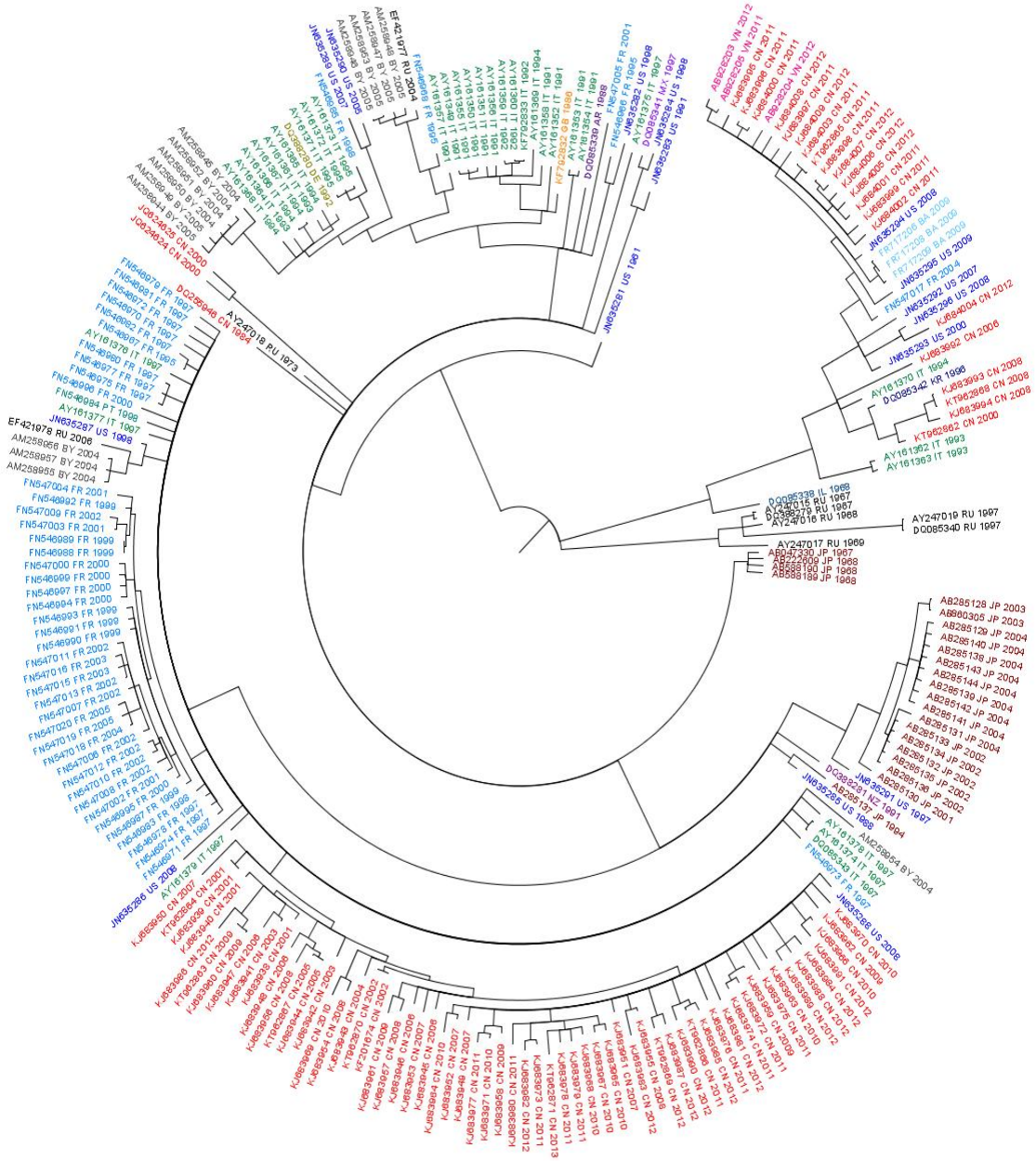

Figure 10. Circular Maximum Clade Credibility (MCC) tree of the full E1 dataset. Taxa are scaled to time of isolation. Branches showing less than 0.9 probability were collapsed and were color coded by country (also given by two letter ISO country code): Argentina Belarus Bosnia and Herzegovina China France Germany Great Britain Israel Italy Japan Korea Mexico New Zealand Portugal Russia United States Vietnam

3.6 Selection Analysis:

Of the three selection analyses used, no significant positive/diversifying selection was found in any of the three methods. SLAC analysis also showed that 0 sites were undergoing positive

selection within the E1 gene and 254 sites were undergoing negative selection. FUBAR analysis revealed 0 sites undergoing positive selection and 382 sites undergoing negative selection, confirming pervasive purifying selection along the E1 gene. Only MEME analysis showed that 1 site was undergoing positive/diversifying selection (254th codon in these alignments, 263rd in the E1 gene), the only indicator that any positive selection is occurring in recent rubella virus evolution. The evolution of the E1 protein is dominated by purifying selection.

Section 4: Discussion

The mean nucleotide substitution rates of 1.60 x 10^{-3} ssy for the whole genome and 1.50 x 10^{-3} ssy for the E1 gene reveal that the whole rubella virus genome – with both more variable non-coding regions and highly conserved genes such as for the RNA polymerase – is evolving at the same rate as the gene responsible for interaction with the mammalian immune system. This result is confirmed by the E1 genes from only the 46 whole genome isolates having an identical substitution rate to the whole genome (1.64 x 10^{-3} ssy). This similarity is understandable because the E1 gene isn't experiencing much positive selection over the studied timeframe; E1 isn't experiencing more substitutions per site per year than the other genes in rubella virus. RNA virus substitution rates are typically between 10^{-2} to 10^{-5} ssy (Hicks and Duffy, 2014) putting the evolutionary rate of the E1 gene and whole genome of the rubella virus well within that range.

The genome of rubella virus and the E1 gene in particular are both evolving at a faster rate than previously thought. Recently, a portion of the E1 gene was found to be evolving at around 1.19 x 10^{-3} nucleotides per year and the whole genome evolving at a rate of around 0.70 x 10^{-3} ssy (Cloete et al, 2014, Figure 6). Going back even further in the literature, the same portion of the E1 gene was found to be evolving at a rate of 0.61 x 10^{-3} ssy (Jenkins et al, 2002, Figure 6). This is showing a trend of faster estimates of E1 evolution over time. The nucleotide substitution rate

found within this study for the E1 gene is however much more in line with a finding that showed the rate of evolution of a couple of E1 genotypes over an eight-year period, in China alone, is around 1.6 x 10^{-3} ssy (Zhu et al, 2012). The faster E1 substitution rate in this study may be due, at least in part, to the combination of these faster evolving Chinese sequences with the global and historical distribution of rubella virus sequences. This would not explain, however, the faster estimate of the whole genome's substitution rate. While only one group has previously looked at the evolution of the whole rubella virus genome, this study's estimate is statistically higher than their rate (also estimated with a relaxed molecular clock, Figure 6). Some of the whole genome sequences used in the previous study did not pass the filtering process in this study: some sequences were used were from commercial vaccines and were extensively passaged in the laboratory. Including these lab-derived and lab-adapted sequences that are not from the natural distribution of rubella viruses may have affected their estimates. It is known that BEAST estimates with shallow TMRCAs have artifactually higher substitution rates (*e.g.*, the Chinese only study (Zhu et al, 2012)), but nothing in the literature suggests that adding additional years of sequences, as happened here in this update of the rate of rubella virus substitution rate, would affect estimates of viral substitution rates that coalesce 100 years or more ago (O'Brien et al, 2008). This study provides strong evidence that the rubella virus genome is continuously experiencing a higher substitution rate than previously thought. Rubella genomes, however, are still evolving more slowly than the other viruses in the MMR vaccine: measles (0.78 x 10^{-2}, Kuhne et al, 2006) and mumps (1.86 x 10^{-2}, Cui et al, 2009).

In response to global vaccination efforts seeking to eradicate rubella, one might think that the population size of the virus might decline, or alternatively, rapidly accelerate the speed in which the genome is evolving as rubella virus experienced diversifying selection to overcome vaccine-induced immunity. Potentially both could occur – a population thinning that leaves only viruses

with a more derived E1 gene surviving. Vaccination in this sense can be a double-edged sword because it means that although immunity is conferred via vaccination, it puts pressure on the wild type versions of the virus to evolve faster or risk dying out (Hanley, 2011). This was not the case in this study, with the whole genome and E1 gene evolving at the same rate – E1 has not been experiencing diversifying selection or evolving faster. Additionally, the effective population sizes (or more precisely the effective number of genetically distinct individuals) of the rubella virus have been shown not to be decreasing in any significant manner. While the global number of cases of rubella has declined (WHO, 2018), effective population size cannot be accurately determined by the number of people with the infection (Frost and Volz, 2010). Instead, these results show that the genetic diversity of rubella virus genomes causing infections has been fairly constant over time. The E1 effective population size shows a bit more variation, though no statistically significant changes. There may have been an insignificant decrease in E1 diversity around 20-25 years ago when China first introduced the rubella vaccine, but this is more than compensated for an estimated increase in the most recent years.

After excluding one recombinant genomic sequence that was previously identified (Cloete et al, 2014, Abernathy et al, 2013, Vauloup-Fellous et al, 2010), this study found no evidence of recombination in rubella virus. RNA viruses often recombine at rapid rates in order to proliferate diversity (Lai, 1992). While the low levels of recombination over the last hundred years do not mean that more rubella virus couldn't experience higher rates of recombination in the future, it implies that rubella relies mostly on mutation to increase its genetic diversity.

A methodological issue revealed in this study is that there were substantially different rates of evolution calculated for the E1 and WGE1 datasets when different clock models were used. Using different clock models focusing on the rate of evolution typically yield results that are well within the 95% HPD intervals for each other, meaning that normally the choice of clock model

isn't consequential for substitution rate estimation (Harrison et al, 2011, Brown and Yang, 2011). While few researchers publish BEAST results from multiple different sets of priors in a paper, a survey of virus evolution papers revealed only one published example of a similar situation, where the choice of clock prior affected the results such that there were non-overlapping HPDs (Brown and Yang, 2011). While path sampling showed that a strict clock was somewhat better fit to the E1 datasets than a relaxed clock, this was in opposition to all previous BEAST analyses of the partial E1 gene (Padhi and Ma, 2014, Cloete et al, 2014, Zhu et al, 2012, Jenkins 2002). The internal confirmation of the distribution of CoV excluding a strict clock (as signified by zero variation in rates among branches) further confirmed the appropriateness of using a relaxed clock prior to model the evolution of rubella virus E1. This is a cautionary result for the virus evolution community that path sampling may not always produce the most appropriate model priors for analysis. This result highlights that path sampling is not infallible and determining best fit models is not trivial. Model comparison in BEAST analyses was previously done using likelihood ratio tests, the Akaike information criterion (Akaike, 1974) or Harmonic Mean estimation (Newton and Raftery, 1994), all of which compare models to each other given a dataset while penalizing the more complex models. Path sampling has been shown to be more accurate than all of these methods (Baele et al, 2013), but clearly it is not a perfect method, and it may be improved upon in the future.

Section 5: Conclusions

The rubella genome is evolving more quickly than previously reported. Vaccination does not appear to be driving a change in the substitution rate of the antigenic E1 gene, nor changing the effective population size of rubella virus. This persistent rapid evolution shows that rubella has high evolutionary potential and could rapidly change and adapt in the future. This has important global health implications, especially with falling vaccine coverage in developed countries and

many countries still lacking routine vaccination against rubella. While this study suggests that current vaccines against rubella remain effective, increased knowledge of the evolution of the rubella virus can help devise next generation strategies that might be necessary to combat a potential resurgence of rubella. These include targeting the most emergent genotypes in both under vaccinated developed and developing countries. The importance of rubella evolution will allow for better preparation in a future world where rubella is found once again in developed countries.

References

Abernathy, E., Chen, M., Bera, J., Shrivastava, S., Kirkness, E., & Zheng, Q. et al. (2013). Analysis of whole genome sequences of 16 strains of rubella virus from the United States, 1961–2009. *Virology Journal*, *10*(1), 32. doi: 10.1186/1743-422x-10-32

Abernathy, E., Hübschen, J., Muller, C., Jin, L., Brown, D., & Komase, K. et al. (2011). Status of Global Virologic Surveillance for Rubella Viruses. *The Journal Of Infectious Diseases*, *204*(suppl_1), S524-S532. doi: 10.1093/infdis/jir099

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions On Automatic Control*, *19*(6), 716-723. doi: 10.1109/tac.1974.1100705

Baele, G., Li, W., Drummond, A., Suchard, M., & Lemey, P. (2013). Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. *Molecular Biology And Evolution*, *30*(2), 239-243. doi: 10.1093/molbev/mss243

Béraud, G., Abrams, S., Beutels, P., Dervaux, B., & Hens, N. (2018). Resurgence risk for measles, mumps and rubella in France in 2018 and 2020. *Eurosurveillance*, *23*(25). doi: 10.2807/1560-7917.es.2018.23.25.1700796

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C., & Xie, D. et al. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *Plos Computational Biology*, *10*(4), e1003537. doi: 10.1371/journal.pcbi.1003537

Brown, R., & Yang, Z. (2011). Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evolutionary Biology*, *11*(1), 271. doi: 10.1186/1471-2148-11-271

Brown, R., & Yang, Z. (2011). Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evolutionary Biology*, *11*(1), 271. doi: 10.1186/1471-2148-11-271

CDC. (2015). Rubella | About Rubella | CDC. Retrieved from https://www.cdc.gov/rubella/about/index.html

Chaye, H., Chong, P., Tripet, B., Brush, B., & Gillam, S. (1992). Localization of the virus neutralizing and hemagglutinin epitopes of E1 glycoprotein of rubella virus. *Virology*, *189*(2), 483-492. doi: 10.1016/0042-6822(92)90572-7

ChildVaxView | 2013-2017 Childhood MMR Vaccination Coverage Trend Report | CDC. (2017). Retrieved from https://www.cdc.gov/vaccines/imz-managers/coverage/childvaxview/data-reports/mmr/trend/index.html

Chojnacki, S., Cowley, A., Lee, J., Foix, A., & Lopez, R. (2017). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Research*, *45*(W1), W550-W553. doi: 10.1093/nar/gkx273

Cloete, L., Tanov, E., Muhire, B., Martin, D., & Harkins, G. (2014). The influence of secondary structure, selection and recombination on rubella virus nucleotide substitution rate estimates. *Virology Journal*, *11*(1), 166. doi: 10.1186/1743-422x-11-166

Cooper, L. (1985). The History and Medical Consequences of Rubella. *Clinical Infectious Diseases*, *7*(Supplement_1), S2-S10. doi: 10.1093/clinids/7.supplement_1.s2

CUI, A., MYERS, R., XU, W., & JIN, L. (2009). Analysis of the genetic variability of the mumps SH gene in viruses circulating in the UK between 1996 and 2005. *Infection, Genetics And Evolution*, *9*(1), 71-80. doi: 10.1016/j.meegid.2008.10.004

Darriba, D., Taboada, G., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 772-772. doi: 10.1038/nmeth.2109

Delport, W., Poon, A., Frost, S., & Kosakovsky Pond, S. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, *26*(19), 2455-2457. doi: 10.1093/bioinformatics/btq429

Dos Reis, M., Gunnell, G., Barba-Montoya, J., Wilkins, A., Yang, Z., & Yoder, A. (2018). Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case. *Systematic Biology*, *67*(4), 594-615. doi: 10.1093/sysbio/syy001

Fox, M., & BORTIN, M. (1946). RUBELLA IN PREGNANCY CAUSING MALFORMATIONS IN NEWBORN. *Obstetrical & Gynecological Survey*, *1*(3), 332-333. doi: 10.1097/00006254-194606000-00028

Frey, T. (1994). Molecular Biology of Rubella Virus. *Advances In Virus Research*, 69-160. doi: 10.1016/s0065-3527(08)60328-0

Frost, S., & Volz, E. (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *365*(1548), 1879-1890. doi: 10.1098/rstb.2010.0060

Geneious | Bioinformatics Software for Molecular Sequence Data Analysis. (2019). Retrieved from https://www.geneious.com/

Gibbs, M., Armstrong, J., & Gibbs, A. (2000). Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, *16*(7), 573-582. doi: 10.1093/bioinformatics/16.7.573

Grant, G., Reef, S., Patel, M., Knapp, J., & Dabbagh, A. (2017). Progress in Rubella and Congenital Rubella Syndrome Control and Elimination — Worldwide, 2000–2016. *MMWR. Morbidity And Mortality Weekly Report*, *66*(45), 1256-1260. doi: 10.15585/mmwr.mm6645a4

Gregg, N. (1941). Congenital Cataract Following German Measles in the Mother. *Epidemiology And Infection*, *107*(01), iii-xiv. doi: 10.1017/s0950268800048627

Hanley, K. (2011). The Double-Edged Sword: How Evolution Can Make or Break a Live-Attenuated Virus Vaccine. *Evolution: Education And Outreach*, *4*(4), 635-643. doi: 10.1007/s12052-011-0365-y

Harrison, A., Lemey, P., Hurles, M., Moyes, C., Horn, S., & Pryor, J. et al. (2011). Genomic Analysis of Hepatitis B Virus Reveals Antigen State and Genotype as Sources of Evolutionary Rate Variation. *Viruses*, *3*(2), 83-101. doi: 10.3390/v3020083

HESS, A. (1914). GERMAN MEASLES (RUBELLA): AN EXPERIMENTAL STUDY. *Archives Of Internal Medicine*, *XIII*(6), 913. doi: 10.1001/archinte.1914.00070120075007

Huelsenbeck, J. (2001). Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science*, *294*(5550), 2310-2314. doi: 10.1126/science.1065889

Hviid, A., Hansen, J., Frisch, M., & Melbye, M. (2019). Measles, Mumps, Rubella Vaccination and Autism. *Annals Of Internal Medicine*. doi: 10.7326/m18-2101

Jenkins, G., Rambaut, A., Pybus, O., & Holmes, E. (2002). Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. *Journal Of Molecular Evolution*, *54*(2), 156-165. doi: 10.1007/s00239-001-0064-3

Jose, J., Snyder, J., & Kuhn, R. (2009). A structural and functional perspective of alphavirus replication and assembly. *Future Microbiology*, *4*(7), 837-856. doi: 10.2217/fmb.09.59

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal Of Molecular Evolution*, *16*(2), 111-120. doi: 10.1007/bf01731581

Kühne, M., Brown, D., & Jin, L. (2006). Genetic variability of measles virus in acute and persistent infections. *Infection, Genetics And Evolution*, *6*(4), 269-276. doi: 10.1016/j.meegid.2005.08.003

Lai, M. (1992). RNA Recombination in Animal and Plant Viruses. *Microbiology Reviews*, *56*(1), 61-79.

Lam, H., Ratmann, O., & Boni, M. (2018). Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Molecular Biology And Evolution*, *35*(1), 247-251. doi: 10.1093/molbev/msx263

Lanier, H., & Knowles, L. (2012). Is Recombination a Problem for Species-Tree Analyses?. *Systematic Biology*, *61*(4), 691-701. doi: 10.1093/sysbio/syr128

Lanzieri, T., Redd, S., Abernathy, E., & Icenogle, J. (2018). Chapter 15: Congenital Rubella Syndrome. *VPD Surveillance Manual, CDC*.

Lee, J., & Bowden, D. (2000). Rubella Virus Replication and Links to Teratogenicity. *Clinical Microbiology Reviews*, *13*(4), 571-587. doi: 10.1128/cmr.13.4.571-587.2000

Lee, J., & Bowden, D. (2000). Rubella Virus Replication and Links to Teratogenicity. *Clinical Microbiology Reviews*, *13*(4), 571-587. doi: 10.1128/cmr.13.4.571-587.2000

Mackenzie, I., Prior, A., & Holzel, A. (1948). Congenital Defects following Rubella: Reports of Two Cases, One of which Shows a Hitherto Undescribed Lesion. *Journal Of Clinical Pathology*, *1*(5), 302-305. doi: 10.1136/jcp.1.5.302

Manual for the Laboratory-based Surveillance of Measles, Rubella, and Congenital Rubella Syndrome. (2015). Retrieved from https://www.who.int/immunization/monitoring_surveillance/burden/laboratory/manual_section6.5/en/

Martin, D., & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, *16*(6), 562-563. doi: 10.1093/bioinformatics/16.6.562

Martin, D., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1). doi: 10.1093/ve/vev003

Martin, D., Posada, D., Crandall, K., & Williamson, C. (2005). A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints. *AIDS Research And Human Retroviruses*, *21*(1), 98-102. doi: 10.1089/aid.2005.21.98

Measles and Rubella Surveillance Data. (2018). Retrieved from https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/active/measles_monthlydata/en/

Miller, M., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop (GCE)*. doi: 10.1109/gce.2010.5676129

MMR Vaccination | What You Should Know | Measles, Mumps, Rubella | CDC. (2019). Retrieved from https://www.cdc.gov/vaccines/vpd/mmr/public/index.html

Mumps | Cases and Outbreaks | CDC. (2018). Retrieved from https://www.cdc.gov/mumps/outbreaks.html

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S., & Scheffler, K. (2013). FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology And Evolution*, *30*(5), 1196-1205. doi: 10.1093/molbev/mst030

Murrell, B., Wertheim, J., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. (2012). Detecting Individual Sites Subject to Episodic Diversifying Selection. *Plos Genetics*, *8*(7), e1002764. doi: 10.1371/journal.pgen.1002764

Newton, M., & Raftery, A. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal Of The Royal Statistical Society: Series B (Methodological)*, *56*(1), 3-26. doi: 10.1111/j.2517-6161.1994.tb01956.x

Njagi, C. (2018). *EVOLUTION OF INDIVIDUAL GENES OF SUGARCANE MOSAIC VIRUS*(Master's). Rutgers University.

Ober, R., Horton, R., & Feemster, R. (1947). Congenital Defects in a Year of Epidemic Rubella. *American Journal Of Public Health And The Nations Health*, *37*(10), 1328-1333. doi: 10.2105/ajph.37.10.1328

O'Brien, J., She, Z., & Suchard, M. (2008). Dating the time of viral subtype divergence. *BMC Evolutionary Biology*, *8*(1), 172. doi: 10.1186/1471-2148-8-172

O'Connell, L. (2019). "Congenital Cataract following German Measles in the Mother" (1941), by Norman McAlister Gregg | The Embryo Project Encyclopedia. Retrieved from http://embryo.asu.edu/handle/10776/6888

Padhi, A., & Ma, L. (2014). Molecular Evolutionary and Epidemiological Dynamics of Genotypes 1G and 2B of Rubella Virus. *Plos ONE*, *9*(10), e110082. doi: 10.1371/journal.pone.0110082

Padidam, M., Sawyer, S., & Fauquet, C. (1999). Possible Emergence of New Geminiviruses by Frequent Recombination. *Virology*, *265*(2), 218-225. doi: 10.1006/viro.1999.0056

Phadke, V., Bednarczyk, R., Salmon, D., & Omer, S. (2016). Association Between Vaccine Refusal and Vaccine-Preventable Diseases in the United States. *JAMA*, *315*(11), 1149. doi: 10.1001/jama.2016.1353

Pinkbook | Rubella | Epidemiology of Vaccine Preventable Diseases | CDC. (2019). Retrieved from https://www.cdc.gov/vaccines/pubs/pinkbook/rubella.html

Pond, S., & Frost, S. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, *21*(10), 2531-2533. doi: 10.1093/bioinformatics/bti320

Posada, D., & Crandall, K. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings Of The National Academy Of Sciences*, *98*(24), 13757-13762. doi: 10.1073/pnas.241370698

Posada, D., & Crandall, K. (2002). The Effect of Recombination on the Accuracy of Phylogeny Estimation. *Journal Of Molecular Evolution*, *54*(3), 396-402. doi: 10.1007/s00239-001-0034-9

Rambaut, A., Drummond, A., Xie, D., Baele, G., & Suchard, M. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901-904. doi: 10.1093/sysbio/syy032

Rambaut, A., Lam, T., Max Carvalho, L., & Pybus, O. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), vew007. doi: 10.1093/ve/vew007

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology*, *25*(9), 1911-1924. doi: 10.1111/mec.13586

Rivailler, P., Abernathy, E., & Icenogle, J. (2017). Genetic diversity of currently circulating rubella viruses: a need to define more precise viral groups. *Journal Of General Virology*, *98*(3), 396-404. doi: 10.1099/jgv.0.000680

Robertson, S., Featherstone, D., Gacic-Dobo, M., & Hersh, B. (2003). Rubella and congenital rubella syndrome: global update. *Revista Panamericana De Salud Pública*, *14*(5). doi: 10.1590/s1020-49892003001000005

Rubella. (2018). Retrieved from https://www.who.int/immunization/diseases/rubella/en/

Rubella and Congenital Rubella Syndrome (CRS). (2019). Retrieved from https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/passive/rubella/en/

Rubivirus ~ ViralZone page. (2019). Retrieved from https://viralzone.expasy.org/626?outline=all_by_species

Schierup, M., & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, *156*(2), 879-891.

Simon-Loriere, E., & Holmes, E. (2011). Why do RNA viruses recombine?. *Nature Reviews Microbiology*, *9*(8), 617-626. doi: 10.1038/nrmicro2614

Smith, J. (1992). Analyzing the mosaic structure of genes. *Journal Of Molecular Evolution*, *34*(2). doi: 10.1007/bf00182389

Stern, A., & Sorek, R. (2010). The phage-host arms race: Shaping the evolution of microbes. *Bioessays*, *33*(1), 43-51. doi: 10.1002/bies.201000071

Su, Q., Ma, C., Wen, N., Fan, C., Yang, H., & Wang, H. et al. (2018). Epidemiological profile and progress toward rubella elimination in China. 10 years after nationwide introduction of rubella vaccine. *Vaccine*, *36*(16), 2079-2085. doi: 10.1016/j.vaccine.2018.03.013

Takkinen, K., Vidgren, G., Ekstrand, J., Hellman, U., Kalkkinen, N., Wernstedt, C., & Pettersson, R. (1988). Nucleotide Sequence of the Rubella Virus Capsid Protein Gene Reveals an Unusually High G/C Content. *Journal Of General Virology*, *69*(3), 603-612. doi: 10.1099/0022-1317-69-3-603

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures On Mathematics In The Life Sciences*, *17*.

Taylor, B., Miller, E., Farrington, C., Petropoulos, M., Favot-Mayaud, I., Li, J., & Waight, P. (2000). Autism and measles, mumps, and rubella vaccine: No epidemiological evidence for a causal association. *The Journal Of Pediatrics*, *136*(1), 125-126. doi: 10.1016/s0022-3476(00)90067-2

The International Medical Congress. (1881). *The British Journal Of Psychiatry*, *27*(119), 403-405. doi: 10.1192/bjp.27.119.403

Uzzell, T., & Corbin, K. (1971). Fitting Discrete Probability Distributions to Evolutionary Events. *Science*, *172*(3988), 1089-1096. doi: 10.1126/science.172.3988.1089

Vauloup-Fellous, C., Hubschen, J., Abernathy, E., Icenogle, J., Gaidot, N., & Dubreuil, P. et al. (2010). Phylogenetic Analysis of Rubella Viruses Involved in Congenital Rubella Infections in France between 1995 and 2009. *Journal Of Clinical Microbiology*, *48*(7), 2530-2535. doi: 10.1128/jcm.00181-10

Veale, H. (1866). History of an Epidemic of Rötheln, with Observations on Its Pathology. *Edinburgh Medical Journal*, *12*(5), 404-414.

Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., & Malik, M. et al. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637-641. doi: 10.1016/s0140-6736(97)11096-0

Weaver, S., Shank, S., Spielman, S., Li, M., Muse, S., & Kosakovsky Pond, S. (2018). Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Molecular Biology And Evolution*, *35*(3), 773-777. doi: 10.1093/molbev/msx335

Wesselhoeft, C. (1947). Rubella (German Measles). *New England Journal Of Medicine*, *236*(25), 943-950. doi: 10.1056/nejm194706192362506

WHO. (2005). Standardization of the nomenclature for genetic characteristics of wild-type rubella viruses. *Weekly Epidemiological Record*.

WHO. (2016). Rubella. Retrieved from https://www.who.int/immunization/diseases/rubella/en/

WHO. (2018). Measles and Rubella Surveillance Data. Retrieved from https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/active/measles_monthlydata/en/

Yalcinkaya, T. (2018). A51 Rubella genotype 1H is still circulating in Turkey. *Virus Evolution*, *4*(suppl_1). doi: 10.1093/ve/vey010.050

Yang, D., Hwang, D., Qui, Z., & Gillam, S. (1998). Effects of Mutations in the Rubella Virus E1 Glycoprotein on E1-E2 Interaction and Membrane Fusion Activity. *Journal Of Virology*, *72*(11), 8747-8755.

Zheng, D., Frey, T., Icenogle, J., Katow, S., Abernathy, E., & Song, K. et al. (2003). Global Distribution of Rubella Virus Genotypes. *Emerging Infectious Diseases*, *9*(12), 1523-1530. doi: 10.3201/eid0912.030242

Zhou, Y., Chen, X., Ushijima, H., & Frey, T. (2012). Analysis of base and codon usage by rubella virus. *Archives Of Virology*, *157*(5), 889-899. doi: 10.1007/s00705-012-1243-9

Zhu, Z., Cui, A., Wang, H., Zhang, Y., Liu, C., & Wang, C. et al. (2012). Emergence and Continuous Evolution of Genotype 1E Rubella Viruses in China. *Journal Of Clinical Microbiology*, *50*(2), 353-363. doi: 10.1128/jcm.01264-11

Zhu, Z., Rivailler, P., Abernathy, E., Cui, A., Zhang, Y., & Mao, N. et al. (2015). Evolutionary analysis of rubella viruses in mainland China during 2010–2012: endemic circulation of genotype 1E and introductions of genotype 2B. *Scientific Reports*, *5*(1). doi: 10.1038/srep07999

Appendix 1. Isolates used in the whole genome analysis.

Accession Number	*Country Isolated In*	*Year of Isolation*	*Genotype*
JN635292	USA	2007	2B
JN635295	USA	2009	2B
JN635294	USA	2008	2B
JN635281	USA	1961	1a
JN635291	USA	1997	1J
JN635293	USA	2000	2B
JN635296	USA	2008	2B
JN635290	USA	2005	1G
JN635285	USA	1988	1D
JN635284	USA	1998	1C
JN635286	USA	2008	1E
JN635289	USA	2007	1G
JN635282	USA	1998	1B
JN635288	USA	2008	1E
JN635287	USA	1998	1E
JN635283	USA	1991	1C
AB928204	Vietnam	2012	2B
AB928203	Vietnam	2012	2B
AB928205	Vietnam	2011	2B
KT962865	China	2011	2B
JQ624624	China	2000	1F
JQ624625	China	2000	1F
KT962871	China	2013	1E
KT962869	China	2012	1E
KT962866	China	2011	1E
KT962863	China	2009	1E
KT962867	China	2005	1E
KT962870	China	2002	1E
KF201674	China	2002	1E
KT962864	China	2001	1E
KT962862	China	2000	2B
KT962868	China	2008	2B
DQ085340	Russia	1997	2C
DQ388279	Russia	1967	2C
AB860305	Japan	2003	1J
AB588189	Japan	1968	-
AB222609	Japan	1968	1a
AB047330	Japan	1967	1a
AB588190	Japan	1968	-
DQ388280	Germany	1992	1G
DQ085339	Argentina	1988	1B
DQ085341	Mexico	1997	1C
DQ085342	Korea	1996	2B
DQ085343	Italy	1997	1E
DQ388281	New Zealand	1991	1D
DQ085338	Israel	1968	2B

Appendix 2. Isolates used in the E1 analysis, including of the whole genome.

Accession Number	*Country Isolated In*	*Year of Isolation*	*Genotype*
AB285128	Japan	2003	-
AB285129	Japan	2004	-
AB285140	Japan	2004	-
AB285138	Japan	2004	-
AB285143	Japan	2004	-
AB285144	Japan	2004	-
AB285139	Japan	2004	-
AB285142	Japan	2004	-
AB285141	Japan	2004	-
AB285131	Japan	2004	-
AB285133	Japan	2002	-
AB285134	Japan	2002	-
AB285132	Japan	2002	-
AB285135	Japan	2002	-
AB285136	Japan	2002	-
AB285130	Japan	2001	-
AB285137	Japan	1994	-
AY161378	Italy	1997	1E
AY161374	Italy	1997	1E
AY161379	Italy	1997	-
AY161376	Italy	1997	-
AY161368	Italy	1994	1G
AY161364	Italy	1993	1G
AY161366	Italy	1994	1G
AY161367	Italy	1994	1G
AY161361	Italy	1993	1G
AY161365	Italy	1994	1G
AY161371	Italy	1995	1G
AY161372	Italy	1995	1G
AY161373	Italy	1995	1G
AY161357	Italy	1991	-
AY161349	Italy	1991	-
AY161355	Italy	1991	-
AY161350	Italy	1991	-
AY161351	Italy	1991	-
AY161356	Italy	1991	-
AY161359	Italy	1992	-
AY161360	Italy	1992	1I
KF792833	Italy	1992	1I
AY161369	Italy	1994	-
AY161358	Italy	1991	-
AY161352	Italy	1991	-
AY161353	Italy	1991	-
AY161354	Italy	1991	-
AY161375	Italy	1997	-
AY161370	Italy	1994	2B
AY161362	Italy	1993	2B
AY161363	Italy	1993	2B
FN546973	France	1997	1E
FN546971	France	1997	1E
FN546974	France	1997	1E
FN546978	France	1997	1E

Accession Number	*Country Isolated In*	*Year of Isolation*	*Genotype*
FN546983	France	1998	1E
FN546987	France	1999	1E
FN546995	France	2000	1E
FN547002	France	2001	1E
FN547008	France	2002	1E
FN547010	France	2002	1E
FN547012	France	2002	1E
FN547006	France	2002	1E
FN547018	France	2004	1E
FN547019	France	2005	1E
FN547020	France	2005	1E
FN547007	France	2002	1E
FN547013	France	2002	1E
FN547015	France	2003	1E
FN547016	France	2003	1E
FN547011	France	2002	1E
FN546990	France	1999	1E
FN546991	France	1999	1E
FN546993	France	1999	1E
FN546994	France	2000	1E
FN546997	France	2000	1E
FN546990	France	2000	1E
FN547000	France	2000	1E
FN546988	France	1999	1E
FN546989	France	1999	1E
FN547003	France	2001	1E
FN547009	France	2002	1E
FN546992	France	1999	1E
FN547004	France	2001	1E
FN546996	France	2000	1E
FN546975	France	1997	1E
FN546977	France	1997	1E
FN546980	France	1997	1E
FN546967	France	1995	1E
FN546982	France	1997	1E
FN546970	France	1997	1E
FN546972	France	1997	1E
FN546981	France	1997	1E
FN546979	France	1997	1E
FN546985	France	1998	1G
FN546968	France	1995	1H
FN547005	France	2001	1B
FN546966	France	1995	1B
FN547017	France	2004	2B
KJ683970	China	2010	1E
KJ683962	China	2009	1E
KJ683966	China	2010	1E

Accession Number	*Country Isolated In*	*Year of Isolation*	*Genotype*
KJ683991	China	2012	1E
KJ683984	China	2012	1E
KJ683988	China	2012	1E
KJ683989	China	2012	1E
KJ683963	China	2010	1E
KJ683975	China	2011	1E
KJ683959	China	2009	1E
KJ683972	China	2011	1E
KJ683974	China	2011	1E
KJ683981	China	2012	1E
KJ683976	China	2011	1E
KJ683985	China	2012	1E
KJ683990	China	2012	1E
KJ683987	China	2012	1E
KJ683955	China	2008	1E
KJ683983	China	2012	1E
KJ683951	China	2007	1E
KJ683965	China	2010	1E
KJ683967	China	2010	1E
KJ683968	China	2010	1E
KJ683979	China	2011	1E
KJ683978	China	2011	1E
KJ683973	China	2011	1E
KJ683982	China	2012	1E
KJ683980	China	2011	1E
KJ683958	China	2009	1E
KJ683971	China	2010	1E
KJ683977	China	2011	1E
KJ683949	China	2007	1E
KJ683952	China	2007	1E
KJ683964	China	2010	1E
KJ683945	China	2006	1E
KJ683953	China	2007	1E
KJ683946	China	2006	1E
KJ683957	China	2008	1E
KJ683961	China	2009	1E
KJ683943	China	2004	1E
KJ683954	China	2008	1E
KJ683969	China	2010	1E
KJ683942	China	2003	1E
KJ683944	China	2005	1E
KJ683956	China	2006	1E
KJ683948	China	2006	1E
KJ683938	China	2001	1E
KJ683941	China	2003	1E
KJ683947	China	2006	1E
KJ683960	China	2009	1E
KJ683986	China	2012	1E
KJ683940	China	2001	1E
KJ683939	China	2001	1E

Accession Number	*Country Isolated In*	*Year of Isolation*	*Genotype*
KJ683950	China	2007	1E
DQ255946	China	1984	-
KJ683995	China	2011	2B
KJ683996	China	2011	2B
KJ684000	China	2011	2B
KJ684008	China	2012	2B
KJ683997	China	2011	2B
KJ684009	China	2012	2B
KJ684003	China	2011	2B
KJ683998	China	2011	2B
KJ684007	China	2012	2B
KJ684006	China	2012	2B
KJ684005	China	2012	2B
KJ684001	China	2011	2B
KJ683999	China	2011	2B
KJ684002	China	2011	2B
KJ684004	China	2012	2B
KJ683992	China	2008	2B
KJ683993	China	2008	2B
KJ683994	China	2008	2B
AM258954	Belarus	2004	1E
AM258955	Belarus	2004	1E
AM258957	Belarus	2004	1E
AM258956	Belarus	2004	1E
AM258944	Belarus	2005	1G
AM258949	Belarus	2005	1G
AM258950	Belarus	2004	1G
AM258951	Belarus	2004	1G
AM258952	Belarus	2004	1G
AM258945	Belarus	2004	1G
AM258946	Belarus	2005	1H
AM258953	Belarus	2005	1H
AM258947	Belarus	2005	1H
AM258948	Belarus	2005	1H
EF421978	Russia	2006	1E
AY247018	Russia	1973	-
EF421977	Russia	2004	1H
AY247015	Russia	1967	2C
AY247016	Russia	1968	2C
AY247019	Russia	1997	2C
FR717206	Bosnia	2009	2B
FR717208	Bosnia	2009	2B
FR717209	Bosnia	2009	2B
KF792832	United Kingdom	1986	1I
FN546984	Portugal	1998	1E