



Analysis of the Impact on Phylogenetic Inference of Non-Reversible Nucleotide Substitution Models

By

Rita Sianga

SNGRIT003

A thesis submitted to the graduate faculty in partial fulfilment of the requirements for the degree of Doctor of Philosophy

PhD Bioinformatics

Faculty of Health Sciences
UNIVERSITY OF CAPE TOWN

February 2023

Supervisor: **Darren Martin**, Department of Integrative Biomedical Sciences,
University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

Copyright page

Declaration

I Rita Sianga declare that this thesis is entirely my own and only aided by the guidance of my supervisor and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature:

Date: 12th February 2023.

Dedication

To my beloved mother Virginia Mutinta Sianga, for your lifetime of love for me and dedication towards ensuring that I get educated, to being my number one cheerleader ever cheering me on. I appreciate and thank you for showing the world that life problems should never dictate your ending. My hard-working single parent your hard work and prayers have just earned you a whole Ph.D. qualified daughter.

To Mete Banda, husband of my youth for the unwavering support, for being with me during every victory and sacrifice of my Ph.D. journey, to the sleepless nights and busy days because you needed to keep the home sane while I ran my never-ending research simulations. To the many times that I felt I couldn't carry on with my studies and you had to convince me that am not a quitter. Pursuing my studies while raising a family was surely made possible because of you by my side. Cheers to meeting a ghetto girl and making it your life purpose that she achieves her dreams.

To my daughters, Tirzah and Tadala. To the many times I was glued to my laptop and could not give my attention to you and yet that never stopped you both from loving me to the fullest. You girls have been a source of inspiration for me to complete my thesis.

Acknowledgements

I would like to first express my profound gratitude to my supervisor, Professor Darren Martin for offering to supervise me and for his willingness to provide great advice and support throughout my Ph.D. journey. I am indebted to him for the confidence he gave me to undertake a challenging program and for being my mentor academically and for teaching me the value of family. I thank him for teaching me what it means to be a scientific researcher, for pushing me to think outside the box and independently, and for ensuring that I did my very best and never settled for less.

I would also like to acknowledge The South African Centre for Epidemiological Modelling and Analysis (SACEMA) for believing in me and according to me with the bursary to pursue my studies. Many thanks to the DSI National Research Authority (NRF) for the continuous financial support throughout my studies.

Additionally, I would like to also thank all my fellow students at SACEMA and CBIO for creating a friendly and conducive environment during our meetings and conferences. I would also like to convey my gratitude to Gerrit Botha and Suresh Maslamoney from the IT department in the department of biomedical sciences for their valuable help and guidance during the creation of the web application RpNRM: Rooting Phylogenetic Trees Using a Non-Reversible Nucleotide Substitution.

To my late grandmother, Addie Muchimba-Sianga, the first educated woman in the family, taught me that school sets one apart, she always prayed for my success, and I know her prayers took me places. To family and friends that have always been there to celebrate my successes and always cheer me on.

To the book *Gifted hands*, to Sonya Carson, a woman I never met but whose words of perseverance to her sons turned me into believing I can do anything I set my mind at. I will forever be grateful for the inspiration provided to the 18-year-old me turning me into a fighter.

To myself, Rita Sianga A.K.A Rita Mete, to the young girl who couldn't label the alimentary canal in English in her fifth grade, to the teenager who cried for days when

she couldn't pursue a law degree, to the teenager who could have never imagined that a masters in mathematical sciences awaited her in the future worse off a Ph.D. in Bioinformatics, to the young woman who fought so hard to achieve her dreams, to the young mother and wife who juggled through family life, a demanding career and thesis. To Rita, girl thank you for daring to be different, for always being there for you, for setting the bar so high, and for ensuring that you get there.

Finally, and above all, I acknowledge my Saviour and Redeemer, my Everlasting Father and my Prince of Peace. To Him that has done exceedingly, abundantly, and above all I could have ever imaged. To Him that indeed doeth all things well. In the depth of my confusing life handling family, pregnancies, school and many more, I looked up to the hills and indeed I can testify that my help came from you. Thank you, Lord!

Table of Contents

Declaration	i
Dedication	ii
Acknowledgements	iii
List of Tables	vii
List of Figures	viii
List of Abbreviations	x
Ethical Approval and Funding	xii
Thesis Abstract	1
Background.....	1
Aims of the Thesis	1
Methods.....	2
Results.....	3
Conclusion.....	3
Chapter 1: Introduction and Literature Review	4
1.1 General Introduction	4
1.2 The Classification of Viruses and their Diverse Replication Cycles ..	6
1.3 A Review on the Dynamics of Mutations and their Impacts on the Biological Characteristics of Viruses.....	12
1.4 A Review on Modelling Nucleotide Substitution Processes	15
1.5 Thesis Organization	26
Chapter 2	26
Chapter 3	27
Chapter 4	27
Chapter 5	28
Chapter 2: Viral Genome Sequence Datasets Display Pervasive Evidence of Strand-Specific Substitution Biases That Are Best Described Using Non- Reversible Nucleotide Substitution Models	29
2.1 Introduction	29
2.2 Materials and Methods.....	32
2.2.1 Virus Sequence Datasets and Phylogenetic Trees.....	32
2.2.2 Model Testing.....	33
2.3 Results and Discussion.....	35
2.4 Conclusion	44

Chapter 3: Assessing the Impact of Model Misspecification on the Accuracy of Phylogenetic Inference.....	46
3.1 Introduction	46
3.2 Materials and Methods.....	47
3.2.1 Defining the Degree of Non-Reversibility	49
3.2.2 Confirming the Kolmogorov conditions set on the irreversibility indices	49
3.2.3 Simulation workflow.....	50
3.2.4 Quantifying the Accuracy of Phylogenetic Inferences.....	52
3.2.5 Statistical Analysis.....	53
3.3 Results and Discussion.....	53
3.3.1 Assessing the Impacts of Model Misspecification on Phylogenetic Tree Inference 53	53
3.4 Conclusion	56
Chapter 4: RpNRM: Web Application for Rooting Phylogenetic Trees Using a Non-Reversible Nucleotide Substitution Model (NREV12)	58
4.1 Introduction	58
4.2 Methods and Materials.....	61
4.2.1 The Software	61
4.2.2 Experimental Design	63
4.2.3 Quantifying the Accuracy of Phylogenetic Inferences.....	69
4.2.4 Statistical Analysis.....	69
4.3 Results and Discussion.....	69
The Outgroup Rooting Method Outperforms the RpNRM Method on Simulated Data	69
The Midpoint Rooting Method Outperforms the RpNRM Method on Empirical Data.....	73
4.4 Conclusion	77
Chapter 5: General Conclusion	79
5.1 General Discussion.....	79
5.2 Limitation of Study	82
5.3 Future Works	82
References.....	84
Appendices.....	99
Supplementary files for Chapter 3	99

List of Tables

Table 1 Number of virus species per genome composition according to the ICTV; Master Species Lists (2021).....	9
Table 2 Full details of the datasets used in the study.....	22
Table 3 AIC Scores and LRT results for double-stranded DNA virus datasets. The lowest AIC scores indicating the best-fitting models are in bold.	37
Table 4 AIC Scores and LRT results for double-stranded RNA datasets. The lowest AIC scores indicating the best fitting models are in bold.	39
Table 5 AIC Scores and LRT results for single-stranded DNA datasets. The lowest AIC scores indicating the best fitting models are in bold.	40
Table 6 AIC Scores and LRT results for single-stranded RNA datasets. The lowest AIC scores indicating the best fitting models are in bold.	43
Table 7 Scaled branch lengths for the five phylogenetic trees used as simulation templates representing sequences with approximately 95%, 90%, 85%, 80% and 75% API.	48
Table 8 Relative rate change for C to A, G to A, A to T, G to C, T to G and C to T mutations under the 11 degrees of non-reversibility alongside the maintained rates for A to C, A to G, T to A, C to G, G to T, and T to C	50
Table 9 Relative mutation rates that were used during the simulation of datasets under low DNR (0.442) and high DNR (2.193).....	67
Table 10 Details of the 28 empirical datasets used in the study.....	68
Table 11 Normalized Robinson Foulds distances between the outgroup determined root locations and the 28 empirical datasets used in the study.	74

List of Figures

Figure 1 The current existing classification of viruses which includes the Baltimore classification (I-VII) and the seven other classifications recognized by the ICTV. The square braces [] represent the intermediate change before formation of mRNA so as to aid in the identification of genomes that have the same genome type e.g., ssDNA with a dsDNA intermediate and ssDNA with a –ssDNA intermediate. 8

Figure 2 the six steps of the virus lifecycle: (1). viral cell entry from 1(a) attachment and 1(b) penetration and uncoating of the viral genome; (2). Gene expression from 2(a) mRNA synthesis (transcription) and 2(b) protein synthesis (translation); (3). Replication of viral genomes (illustration focused on Baltimore classification only); (4). Assembly of progeny viruses; and (5). Release for infection of new cells..... 11

Figure 3 The impact that APOBEC3G-mediated cytidine deamination has on HIV-1 infection is the mediation of C to U deamination which ultimately causes G to A hypermutation of the coding strand. 14

Figure 4 Ternary plots illustrating the relative fit of the NREV12, NREV6, and GTR nucleotide substitution models based on weighted AIC scores for 30 dsDNA, 31 dsRNA, 33 ssDNA, and 47 ssRNA virus nucleotide sequence datasets. These plots were produced using the Akaike weights function (where Akaike weights is the relative likelihood of the model) with an overlaid density function (implemented in the qpcR package of RStudio (Ritz & Spiess, 2008) to indicate point densities. Each model is represented by a corner of the triangles, and each circle represents the relative fit of each of the three models to a single nucleotide sequence dataset. The sides of the triangle represent model support axes ranging from 0-100%, with the position of a circle in relation to each of the sides of the triangle indicating the probability of models best describing the nucleotide sequence dataset that is represented by that point. Whereas strong red colours represent a very high density of nucleotide sequence datasets that favour a particular model, bluer colours indicate a lower, but still substantial, density of datasets that favour a model. 36

Figure 5 Phylogenetic tree inferred from an alignment of real sequences (Avian Leukosis virus) that was used to simulate datasets with DNRs varying from 0 to 20. The alignment of Avian Leukosis virus had an average pairwise sequence identity (API) of ~90% and the branches of this tree were scaled to produce four other trees reflecting branch tip sequences with approximate pairwise identities of ~75%, ~80%, ~85% and ~95%..... 48

Figure 6 Simulation workflow to assess the impact of model misspecification. 51

Figure 7 Weighted Robinson-Foulds distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility and different average pairwise sequence identities (APIs) (~75%, ~80%, ~85%, ~90% and ~95%). "ns" above a pair of box and whisker plots indicates a paired t-test adjusted p-value of greater than or equal to 0.05 and "*" indicate a paired t test adjusted p-value of <0.05 55

Figure 8 Graphical view of the web application RpNRM displaying the user interface features. <http://rpnrm.ilifu.ac.za/> 62

Figure 9 The three actual phylogenetic trees used during the simulation study. These trees differ on balance measures with tree one having colless 3, tree two having colless 10 and tree 3 having colless 20. 65

Figure 10 Normalized RF distances between true and inferred tree root locations under different rooting methods and DNR. The first two rows display results for low DNR (0.442) while the last two rows display results for high DNR (2.193)..... 71

Figure 11 Comparison of normalized RF distances between the true and inferred roots under different rooting methods for high DNR inferred trees versus the low DNR inferred trees. Each row represents results for each rooting method while each column represents results for each tree balance level. Because the size of the tree affects root estimation positively (Figure 10), this comparison assessment was done using the results for the simulation of 4000 sites 10 taxa. The p. values indicate the results of the Wilcoxon test to test for significant differences between normalized RF distances for trees inferred under the assumption of low DNR versus high DNR. ... 72

Figure 12 The association between the normalized Robinson-Faulds distance and the degree of non-reversibility, and the association between level of tree balance in inferred trees and the normalized RF distance. The green histogram represents the distribution of the DNR and Colless levels respectively while the orange histogram represents the distribution of the normalized RF distance..... 77

Figure 13 DNR values for data sets under each respective model. The dotted red line indicating 0.25 DNR threshold.111

List of Abbreviations

A	Adenine
APT	Actual Phylogenetic Tree
API	Average Pairwise Identity
AIDS	Acquired Immunodeficiency Syndrome
APOBE	Apolipoprotein B mRNA editing enzyme catalytic
APOBEC3G	Apolipoprotein B mRNA-editing catalytic polypeptide-like 3G
BC	Baltimore Classification
BCI	Baltimore Classification Group One
BCII	Baltimore Classification Group Two
BCIII	Baltimore Classification Group Three
BCIV	Baltimore Classification Group Four
BCV	Baltimore Classification Group Five
BCVI	Baltimore Classification Group Six
BCVII	Baltimore Classification Group Seven
C	Cytosine
DdDp	DNA-dependent DNA polymerases
DNA	Deoxyribonucleic acid
DNR	Degree of Non-Reversibility
dsDNA	Double stranded Deoxyribonucleic Acid
dsRNA	Double stranded Ribonucleic Acid
F81)	Felsenstein
G	Guanine
GTR	General Time-reversible Nucleotide Substitution Model
HIV	Human immunodeficiency Virus
ICTV	The International Committee on Taxonomy of Viruses
IPT	Inferred Phylogenetic Tree
IRI1	Irreversibility Index 1
IRI2	Irreversibility Index 2
IRI3	Irreversibility Index 3
JC	Jukes and Cantor

mRNA	Messenger ribonucleic acid
NNI	Nearest-neighbour interchange
NREV12	12-Rate Non-reversible Nucleotide Substitution Model
NREV6	6-rate non-reversible Nucleotide Substitution Model
RF	Robinson and Fould
RdRp	RNA-dependent RNA polymerases
s/n/c	substitutions per nucleotide site per cell infection
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
ssDNA	Single stranded Deoxyribonucleic Acid
ssRNA	Single stranded Ribonucleic Acid
SPR	Subtree Pruning and Regrafting
T	Thymine
TBR	Tree Bisection and Reconnection
U	Uracil
VOC	Variants of Concern
wRF	Weighted Robinson and Fould

Ethical Approval and Funding

The need for no Ethical clearance was obtained from the Faculty of Health Sciences' Human Research Ethics Committee at the University of Cape Town due to all data used in this study being freely accessible over the internet in genome banks.

This research was supported by funding from the National Research Foundation of South Africa through the South African Centre for Epidemiological Modelling and Analysis (SACEMA).

Thesis Abstract

Background

The vast majority of phylogenetic trees are inferred using time-reversible evolutionary models that assume that the relative rates of X to Y substitution is the same as the relative rate of Y to X substitution for any pair of characters such that evolution is modelled as if it occurs the same way both forward and backward in time. However, there is no reason to assume that the underlying biochemical mutational processes or evolutionary processes that lead to the fixation of substitutions are similarly symmetric. The computational efficiency of reversible nucleotide substitution models during phylogenetic inference has led to their widespread use irrespective of their inability to yield any information relating to the direction of evolution within phylogenetic trees. To accurately interpret the direction of evolution in phylogenetic trees, the root position must be known. I consider two non-reversible nucleotide substitution models: (1) a 6-rate non-reversible model (NREV6) in which Watson-Crick complementary substitutions occur at identical rates that is applicable to analysing mutational processes in double-stranded (ds) RNA or dsDNA genomes; and (2) a 12-rate non-reversible model (NREV12) in which all substitution types are free to occur at different rates that is applicable to analysing mutational processes in single-stranded (ss) RNA or ssDNA genomes.

Aims of the Thesis

The aim of this thesis was therefore to contribute to the accuracy with which phylogenetic trees can be inferred and rooted to reflect the actual direction and patterns of past evolution in viruses. The objectives of each specified chapter that addressed the aim were:

- To confirm the non-reversibility of evolution in diverse viral families by testing the goodness of-fit of GTR, NREV6 and NREV12 models to actual viral genome sequence data (**Chapter 2**).

- To assess how accurately maximum phylogenies are inferred when using GTR and NREV12 on simulated datasets generated using the NREV12 model and to determine the association between degrees of evolutionary non-reversibility and the degree of error in phylogenies that are inferred using GTR (**Chapter 3**).
- To determine the accuracy with which non-reversible models can be used to accurately root Phylogenetic trees (**Chapter 4**).

Methods

I evaluated the fit of NREV12, NREV6, and GTR to 141 individual viral genome sequence datasets using a previously published model test formulated in the HyPhy scripting language and used likelihood ratio and Akaike Information Criterion-based model tests to determine the best fit model of each dataset.

Further, a total of 5500 DNA datasets were simulated along a phylogenetic tree using non-reversible relative rate of nucleotide substitution at varying degrees of non-reversibility in order to assess the impact of model misspecification by assessing how accurately phylogenetic trees were inferred from the simulated datasets when a non-reversible model and reversible model were used.

I further present a web application, RpNRM (<http://rpnrm.ilifu.ac.za/>), comprising web-based front-end R shiny interface, a model-testing, and likelihood calculator implemented in HyPhy. Given a nucleotide alignment and a phylogenetic tree as inputs the application calculates the most likely root position on the tree using a 12 rate non-reversible nucleotide substitution model. I use 800 simulated DNA datasets and 28 real viral genome datasets to compare the rooting accuracy of RpNRM relative to outgroup, midpoint and IQ-TREE based rooting methods.

Results

I show that there is abundant evidence of evolutionary non-reversibility within diverse viral genome sequence datasets. Surprisingly, NREV12 provided a significantly better fit to 21/31 dsRNA and 20/30 dsDNA datasets than did the General Time Reversible (GTR) and NREV6 models with NREV6 providing a better fit than NREV12 and GTR in only 5/30 dsDNA and 2/31 dsRNA datasets. As expected, NREV12 provided a significantly better fit to 24/33 ssDNA and 40/47 ssRNA datasets.

I find that, based on tree branch lengths and topologies, phylogenetic trees were more accurately inferred when a non-reversible model (NRV12) was used as compared to the commonly used reversible GTR model. I show that increasing degrees of strand-specific substitution bias decrease the accuracy of phylogenetic inference irrespective of whether GTR or NREV12 is used to describe mutational processes. However, in cases where strand-specific substitution biases are extreme (such as in SARS-CoV-2 and Torque teno sus virus datasets) NREV12 tends to yield more accurate phylogenetic trees than those obtained using GTR.

I further show that using simulated data, contingent on sufficiently high degrees of non-reversibility in mutational processes, RpNRM can be useful at identifying root locations but not as accurate as an out-group based rooting approach.

Conclusion

I show that regardless of the high evidence of non-reversibility NREV12 should, therefore, be seriously considered during the model selection phase of phylogenetic analyses involving viral genomic sequences. RpNRM will be useful for rooting phylogenetic trees in instances where suitable outgroup sequences are unavailable and there is strong-enough support for NREV12 fitting the supplied sequence data better than reversible nucleotide substitution models.

Chapter 1: Introduction and Literature Review

Dare to be Different

1.1 General Introduction

In the kingdom of life, viruses and viroids have by far the highest rates of mutations (Duffy, Shackelton, & Holmes, 2008). This, coupled with very short generation times, the absence of a repair mechanism during the RNA replication step and large population sizes, underlies the extreme degrees of genome sequence diversity observed within viral populations (Stern & Andino, 2016), (Duffy, Shackelton, & Holmes, 2008). The ability for viruses to mutate and adapt to host environments has been the leading cause of drug resistance, immune or vaccine escape, and increased pathogenicity (Morens, Folkers, & Fauci, 2004), (Rambaut, Posada, Crandall, & Holmes, 2004), (Sanjuán & Domingo-Calap, 2016).

As is perhaps best exemplified by acquired immunodeficiency syndrome (AIDS), the disease caused by the human immunodeficiency virus (HIV), its high viral mutation rates have been the cause for its fast evolutionary change (Lemey, Rambaut, & Pybus, 2006) which has been a particularly formidable obstacle to the development of effective vaccines and therapeutic drugs (Li, et al., 2015) (Gaschen, et al., 2002) (Worobey & Han, 2012). Despite concerted and well-funded global efforts to produce vaccines for HIV since the virus was discovered in the 1980s, 40 years, over 60 million infections (Sharp & Hahn, 2011) and over 36.3 million deaths later (WHO, 2021), we still do not have anything even vaguely approaching an effective vaccine (Shin, 2016), (Ajbani, 2016). Although efforts to produce antiretroviral (ARV) therapeutics for HIV have been comparatively successful, there remains a persistent issue with the evolution of HIV variants displaying resistance to multiple different ARVs (van Zyl, Bale, & Kearney, 2018), (Bretscher, Althaus, Müller, & Bonhoeffer, 2004).

Similarly, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) pandemic (Rochman, et al., 2021) has seen the rapid rise of novel variants of concern (VOCs) primarily characterized by mutations that enhance transmissibility and/or

enable the evasion of human immunity (Naqvi, et al., 2020), (Faria, et al., 2021), (Tegally, et al., 2021), (Viana, et al., 2022), (Meng, et al., 2022). For example, the Alpha VOC first discovered in the UK and responsible for surges in cases in the UK, Europe and the USA had multiple novel mutations in the viral S-gene that encodes a glycoprotein, called spike, that is both targeted by host immunity and is used by the virus to enter host cells (Brief, 2020), (Choi & Smith, 2021), (Tao, et al., 2021). These S-gene mutations and others throughout the genome of Alpha variant viruses made them twice as transmissible as previous variants (Brief, 2020), (Meng, et al., 2021), (Tegally, et al., 2021). Similarly, the Beta and Gamma VOCs contained multiple spike mutations (some shared with the Alpha variant) that, besides enhancing transmissibility, also enabled escape from host immunity acquired through natural infection (Brüssow, 2021). The same i.e., large number of mutations also holds for other VOCs such as Delta and Omicron. It is important that the molecular processes that yielded the transmission enhancing and immune escape mediating mutations in these lineages are intensively studied since they could seriously undermine our capacity to bring the COVID-19 pandemic under control (Pachetti, et al., 2020), (Davis, et al., 2021), (Zhou, et al., 2021), (Hossain, Hassanzadeganroudsari, & Apostolopoulos, 2021).

In the case of Hepatitis C virus (HCV), arising mutations undermined our ability to treat the virus using the drug, Ribavirin (Pawlotsky, 2011). Ribavirin has in the past been used as a nucleotide analogue that inhibits the replication of various RNA viruses and has been widely used since 1986 to treat HCV (Patterson & Fernandez-Larsson, 1990). Prolonged exposure of HCV to the drug has yielded resistant HCV mutants that are capable of efficient replication even in the presence of high drug concentrations (Mejer, et al., 2020).

Antiviral drug resistance and increasingly diverse viral populations are not issues restricted to HIV, SARS-COV and HCV: the evolutionary dynamics of these viruses are merely subplots in the unimaginably bigger and more complicated story of virus evolution. Viruses are the most abundant biological entities on earth, constraining the proliferation of more complex organisms and impacting the biogeochemical cycles of the entire planet (Krupovic & Bamford, 2011).

To fully understand the dynamics of viral evolution, it is important to understand both the physical and biochemical processes that cause the different types of mutations that occur, and the ways in which individual mutations at particular genome sites impact and the biological characteristics of viruses (Peter & Pakorn, 2018), (Holmes E. C., 2009), (Moya, Holmes, & González-Candelas, 2004). In the following subsections I review the classification of viruses and their diverse replication cycles. I further review the dynamics of mutations, the impact that mutations can have on and the biological characteristics of viruses. Given that the ultimate focus of the thesis is on the modelling viral evolution, I review how mutation processes are modelled, and the shortcomings of these models.

1.2 The Classification of Viruses and their Diverse Replication Cycles

In Latin the word virus means 'venom' or 'poisonous fluid', (Lwoff, 1957). Viruses are known to infect literally all forms of organisms (Mihara, et al., 2016), (Gergerich & Dolja, 2011). The strategies used to do this are many and varied but all which require the entry of viral genomes into either the cytoplasm or nucleus of host cells (Cann, 2008), (Roossinck & Witzany, 2012), (Whittaker, Kann, & Helenius, 2000).

The International Committee on Taxonomy of Viruses (ICTV; <https://talk.ictvonline.org/>) is responsible for the assignment of all known viruses to a hierarchy of taxonomic levels (Simmonds & Aiewsakun, 2018), (International Committee on Taxonomy of Viruses Executive, 2020). Based on this classification scheme, the ICTV presently recognizes over 233 families, 168 subfamilies, 2605 genera and 10,434 species (ICTV; [Master Species Lists | ICTV](#), 2021). Crucially these recognized species represent only a small fraction of the over 10^{32} virus particles that are estimated to exist within the Earth's biosphere (Hendrix, Hatfull, Ford, Smith, & Burns, 2002), (Mushegian, 2020), (Krupovic & Bamford, 2011) (Bamford, Grimes, & Stuart, 2005).

At the highest-level viruses are classified according to whether their genomes are comprised of single stranded (ss) or double stranded (ds) DNA or RNA. Referred to

as the Baltimore classification (BC) system (Baltimore, 1971) (Sanjuán & Domingo-Calap, 2016), this classification system is based on the flow of information from viral genomes to expressed proteins which divides viruses into seven categories (Figure 1): (1) double-stranded DNA viruses (dsDNA) (BCI) where one or both strands is used for messenger RNA (mRNA) synthesis (e.g. papillomaviruses, herpesviruses, adenoviruses, poxviruses); (2) positive sense single-stranded DNA viruses (+ssDNA) (BCII) with a dsDNA intermediate where the positive or negative strand is used for mRNA synthesis (e.g. parvoviruses, geminiviruses, microviruses); (3) double-stranded RNA viruses (dsRNA) (BCIII) where the + strand is used as mRNA (rotaviruses, bursal disease virus); (4) positive sense single stranded RNA viruses (+ssRNA) (BCVI) where the strand is directly used as the mRNA (e.g., picornavirus); (5) negative sense single stranded RNA viruses (-ssRNA) (BCV) where the strand is used as a template for mRNA synthesis (e.g. influenza viruses, Ebola virus, rabies virus); (6) retro-transcribing positive sense single stranded RNA viruses (+ssRNA-RT) (BCIV) that express a reverse transcriptase enzyme and have a double stranded DNA intermediate (e.g. HIV, human T cell leukaemia virus); (7) gapped DNA viruses with a dsDNA intermediate (BCIIV) (e.g. Hepatitis B viruses) (the gapped DNA viruses were initially not a part of the Baltimore classification system (Galibert, Mandart, Fitoussi, Tiollais, & Charnay, 1979), (Kay & Zoulim, 2007), (Bruslind, 2020), (Koonin, Krupovic, & Agol, 2021).

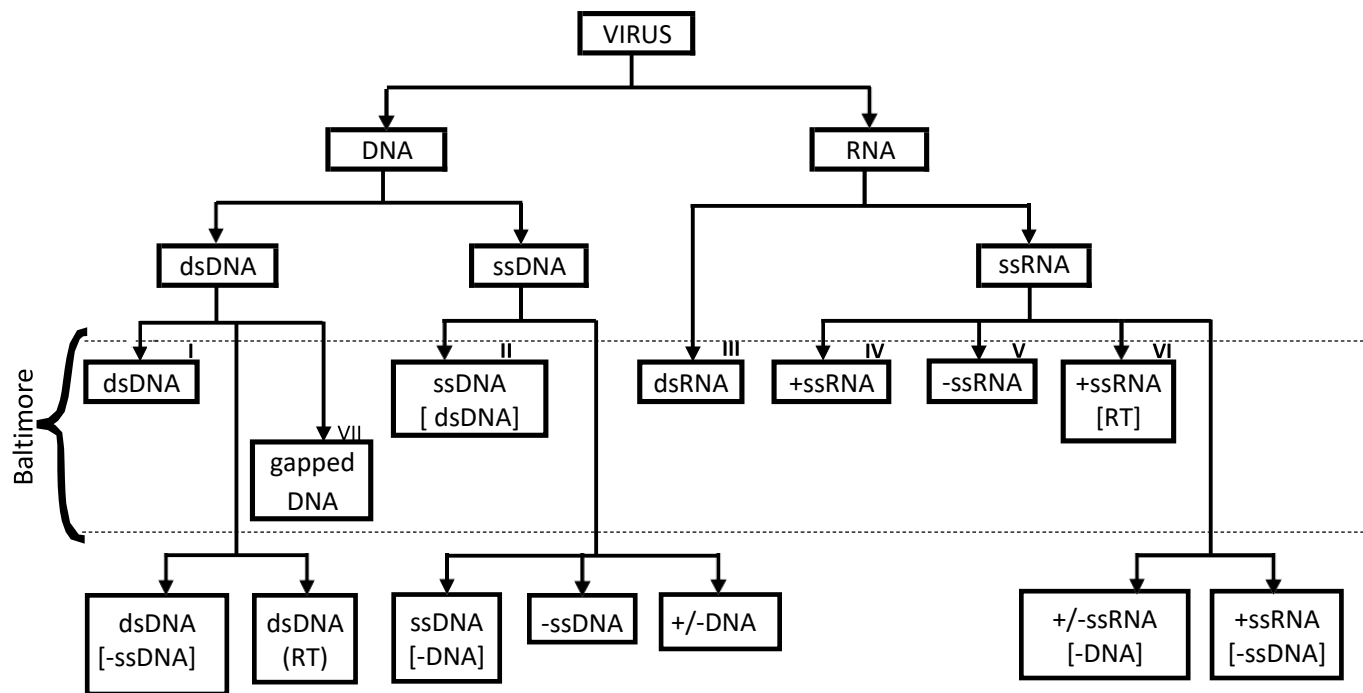


Figure 1 The current existing classification of viruses which includes the Baltimore classification (I-VII) and the seven other classifications recognized by the ICTV. The square braces [] represent the intermediate change before formation of mRNA so as to aid in the identification of genomes that have the same genome type e.g., ssDNA with a dsDNA intermediate and ssDNA with a –ssDNA intermediate.

More recently a more nuanced ICTV-recognized classification has emerged that is based on the DNA/RNA composition of the genome, its stranded-ness and the path of information flow towards protein translation (Koonin, Krupovic, & Agol, 2021) (Figure 1). With this classification we have; (1) positive sense single stranded RNA viruses with a negative sense RNA intermediate (e.g. rhinoviruses, hepatitis C virus, noroviruses, tobacco mosaic virus); (2) negative sense single stranded DNA viruses that is used as template for mRNA synthesis (e.g., Anelloviruses, mink enteritis virus, Torque teno viruses); (3) positive sense single stranded DNA with a negative single stranded DNA intermediate which is used as synthesis for mRNA; (4) double stranded DNA with a ssDNA intermediate; (5) double stranded DNA with reverse transcriptase enzyme (e.g. Tungroviruses) and lastly the antisense genomes (6) single stranded DNA (e.g. circoviruses, smacoviruses, genomiviridae) and (7) single stranded RNA where certain genes are located in one strand while others on the complementary strand (Koonin, Krupovic, & Agol, 2021) (Rosario, Duffy, & Breitbart, 2012), (Krupovic, Ghabrial, Jiang, & Varsani, 2016), (Kaczorowska & van der Hoek, 2020) (Webb, Rakibuzzaman, & Ramamoorthy, 2020).

Table 1 Number of virus species per genome composition according to the ICTV; Master Species Lists (2021).

Genome Composition	Number of Species
dsDNA	4483
dsDNA; ssDNA	2
dsDNA-RT	111
ssDNA	832
ssDNA(-)	156
ssDNA(+)	198
ssDNA(+/-)	389
dsRNA	268
ssRNA	84
ssRNA(-)	930
ssRNA(+)	2676
ssRNA(+/-)	161
ssRNA-RT	144

While some virus particles contain one nucleic acid molecule, others such as those from the Nanoviridae contain as many as 12 different molecules (Varsani, Lefevre, Roumagnac, & Martin, 2018). If multiple nucleic acid molecules are packaged within single particles the viruses are referred to as segmented and if the different molecules are each packaged within different particles the viruses are referred to as multipartite (Varsani, Lefevre, Roumagnac, & Martin, 2018), (O'Carroll & Rein, 2016). Of all the major virus groups, dsDNA viruses encompass by far the greatest swathe of known viral diversity representing the highest number of known virus species (Table 1). dsDNA viruses also have the greatest variety of particle morphologies and sizes (Krupovic & Bamford, 2011) (Bamford & Zuckerman, 2021).

Regardless of differences in genome molecular structure, in order for viruses to replicate and create new infectious viral progeny with the ability to infect other target cells, all viruses employ specific strategies (Figure 2) to:

(1) **Enter appropriate host cells:** This requires attachment and penetration into the target cells frequently with the aid of attachment factors and viral receptors found on the surface layer of target cells (Ryu, 2017). Once the virus is inside the cell (cytoplasm) and exposed to the perinuclear space, the virus will uncoat, exposing its

genetic material to cellular machinery for the process of gene expression and viral genome replication (Sattentau, 2008), (Louten, 2016), (Ryu, 2017).

(2) **Transcribe and translate their proteins.** This requires the conversion of uncoated viral genomes either directly into host-recognized mRNA (usually mediated by an encapsulated viral protein), or into a configuration that permits the host-mediated synthesis of mRNA conversion of transcription of mRNA (usually mediated by a mixture of host and viral encoded proteins) (Sattentau, 2008), (Louten, 2016), (Ryu, 2017).

(3) **Replicate their genomes.** This is usually mediated by a combination of host and viral encoded proteins (Sattentau, 2008), (Louten, 2016), (Ryu, 2017).

(4) **Assemble and package progeny virions.** This generally involves host-encoded components of intra-cellular protein transport systems and the assembly of virus encoded capsid proteins around virial nucleic acids and, in the case of enveloped viruses, may also involve the inclusion of host and viral encoded proteins into host-produced membranes that, within mature viral particles, will sheath the viral capsid (Sattentau, 2008), (Louten, 2016), (Ryu, 2017).

(5) **Move to new uninfected host cells.** These cells can either be within the same host as the originally infected cell (in the case of intra-host movement), or be in a different host organism (in the case of inter-host transmission) (Sattentau, 2008), (Louten, 2016), (Ryu, 2017).

All five of these viral strategies can (and usually do) involve virus-encoded countermeasures to overcome mechanisms and processes within host cells that have evolved specifically to protect them from viral infections (Alcami & Koszinowski, 2000), (Lanier, 2008), (Galluzzi, Brenner, Morselli, Touat, & Kroemer, 2008).

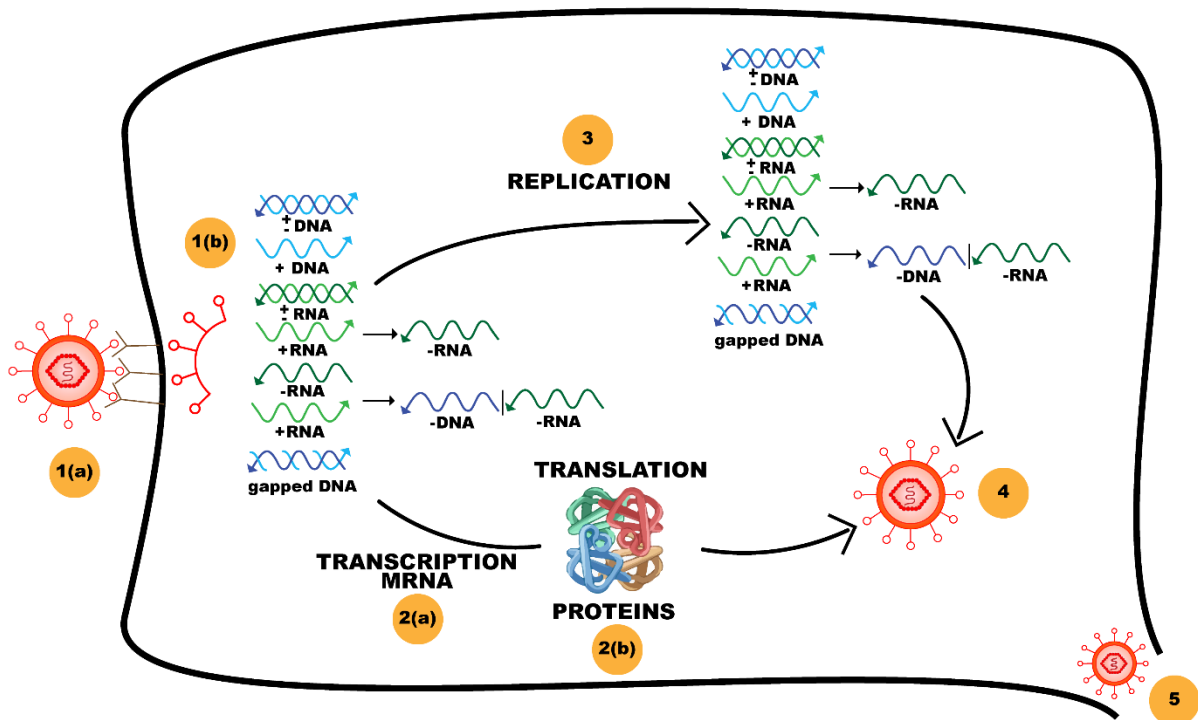


Figure 2 the six steps of the virus lifecycle: (1). viral cell entry from 1(a) attachment and 1(b) penetration and uncoating of the viral genome; (2). Gene expression from 2(a) mRNA synthesis (transcription) and 2(b) protein synthesis (translation); (3). Replication of viral genomes (illustration focused on Baltimore classification only); (4). Assembly of progeny viruses; and (5). Release for infection of new cells.

At various stages during viral life cycles, but especially commonly during replication phases, the information encoded in viral genomes can be altered or mutated (through copying errors (Lemey, Salemi, & Vandamme, 2009), (Vignuzzi & López, 2019) (Gärtner, et al., 2009) or the action of environmental radiation, chemical mutagens or host-encoded nucleic acid editing enzymes (Lemey, Salemi, & Vandamme, 2009), (Henderson, Delaney, Gu, Tannenbaum, & Essigmann, 2002) such that progeny genomes will sometimes have slight genetic differences to their parents. It is the gradual accumulation of these slight differences that is the main underlying cause of the vastness of viral genetic diversity. In the following section, I review the dynamics of viral mutations and their impacts on viral biological characteristics.

1.3 A Review on the Dynamics of Mutations and their Impacts on the Biological Characteristics of Viruses

With the vast array of virus families each having their own life cycles and mutational dynamics, it is unsurprising that mutation rates vary widely between different virus groups (Rodpothong & Auewarakul, 2012), (Grubaugh, et al., 2019), (Worobey & Holmes, 1999). The general trend is that RNA viruses have higher mutation rates (in the range of 10^{-6} – 10^{-4} substitutions per nucleotide site per cell infection (s/n/c)) than DNA viruses (in the range of 10^{-8} – 10^{-6} s/n/c) (Peck & Luring, 2018), (Risso-Ballester & Sanjuán, 2019), (Rodpothong & Auewarakul, 2012), (Sanjuán, Chirico, Mansky, & Belshaw, 2010), (Domingo & Holland, 1997), (Worobey & Holmes, 1999). Also, viruses with single-stranded genomes tend to mutate at faster rates than those with double-stranded genomes, and genomes with smaller sizes mutate at faster rates than those with bigger sizes (Sanjuán & Domingo-Calap, 2016), (Brüssow, 2021), (Peck & Luring, 2018), (Risso-Ballester & Sanjuán, 2019).

There are multiple factors that impact viral mutation rates (Sanjuán & Domingo-Calap, 2016) primary among these being (1) the basal copying fidelity of the polymerases that actually replicate DNA/RNA, (2) the presence/absence of proofreading during the replication process, (3) the mutagenicity of the cellular microenvironments within which viruses replicate, and (4) the access of replicated genomes to post-replicative repair mechanisms (Peck & Luring, 2018), (Sanjuán & Domingo-Calap, 2016), (Combe & Sanjuan, 2014). Therefore, whereas HIV replicates using reverse transcriptase (RT) that lacks 3' exonuclease proofreading activity (Svarovskaia, Cheslock, Zhang, Hu, & Pathak, 2003), (Menéndez-Arias, 2009), coronaviruses such as SARS-CoV-2 replicate using an RNA dependent RNA polymerase (RdRp) complex that is exceptional among known RNA viruses in that it has 3' exonuclease proofreading activity (Domingo, García-Crespo, Lobo-Vega, & Perales, 2021). Primarily as a consequence of this difference, HIV has a mutation rate that is almost ten times higher than most viruses (Risso-Ballester & Sanjuán, 2019) (Pachetti, et al., 2020), (Fitzsimmons, et al., 2018). Due to the rapid rate at which HIV genomes accumulate mutations, it is characterised by a high degree of genome diversity even within an infected individual (Santoro & Perno, 2013) resulting in a genetically complex

HIV population (Smyth, Davenport, & Mak, 2012). Within three weeks following the onset of an infection, an infected person will carry approximately 10^6 to 10^7 virions per ml of cell-free plasma displaying a high degree of genome diversity (Coffin, Hughes, & Varmus, 1997) (Fisher, et al., 2022) (Cuevas, Geller, Garijo, & López-Aldeguer, 2015) (Abram, Ferris, Shao, Alvord, & Hughes, 2010)

High mutation rates can be very disadvantageous for viruses. In some cases, the hosts that viruses infect will actively deploy anti-viral defences that harmfully elevate viral mutation rates. The best studied of these host-produced defensive “hypermutation” proteins are RNA editing enzymes that cause C and/or A deamination such as Apolipoprotein B mRNA-editing catalytic (APOBEC) proteins that will induce mutations within viral genes that will have a high probability of yielding either dysfunctional or truncated proteins. APOBEC belongs to a family of cellular cytidine deaminases that function specifically as an innate cellular defence against viruses such as HIV, HCV, foot and mouth disease and many others. Of particular relevance in the context of HIV infections is the action of APOBEC3G (apolipoprotein B mRNA-editing catalytic polypeptide-like 3G) (Figure 3) which causes an excess of C to U mutations in the complementary strand of the HIV genome which subsequently causes an excess of G to A mutations over A to G mutations in the virion strand (Figure 3). (Yu, et al., 2004), (Sharma, Patnaik, Taggart, & Baysal, 20016): a common consequence of which is the introduction of deleterious stop codons into viral genes. C to U hyper mutation will also yield within viral genomes substantial base frequency asymmetries such as: $U \gg A > G \gg C$ (Simmonds, 2020) (Simmonds & Ansari, 2021) (Sanjuán & Domingo-Calap, 2016) (Chelico, Pham, Calabrese, & Goodman, 2006), (Smith & Simmonds, 1997) , (Newman, et al., 2005) (Figure 3).

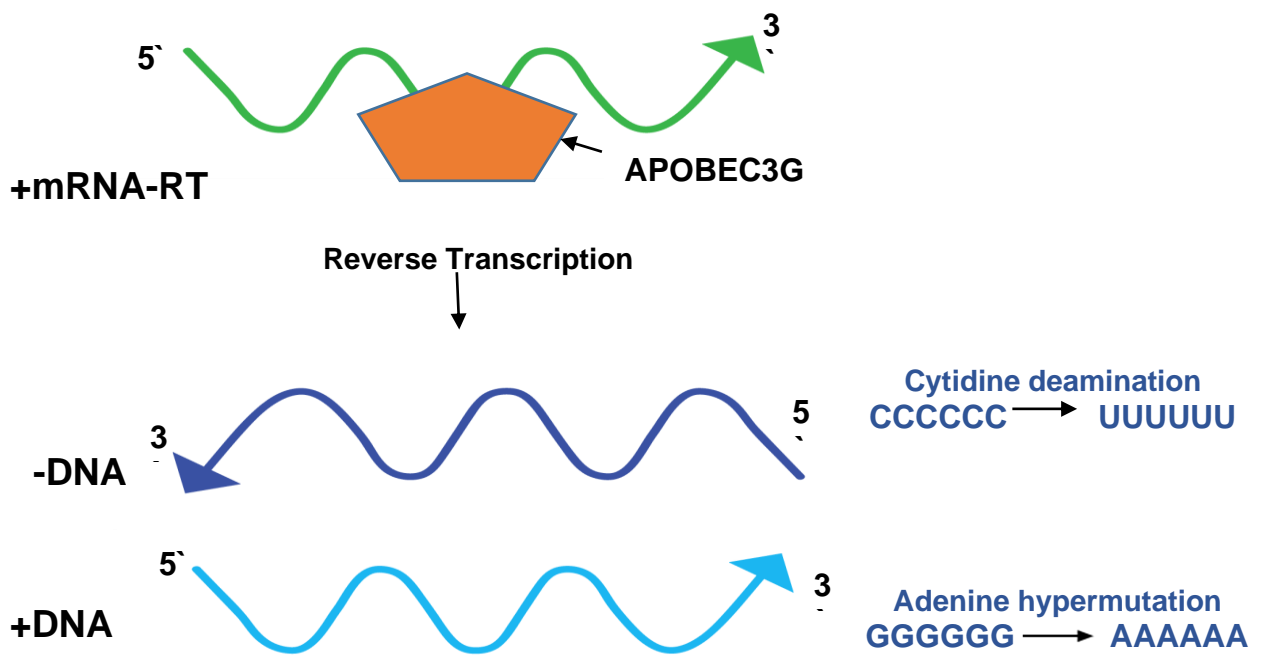


Figure 3 The impact that APOBEC3G-mediated cytidine deamination has on HIV-1 infection is the mediation of C to U deamination which ultimately causes G to A hypermutation of the coding strand.

Another example of a cellular anti-viral defence protein that could yield both large strand-specific imbalances in mutation rates, and skewed equilibrium nucleotide frequencies are ADAR proteins, adenosine deaminases that act on RNA, which promotes A > G mutations (Kustin & Adi, 2021), (Samuel, 2011).

Among the key tools that are used to understand the emergence of mutations responsible for immune/vaccine evasion, increased transmission, and anti-viral drug resistance in viruses such as HIV, SARS-CoV-2 and HCV, are statistical models of nucleotide substitution. The purpose of these models in the context of virus evolutionary studies are to explain the process of evolution accurately enough to infer the evolutionary relationships between virus genome sequences, determine the sequences of ancestral genomes and illuminate the occurrence of mutation trends that are indicative of, for example, adaptively beneficial mutations or mutation rates that are consistent enough that they can be used to infer the divergence times of viral lineages.

Given a sampled set of viral sequences, of primary interest is determining the evolutionary relationships (relatedness) between the sequences: an endeavour referred to as phylogenetics or the construction of phylogenetic trees. Essentially virus phylogenetic trees are an attempt to look backwards in time to determine, as accurately as possible how past mutations events in different viral lineages have yielded the genetic diversity observable within currently sequenced viral genomes. There are numerous ways in which nucleotide substitutions can be modelled, each of which make various assumptions relating to the mutation processes underlying the evolution of sequences (Abadi, Azouri, Pupko, & Mayrose, 2019). With many unobserved past mutations, the goal is to infer these changes using a group of observed sequences (Cavalli-Sforza & Edwards, 1967). The great task has mainly been the development of evolutionary models of nucleotide substitution through the use of statistical inference methods of choice (Cavalli-Sforza & Edwards, 1967). The most common statistical inference methods include the prominent maximum likelihood method (The ML method is a consistent method and it has been introduced for inferring phylogeny (Edwards, Cavalli-Sforza, Heywood, & McNeill, 1964), (Felsenstein, 1981) or the Bayesian inference methods (Simion, Delsuc, & Philippe, 2020). The choice on which method to use depends on the preferences of the researcher and the limitations of the datasets being analysed.

In the following subsections, I provide an in-depth discussion on how evolutionary models of nucleotide substitution are constructed, the history and discovery of widely used models, and the limitations of these models.

1.4 A Review on Modelling Nucleotide Substitution Processes

When modelling evolution, nucleotide substitutions have generally been assumed to follow a stochastic process, $X(t)$ that holds true to the Markov chain property i.e., $P(X_{t+1} = x | X_1 = x_1, X_2 = x_2 \dots X_t = x_t) = P(X_{t+1} = x | X_t = x_t)$ such that $X(t): t \geq 0$ (Baele, Gill, Lemey, & Suchard, 2019), (Strimmer, von Haeseler, & Salemi, 2003). The Markov chain property states that for a given site in a sequence alignment, the

probability that it will change from one state to another is conditioned only on the current state and is not affected by any previous state (Jermini, Jayaswal, Ababneh, & Robinson, 2017). Thus most models of evolutionary change, including those discussed in this thesis, are part of the continuous-in-time (i.e., $X = X_t, 0 \leq T < \infty$) Markov chain family of models where nucleotide bases have one of four states (Adenine, Guanine, Cytosine and Thymine), $K = 4$ i.e., in any given mutation event a change can occur from any of the four bases (A, G, C, or T) to any of the other three bases in the state space where K is the state space (Yang Z. , 1994) , (Rodriguez, Oliver, Marin, & Medina, 1990), (Baele, Gill, Lemey, & Suchard, 2019).

The role of nucleotide substitution models is to model the different possible nucleotide changes that may have occurred at any given sequence site, N , within the ancestral sequences that evolved into the sequences in an analysed sequence alignment (Jermini, Jayaswal, Ababneh, & Robinson, 2017). Specifically, the models are vital in the determination of how best the unobserved evolution can be inferred from observed nucleotide sequences i.e., the relationships between existing taxa in an alignment of nucleotide sequences, the evolutionary distances between the sequences, and the relationships between evolutionary distances, and time (Kelchner & Thomas, 2007) (Collins, Boykin, Cruickshank, & Armstrong, 2012), (Lemey, Rambaut, Drummond, & Suchard, 2009). Evolutionary modelling is generally done on a site-by-site basis because it is commonly assumed that the evolution of every genome site is independent of that of all other genome sites (Stern, et al., 2017). Mathematically this means that, given sites i, j on any DNA sequence, the probability that i changes to A, for example, does not affect the probability of what site j changes to.

The stochastic process $X(t)$ at each site, N , is determined by the rate matrix, \mathbf{R} , (equation (1)) where each entry in matrix, \mathbf{R} , is defined as r_{ij} the relative rate of change from state j to state i over a period of time, t , where $r_{ij} = -\sum_{j \neq i} r_{ij}$ and Q the instantaneous rate matrix that includes $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ a frequency distribution matrix representing the frequency of each of the four nucleotide states at equilibrium with the condition that $0 \leq \pi_i \leq 1; \forall; i$ and $\sum_i \pi_i = 1$ for all four nucleotides.

$$\mathbf{R} = \{r_{ij}\} = \begin{pmatrix} r_{AA} & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & r_{CC} & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & r_{GG} & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & r_{TT} \end{pmatrix}$$

$$\mathbf{Q} = \{R\pi_i\} = \begin{pmatrix} r_{AA}\pi_A & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{CA}\pi_A & r_{CC}\pi_C & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{GA}\pi_A & r_{GC}\pi_C & r_{GG}\pi_G & r_{GT}\pi_T \\ r_{TA}\pi_A & r_{TC}\pi_C & r_{TG}\pi_G & r_{TT}\pi_T \end{pmatrix}$$

(1)

Mathematically, the rate matrix, \mathbf{Q} , must satisfy two conditions to be considered valid i.e.:

1. $q_{ij} > 0$ for all $i \neq j$
2. $q_{ii} = -\sum_{j \neq i} r_{ij}$

The instantaneous matrix, \mathbf{Q} , and the frequency distribution matrix, π , together form the probability transition matrix, $P(t) = e^{Qt}$. This matrix defines the probability of changing from one base to another at a site, N . Solving the differential equation of the probability transition matrix $P'(t) = QP(t)$ at an initial condition of $P(0) = i$ results in $P(t) = e^{Qt}$; a probability transition matrix where the position in the ij th entry is the probability that a state i will mutate to a state j during the evolutionary time interval of length, t (Yang Z. , 2003), (Allen & Whelan, 2014) (Strimmer, von Haeseler, & Salemi, 2009).

Due to the complexity in the processes that lead to mutations, the construction of nucleotide substitution models has relied on simplifying assumptions about these processes (Abadi, Azouri, Pupko, & Mayrose, 2019). These assumptions can range from extremely strong (such as mutational processes result in all substitutions occurring at exactly the same rate) to very relaxed (such as variations in mutational processes for different nucleotides result in different substitutions happening at different rates). Starting off with the simplest model, Jukes-Cantor model (Jukes & Cantor, 1969) equation (1), that assumes that the frequencies of the four different nucleotides A, C, G and T are equal on any given genome sequence such that $\pi =$

(0.25,0.25,0.25,0.25) and that the relative rates of all possible nucleotide substitutions are the same: i.e. given i and j are bases, the rate of i to j will be the same for all base changes (i.e., $r_{ij} = \alpha; \forall i, j$; Equation (2)) (Arenas, 2015)

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \alpha\pi & \alpha\pi & \alpha\pi \\ \alpha\pi & - & \alpha\pi & \alpha\pi \\ \alpha\pi & \alpha\pi & - & \alpha\pi \\ \alpha\pi & \alpha\pi & \alpha\pi & - \end{pmatrix} \quad (2)$$

The Kimura 2 parameter model enables transversion substitutions (i.e. those between big bases such as A and G and small bases such as C and T) and transition substitutions (i.e. those from large bases to large bases or small bases to small bases) to occur at different rates (Kimura, 1980) equation (3). In reality transitions do in fact generally occur more commonly than transversions. The Kimura 2 parameter model maintains an equal frequency distribution matrix $\pi = (0.25,0.25,0.25,0.25)$ as does the JC model.

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \beta\pi & \alpha\pi & \beta\pi \\ \beta\pi & - & \beta\pi & \alpha\pi \\ \alpha\pi & \beta\pi & - & \beta\pi \\ \beta\pi & \alpha\pi & \beta\pi & - \end{pmatrix} \quad (3)$$

A more parameterized of the Kimura 2 parameter model, F81 (Felsenstein, 1981) in equation (4), not only permits transitions and transversions to occur at different frequencies but also allows for differences in the equilibrium frequencies for different nucleotides (i.e. all nucleotides are not constrained to remain at a frequency of 0.25) (Yang, Goldman, & Friday, 1994) (Zardoya, 2021) (Arenas, 2015) (Posada & Crandall, 2021).

$$Q = \{q_{ij}\} = \begin{pmatrix} - & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & - & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & - & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & - \end{pmatrix} \quad (4)$$

Multiple models have been formulated to further relax the assumptions of the K81 and

F81 models in pursuit of greater realism. To date the most realistic models that are in widespread use are those of the general time- reversible (GTR) family (Arenas, 2015). As with F81, GTR allows nucleotide frequencies to vary, but GTR additionally allows the changes between the different nucleotide states to all occur at different rates. So, for two given states, *i* and *j*, the two states have the same probability or rate of mutating from one to the other. For example, the rate, *m*, of A changing to G is equal to the rate, *n*, of G changing to A. This property is referred to as time-reversibility and yields a symmetrical rate matrix, *Q* (Equation (5) (Posada D. , 2003), (Strimmer, von Haeseler, & Salemi, 2009).

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & d\pi_T \\ a\pi_A & - & c\pi_G & e\pi_T \\ b\pi_A & c\pi_C & - & f\pi_T \\ d\pi_A & e\pi_C & f\pi_G & - \end{pmatrix} \quad (5)$$

Additional modifications have been made to GTR and other simpler models to, for example, account for (1) the occurrence of invariant nucleotide sites (which are common in datasets of closely related sequences) and (2) rate heterogeneity where different nucleotides sites across an analysed set of sequences are evolving at different rates (Zardoya, 2021).

The decision as to which nucleotide substitution model is the most appropriate for any given dataset depends on which models most accurately describe the actual mutational processes that yielded the sequences in that dataset (Arenas, 2015). It is important that the computationally convenient assumptions that are made by models – such as that the mutational process is time-reversible – are justifiable.

There is, in fact, good reason to believe that the time reversibility assumption of the most used nucleotide substitution models is biologically implausible (Cherlin, et al., 2018). In the case of double stranded DNA viruses which have two genomic DNA strands that are complementary to each other (Forsdyke, 2016), the rates of complementary substitutions (such as A to G and T to C) might be expected to be identical but not necessarily the rates of reverse substitutions (such as A to G and G to A). It is expected that mutation rates of viruses with double stranded DNA and RNA

genomes will be different from those of viruses with single stranded DNA or RNA genomes (Duffy, Shackelton, & Holmes, 2008), (Sanjuán & Domingo-Calap, 2016). Double stranded genomes are expected to not experience strand-biased mutations and their two genome strands should be similarly mutable (Frank & Lobry, 1991). Possibly because of such mutation rate symmetry, similar base compositions are commonly found in two complementary strands of double stranded viral genomes (Baisnée, Hampson, & Baldi, 2002).

Since nucleotides are always paired A with T and G with C, in the case of no strand bias, it is expected that there would be six relative substitution rates (Sueoka, 1995) such that the frequencies of A nucleotides will equal the frequency of T nucleotides and the frequency of G nucleotides will equal the frequency of C nucleotides. This holds true to the first parity rule that the total number of As will equal total number of Ts on a given piece of double stranded DNA thus causing the symmetry of complimentary substitutions (Fariselli, Taccioli, Pagani, & Maritan, 2021) (Forsdyke & Mortimer, Chargaff's legacy, 2000). When a G to A substitution occurs on one strand it is automatically expected that the substitution on the complementary strand will be C to T (Watson & Crick, 1953). Thus, for all cellular organisms and viruses with double stranded DNA or RNA genomes, a six-rate non-reversible nucleotide substitution model with a different substitution rate category for each of the six possible pairs of complementary nucleotide substitutions would best describe their evolutionary process i.e., CA=GT, GA=CT, TC=AG, TG=AC, GC=CG and TA=AT.

Non-reversibility considerations should even be more important for viruses with single stranded DNA or RNA genomes because in such viruses, mutational processes operating on the genome strand that is in existence for the longest time should drive most of the observed mutations: a dynamic that will yield a degree of asymmetry in the rate matrix Q (Lobry & Lobry, 1999). The genome strand of ssDNA and ssRNA viruses that is packaged within virions is expected to be in existence, and therefore exposed to mutagenic processes, for far longer periods of time than the complementary genome strand that is not packaged and is only expected to be in existence during replication and/or gene expression (Duffy & Holmes, 2008), (Simmonds & Ansari, 2021).

Thus, for viruses with single stranded DNA or RNA genomes, a 12-rate non-reversible nucleotide substitution model with a different substitution rate category for each of the 12 possible pairs of nucleotide substitutions would best describe their evolutionary process i.e., a specific mutation rate for mutations from CA, AC, GT, TG, AG, GA, AT, TA, CG, GC, TC and CT.

This thesis focuses on two non-reversible models of nucleotide substitution: (1) a non-reversible nucleotide substitution model with a different substitution rate category for each of the six possible pairs of complementary nucleotide substitutions (hereafter referred to as NREV6) and, (2) a non-reversible nucleotide substitution model with a different substitution rate category for each of the 12 possible pairs of nucleotide substitutions (hereafter referred to as NREV12) Equation (5). Both models assume stationarity such that nucleotide frequencies are in equilibrium throughout the process of evolution. These models have in the past been incorporated in model testing which is formulated in the HyPhy scripting language (Pond, Frost, & Muse, 2005; a standalone version of this script can be obtained from <https://github.com/veg/hyphy-analyses/tree/master/NucleotideNonREV>)

I hypothesize that for viruses with double-stranded DNA or RNA genomes both strands might be in existence for equal/similar amounts of time and hence that complementary substitutions might occur at the same rates (Baele, Van de Peer, & Vansteelandt, 2010). NREV6 might therefore be expected to best describe the evolution of nucleotide sequences from organisms with double-stranded DNA or RNA genomes. For viruses with single-stranded DNA or RNA genomes we might expect strand-specific mutation biases to occur such that NREV12 might best describe the evolution of sequences sampled from these organisms.

In this thesis, I analyse 141 viral genome datasets (Table 2) compiled from online sequence databases and test each of these to see which of GTR, NREV12 and NREV6 fit best. Using relative mutation rates from the model test, I calculate and report the degree of non-reversibility (DNR) for each of the 141 viral genome datasets. I assess the impact of model misspecification on phylogenetic inference using phylogenetic trees that are inferred from an alignment of real sequences of Avian Leukosis virus varied across different average pairwise identities (API), and I further

create and test a web application: Rooting Phylogenetic trees using a Non-Reversible Nucleotide Substitution Model (RpNRM), that roots phylogenetic trees using the NREV12 model. I show that: (1) all virus genome types (i.e., ssRNA, ssDNA, dsRNA, and dsDNA) have evidence of strand-specific nucleotide substitution biases such that, in most instances, NREV12 is the best fitting model out of NREV12, GTR, and NREV6; (2) even when sequences have evolved with a high degree of nucleotide substitution non-reversibility the accuracy of phylogenetic inference, irrespective of whether GTR or NREV12 is used to describe mutational processes, decreases; however, NREV12 tends to yield more accurate phylogenetic trees than those obtained using GTR; and (3) even if degrees of non-reversibility is sufficiently high, and phylogenetic trees are highly imbalanced, using a non-reversible nucleotide substitution model to root phylogenetic trees that are constructed using virus genome sequences is, in most cases, unlikely to yield substantially more accurate root locations than the midpoint rooting method. This thesis, therefore, promotes the use of non-reversible nucleotide substitution models both to more accurately uncover the mutational processes that yield the viral genome sequences sampled in nature, and as an alternative means of rooting viral phylogenetic trees in cases where there is no outgroup sequence, and the midpoint is inappropriate in a case of extreme imbalanced trees.

Table 2 Full details of the datasets used in the study.

Genome Type	Virus Family	Virus Genus	Virus Species	Dataset Name	Number of Sequences	Alignment Length
ssDNA	Circoviridae	Circovirus	Beak and feather disease virus	BFDV	20	2059
			Duck circovirus, Goose circovirus	DG_CV	20	10821
			Columbine circovirus	PiCV	34	2436
			Circovirus	CCCC	34	4248
			Bat circovirus	BTC	202	4019
			Porcine Circovirus 2	POCV2	97	2551
			Cyclovirus	CCV	23	3458
	Geminiviridae	Begomovirus	East Africa cassava mosaic virus, South African cassava mosaic virus	Begomo6	20	28412
			Tomato yellow leaf curl virus	Begomo5	252	18247
			Malvastrum yellow vein Yunnan virus, Cotton leaf curl Multan virus, Bhendi yellow vein India virus	Begomo9	20	1379

			Tobacco yellow dwarf virus, Chickpea chlorosis virus, Chickpea yellows virus	Dicot_1	29	3570
		Mastrevirus	Chickpea chlorotic dwarf virus	Dicot_2	20	2585
			Maize streak virus	MSV	144	2837
			Panicum streak virus	PanSV	20	2736
			Wheat dwarf virus	WDV	15	
	Anelloviridae	Anellovirus	Torque teno virus 1	TTV_1	21	
			Torque teno sus virus	TTSV	34	
	Parvoviridae	Aneptorquevirus	minute virus of mice, MVM	MVM	6	5528
		Protoparvovirus	Human parvovirus	HPV	37	6872
			Canine parvovirus	CPV	186	10786
			porcine parvovirus	PPV	149	8251
		Amdoparvovirus	Carnivore amdoparvovirus	CAV_P	24	4295
	Nanoviridae	Babuvirus	Banana bunchy top virus	BBTV_M	117	1123
				BBTV_N	100	
				BBTV_R	113	
				BBTV_S	98	
		Nanovirus	Coconut foliar decay virus	CCDV	37	2287
			Milk vetch dwarf virus full genome	MDV	12	4253
			Pea necrotic yellow dwarf virus	PYDV	160	2027
			Faba bean necrotic stunt virus	FBNS	138	1005
	Microviridae	Microviruses	Microvirus	BMV	76	
	Pleolipoviridae	Pleolipoviruses	Betapleolipovirus	BPV	16	20793
			Alphapleolipovirus	APV	10	12134
ssRNA	Astroviridae	Astroviruses	Human astrovirus	HAV	89	6184
			Bovine astrovirus	BAV	38	7665
			Mamastrovirus	MMV	17	6734
			porcine astrovirus	PAV	20	7275
			chicken astrovirus	CKV	26	8210
			Goose astrovirus	GA	12	
			Canine astrovirus	CAV_A	29	6983
	Bromoviridae	Cucumovirus	Cucumber mosaic virus	CMV_RN A1	29	3603
				CMV_RN A2	24	3141
				CMV_RN A3	27	2347
		Alphavirus	Alphafa mosaic virus	AMS	27	2437
		Cucumovirus	Peanut stunt virus	PSV	62	3691
	Caliciviridae	Lagovirus	Lagovirus	LAV	20	7478
		Coronaviruses	Norovirus	NoV	20	7697
		Vesivirus	Vesivirus	VSV	20	8513

	Closteroviridae	Closterovirus	Citrus tristeza virus	CTV	20	4925
	Flaviviridae	Flavivirus	Dengue virus	DGV_T1	20	10821
			Japanese encephalitis virus	JEV	138	11001
	Hepeviridae	Hepevirus	Hepatitis E virus	HPVE1	74	7420
			Hepatitis E2 virus	HPVE2	58	7420
	Picornaviridae	Enterovirus	Human Rhinovirus A	HRV_A	107	7552
			Enterovirus A	ENV_A	269	8817
		Teschovirus	Techovirus	TCV	24	8251
		Aichivirus	Aichivirus	AiV	17	8563
		Aphthovirus	Foot and mouth disease virus	FMDV	222	8606
			Avihepatovirus	AHP	134	7978
		Cardiovirus	Encephalomyo carditis virus	ECV	34	8080
	Cardiovirus		CDV	32	8271	
	Fusariviridae	Fusarivirus	Fusariviruses	FRV	44	11598
	Retroviridae	Lentivirus	Human immuno-deficiency virus 1	HIV1_set A	99	11001
				HIV1_M	132	11096
				HIV1_set C	40	10353
				HIV1_set D	77	10663
				HIV1_set E	66	10372
				HIV1_set F	39	10352
			Simian immuno-deficiency virus	SIV	23	5441
			Bovine immunodeficiency	BIV	19	8482
			Feline immunodeficiency	FIV	11	39907
			equine infectious anemia virus	EIV	16	5324
	caprine arthritis encephalitis virus	CAV	70	10678		
	Orthomyxo-viridae	Influenzavirus	Influenza virus A	FluA_2	25	2322
			Influenza virus B	FluB_1	25	2555
	Filoviridae	Ebolavirus	Ebola virus	Ebola_2	31	18962
	Coronaviridae	Merbecovirus	Middle East respiratory syndrome	MERS-COV	100	30123
		Sarbecoviruses	Severe acute respiratory syndrome coronavirus 1	SARS-COV1	21	30184
			Severe acute respiratory syndrome coronavirus 2	SARS-COV2	100	3344
			Sarbecoviruses	SARB	68	30927
dsDNA	Papillomaviridae	Alphapapillomavirus	Alphapapillomavirus 6	APPV 6	16	
			Alphapapillomavirus 7	HPV18_2	24	7857
				HPV45_2	13	7858
		Alphapapillomavirus 9	HPV16_2	67	8017	

				HPV31	17	7970
			Alphapapillomavirus 10	HPV6_1	126	8155
			Bovine papillomavirus	BPV	96	68621
			Lambda papillomavirus	LPV	7	8611
			Deltapapillomavirus	DPV	67	8319
			Xipapillomavirus	XPV	16	8256
	Polyomaviridae	Polyomavirus	BK polyomavirus	BK_2	71	5699
			JC polyomavirus	JC_2	87	5229
			Bat polyomavirus	BPV	26	7982
			Simian virus 40	SMV_40	222	8606
	Caulimoviridae	Caulimovirus	cauliflower mosaic virus	CMV	28	
			Cacao swollen shoot virus	CSSV	48	8766
			Strawberry vein banding virus	SVBV	15	7942
			Dioscorea bacilliform AL virus	DBAV	15	7966
			Rice tungro bacilliform virus	RTBV	13	8160
			Badnavirus	BDV	30	12255
	Siphoviridae	Escherichia virus Lambda	coliphage lambda	CLV	21	3901
	Tectiviridae	Tectivirus	Tectivirus	TTIV	33	18266
	Adenoviridae	Aviadenovirus	Fowl aviadenovirus C	FAV_C	57	41784
			Fowl aviadenovirus E	FAV_E	32	13400
			Fowl aviadenovirus A	FAV_A	11	39907
			Fowl aviadenovirus D	FAV_D	25	43915
		Mastadenovirus	Human mastadenovirus B	HMAV_B	30	10261
			Human mastadenovirus D	HMAV_D	33	35257
			Human mastadenovirus C	HMAV_C	31	36100
			Human mastadenovirus E	HMAV_E	37	35994
dsRNA	Birnaviridae	Avibirnavirus	Gumburo virus_setA	GBV_A	100	3639
			Gumburo virus_setB	GBV_B	87	3405
		Aquabirnavirus	Infectious pancreatic necrosis virus	IPNV	22	2732
			Aquabirnavirus	AQBV	36	3363
	Reoviridae	Orbivirus	Bluetongue_virus_setA	BTV_A	85	1984
			Bluetongue_virus_setB	BTV_B	82	2865
			Bluetongue_virus_setC	BTV_C	83	1784
			Bluetongue_virus_setD	BTV_D	90	1171
			Bluetongue_virus_setF	BTV_F	107	1081
			Bluetongue_virus_setG	BTV_G	87	3944
			Bluetongue_virus_setH	BTV_H	100	1137
			Bluetongue_virus_setI	BTV_I	148	825
		Rotavirus	Bovine_rotavirus_A_set C	BRVA_C	135	1072

		Human_rotavirus_A_set A	HRVA_A	43	2760	
		Human_rotavirus_A_set B	HRVA_B	40	1356	
		Human_rotavirus_A_set C	HRVA_C	40	3302	
		Human_rotavirus_A_set D2	HRVA_D2	39	1379	
		Human_rotavirus_A_set E	HRVA_E	32	1066	
		Human_rotavirus_A_set F	BRVA_F	73	751	
		Human_rotavirus_A_set G	HRVA_G	47	2591	
		Human_rotavirus_A_set H	HRVA_H	51	2361	
		Porcine_rotavirus_A_set A	PRVA_A	57	1062	
		Porcine_rotavirus_A_set B	PRVA_B	55	1356	
		Human_rotavirus_C_set A	HRVC_A	55	1073	
		Orthoreovirus	Pteropine orthoreovirus	PTOV	66	2397
		Fijivirus	Fijivirus_setB	FJV_B	145	1803
	Totiviridae	Totivirus	Totivirus	TTV	49	5090
		Giardiavirus	Giardiavirus	GDV	9	6277
	Hypoviridae	Hypovirus	Hypovirus	HPV	63	11444
	Endornaviridae	Endornavirus	Endornavirus	EDV	107	21587
		Alphaendornavirus	Bell pepper alphaendornavirus	BPAV	15	1562

1.5 Thesis Organization

This thesis is organized into five chapters. Chapters 2 to 4 drawn from manuscripts that have been submitted to peer reviewed journals for publication. Chapter 5 is a general discussion on the significance and impact of the studies presented in Chapters 2 to 4.

Chapter 2

I present two non-reversible nucleotide substitution models: NREV6 and NREV12. I further conduct and present model testing results under NREV12, NREV6 and GTR

for selected viral genome sequence datasets (including selections of ssDNA, ssRNA, dsDNA and dsRNA viruses). The goal of these model tests was to assess how well GTR, NREV6, and NREV12 describe mutational processes that operate during the evolution of viruses with ssDNA, ssRNA, dsDNA and dsRNA genomes. Specifically, I tested the relative fit of NREV12, NREV6 and GTR to 141 individual empirical viral genome sequence datasets. I show that, for most viral sequence datasets (but especially for those of ssDNA and ssRNA viruses) NREV12 is a significantly better descriptor of mutational processes than either GTR or NREV6.

Chapter 3

I present an assessment of the impacts of model misspecification on phylogenetic tree inference. Specifically, I test what impact the routine misspecification of reversible models would have on the accuracy of phylogenetic inference of sequences with strand-specific nucleotide substitution biases. Specifically, I compared the accuracy with which known phylogenetic trees are inferred for datasets simulated with varying degrees of non-reversibility and average pairwise identity levels and I reveal that even when degrees of non-reversibility are extreme, these have only a minor impact on the accuracy of phylogenetic inference using GTR as a nucleotide substitution model when compared to the non-reversible model. Here I show that using NREV12 a non-reversible nucleotide substitution model against the reversible model to construct a phylogenetic tree from virus genome sequences that display strand specific nucleotide substitution biases does not guarantee one to yield a more accurate phylogenetic tree.

Chapter 4

I present and assess the utility of a web application “Rooting phylogenetic trees using a non-reversible nucleotide substitution model (RpNRM)” for rooting phylogenetic trees using a 12 rate non-reversible model, the NREV12 model. Here I show that an increase in DNR does not affect the rooting accuracy of the outgroup method and does not consistently guarantee a better root position placement by the non-reversible methods when compared to the midpoint method. I further assess the impact that the

degree of tree balance has on the accuracy of the RpNRM and show that tree imbalance greatly impacts the accuracy of phylogenetic inference with the methods employing non-reversible nucleotide substitution models being only marginally less impacted than the midpoint rooting method while the outgroup method being completely unaffected.

Chapter 5

I provide a conclusion to this thesis and recommendations for how non-reversible nucleotide substitutions models such as NREV6 and NREV12 might be productively used in future viral molecular evolution studies.

Chapter 2: Viral Genome Sequence Datasets Display Pervasive Evidence of Strand-Specific Substitution Biases That Are Best Described Using Non-Reversible Nucleotide Substitution Models

2.1 Introduction

Modelling the nucleotide substitution processes that underly the diversification of virus genome sequences is at the heart of many viral evolutionary analyses. The most widely used nucleotide substitution models belong to the general time reversible (GTR) family (Tavaré, 1986) that assume that the Markov process of evolution will occur in the same way both forward and backward in time such that, when the arrow of time is inverted, the forward process cannot be distinguished from the backward process (Hoff, Orf, Riehm, Darriba, & Stamatakis, 2016), (Lio & Goldman, 1998), (Tavaré, 1986).

The essence of the GTR model is captured in the definition of its instantaneous rate matrix in equation (6); a matrix that models the rates at which the four different nucleotides {A,C,G,T} are exchanged, ensuring that the detailed reversibility balance condition: $Q_{ji}\pi_i = Q_{ij}\pi_j$ (where Q_{ji} is the instantaneous rate of change from j to i and π_i is the equilibrium probability of state i) is met (Squartini & Arndt, 2008), (Posada D. , 2003). The instantaneous rate matrix of the GTR model consists of six parameters (a, b, c, d, e, and f) which refer to the nucleotide substitution rates in alphabetical order and indicates the relative rates from base i to j in the state space {A, C, G, T} and the π_i 's are the equilibrium frequencies of each base.

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & d\pi_T \\ a\pi_A & - & c\pi_G & e\pi_T \\ b\pi_A & c\pi_C & - & f\pi_T \\ d\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}$$

(6)

The rate matrix in equation (6) is symmetrical in that, for example, the relative rate at which A changes to G is the same as the relative rate at which G changes to A.

Time reversible nucleotide substitution models such as GTR form the basis of almost all nucleotide sequence-focused evolutionary analyses (including those involving eukaryotes, prokaryotes, and viruses) (Lefort, Longueville, & Gascuel, 2017), (Posada & Crandall, 2001), (Minin, Abdo, Joyce, & Sullivan, 2003).

The reliability of a phylogenetic tree constructed using a particular nucleotide sequence dataset should be maximized when the evolutionary models used to construct the tree accurately reflect the evolutionary processes that yielded the nucleotide sequence dataset (Buckley & Cunningham, 2002), (Ripplinger & Sullivan, 2008), (Hoff, Orf, Riehm, Darriba, & Stamatakis, 2016). The suitability of different models for describing the evolution of DNA or RNA sequences is, therefore, expected to depend to some degree on the biological and environmental contexts of the sequences being analysed.

Mutations in viral genomes arise due to diverse biotic (such as replication enzyme infidelities, RNA/DNA editing enzymes) and abiotic (such as ionizing radiation, inorganic oxidizers and chemical mutagens) factors (Sanjuán & Domingo-Calap, 2016). Mutagenic chemical reactions or types of radiation that, for example, cause G to A or C to U mutations in DNA or RNA, are not the same as those that cause A to G or U to C mutations (Cheng, Cahill, Kasai, Nishimura, & Loeb, 1992), (Nguyen, et al., 1992), (Chelico, Pham, Calabrese, & Goodman, 2006), (Sharma, Patnaik, Taggart, & Baysal, 20016). It should not be expected, therefore, that the relative rates of G to A substitution will equal those of A to G substitution. Instead, in evolving double-stranded (ds) DNA and dsRNA molecules where both strands of the genome are in existence for similar amounts of time, both G to A and C to T substitutions should occur at similar rates. Therefore, for nucleotide sequence datasets derived from any organisms with

dsDNA or dsRNA genomes, a non-reversible nucleotide substitution model with a different relative substitution rate category for each of the six possible pairs of complementary nucleotide substitutions (e.g. NREV6 in equation (7), might plausibly provide a better description of mutational processes than GTR (Baele, Van de Peer, & Vansteelandt, 2010), (Wickner, 1993).

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ f\pi_A & - & d\pi_G & e\pi_T \\ e\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & b\pi_C & a\pi_G & - \end{pmatrix} \quad (7)$$

In the case of single-stranded (ss) RNA viruses, ssDNA viruses, retroviruses, and dsRNA/dsDNA viruses where the two complementary genome strands do not exist for equal amounts of time (Yu, et al., 2004), a model where all twelve different substitutions are free to occur at different rates might be best. Specifically, with ssRNA viruses, ssDNA viruses, and retroviruses, only one of the genome strands (called the virion strand) is packaged into viral particles for transmission and, in many dsRNA viruses, the genome strand that is translated into proteins (called the + strand) exists for longer during the life cycle than does the complementary (or –) strand (Bruslind, 2020), (Onwubiko, et al., 2020). In all these viruses, some degree of strand-specific substitution bias is expected to occur (Van Der Walt, Martin, Varsani, Polston, & Rybicki, 2008), (Polak & Arndt, 2008) such that NREV6 might be anticipated to provide a poorer description of mutational processes than a model such as NREV12 (equation (8)) where each of the twelve different types of substitution has a different relative rate (Baele, Van de Peer, & Vansteelandt, 2010).

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ g\pi_A & - & d\pi_G & e\pi_T \\ h\pi_A & i\pi_C & - & f\pi_T \\ j\pi_A & k\pi_C & l\pi_G & - \end{pmatrix} \quad (8)$$

2.2 Materials and Methods

2.2.1 Virus Sequence Datasets and Phylogenetic Trees

I obtained nucleotide sequences from the National Centre for Biotechnology Information Taxonomy database (<http://www.ncbi.nlm.nih.gov/taxonomy>) and also obtained from the Los Alamos National Laboratory HIV sequence database (<https://www.hiv.lanl.gov/content/index>). These included gene and whole-genome sequences for viruses with ssRNA, ssDNA, dsRNA, and dsDNA genomes (datasets are summarized in Table 2). An outgroup sequence from a closely related virus species was added to each dataset to help root phylogenetic trees accurately. The sequences in each of the datasets were aligned using MUSCLE (Edgar, 2004) implemented in Aliview (Larsson, 2014).

To obtain a phylogenetic tree, among many methods of phylogenetic tree reconstruction, of use in this research is the widely used maximum likelihood method (Schmidt, Strimmer, Vingron, & Von Haeseler, 2002) where phylogenetic trees were constructed from each alignment using RAxML v8.2 (Stamatakis, 2016).

Felsenstein (1981) describes maximum likelihood methods in tree reconstruction as the process of finding the evolutionary tree that yields the highest probability of having evolved the data that is being observed in the alignment. Given the alignment sequences as data D , the model M : containing transition probabilities describing the process of evolution at each given site, the parameters ϕ consisting of tree topology T , branch lengths l , the rates in the rate matrix Q which are gamma distributed with scale parameter α , the likelihood of the tree is defined as: $L = P(D|M, \phi)$. Using RAxML, a starting tree is selected which consists of leaves/nodes as observed in the data set D . The tree will consist of internal nodes called ancestral nodes but whose sequences are unknown. In RAxML, you can either set the sequence that should be considered as outgroup hence rooted on or allow it to hypothetically root at a node that seems convenient. Since only reversible models are used to reconstruct trees, it means the root selected will not influence the likelihood value due to the assumption of reversibility.

The likelihood calculation is done for each site because it is generally assumed that site evolution is independent of all other sites on a DNA sequence (Schöniger & Von Haeseler, 1994). Mathematically this means given sites i, j on a DNA sequence, the probability that i changes to A for example does not affect the probability of what site j changes to. This means we can find the likelihood of each site without being affected by other sites. To find the likelihood of one site, the steps are described below as done by Strimmer (1997). To begin, the prior probability (equilibrium frequency) of the state i at the root r is determined. Then the transition probabilities for each branch length from the root all the way till the external nodes (the current tree leaves or taxa) are calculated for each possible state and added. Given that there exists n internal nodes including the root, the likelihood for one site is calculated then the product of the probabilities (likelihoods) for each site gives the likelihood of the entire dataset D . To then find the likelihood of the entire sequence, we find the product of all sums of all possible transition probabilities at each site. When setting the rates of the substitutions for sites, we take into account that rates are not the same for all sites but rather each site rate is picked independent of other sites from a distribution of rates. The gamma distribution has been and still is the widely used choice of rates distribution.

2.2.2 Model Testing

I evaluated the fit of NREV12, NREV6, and GTR to the 141 individual sequence datasets using a previously published model test (Harkins, et al., 2009) formulated in the HyPhy scripting language (Pond, Frost, & Muse, 2005). This script (obtainable from <https://github.com/veg/hyphy-analyses/tree/master/NucleotideNonREV>) took as input a rooted maximum likelihood phylogenetic tree (minus the rooting sequence) and its corresponding nucleotide sequence alignment. The first step of the model testing process involved the harvesting of nucleotide sequences $\left(\frac{\pi_i}{\sum_{i=1}^4 \pi_i}\right)$ from the sequence alignment into a vector of frequencies $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ called the frequency distribution matrix containing three free parameters with that of nucleotide T conditioned at absolute $(1 - \pi_A - \pi_C - \pi_G)$.

Once the frequencies were harvested, the first stochastic rate matrix consisting only

of the relative rates picked from a gamma distribution was defined to satisfy the reversibility conditions of relative rates being equal in reverse i.e., $r_{AG}=r_{GA}$, $r_{AC}=r_{CA}$, $r_{AT}=r_{TA}$, $r_{CG}=r_{GC}$, $r_{TG}=r_{GT}$, $r_{CT}=r_{TC}$. Thereafter the instantaneous rate matrix Q was calculated by multiplying the relative substitution rates by the appropriate nucleotide frequencies which were used to form the GTR model probability transition matrix, P by $P(t) = e^{Qt}$. The role of the GTR model during model testing was to model mutations along the branches of tree, T . Given parameters of the GTR model describing (1) equilibrium nucleotide frequencies, and (2) the nucleotide substitution process, the likelihood, L , of the observed data, D , was calculated with the values of all the independent parameters, Θ , being investigated to find the combination that maximized the value of L (maximizing $L(\Theta|D, T)$). The value of the log likelihood ($\ln L$) under the GTR model was, at this point, stored for future comparisons with those of the NREV6 and NREV12 models.

The model was then changed to one satisfying the complementary relative reversibility conditions of the NREV6 model: i.e. $r_{AG}=r_{TC}$, $r_{AC}=r_{TG}$, $r_{AT}=r_{TA}$, $r_{CG}=r_{GC}$, $r_{CA}=r_{GT}$, $r_{CT}=r_{GA}$. The $\ln L$ of observing the data, D , given the tree, T , under the NREV6 model was calculated and stored for later comparisons. The model was then changed to the NREV12 model for which relative rates were defined such that it satisfied complete non-reversibility.

For each dataset I then used the $\ln L$ scores and numbers of free parameters in the three models for likelihood ratio tests (LRTs; (Anisimova, Bielawski, & Yang, 2001)) to determine whether (1) NREV12 fitted the data significantly better than GTR and (2) whether NREV12 fitted the data significantly better than NREV6. Specifically, for the NREV12 vs GTR comparison we calculated the LRT statistic as $2(\ln L_{NREV12} - \ln L_{GTR})$ with the p value being calculated as $1 - \text{chi}(\text{LRT}, df_{NREV12} - df_{GTR})$. For the NREV12 vs NREV6 comparison we calculated the LRT statistic as $2(\ln L_{NREV12} - \ln L_{NREV6})$ with the p value being calculated as $1 - \text{chi}(\text{LRT}, df_{NREV12} - df_{NREV6})$.

Further, the $\ln L$ scores and numbers of free parameters for each model were used to calculate AIC scores for each of the models (equation (9)) (Posada & Buckley, 2004) which enabled us to identify which of the three models fit the data best. The model

with the lowest AIC score was selected as the best fitting model with the AIC scores for the different models being calculated as follows:

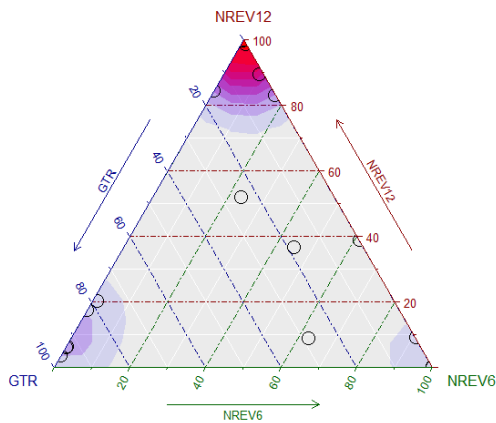
$$\begin{aligned}AIC_{GTR} &= 2(df_{GTR}) - 2(\ln L_{GTR}) \\AIC_{NREV6} &= 2(df_{NREV6}) - 2(\ln L_{NREV6}), \\AIC_{NREV12} &= 2(df_{NREV12}) - 2(\ln L_{NREV12}).\end{aligned}\tag{9}$$

Where degrees of freedom for each model is number of free parameters i.e. five substitution rates and three base frequencies for the GTR and NREV6 models, and eleven free parameters and three base frequencies for the NREV12 model.

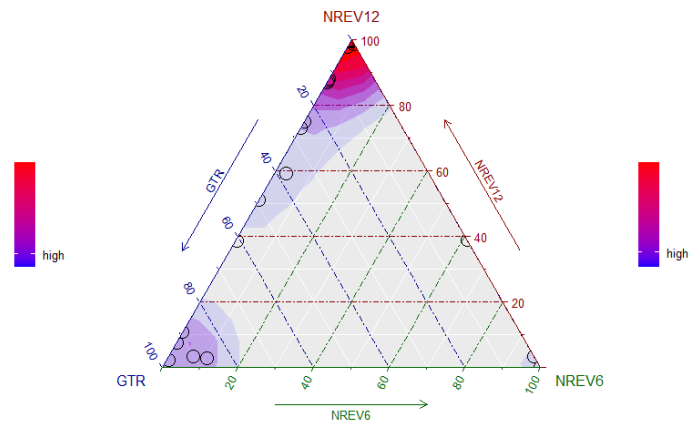
2.3 Results and Discussion

I tested for evidence of non-reversibility in the nucleotide substitution process in 141 virus sequence datasets (**Table 2**) (33 ssDNA virus datasets, 30 dsDNA virus datasets, 31 dsRNA virus datasets, and 47 ssRNA virus datasets, all consisting of either full genome sequences (for unsegmented viruses), or complete genome component sequences (for viruses with segmented genomes). Specifically, for each dataset, I compared the goodness-of-fit of the GTR+G, NREV6+G, and NREV12+G models (where G represents gamma-distributed rates).

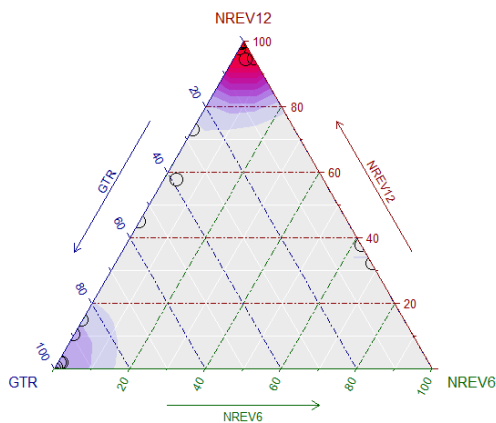
dsDNA Viruses



dsRNA Viruses



ssDNA Viruses



ssRNA Viruses

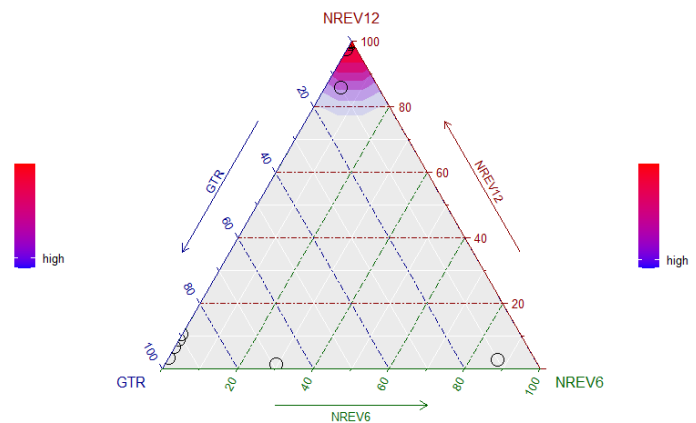


Figure 4 Ternary plots illustrating the relative fit of the NREV12, NREV6, and GTR nucleotide substitution models based on weighted AIC scores for 30 dsDNA, 31 dsRNA, 33 ssDNA, and 47 ssRNA virus nucleotide sequence datasets. These plots were produced using the Akaike weights function (where Akaike weights is the relative likelihood of the model) with an overlaid density function (implemented in the qpcR package of RStudio (Ritz & Spiess, 2008) to indicate point densities. Each model is represented by a corner of the triangles, and each circle represents the relative fit of each of the three models to a single nucleotide sequence dataset. The sides of the triangle represent model support axes ranging from 0-100%, with the position of a circle in relation to each of the sides of the triangle indicating the probability of models best describing the nucleotide sequence dataset that is represented by that point. Whereas strong red colours represent a very high density of nucleotide sequence datasets that favour a particular model, bluer colours indicate a lower, but still substantial, density of datasets that favour a model.

Given that dsDNA viruses such as adenoviruses, papillomaviruses and herpesviruses have both their DNA strands in existence for similar amounts of time before DNA-dependant-DNA polymerase enzymes copy both their + and – DNA strands during replication (Hanson, 2009), we had anticipated that the best fitting substitution model for sequence datasets of these viruses would be NREV6. Using weighted AIC scores

to reveal trends of model support (Figure 4), it is surprising that NREV12 was overall the best-supported model (illustrated by the redder hues around the top corner of the dsDNA plot in Figure 2. Out of the 30 dsDNA datasets considered, we found that NREV6 provided a better fit to 5/30 dsDNA datasets i.e., HPV18, HPV45, HPV16, HPV6, BPV, and SV40 (**Table 3**) and in similar manner GTR provided a better fit to 5/30 dsDNA datasets i.e. Alphapapilloma virus 6, JC polyomavirus, DPV, RTBV, and DBAV while NREV12 was the best fitting model for the remainder (20) (Table 3) Further, likelihood ratio tests (LRTs) revealed strong overall support for NREV12, with this model providing a significantly better fit ($p < 0.05$) than NREV6 for 25/30 of the dsDNA datasets. Similarly, based on LRTs, NREV12 provided a significantly better fit than GTR in 24/30 dsDNA virus datasets.

Table 3 AIC Scores and LRT results for double-stranded DNA virus datasets. The lowest AIC scores indicating the best-fitting models are in bold. The degree of non-reversibility (DNR) column is the absolute difference between the relative rate differences of two nucleotide pairs within the rate matrix, Q .

Virus Family	Dataset	AIC Score GTR	AIC Score NREV-6	AIC Score NREV-12	P-Value GTR vs NREV-12	P-Value NREV-6 vs NREV-12	DNR
Papillomaviridae	APPV 6	35099.5	35108.0	35102.2	>0.05	0.007	0.089
	HPV18_2	25202.9	25174.6	25179.2	<0.001	>0.05	0.323
	HPV45_2	23600.6	23599.0	23602.9	>0.05	>0.05	0.285
	HPV16_2	29734.0	29664.5	29665.4	<0.001	>0.05	0.371
	HPV31	24681.4	24677.3	24672.8	0.002	0.01	0.165
	HPV6_1	31199.1	31150.0	31141.2	<0.001	<0.001	0.451
	LPV	67165.7	67188.1	67145.5	<0.001	<0.001	0.42
	DPV	69829.7	69889.2	69835.1	>0.05	<0.001	0.056
	XPV	95455.6	95617.1	95452.2	<0.001	<0.001	0.072
BATV	134821	134511	133322	<0.001	<0.001	0.402	
Polyomaviridae	JC_2	51806.7	51819.6	51812.0	>0.05	0.003	0.089
	BK_2	21472.6	21472.7	21471.1	0.03	0.03	0.244
	SV40	16859.8	16858.0	16858.4	0.037	>0.05	0.567
	BPV	148614.9	148573.8	148585.2	<0.001	>0.05	0.064
Caulimoviridae	CMV	124083.9	124221.0	123888.6	<0.001	<0.001	0.351
	CSSV	145327.0	146575	145202	<0.001	<0.001	0.158
	SVBV	138575	138488.1	138464.7	<0.001	<0.001	0.174
	DBAV	46495.5	46514.1	46502.0	>0.05	<0.001	0.0335
	RTBV	54987.9	55350.1	54991	>0.05	<0.001	0.082
	BDV	376325.2	376647.6	376029.9	<0.001	<0.001	0.140
Siphoviridae	CLV	237362.3	237351.8	237348.6	<0.001	<0.01	0.070
Tectiviridae	TTIV	913864.9	913915.4	913773.1	<0.001	<0.001	0.279
Adenoviridae	FAV_C	3074086.7	3074207.5	3073739.1	<0.001	<0.001	0.169
	FAV_E	103482.3	103222.7	102636.7	<0.001	<0.001	0.357

	FAV_D	2326925.6	2325719.4	2324784.5	<0.001	<0.001	0.551
	FAV_A	705328.5	705436..5	705197.8	<0.001	<0.001	0.645
	HMAV_B	103796.7	103937.44	103753.8	<0.001	<0.001	10.890
	HMAV_D	1748635.2	1749769	1748119.1	<0.001	<0.001	0.646
	HMAV_C	2851144.5	2851357.1	2851133	0.006	<0.001	0.0225
	HMAV_E	1915044.8	1915065.3	1914998	<0.001	<0.001	0.049

As NREV6 was not the best fitting model for most of the dsDNA virus datasets I infer that, in most dsDNA virus species, strand-specific substitution biases are not irrelevant. Further, the datasets where NREV6 was not the best fit are from species in families containing other species where NREV6 was the best fit, indicating that such strand-specific substitution biases are unlikely to be a consequence of some broadly conserved feature of viral life cycles in these families (such as, for example, ssDNA replicative intermediates). It is instead plausible that these differences may relate to:

- (i) differences in replication fidelity and/or proofreading efficiency on the leading and trailing DNA strands in some dsDNA virus species (Grigoriev, 1999): These differences are common in eukaryotes (Youri, Newlon, & KunkelThomas, 2002) (Furusawa, 2012) and prokaryotes (Fijalkowska, Jonczyk, Tkaczyk, Bialoskorska, & Schaaper, 1998) and, considering that the replication processes of dsDNA viruses analysed here mirror those of their eukaryote hosts, it is perhaps unsurprising that most of these viruses also display some evidence of strand-specific substitution biases
- (ii) extra exposure to DNA damage of displaced template strands during unidirectional rolling circle replication in some papillomavirus species such as HPV16 could be a contributor to strand-specific nucleotide substitution biases (Kusumoto-Matsuo, Kanda, & Kukimoto, 2011).
- (iii) extra time spent by non-coding strands in single-stranded dissociated states during RNA transcription in some papillomavirus and polyomavirus species (Fernandes & de Medeiros Fernandes, 2012): during transcription processes, the dissociated non-coding strand is transiently more exposed to damage than the coding strand (Wei, et al., 2010) which might also contribute to strand-specific substitution biases.
- (iv) other unmodeled evolutionary processes which manifest as a preference for NREV12

Similarly, and equally surprising, I found that NREV12 was overall the best supported model for dsRNA viruses (illustrated by the redder hues around the top corner of the dsRNA plot in Figure 4).

Table 4 AIC Scores and LRT results for double-stranded RNA datasets. The lowest AIC scores indicating the best fitting models are in bold. The degree of non-reversibility (DNR) column is the absolute difference between the relative rate differences of two nucleotide pairs within the rate matrix, Q .

Virus Family	Dataset	AIC score GTR	AIC score NREV-6	AIC score NREV-12	GTR vs NREV-12	NREV-6 vs NREV-12	DNR
Birnaviridae	AQBV	31754.9	31853.3	31721.9	<0.001	<0.001	0.219
	GBV_A	47176.9	47347.2	47154.8	<0.001	<0.001	0.142
	IPNV	79186.2	79221.9	79182.4	0.0145	<0.001	0.162
	GBV_B	39313.7	39062.8	38938.7	<0.001	<0.001	0.201
Reoviridae	BTV_A	34803.5	34895.1	34801.3	0.03	<0.001	0.042
	BTV_B	48849.9	48893.	48837.1	<0.001	<0.001	0.043
	BTV_C	28350.9	28386.5	28350.8	>0.05	<0.001	0.061
	BTV_D	24969.1	24947.3	24894.0	<0.001	<0.001	0.191
	BTV_F	20622.7	20708.5	20610.2	<0.001	<0.001	0.067
	BTV_G	63349.9	63485.0	63345.9	0.00426	<0.001	0.040
	BTV_H	20596.7	20685.5	20586.1	<0.001	<0.001	0.118
	BTV_I	17592.7	17622.5	17588.8	0.01	<0.001	0.095
	BRVA_C	41206.7	41187.4	41137.1	<0.001	<0.001	0.128
	HRVA_A	17030.5	17043.2	17035.5	>0.05	0.003	0.036
	HRVA_B	8275.1	8280.3	8281.7	>0.05	>0.05	0.087
	HRVA_C	12815.1	12842.6	12807.6	0.003	<0.001	0.132
	HRVA_D2	8036.8	8041.0	8043.7	>0.05	>0.05	0.057
	HRVA_E	7045.9	7056.1	7053.3	>0.05	0.02	0.102
	HRVA_F	7046.0	7056.7	7053.4	>0.05	0.02	0.0710
	HRVA_G	18424.2	18434.0	18425.1	>0.05	<0.001	0.123
	HRVA_H	20431.4	20413.87	20420.5.6	0.002	>0.05	0.163
	PRVA_A	28540.7	28441.9	28398.7	<0.001	<0.001	0.204
	PRVA_B	14757.7	14775.5	14732.6	<0.001	<0.001	0.351
	HRVC_A	6713.2	6718.2	6712.3	0.045	0.007	0.124
	PTOV	202011.3	202106.5	201878.5	<0.001	<0.001	0.039
FJV_B	9274.1	9250.0	9250.9	<0.001	>0.05	0.194	
Endornaviridae	EDV	1771992.8	1772689.1	1771950.6	<0.001	<0.001	0.121
	BPAV	70386.5	70540.2	70390.7	>0.05	0.00	0.047
Totiviridae	TTV	617302.6	617462.6	617172.9	<0.001	<0.001	0.052
	GDV	80435.8	80396.5	80387.7	<0.001	0.002	0.109
Hypoviridae	HPV	66859.8	66899.8	66857.8	0.03	<0.001	0.057

NREV6 fit only two (Human rotavirus A set H and Fiji virus) of the 31 dsRNA datasets better than both NREV12 and GTR. NREV12 was found to be the best fitting model

for 21/31 datasets and GTR was the best fitting of 8/31 datasets (Table 4). In all three Birnaviridae family datasets (which contains virus species with two genome segments) and in 17/22 of Reoviridae family datasets (which contain virus species with 10-12 genome segments) the NREV12 model provided the best fit. Based on the LRTs, strong overall support for NREV12 was found, with this model providing a significantly better fit ($p < 0.05$) compared to NREV6 to 27/31 dsRNA virus datasets, while NREV12 when compared to the GTR model, provided a significantly better fit to 23/31 of the datasets.

I anticipated that NREV12 might fit many of these dsRNA datasets better than NREV6 simply because, during their infection cycles, the coding +strand of dsRNA viruses (the one from which protein translation occurs) tends to exist for longer periods within an infected cell than the non-coding –strand. Specifically, there are three stages during double-stranded RNA virus replication (Wickner, 1993). Firstly, transcription, which is the synthesis of the viral +strands from a dsRNA template which takes place in the cytoplasm within the viral particles. These +strands exist within the cell for prolonged periods in the absence of complementary -strands and are used as templates for translation of viral proteins. In the second step the +strands remaining after translation then act as templates for -strand synthesis during the formation of new dsRNA molecules. The +strands of dsRNA viruses are therefore likely more impacted by mutational processes, which in turn could explain the pervasive strand specific substitution biases seen in this group of viruses.

Table 5 AIC Scores and LRT results for single-stranded DNA datasets. The lowest AIC scores indicating the best fitting models are in bold. The degree of non-reversibility (DNR) column is the absolute difference between the relative rate differences of two nucleotide pairs within the rate matrix, Q .

Virus Family	Dataset	AIC Score GTR	AIC Score NREV-6	AIC Score NREV-12	P-Value GTR vs NREV-12	P-Value NREV-6 vs NREV-12	DNR
Nanoviridae	BBTV M	15044.3	15207.9	14984.4	<0.001	<0.001	0.662
	BBTV N	10605.6	10686.2	10595.2	<0.001	<0.001	0.533
	BBTV R	18484.5	18544	18480.8	>0.05	<0.001	0.609
	BBTV S	12718.9	12757.2	12707.3	<0.001	<0.001	0.728
	CCDV	38622.7	38632.0	38630.5	>0.05	0.03	0.050
	MDV	36232.8	36063	36064	<0.001	>0.05	0.142
	PYDV	56138.4	56076.6	56056.4	<0.001	<0.001	0.187
	FBNS	100153.6	100135.6	100120.5	<0.001	<0.001	0.098
Geminiviridae	Begomo 5	28192.1	28311.9	28192.5	>0.05	<0.001	0.1995

	Begomo 6	16743.0	16722.6	16724.1	<0.001	>0.05	0.214
	Begomo 9	8517.6	8540.8	8515.6	0.03	<0.001	0.312
	Dicot 1	44730.7	44594.3	44583.3	<0.001	<0.001	0.200
	Dicot 2	39909.9	39919.8	39917.9	>0.05	<0.001	0.100
	MSV	252645.3	254347.5	254347.5	<0.001	<0.001	0.144
	PanSV	94601.2	94600.3	94593.7	<0.001	<0.001	0.182
	WDV	35301.7	35313.2	35253.8	<0.001	<0.001	0.1033
Circoviridae	BFDV	17256.7	17262.7	17246.7	<0.001	<0.001	0.224
	DG_CV	12754.8	12779.5	12758.3	>0.05	<0.001	0.116
	PICV	19180.5	19192.5	19191.0	>0.05	0.04	0.117
	CCCC	84435.7	84377.4	84315.3	<0.001	<0.001	0.132
	BTC	262910.4	262060.1	261985.4	<0.001	<0.001	0.178
	POCV2	24940.9	24953.8	24915.8	<0.001	<0.001	0.162
	CCV	90307.9	90301.5	90285.9	<0.001	<0.001	0.114
Anelloviridae	TTV_1	825811	826800	825292	<0.001	<0.001	0.513
	TTSV	332287.9	332397.4	332258.2	<0.001	<0.001	1.560
Parvoviridae	MVM	26756.3	26743.9	26686.9	<0.001	<0.001	0.148
	HPV	67051.2	67080.1	67001.8	<0.001	<0.001	0.235
	CPV	85731	85695	85689.3	<0.001	0.007	0.062
	PPV	163006.8	163090.7	162995.9	<0.001	<0.001	0.143
	CAV_P	37073.3	37115.5	37065.7	<0.001	<0.001	0.162
Microviridae	BMV	31175.3	31164.8	31147.3	<0.001	<0.001	0.188
Pleolipoviridae	APV	85700.2	85617.4	85402.8	<0.001	<0.001	0.204
	BPV	204797.5	204802.3	204796.7	0.04	0.007	0.064

For the ssRNA and ssDNA viruses where one genome strand exists during the virus life cycle for far longer periods of time than the other such that complementary substitutions would not be expected to occur at similar rates, we anticipated that NREV12 should provide a better fit than both NREV6 and GTR. Indeed, for ssRNA viruses, NREV12 has a better AIC score than NREV6 and GTR for 40/47 of the ssRNA datasets and 24/33 of the ssDNA virus datasets (Figure 4). Of the nine ssDNA virus datasets where NREV12 was not the best model, 7/9 were best described by GTR and 2/9 were best described by NREV6 model (Table 5). Of the seven ssRNA datasets (Table 5) where NREV12 was not the best model, 6/7 were best described by the GTR model and 1/7 was best described by the NREV6 model.

Based on the LRTs, strong overall support for NREV12 was found, with this model providing a significantly better fit ($p < 0.05$) than NREV6 for 45/47 of the ssRNA virus

datasets (Table 6) and 31/33 of the ssDNA virus datasets (Table 5). Similarly, based on LRTs, NREV12 provided a significantly better fit than GTR for 40/47 of the ssRNA virus datasets (Table 6) and 27/33 of the ssDNA virus datasets (Table 5). More support for GTR than NREV6 was found in ssRNA and ssDNA virus datasets hence making GTR the second-best performing model for these viruses.

I further found that DNR estimates alone did not cleanly differentiate between datasets for which NREV12 was or was not best supported (Table 3, Table 4, Table 5 and Table 6). For the 107 nucleotide sequence datasets with a model preference of NREV12, ten had estimated DNRs that were greater than 0.5, 13 had DNRs between 0.25 and 0.5 and 84 had DNRs between 0.0225 and 0.25. For the ten nucleotide sequence datasets with a model preference of NREV6, one had an estimated DNR greater than 0.5, four had estimated DNRs between 0.25 and 0.5 and five had estimated DNRs between 0.064 and 0.25. For the 24 nucleotide sequence datasets with a model preference of GTR, none had estimated DNRs greater than 0.5, one had an estimated DNR between 0.25 and 0.5 and the remainder had estimated DNRs between 0.0335 and 0.25.

The dsDNA virus dataset with the highest DNR was Human mastadenovirus D (DNR = 0.646), the dsRNA virus dataset with the highest estimated DNR was Porcine_rotavirus_B (0.351), the ssRNA virus dataset with the highest DNR was SARS-CoV-2 (DNR = 1.536) and the ssDNA virus dataset with the highest DNR was Torque teno sus virus (DNR = 1.56).

Therefore, while NREV12 appears to be generally more appropriate than either NREV6 or GTR for describing mutational processes in ssRNA, ssDNA, dsDNA, and dsRNA viruses, this might only be particularly relevant from a practical perspective when datasets of these viruses yield DNR estimates that are greater than 0.25. For such datasets NREV12 (and possibly NREV 6 in some instances) might be especially useful for both determining the direction of evolution across phylogenetic trees (i.e., it could potentially be used to root these trees) and for quantifying genomic strand-specific nucleotide substitution biases (Harkins et al., 2009).

Table 6 AIC Scores and LRT results for single-stranded RNA datasets. The lowest AIC scores indicating the best fitting models are in bold. The degree of non-reversibility (DNR) column is the absolute difference between the relative rate differences of two nucleotide pairs within the rate matrix, Q .

Virus Family	Dataset	AIC Score GTR	AIC Score NREV-6	AIC Score NREV-12	P-Value GTR vs NREV-12	P-Value NREV-6 vs NREV-12	DNR
Astroviridae	HAV	94580.7	94926.3	94548.1	<0.001	<0.001	0.096
	BAV	188307.1	188572.9	188144.9	<0.001	<0.001	0.108
	MMV	281072.2	281094.5	281076.9	>0.05	<0.001	0.072
	PAV	150626.88	150827.6	150609.5	<0.001	<0.001	0.069
	CKV	90902.3	91233.1	90873.0	<0.001	<0.001	0.083
	GA	64998.5	65223.9	64975.9	<0.001	<0.001	0.110
	CAV_A	85558.8	85617.4	85547.3	<0.01	<0.001	0.076
Bromoviridae	CMV RNA1	34197.5	34198.8	34147.7	<0.001	<0.001	0.124
	CMV RNA2	31398.2	31455.9	31388.7	<0.001	<0.001	0.091
	CMV RNA3	24337.2	24360.3	24343.9	>0.05	<0.001	0.073
	AMS	24337.2	24360.3	24343.9	>0.05	<0.001	0.073
	PSV	67707	67786.5	67691	<0.001	<0.001	0.048
Caliciviridae	LAV	73042.8	73102.4	72984.6	<0.001	<0.001	0.120
	NoV	207667.2	207777.5	207660	<0.001	<0.001	0.047
	VSV	235936.4	236051.4	235913.3	<0.001	<0.001	0.046
Closteroviridae	CTV	30062.2	29980.4	29960.1	<0.001	<0.001	0.272
Flaviviridae	DGV_T1	69771.9	70030.5	69776.2	>0.05	<0.001	0.063
	JEV	146920.8	148101.5	146885.5	<0.001	<0.001	0.091
Hepeviridae	HPVE1	200439.5	200863.8	200179.8	<0.001	<0.001	0.073
	HPVE2	155709.1	155983.8	155518.6	<0.001	<0.001	0.088
Picornaviridae	ENV_A	552287.9	553535.5	551794.1	<0.001	<0.001	0.061
	HRV_A	102218.7	102267.0	101550.7	<0.001	<0.001	0.285
	AIV	101073.1	101136.7	101052.2	<0.001	<0.001	0.093
	AHP	139635.7	140119.6	139506.9	<0.001	<0.001	0.170
	ECV	82078.9	82181.0	82065.8	<0.001	<0.001	0.066
	CDV	130551.3	130896.7	130478.3	<0.001	<0.001	0.086
	TCV	53027.3	53029	53023	0.0151	0.0422	0.033
	FMDV	455180.6	455582.6	454806.1	<0.001	<0.001	0.117
Fusariviridae	FRV	52413.1	52470.6	52418.4	>0.05	<0.001	0.076
Retroviridae	HIV1_setA	344014.4	344295.1	343669.7	<0.001	<0.001	0.237
	HIV1_M	80764.1	80829.5	80668.1	<0.001	<0.001	0.442
	HIV1_setC	180575.0	180702.3	180494.4	<0.001	<0.001	0.107
	HIV1_setD	298489.9	298695.3	298260.6	<0.001	<0.001	0.133
	HIV1_setE	289111.3	289292.1	288941.9	<0.001	<0.001	0.112
	HIV1_setF	214375.9	214692.2	214289.4	<0.001	<0.001	0.148
	EIV	126149	126365.4	125300	<0.001	<0.001	0.192
	BIV	24505.2	24506.9	24513.2	>0.05	>0.05	0.15
	FIV	164542.1	164487.9	164260.4	<0.001	<0.001	0.114
	CAV	351329.9	351871.5	350721.9	<0.001	<0.001	0.174

	SIV	110731.2	110816	110663.3	<0.001	<0.001	0.144
Filoviridae	Ebola_2	53147.3	53143.0	53149.9	>0.05	>0.50	0.264
Orthomyxo- viridae	Flu A 2	82872.8	83010.2	82849.7	<0.001	<0.001	0.27
	Flu B 1	50090.4	50144.1	50060.9	<0.001	<0.001	0.311
Coronaviridae	SARS- COV1	214715.3	214968.5	214644.39	<0.001	<0.001	0.198
	SARS- COV2	15715.4.2	15715.6	15696.7	<0.001	<0.001	1.536
	SARB	573966.3	573815.1	572517.0	<0.001	<0.001	0.301
	MERS-COV	516683.2	516983.4	516608.9	<0.001	<0.001	0.169

2.4 Conclusion

Based on results obtained using a model testing script implemented in the HyPhy programming language, the non-reversible nucleotide substitution model NREV12 has been found to generally provide a substantially better fit to virus nucleotide sequence datasets than does the GTR model; a widely used reversible substitution model. Even though it was expected that a six-rate model would best fit double stranded RNA and DNA viruses NREV12 has been found to also generally provide a better fit to virus nucleotide sequence datasets than does NREV6; a model that would be expected to best describe the evolution of sequences without strand-specific substitution biases.

This suggests that strand-specific nucleotide substitution biases are common during viral evolution irrespective of genome type unlike what we had expected. In fact, such biases should be expected for any viruses because ultimately each virus is still biologically expected that one of its genome strands will either be in existence for substantially longer periods of time than the other or the strand will be more exposed to mutagenic processes than the other during transmission, replication, or gene expression. Specifically in the case of dsRNA viruses, where the life cycle is expected to move from dsRNA to mRNA since the protein translation occurs from the +strand and then to viral proteins, it is expected then that the +strand will exist for much longer periods within an infected cell than the -strand will. Initially we expected that both strands will be in existence for equal amounts of time thus making complementary mutations to occur at the same rate.

I further observed that regardless of the NREV12 model providing a best fit to most datasets in all genome types, the GTR model was second best performing because it still attained a degree of influence during model testing. This behaviour was expected in that these sequence trees used during model testing were reconstructed using a GTR+G model thus the reason why the GTR model provided a better fit after the NREV12 model in all genome types.

The lack of information regarding the proportion of time of the viral life cycle that is spent in single stranded or double stranded state makes it difficult to comment on the biological validity of the NREV6 model and its poor performance in comparison to NREV12. Nevertheless, it is clearly apparent that, when information is required on underlying mutation processes in viruses, models such as NREV12 should yield more valuable insights than GTR.

Chapter 3: Assessing the Impact of Model Misspecification on the Accuracy of Phylogenetic Inference

3.1 Introduction

Phylogenetic inference is an analytical approach to illuminating how sets of homologous sequences have evolved since the time of their most recent common ancestors. Insights from phylogenetic and other analytical techniques that derive additional power from accurately inferred phylogenetic trees have proved vital in making sense of viral evolution. This has been particularly true during the viral outbreaks, epidemics and pandemics that have occurred over the past 20 years, during which these tools have enabled the rapid identification of new potentially important viral variants ((Pater, et al., 2021), (Pybus & Rambaut, 2009), (Sridhar, Teng, Chiu, Lau, & Woo, 2017)), aided in identifying the temporal and geographical origins of outbreaks ((Worobey, Cox, & Gill, 2019), (Ciccozzi, et al., 2019), (Holmes, et al., 2021), (Gilbert, et al., 2007), (Huang, et al., 2013), (Hovmöller, Alexandrov, Hardman, & Janies, 2010), (Holmes, et al., 2021),), helped uncover the risk factors for infection and transmission ((Klinkenberg, Backer, Didelot, Colijn, & Wallinga, 2017), (Volz & Frost, 2013), (Organization, 2012)), enabled the tracking the spread of viruses in localized outbreaks (Klinkenberg, Backer, Didelot, Colijn, & Wallinga, 2017) and throughout the world ((Grubaugh, et al., 2019), (Pybus & Rambaut, 2009), (Popa, et al., 2020)), helped identify drug resistance and immune escape mutations ((Venkatesan, et al., 2018), (Akand & Downard, 2018), (Brumme, et al., 2009)), and informed the development of vaccines and the timings of vaccine updates ((Agor & Özaltın, 2018), (Hampson, et al., 2017), (Morris, et al., 2018), (Dearlove, et al., 2020)). However, the commonly used reversible nucleotide substitution models that underly modern phylogenetic inference are, as I have demonstrated in Chapter 2, not accurate descriptors of how most viruses evolve. It is therefore of interest to determine whether the overwhelming past use during phylogenetic inference of (less appropriate) reversible models instead of (more appropriate) non-reversible models is actually a

problem and, if so, under what specific conditions has it been a problem.

Ultimately, most violations of the assumptions of nucleotide substitution models – such as assuming reversibility when in fact sequences have been evolving in an obviously non-reversible way – might, but are not guaranteed to, undermine the accuracy with which phylogenetic tree topologies (i.e. the arrangement of tree branches) and branch lengths are inferred: both of which could undermine the accuracy and power of downstream statistical analyses that assume inferred phylogenetic trees are accurate reflections of past sequence evolution (Kapli, Flouri, & Telford, 2021). (Naser-Khdour, Minh, Zhang, Stone, & Lanfear, 2019) (Simion, Delsuc, & Philippe, 2020).

Here I use simulations to demonstrate that whereas strand-specific nucleotide substitution biases reduce the accuracy of phylogenetic inference, this reduced accuracy occurs irrespective of whether the GTR or NREV12 models are used for phylogenetic inference. Only when these substitution biases become more extreme than naturally occurs in real virus sequence datasets does use of NREV12 yield significantly more accurate phylogenetic trees than does use of GTR.

3.2 Materials and Methods

I tested the accuracy of phylogenetic tree inference under reversible and non-reversible models using simulated datasets with varying average pairwise nucleotide sequence identities (APIs) evolved under the NREV12 model with different degrees of non-reversibility (DNR). The goal of these tests was not to exhaustively evaluate model misspecification issues during phylogenetic tree inference but rather to check, in instances where viral taxa are known to be evolving in a detectably non-reversible manner (i.e. where NREV12 or NREV6 fits the data better than GTR), whether not accounting for this might decrease the accuracy of phylogenetic inference. Using IQTREE, a phylogenetic inference program that has the option to apply an NREV12-like model (referred to in IQTREE as the UNREST model), a phylogenetic tree was inferred from an alignment of real sequences (Avian Leukosis virus) with an average API of ~90% (Figure 5). The branch lengths on this tree were then scaled (Table 7) to create four other phylogenetic trees representing sequences with approximately 95%,

85%, 80% and 75% API. These five trees are hereafter referred to as “true” trees and each individual tree was used as the starting point of a different set of simulations. With an increase in API, the tree branch lengths reduce. This reduced branch length is an indication that when reconstructing a phylogenetic tree for sequences that are more related genetically, shorter evolutionary distances are expected.

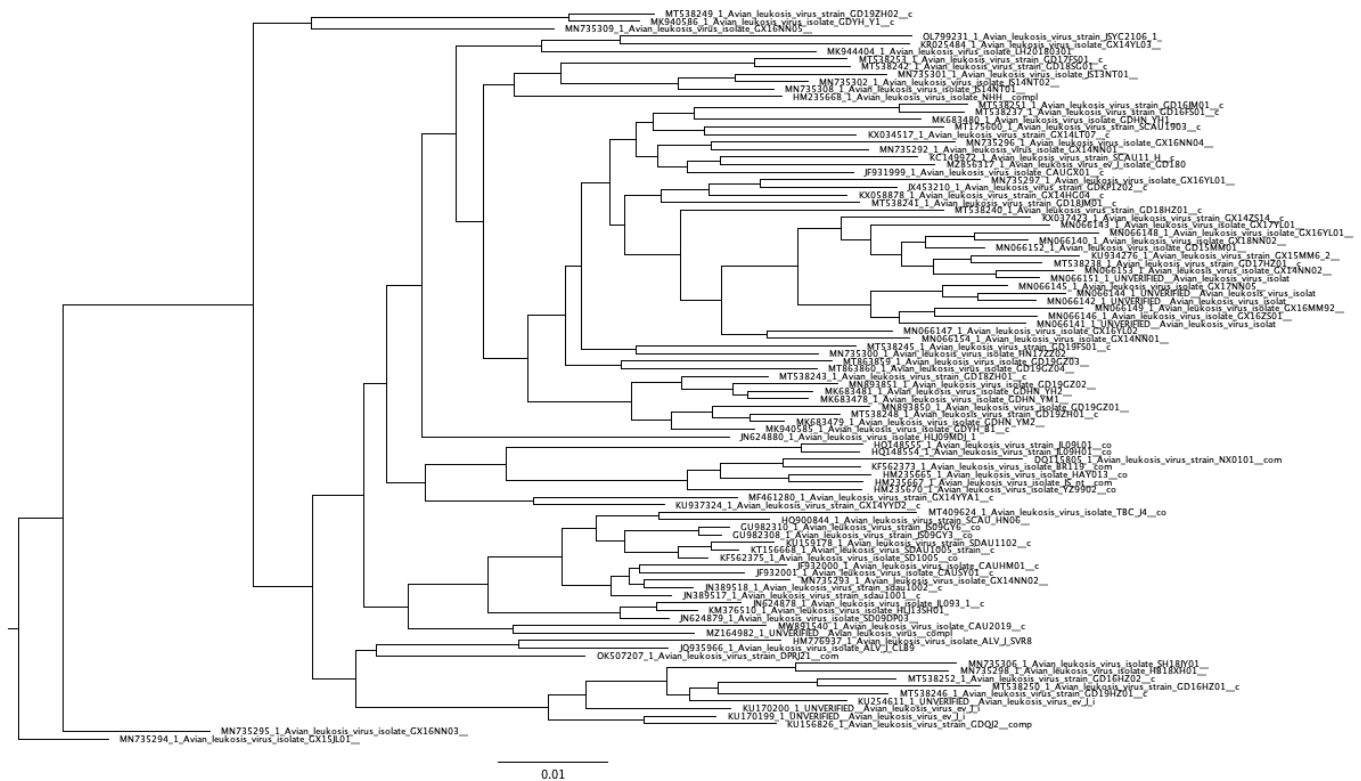


Figure 5 Phylogenetic tree inferred from an alignment of real sequences (Avian Leukosis virus) that was used to simulate datasets with DNRs varying from 0 to 20. The alignment of Avian Leukosis virus had an average pairwise sequence identity (API) of ~90% and the branches of this tree were scaled to produce four other trees reflecting branch tip sequences with approximate pairwise identities of ~75%, ~80%, ~85% and ~95%.

Table 7 Scaled branch lengths for the five phylogenetic trees used as simulation templates representing sequences with approximately 95%, 90%, 85%, 80% and 75% API.

API	Minimum branch length	Maximum branch length	Average branch length
75	0.001199	0.033522	0.010844
80	0.001124	0.031427	0.010166
85	0.001058	0.029579	0.009568
90	0.000992	0.027935	0.009037
95	0.000946	0.026465	0.008561

3.2.1 Defining the Degree of Non-Reversibility

To conduct the simulations, I defined the degree of non-reversibility (DNR) of a nucleotide substitution model as the absolute difference between the relative rate differences of two nucleotide pairs within the rate matrix, R : i.e. for two nucleotides, i and j , there exists a relative rate of i to j substitutions that I will refer to as r_{ij} , and a relative rate j to i substitutions that I will refer to as r_{ji} . Under the NREV12 model, the degree of non-reversibility (DNR) between i and j is defined simply as the absolute difference between r_{ij} and r_{ji} : ($|r_{ij} - r_{ji}|$). $DNR = \sum_{\forall i,j,j \neq i} |r_{ij} - r_{ji}|$. I use DNR as a mathematical representation of the degree to which the rates of all pairs of reverse substitutions differ from one another. For each of the 141 individual real viral sequence datasets analysed in Chapter 2 we calculated the average DNR using the relative rate estimates inferred using the NREV12 model.

3.2.2 Confirming the Kolmogorov conditions set on the irreversibility indices

It was necessary to confirm the existence of non-reversibility in the alignment of real sequences used in this study (Avian Leukosisvirus) using the Kolmogorov conditions set on the irreversibility indices (IRI1, IRI,2 and IRI3; Equations (10), (11) and (12)). (Squartini & Arndt, 2008) where Q_{ji} is the instantaneous rate of change from j to i . As an already existing standardized measure of non-reversibility, it is expected that under DNR=0 all three indices should be approximately zero, whereas it is expected that when DNR > 0 the indices should all be different from zero.

$$IRI1 = \frac{Q_{AG}Q_{GT}Q_{TC}Q_{CA} - Q_{AC}Q_{CT}Q_{TG}Q_{GA}}{Q_{AG}Q_{GT}Q_{TC}Q_{CA} + Q_{AC}Q_{CT}Q_{TG}Q_{GA}} \quad (10)$$

$$IRI2 = \frac{Q_{AT}Q_{TG}Q_{GC}Q_{CA} - Q_{AC}Q_{CG}Q_{GT}Q_{TA}}{Q_{AT}Q_{TG}Q_{GC}Q_{CA} + Q_{AC}Q_{CG}Q_{GT}Q_{TA}} \quad (11)$$

$$IRI3 = \frac{Q_{AT}Q_{TC}Q_{CG}Q_{GA} - Q_{AG}Q_{GC}Q_{CT}Q_{TA}}{Q_{AT}Q_{TC}Q_{CG}Q_{GA} + Q_{AG}Q_{GC}Q_{CT}Q_{TA}} \quad (12)$$

Accordingly, whereas for simulated datasets where DNR was 0 the IRI1, IRI2 and IRI3 indices were all approximately zero indicating that the sequences in these datasets had (as expected), evolved in a time-reversible manner, for datasets where DNR was greater than 0, the IRI1, IRI2 and IRI3 indices were all different from zero indicating that the sequences in these datasets had indeed evolved in a time non-reversible manner.

3.2.3 Simulation workflow

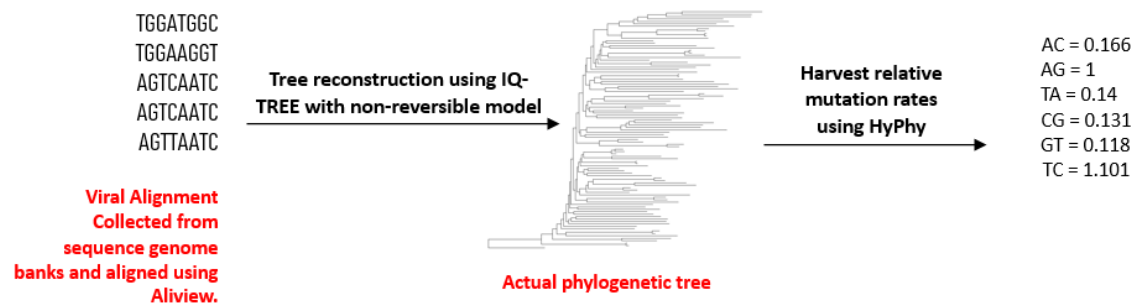
To test whether failure to account for non-reversibility might decrease the accuracy of phylogenetic inference we simulated the evolution of 5,500 nucleotide sequence alignments evolved non-reversibly under varying DNR along the five true phylogenetic trees: 100 datasets per true tree per simulated degree of non-reversibility (DNR). Specifically, simulations were done using HyPhy (Pond & Muse, 2005) with relative rates ranging from a completely reversible matrix i.e., $CA = AC = 0.166$, $GA = AG = 1$, $AT = TA = 0.14$, $GC = CG = 0.131$, $TG = GT = 0.188$ and $CT = TC = 1.101$ – representing DNR=0 – through matrices with DNR=2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 (Table 8). These baselines simulated substitution rates are reflective of those seen in empirical viral nucleotide sequence datasets.

Table 8 Relative rate change for C to A, G to A, A to T, G to C, T to G and C to T mutations under the 11 degrees of non-reversibility alongside the maintained rates for A to C, A to G, T to A, C to G, G to T, and T to C

Degree of Non-Reversibility (DNR)	Relative rates of different nucleotide substitution types (from-to)											
	C-A	A-C	G-A	A-G	A-T	T-A	G-C	C-G	T-G	G-T	C-T	T-C
0	0.166	0.166	1	1	0.14	0.14	0.131	0.131	0.118	0.118	1.101	1.101
2	2.166	0.166	3	1	2.14	0.14	2.131	0.131	2.118	0.118	3.101	1.101
4	4.166	0.166	5	1	4.14	0.14	4.131	0.131	4.118	0.118	5.101	1.101
6	6.166	0.166	7	1	6.14	0.14	6.131	0.131	6.118	0.118	7.101	1.101
8	8.166	0.166	9	1	8.14	0.14	8.131	0.131	8.118	0.118	9.101	1.101
10	10.166	0.166	11	1	10.14	0.14	10.131	0.131	10.118	0.118	11.101	1.101
12	12.166	0.166	13	1	12.14	0.14	12.131	0.131	12.118	0.118	13.101	1.101

14	14.166	0.166	15	1	14.14	0.14	14.131	0.131	14.118	0.118	15.101	1.101
16	16.166	0.166	17	1	16.14	0.14	16.131	0.131	16.118	0.118	17.101	1.101
18	18.166	0.166	19	1	18.14	0.14	18.131	0.131	18.118	0.118	19.101	1.101
20	20.166	0.166	21	1	20.14	0.14	20.131	0.131	20.118	0.118	21.101	1.101

Further, it should be noted that all simulations under NREV12 were performed under the stationarity criterion: $\pi e^{Qt} = \pi$ (where Q is the rate matrix and π is the nucleotide frequency distribution and $t \geq 0$).



Degree of Non-Reversibility (DNR)	0	2	4	6	8	10	12	14	16	18	20
CA	0.166	2.166	4.166	6.166	8.166	10.166	12.166	14.166	16.166	18.166	20.166
GA	1	3	5	7	9	11	13	15	17	19	21
AT	0.14	2.14	4.14	6.14	8.14	10.14	12.14	14.14	16.14	18.14	20.14
GC	0.131	2.131	4.131	6.131	8.131	10.131	12.131	14.131	16.131	18.131	20.131
TG	0.118	2.118	4.118	6.118	8.118	10.118	12.118	14.118	16.118	18.118	20.118
CT	1.101	3.101	5.101	7.101	9.101	11.101	13.101	15.101	17.101	19.101	21.101

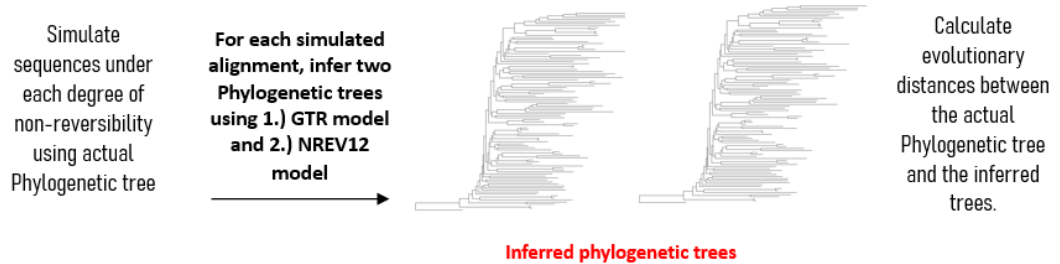


Figure 6 Simulation workflow to assess the impact of model misspecification.

3.2.4 Quantifying the Accuracy of Phylogenetic Inferences

Knowing how accurately the inferred phylogenetic trees were reconstructed requires knowing the level of similarity (or degree of discrepancy) between the actual phylogenetic tree that was used during the simulation stage and the corresponding inferred trees. To do this, there are multiple available algorithms that are used to compare phylogenetic trees such as subtree pruning and regrafting (SPR) (Whidden, Zeh, & Beiko, 2014), (Kuhner, 2015), (Wu, 2009), nearest-neighbour interchange (NNI) (Jahn, Beerenwinkel, & Zhang, 2021), tree bisection and reconnection (TBR) (Kelk & Linz, 2019), Robinson and Foulds distance (RF) and the weighted Robinson and Foulds distance (wRF) (Llabrés, Rosselló, & Valiente, 2021), (Kuhner & Yamato, 2015). The SPR, NNI and TBR are considered as natural methods as they yield a distance, D , reflecting the smallest or minimum number of rearrangements (considered as steps) that would be needed to transform one tree (e.g., the inferred tree) into another tree (e.g., the actual true tree) (Kuhner & Yamato, Practical performance of tree comparison metrics. *Systematic Biology*,, 2015). The RF distance is one of the least complicated measures of topology distance and is most widely used (Pattengale, Gottlieb, & Moret, 2007).

The algorithm used to calculate the RF distance is such that for any two given trees (in this case the actual and the inferred tree), the distance, D , is the total number of branch partitions that appear in the actual tree that are not in the inferred tree. There are cases where two trees have the exact same topology, but their branch lengths differ, and in such cases the RF distance can provide insights into more subtle differences between trees. For this study a modification of the RF distance, called the weighted RF (wRF) distance, was preferred as it explicitly considers both topological, and branch length differences between trees. For the wRF method, for each branch in the actual tree, the measure looks for the exact same branch in the corresponding inferred tree with respect to the tip labels and calculates the absolute branch length difference between the two (Kuhner & Yamato, 2015), (Kuhner, 2015), (Robinson & Foulds, 1981). In a case where the branch only appears in either the actual or the inferred tree, it is considered to have zero length in the tree that it does not occur in. Finally, all the absolute differences are summed, and this value, is the wRF distance

between the actual and inferred trees (Wade T. , Rangel, Kundu, Fournier, & Bansal, Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families, 2020), (Robinson & Foulds, 1981).

3.2.5 Statistical Analysis

To both assess the wRF and SPR distances between inferred and true trees and make a comparative assessment of how use of GTR (the mis-specified model) instead of NREV12 (the best fitting model) impacted the accuracy of trees inferred from simulated alignments, the phangorn package (Schliep, 2011) in R version 4.2 was used to compare SPR and wRF tree distance data. The SPR distances were assessed as categorical variables and described as relative frequencies (%) were the Wilcoxon signed rank test was used to compare tree distances. The wRF distances were assessed as continuous variables. The wRF data was represented as means and standard deviations and for each of the analysed DNRs a paired t-test (correlated t-test) was then used to compare whether the wRF mean scores of trees inferred using GTR and NREV12 under each API and DNR level were significantly different. All statistical tests in this research were considered significant at $p.value < 0.05$.

3.3 Results and Discussion

3.3.1 Assessing the Impacts of Model Misspecification on Phylogenetic Tree Inference

To determine whether it might sometimes be worthwhile using NREV12 rather than GTR for phylogenetic inference when NREV12 is the best fitting nucleotide substitution model, I used simulated datasets to compare the accuracy of phylogenetic trees inferred using these models. Specifically, I simulated datasets with DNRs varying from 0 to 20 along known “true” phylogenetic trees with branches scaled to reflect branch tip sequences with APIs of ~75%, ~80%, ~85%, ~90% and ~95%. For each of five API levels, we therefore simulated 5500 datasets (comprising 100 datasets for each DNR = 0, 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20).

Phylogenetic trees were inferred from these 5500 simulated datasets (i.e. the inferred trees) and compared to the phylogenetic trees used to simulate the datasets (i.e., the true trees) using wRF and SPR distances to assess the impact of varying DNR on the accuracy of phylogenetic inference. I further tested whether the accuracy of phylogenetic inference could be improved for sequences that had evolved under $DNR > 0$ by using NREV12 instead of GTR. Specifically, for every simulated dataset a phylogenetic tree was inferred using GTR and another using NREV12 and the wRF and SPR distances of each of these trees to the true tree was determined. I was particularly interested in determining whether trees inferred using a mis-specified model (i.e. GTR in this case) would be less accurate than trees inferred with a correctly specified model (i.e. NREV12).

I found that irrespective of dataset diversity and the nucleotide substitution model used, phylogenetic inference tended to become less accurate (i.e. wRF scores increased) as DNR increased (Figure 7). This tendency was, however, slightly more pronounced when using a (miss-specified) GTR model than when using a (correctly specified) NREV12 model with, for any given dataset having $DNR > 0$, the use of NREV12 tending to yield slightly more accurate phylogenetic trees than when GTR was used. There were, however, only statistically significant improvements ($p < 0.05$ paired t-test) in the accuracy of phylogenetic trees inferred using NREV12 relative to those inferred using GTR in lower diversity datasets (i.e. those with APIs of 85%, 90%, and 95%); and then only for $DNR > 8$. This behaviour was expected for the GTR model with increasing DNR, but I expected trees to be consistently better inferred under the NREV12 model as DNR increased.

It is also expected that a better-fitting nucleotide substitution model should yield a better phylogenetic inference accuracy (Posada & Crandall, Selecting the best-fit model of nucleotide substitution, 2001). However, the wRF metric combines measures of both the accuracy with which branch lengths are inferred and the accuracy with which tree topologies are inferred. In instances where the true and inferred trees differ substantially it may be more appropriate to disregard branch length information and only consider the topologies of the trees being compared (Kuhner & Yamato, 2015). I was therefore interested in determining whether the observed decreases in phylogenetic inference accuracy when DNR increased above 2 (i.e. higher wRF

scores in Figure 3) were attributable to branch lengths being inaccurately inferred, or whether they were attributable to topologies being inaccurately inferred. To do this I considered the SPR metric of tree dissimilarity, which considers only topologies.

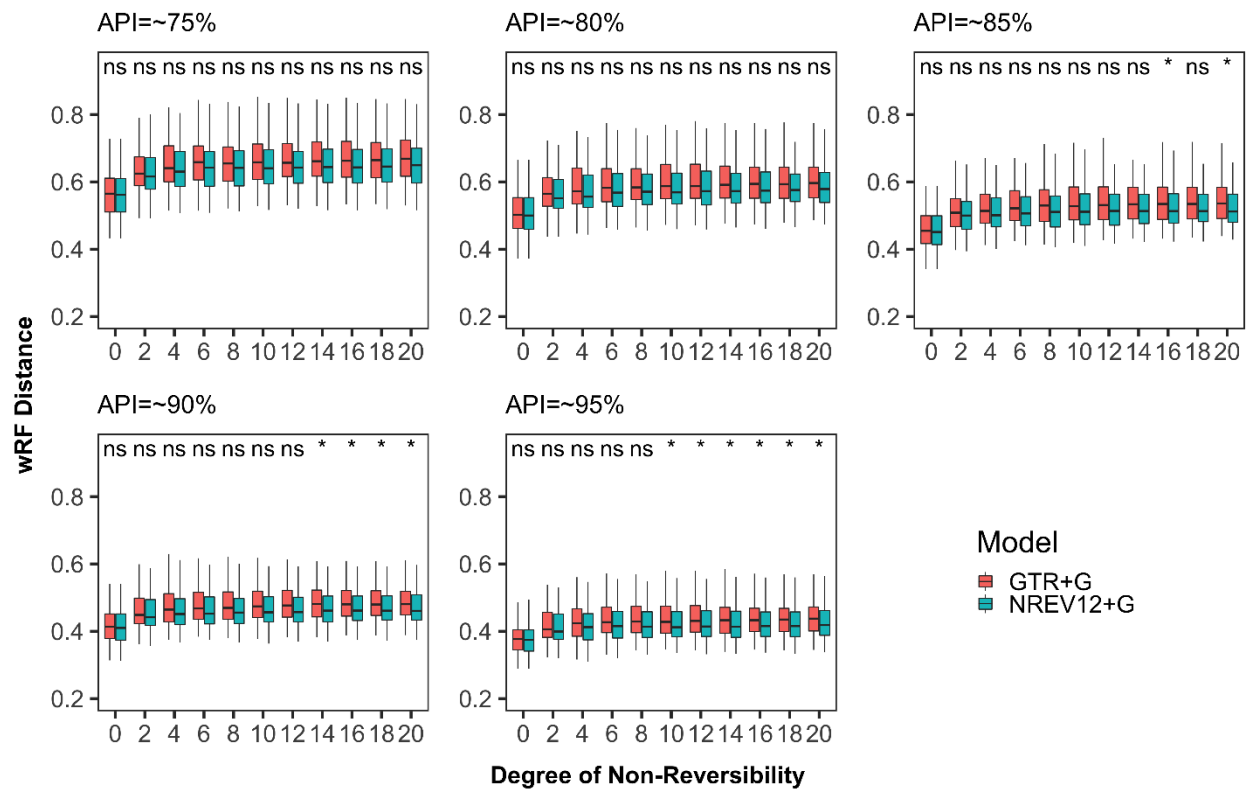


Figure 7 Weighted Robinson-Foulds distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility and different average pairwise sequence identities (APIs) (~75%, ~80%, ~85%, ~90% and ~95%). "ns" above a pair of box and whisker plots indicates a paired t-test adjusted p-value of greater than or equal to 0.05 and "*" indicate a paired t test adjusted p-value of <0.05

Adjudged by SPR (Supplementary Figures 1, 2, 3, 4 and 5), out of the 5,500 simulated trees inferred using the NREV12 model, 5,466 were accurately inferred (i.e. these yielded a SPR score of 0 when comparing inferred and true trees) while 34 needed one subtree prune-and-regraft operation (i.e. a SPR score of 1) to transform the inferred tree into the true tree. Similarly, out of the 5,500 simulated trees inferred using the GTR model 5,480 were accurately inferred (i.e. with SPR=0) while 20 needed one subtree prune-and-regraft operation (i.e. SPR=1) to transform the inferred tree into the true tree. Based on a goodness of fit test, I found that the proportion of trees with inaccurate tree topologies under the GTR model was not significantly different from the proportion with inaccurate tree topologies under the NREV12 model ($p = 0.37$,

0.78, 0.25, 0.26, and 0.16 for the API ~75%, ~80%, ~85%, ~90%, and ~95% datasets respectively). Further, there was no clear association between increasing DNR and the alignments that yielded trees with SPR scores of > 0 .

When considering the wRF and SPR metrics together it is apparent that irrespective of whether GTR or NREV12 is used for phylogenetic inference, increasing DNR decreases the accuracy with which branch lengths are estimated but does not have a strong impact on the accuracy with which tree topologies are inferred.

It is important to note that the DNRs of 3 of real virus genome sequence datasets were < 1 (where one i.e., Human Rotavirus A had an extreme DNR OF 10) and none of the datasets that I examined in Chapter 2 had a DNR greater than 1.56 except for HRVA_A. At these “biologically realistic” DNR levels there were no significant differences between the wRF or SPR scores of trees inferred with GTR and NREV12: an indication that utilizing GTR (the more convenient but wrong model) rather than NREV12 (the less convenient but correct model) during phylogenetic inferences with these real datasets would likely have no significant impact on the accuracy of inferred trees. This is to also show that getting branch lengths correct is vital in tree accuracy as inaccurate branch lengths would result in either underestimated or overestimated evolutionary genetic distances which would result in wrong interpretation of passage of time between variants in a case of viruses.

3.4 Conclusion

I had anticipated that given evidence of real-world viral genome sequences evolving both non-reversibly and with strand-specific substitution biases (see chapter 2), inferring trees using a model such as NREV12 that appropriately accounts for this might: (1) minimize the impact of high DNR on the accuracy of phylogenetic inference (i.e. wRF scores presented in blue in Figure 7 were naïvely expected to not increase with increasing DNR), and (2) yield significantly more accurate phylogenetic inferences than when using GTR for all datasets where NREV12 was the most appropriate model and DNRs were greater than zero. However, even when using NREV12, increasing DNR clearly increased the branch length differences between

inferred and actual trees, thus decreasing the accuracy of phylogenetic inference, and, for datasets where DNRs were greater than zero, using GTR did not consistently yield significantly less accurate phylogenetic inferences than those attained using NREV12. Based on the SPR metric, tree topologies were accurately inferred both when using NREV12 and GTR for all datasets irrespective of DNR and API levels.

From a practical perspective, choosing a non-reversible nucleotide substitution model to construct phylogenetic trees from virus genome sequences that display strand specific nucleotide substitution biases is not guaranteed to yield more accurate phylogenetic trees. Nevertheless, in instances where strand-specific substitution biases yield DNR levels that are higher than ~ 0.5 (such as are found in our SARS-CoV-2, Torque teno sus virus and Banana bunchy top virus datasets) it may be prudent to select a model such as NREV12 (such as is implemented in programs like IQTREE) over GTR as the better of two suboptimal choices.

The lack of available data regarding the proportions of viral life cycles during which genomes exist in single and double-stranded states makes it difficult to rationally predict the situations where the use of models such as GTR and NREV12 might be most justified: particularly in light of the poor overall performance of GTR relative to NREV12 with respect to describing mutational processes in viral genome sequence datasets. I, therefore, recommend case-by-case assessments of NREV12 vs GTR model fit when deciding whether it is appropriate to consider the application of non-reversible models for phylogenetic inference and/or more sophisticated phylogenetic model-based analyses such as those intended to test for evidence of natural selection or the existence of molecular clocks.

Chapter 4: RpNRM: Web Application for Rooting Phylogenetic Trees Using a Non-Reversible Nucleotide Substitution Model (NREV12)

4.1 Introduction

The correct interpretation of a phylogenetic tree frequently requires the identification of its root: the point along the branches of the tree that represents the most recent common ancestor (MRCA) of all the taxa represented within the tree (Williams, et al., 2015), (Tian & Kubatko, 2017), (Graham, Olmstead, & Barrett, 2002). Given the usefulness of rooted phylogenies, various approaches are used to identify the branch in a tree along which the root node (i.e. the point representing the MRCA) occurs (Kinene, Wainaina, Maina, & Boykin, 2016), (Graham, Olmstead, & Barrett, 2002). Among the most commonly used approaches for finding the location of this root node are midpoint rooting, molecular clock rooting, and outgroup rooting methods (Huelsenbeck, Bollback, & Levine, 2002).

The midpoint rooting method simply infers the root-node location as the midpoint of the longest branch-path in the tree between any two sequences (or taxa) represented within the tree (Hess & De Moraes Russo, 2007), (Wade T. , Rangel, Kundu, Fournier, & Bansal, 2020), (Boykin, Kubatko, & Lowrey, 2010), (Tria, Landan, & Dagan, 2017). The validity of this rooting approach hinges on the (frequently incorrect) assumption that the rate of evolution does not vary along the longest branch path separating the two sequences that sit at opposite ends of the longest branch-path (Hess & De Moraes Russo, 2007), (Hendy & Penny, 1989).

The molecular clock method reasonably assumes the degree of genetic difference between any two sequences represented within a tree is a function of time since the two sequences last shared a common ancestor. With this rooting method, the root is placed at a position along one of the tree branches that either minimizes the variance among the leaf nodes in their distance to the root location, or, when sequences

represented in the tree were sampled over a long-enough timeframe, produces the best linear regression of the root to sampled sequence (i.e. terminal branch tip or just tip) distances against sampling times (Mai, Sayyari, & Mirarab, 2017), (Kinene, Wainaina, Maina, & Boykin, 2016), (Graham, Olmstead, & Barrett, 2002). For distantly related species, this method may be problematic due to the lack of linearity between genetic distances and divergence times since MRCAs, and it is therefore necessary to test if sampling times of sequences within datasets are significantly correlated with the genetic distances between the sequences and the inferred root node: a so-called strict molecular clock test. If there was evidence of such a correlation then this would warrant the use of phylogenetic inference techniques that are amenable to the use of explicit molecular clock models which, if used correctly, can accurately calibrate the relationship between rates of genetic divergence in different parts of the tree and the progression of time (Kinene, Wainaina, Maina, & Boykin, Rooting trees, methods for. Encyclopedia of Evolutionary Biology, 2016), (Wade T. , Rangel, Kundu, Fournier, & Bansal, 2020), (Smith & Peterson, 2002). Whereas strict molecular clock models assume that the relationship between genetic divergence and time is constant across all branches in a tree, relaxed molecular clock models permit this relationship to vary somewhat between branches.

The most commonly used, and likely the most accurate (Hess & De Moraes Russo, 2007), (Lyons-Weiler, Hoelzer, & Tausch, 1998) of the widely used phylogenetic tree rooting methods is outgroup rooting. This approach requires the existence of a sequence of a taxon that is more distantly related to all the taxa of interest than they are to one another, but is still closely related enough to these taxa of interest that they all share a MRCA that is only slightly older than the MRCA shared by all the taxa of interest (Wade T. , Rangel, Kundu, Fournier, & Bansal, 2020). With this method, the root is placed on the branch that connects the outgroup sequence and the rest of the tree. Though the outgroup method is widely used, in cases where the outgroup is distantly related to the ingroup sequences, there will be a high probability of inaccurate rooting (Graham, Olmstead, & Barrett, 2002) (Huelsenbeck, Bollback, & Levine, 2002), (Mai, Sayyari, & Mirarab, 2017) (Mavian, et al., 2020). Other challenges with the outgroup method are that it is sometimes unclear whether presumed outgroup sequences do in fact share older MRCAs with the ingroup sequences than the ingroup sequences share with one another. Therefore, the success of the outgroup rooting

method hinges on the appropriateness of the selected outgroup sequence.

In many instances it remains difficult to root phylogenetic trees using these various approaches; either different lineages in a phylogeny do not all evolve at the same rate such that midpoint rooting is unreliable (Tabatabaee, Sarker, & Warnow, 2022), the sequences represented in a phylogeny have not been sampled over a long-enough period to obtain enough support for an explicit clock model, or no appropriate outgroup sequence is available (Roger & Hug, 2006). In such cases phylogenetic trees could potentially still be rooted using nucleotide substitution methods that do not make a reversibility assumption (Yap & Speed, Rooting a phylogenetic tree with nonreversible substitution models, 2005), (Boykin, Kubatko, & Lowrey, 2010) i.e., models that do not assume that rates of substitution from nucleotide X to nucleotide Y are the same as those from nucleotide Y to nucleotide X. Such non-reversible models can in many (if not most) cases more realistically describe actual substitution processes than reversible models in that the mutagenic processes that cause nucleotide X to change to nucleotide Y are not the same as those that cause nucleotide Y to change to nucleotide X ((Nguyen, et al., 1992), (Cheng, Cahill, Kasai, Nishimura, & Loeb, 1992)).

Using non-reversible models of nucleotide substitution can capture the inherent non-reversibility of evolution and derive additional power to detect the direction of evolution from DNA/RNA strand-specific substitution biases (Naser-Khdour, Quang Minh, & Lanfear, 2022), (Yap & Speed, 2005), (Tria, Landan, & Dagan, 2017). When significant evidence of strand specific substitution biases exists, such as when model testing reveals that non-reversible nucleotide substitution models are strongly favoured over reversible ones for a given datasets, then the likelihood of any given rooted phylogenetic tree derived from that dataset should be strongly impacted by the root position.

Phylogenetic inference using non-reversible models can, however, be computationally expensive because these models render several commonly used algorithmic techniques for efficient likelihood computation such as Felsenstein's Pulley inapplicable (Boussau & Gouy, 2006) (Mai, Sayyari, & Mirarab, 2017): a factor which may explain why most of the popular phylogenetic inference software packages do not implement

such models and why, therefore, they are not more commonly used (Yap & Speed, Rooting a phylogenetic tree with nonreversible substitution models, 2005).

Regardless of the computational difficulty, there are a number of emerging applications that enable the rooting phylogenetic trees using non-reversible models. One of these applications is IQTREE, a phylogenetic inference program that has the option to apply a twelve-rate non-reversible model (Minh, et al., 2020), (Naser-Khdour, Quang Minh, & Lanfear, Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. 71(4),, 2022).

Here I introduce a web application, called RpNRM, which can be accessed via the university of Cape Town computational biology division server <http://rpnrm.ilifu.ac.za/> . RpNRM works such that, given a phylogenetic tree and the nucleotide alignment used to generate the tree, will attempt to identify the root branch using a stationary 12 rate non-reversible (NREV12) model. Specifically, the application exhaustively reroots the tree on each branch and identifies the branch along which the root node resides as the rooting branch of the tree with the highest likelihood. The application also indicates whether there is sufficient evidence of non-reversibility in the substitution process to warrant using NREV12 to identify the root location. I compare the accuracy of tree rooting's obtained with RpNRM with those obtained using rooted phylogenetic trees inferred using non-reversible models in IQTREE, midpoint rooting and outgroup rooting.

4.2 Methods and Materials

4.2.1 The Software

RpNRM takes as input a phylogenetic tree in newick format and the fasta-formatted multiple sequence alignment used to make the tree. The RpNRM interface is written in, R (Ihaka & Gentleman, 1996) using the shiny package (Chang, Cheng, Allaire, Xie, & McPherson, 2017) and the HyPhy (Pond, Frost, & Muse, 2005) scripting language. HyPhy enables the designation and fitting of diverse discrete and continuous-in-time Markov chain models of sequence evolution such as NREV12.

RpNRM has four features that a user can interact with: (1) upload a tree; (2) upload an alignment; (3) root a tree; and (4) download a rooted tree. Once the data has been uploaded and rooting is completed the application displays (1) a rooted phylogenetic tree (which can be downloaded in newick format); (2) the total number of rooted trees that were tested; and (3) the result of a model test comparing the goodness of fit between NREV12 and the general time-reversible model.

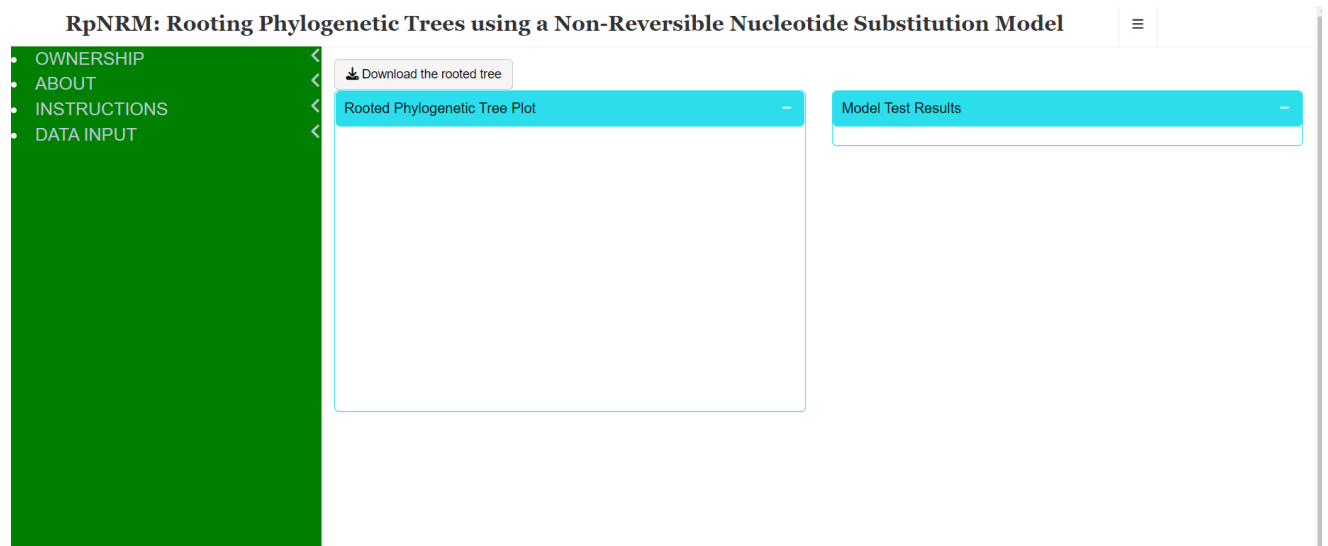


Figure 8 Graphical view of the web application RpNRM displaying the user interface features. <http://rpnrm.ilifu.ac.za/>

4.2.1.1 Phase One: Model Testing

A key feature of RpNRM that is implemented before any rooting is done is a model test to ensure that there is a significant degree of nucleotide substitution non-reversibility in the provided dataset to ensure that it contains a sufficient signal of evolution directionality to yield a reliable rooting using NREV12. The model test is conducted using a model testing script formulated in the HyPhy scripting language (Pond, Frost, & Muse, 2005; a standalone version of this script can be obtained from <https://github.com/veg/hyphy-analyses/tree/master/NucleotideNonREV>). The primary role of the script is to assess whether the evolution of the sequences in the input alignment is better described by the non-reversible NREV12 or NREV6 models than by the reversible GTR model and involves:

1. Harvesting of nucleotide sequences $\left(\frac{\pi_i}{\sum_{i=1}^4 \pi_i}\right)$ from the sequence alignment into a vector of frequencies $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$.
2. Defining three stochastic instantaneous rate matrices, Q , by multiplying the relative substitution rates by the appropriate nucleotide frequencies i.e.: (i) General time reversible (GTR) matrix with relative rates conditioned at $r_{AT}=r_{TA}$, $r_{CG}=r_{GC}$, $r_{TG}=r_{GT}$, $r_{CT}=r_{TC}$ $r_{AG}=r_{GA}$, $r_{AC}=r_{CA}$; or (ii) 12 rate non-reversible (NREV12) matrix with relative rates defined such that the matrix satisfies complete non-reversibility
3. Defining the three model probability transition matrices P by $P(t) = e^{Qt}$
4. Calculating $L(\theta|D, T)$ where D is the input alignment, T is the input tree, and θ are the independent parameters maximizing $L(\theta|D, T)$
5. Performing likelihood ratio tests to determine whether (1) NREV12 fitted the data significantly better than GTR with LRT statistic as $2(\ln L_{NREV12} - \ln L_{GTR})$ with p value being calculated as $1 - \text{chi}(LRT, df_{NREV12} - df_{GTR})$

4.2.1.2 Phase Two: Rooting at Every Branch and Maximum Likelihood Assessment

RpNRM uses the function “reroot” in the phytools R package (Revell, 2012) to successively re-root the input phylogenetic tree at every branch. For every input tree, T , representing the evolutionary relationships of n sequences there exist $2n - 3$ branches each of which is a potential root branch. The reroot function is used to reroot the tree at all $2n - 3$ branches with the likelihood of every tree under NREV12 being determined and compared to discover the branch yielding the maximum likelihood.

4.2.2 Experimental Design

I describe in this section how the simulation and empirical experiment setup was done and how the rooting error was measured for each of the methods. For both simulated data and empirical data, I measured the topological distances from the estimated roots to the true root (the root position as provided by the out-group method), specifically, using R I calculated the Robinson Foulds distance from the root of the estimated tree

to the root of the true tree then normalized the distances by the number of nodes in the trees.

4.2.2.1 Simulations

To meaningfully compare the rooting accuracy of RpNRM against other methods (i.e., outgroup, midpoint and IQTREE methods), tests on 800 simulated datasets was conducted. I downloaded an HIV-1M nucleotide sequence alignment from the Los Alamos National Laboratory HIV sequence database (<https://www.hiv.lanl.gov/content/index>) with an outgroup sequence of simian immunodeficiency virus chimpanzee (SIVcpz) to use as the true trees and simulation templates. The sequence dataset was aligned using muscle (Edgar, 2004) and a maximum likelihood phylogenetic tree for the actual dataset was constructed using RAxML v8.2 (Stamatakis, 2016) under the GTR+GAMMA model and further rooted using the outgroup method hereby referred to as “actual phylogenetic tree”. The actual phylogenetic tree was further tested for tree balance using the colless metric determined by the R package, castor (Louca & Doebeli, 2018) . I modified the actual tree (with a colless = 3) to have a colless measures of 10 and 20 thus creating three “true” or “actual” phylogenetic trees to use for the simulation process: respectively referred to a balanced (colless = 3), moderately imbalanced (colless = 10) and extremely imbalanced (colless = 20) (Figure 9).

To ensure that the data under study did indeed have evidence of non-reversibility I assessed the Kolmogorov conditions (Squartini & Arndt, 2008) and tested whether all three irreversibility indices (IRI's) were different from zero using relative rates obtained from a model test in hyphy (Table 9). I found IRI1 to be -0.5037454, IRI2 to be 0.550509 and IRI3 to be -0.2783045 hence fulfilling the requirements of the Kolmogorov conditions.

To compare how well trees were rooted among the four different rooting methods, nucleotide sequence alignments were simulated through the three actual phylogenetic trees i.e., balanced, moderately imbalanced, and extremely imbalanced using relative substitution rates obtained from the model test output Table 9). The number of sites

in the simulated alignments was varied such that for each tree balance type, two different simulations were conducted with the first containing 1000 sites and the second 4000 sites. The simulation process is outlined below:

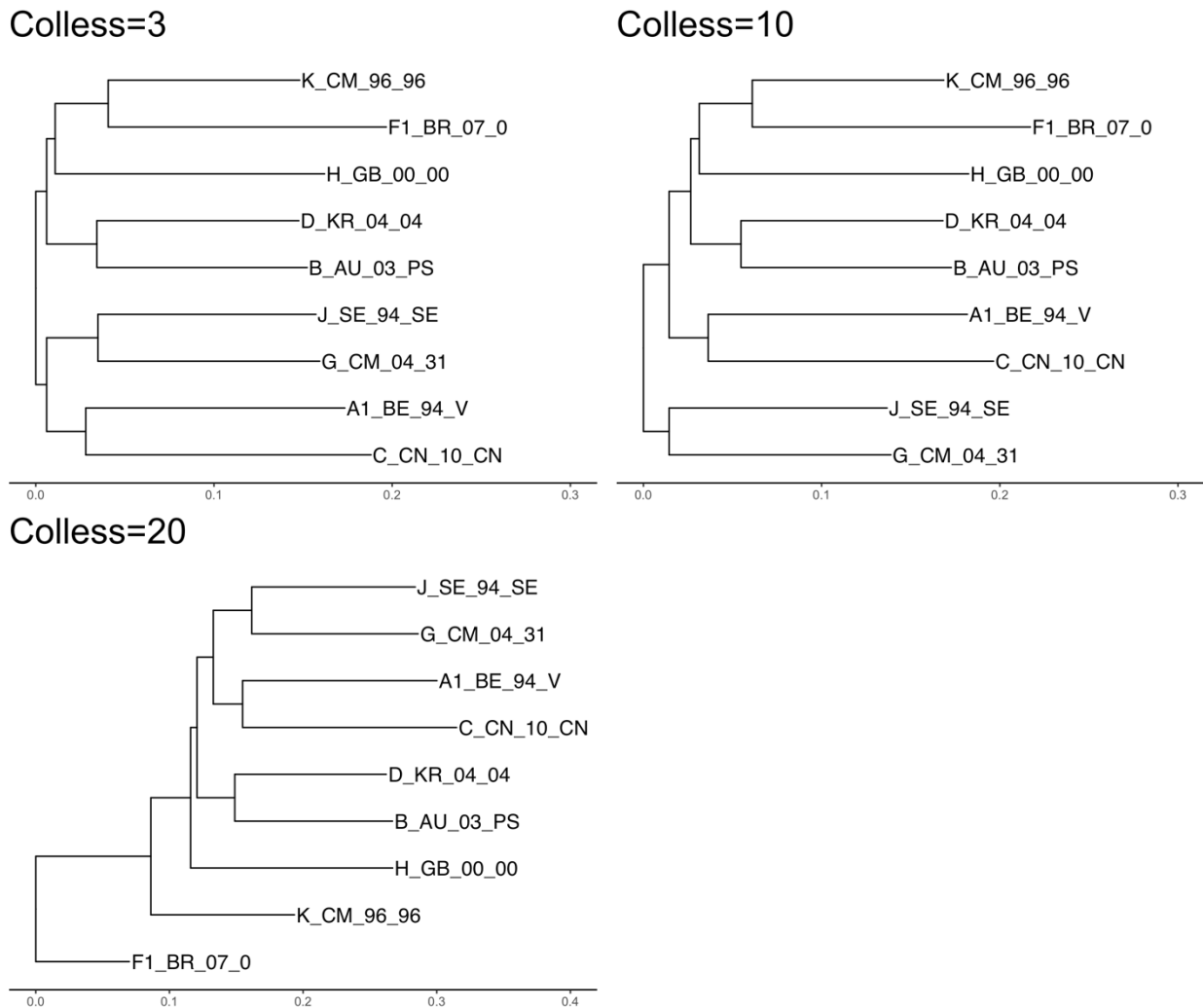


Figure 9 The three actual phylogenetic trees used during the simulation study. These trees differ on balance measures with tree one having colless 3, tree two having colless 10 and tree 3 having colless 20.

Simulation Process for Trees to be Rooted with the Outgroup Method

1. Sequence simulation: Simulate 100 viral alignments containing sequences with 1000 sites using a balanced HIV-1M rooted tree with 10 taxa that contains the outgroup sequence, SIVcpz.
2. Tree reconstruction: Reconstruct trees with the SIVcpz sequence inclusive in the simulated dataset using RAxML hereafter referred to as inferred phylogenetic trees

3. Root the inferred trees using R by placing the root at the point where the outgroup sequence branch joins to the rest of the tree
4. Repeat from step one for the simulation of 10 taxa datasets but where 4000 sites are simulated instead of 1000
5. Change the simulation tree to the moderately imbalanced tree and redo steps one to four
6. Change the simulation tree to the extremely imbalanced tree and redo from steps one to four

Simulation Process for Trees to be Rooted with the RpNRM, IQTREE, and Midpoint Methods

1. Sequence simulation: Simulate 100 viral alignments containing sequences with 1000 sites using an HIV-1M rooted tree with 10 taxa that contains the outgroup sequence, SIVcpz
2. Tree reconstruction: Remove the SIVcpz sequence from the simulated dataset and reconstruct phylogenetic trees using RAXML
3. Root the inferred trees using RpNRM /IQTREE/ Midpoint method using R package phytools (Revell, 2012)
4. Repeat from step one for the simulation of 10 taxa datasets but with 4000 sites instead of 1000
5. Change the simulation tree to the moderately imbalanced tree and redo steps one to four
6. Change the simulation tree to the extremely imbalanced tree and redo from steps one to four

I further assessed whether increasing the degree of non-reversibility in the simulation process would influence the overall performance of the rooting methods. To do this, I increased the relative nucleotide substitution rates during simulations to attain a high (but biologically plausible) DNR of 2.193 (the DNR of the initial set of simulations was 0.442) I repeated the entire simulation process to attain an additional 600 trees and alignments for testing: i.e. the total number of simulated trees and alignments that were examined was 1200 (including 100 trees for each of two dataset sizes, two levels of DNR and three levels of tree imbalance)

Table 9 Relative mutation rates that were used during the simulation of datasets under low DNR (0.442) and high DNR (2.193)

Base Pair	DNR=0.442	DNR=2.193
AC	0.449	0.449
CA	0.724	1.82
AG	1	1
GA	1.509	3.545
AT	0.179	0.179
TA	0.280	0.684
CG	0.298	0.298
GC	0.396	0.788
CT	1.507	6.75
TC	0.172	0.172
GT	0.220	0.220
TG	0.554	1.89

4.2.2.2 Empirical data

I further conducted tests to assess how accurately RpNRM and IQ-TREE rooted 28 empirical virus sequence datasets (Table 10) each containing a DNR of greater than 0.25. Each dataset consisted of in-group sequences and one appropriate outgroup sequence that was expected to accurately indicate the root of the in-group sequences. We downloaded the nucleotide sequences from the National Centre for Biotechnology Information Taxonomy database (<http://www.ncbi.nlm.nih.gov/taxonomy>). The collected sequence datasets were split into two groups i.e., one with an outgroup sequence included and another set without an outgroup sequence. These datasets were aligned using Muscle (Edgar, 2004) as implemented in Aliview (Larsson, 2014) and phylogenetic trees were reconstructed the alignments using RAxML v8.2 (Stamatakis, 2016) under the GTR+GAMMA model. The trees reconstructed using alignments containing an outgroup sequence were then rooted using the published out-group sequences (to identify the “true” root location), while the trees reconstructed using alignments without an outgroup sequence were rooted using RpNRM and IQ-TREE. We used the normalized RF distance to determine how far the RpNRM and IQTREE inferred root locations differed from that inferred using the outgroup rooting method.

Table 10 Details of the 28 empirical datasets used in the study.

Virus	Number of Taxa	Outgroup	Sites	Outgroup Reference
BBTVM	20	Abaca bunchy top virus DNA component M	1052	(Das & Banerjee, 2018)
BBTVN	12	Abaca bunchy top virus DNA component N	1083	
BBTVR	24	Abaca bunchy top virus DNA-component R	1111	
BBTVS	26	Abaca bunchy top virus DNA component S	1075	
BEGOMO9	20	Sida golden yellow vein virus	1379	(Jiang, et al., 2022)
CMV	14	Horseradish latent virus	8210	(Yasaka, et al., 2014)
FAV_A	11	Goose adenovirus	3855	(Goraichuk, Davis, Kulkarni, Afonso, & Suarez, 2021)
FAV_D	26		2271	
FAV_E	21		2716	
HIVM	24	SIV_CPZ	2696	(Korber, et al., 2000)
HPV6	19	Deltapapillomavirus	3018	(Oliveira, Lordello, Zardo, & Bonvicino, 2011)
HPV16	15		2360	
HPV18	13		1771	
HPV45	13		1182	
HRVA_A	43	Avian Rotavirus	2759	(Park, et al., 2011)
HMAV_B	30	Ovine adenovirus7	1089	(Lange, et al., 2019)
HMAV_D	14		1155	
FluA	12	B/Lee/40	2575	(Suarez & Perdue, 1998)
FluB	12	B/Lee/40	2575	(Chen, et al., 2007)
LPV				
PRVA	43	Avian Rotavirus	2759	(Park, et al., 2011)
SARB	23	Beta coronavirus	5992	
Sars-cov-2	30	Bat coronavirus	3298	(Forster, Forster, Renfrew, & Forster, 2020)
SMV_40	23	Simian agent 12	5441	(Teutsch, et al., 2015)
TTIV	11	pigeon Torque teno virus (PTTV)	4333	(Troiano, Bellardi, & Parrella, 2019)
TTSV	11		3325	
TTV	33	Pavovirus Parvoviridae	5725	(Martínez & Sebastián, 2018)
Ebola	21	Bundibugyo Ebolavirus		(Grard, et al., 2011)

4.2.3 Quantifying the Accuracy of Phylogenetic Inferences

To assess which method rooted the tree more accurately, the topological distances from the estimated roots to the true root was determined using the normalized RF measure. Given that the root in the true tree is placed on edge (v_1, v_2) and the root in the inferred tree of the respective rooting method is placed on edge (u_1, u_2) , the RF distance is the number of intermediate nodes from v_1 to u_1 .

4.2.4 Statistical Analysis

To assess the normalized RF distances and make a comparative assessment of how well trees were rooted among the four rooting methods, R version 4.2 was used to compute the nRF distances. The nRF distances were assessed as continuous variables and represented as means and standard deviations (Supplementary Tables 1, 2, 3, 4, 5) and for each tree balance measure, a paired t-test (correlated t-test) was then used to compare whether the nRF mean scores of trees rooted with one method were significantly different from trees rooted with another method i.e., outgroup vs RpNRM, outgroup vs midpoint, outgroup vs IQTREE, RpNRM vs midpoint, RpNRM vs IQTREE and midpoint vs IQTREE. For each rooting method, I further tested for difference in means (Wilcoxon test) between trees simulated at high DNR and trees simulated at low DNR for each respective rooting method. All statistical tests in this experiment were considered significant at $p.value < 0.05$.

4.3 Results and Discussion

The Outgroup Rooting Method Outperforms the RpNRM Method on Simulated Data

To determine whether, given evidence of evolutionary non-reversibility in a dataset, RpNRM can be used on its own as a reliable rooting application, it was necessary that we compared how well RpNRM performed against the other three rooting methods

i.e., outgroup, midpoint, and IQTREE. Specifically, the measure of the normalized RF distance between the estimated roots to the true root was used to indicate how well a rooting method rooted the phylogenetic trees. A normalized RF distance of zero was an indication that the true root was placed at the correct position in the tree and the higher the normalized RF distance the further away the root was placed from the correct position. To compare against low and high DNR, I further assessed whether increasing the degree of non-reversibility in the simulation process would impact the overall performance of the rooting methods. I altered the relative nucleotide substitution rates used during the simulation process from DNR = 0.442 to 2.193.

As evidenced by the normalized RF distances (Figure 10), for all simulation types (i.e., tree balance levels and alignment size), the outgroup method outperformed the midpoint, IQTREE, and RpNRM methods: i.e. normalized RF distances between true and inferred root positions were significantly lower ($p < 0.05$) for the outgroup rooting method than for the other three methods. This confirms previous assessments of rooting methods (Huelsenbeck, Bollback, & Levine, 2002) and affirms that outgroup rooting should remain the “gold-standard” of rooting approaches.

The midpoint rooting method outperformed the RpNRM and IQTREE methods for almost all datasets: the exception being the extremely imbalanced tree datasets where it was either not significantly better/worse than IQtree (for the 1000 site datasets) or was outperformed by IQTREE (for the 4000 site datasets). Overall, the tree rooting method implemented in IQTREE performed better than that implemented in RpNRM, except in the case of moderately imbalanced trees where the accuracy of RpNRM and IQTREE were not significantly different.

The midpoint rooting method involves the placement of the root at the midpoint of the longest branch-path between two sequences in the tree (Hess & De Moraes Russo, 2007), (Wade T. , Rangel, Kundu, Fournier, & Bansal, 2020), (Boykin, Kubatko, & Lowrey, 2010)), and it was therefore expected, that accuracy of this method would be highest for balanced trees and that it would get progressive less accurate with increasing levels of tree imbalance (Mooers & Heard, Inferring evolutionary process from phylogenetic tree shape, 1997), (Simberloff, Heck, McCoy, & Connor, 1981) (Hess & De Moraes Russo, 2007). Although this expectation was confirmed, similar,

less expected, decreases in rooting accuracy with increasing levels of tree imbalance were also observed for the RpNRM and IQTREE rooting methods: particularly so for the RpNRM method. (Figure 10, Figure 11).

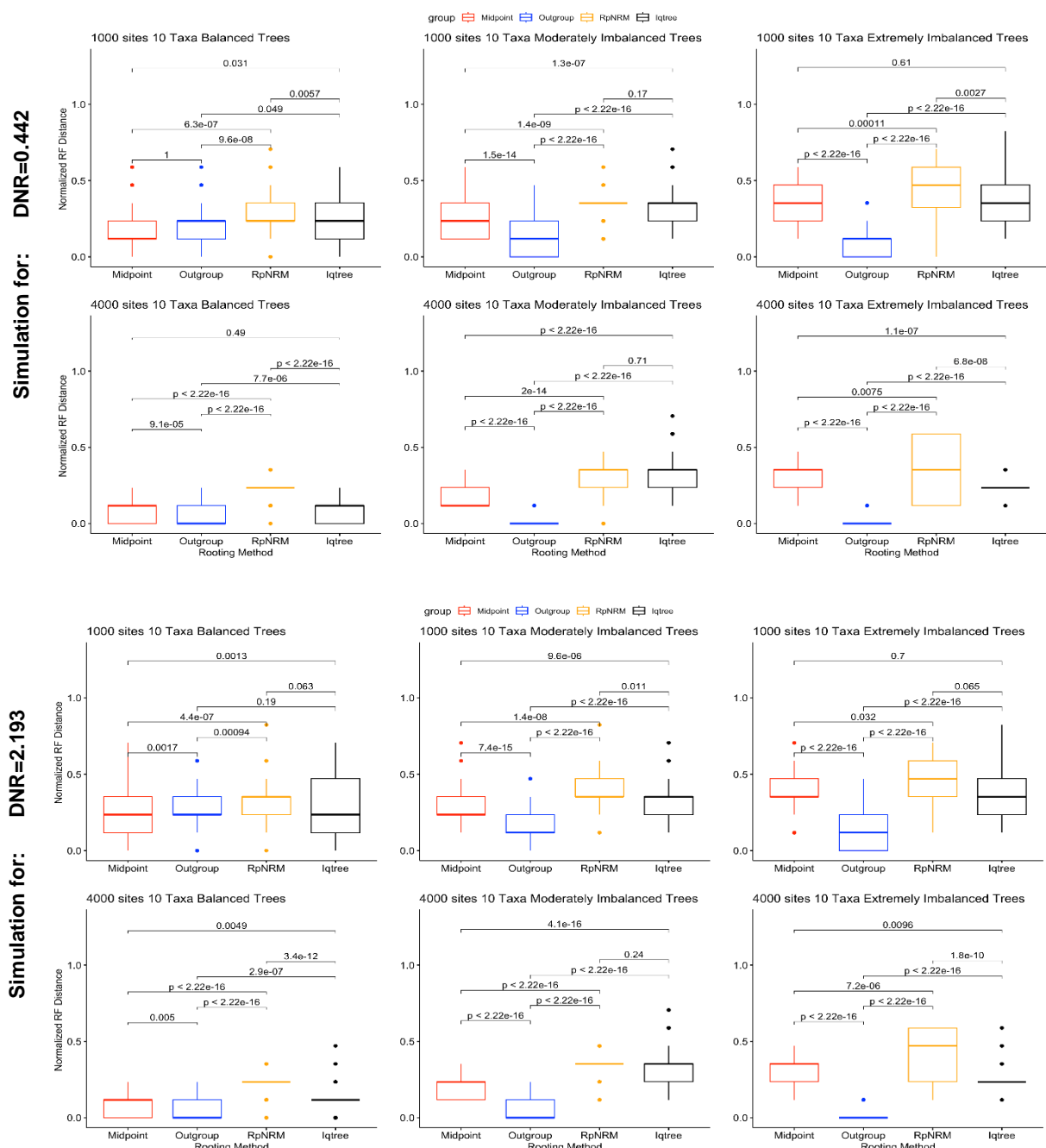


Figure 10 Normalized RF distances between true and inferred tree root locations under different rooting methods and DNR. The first two rows display results for low DNR (0.442) while the last two rows display results for high DNR (2.193).

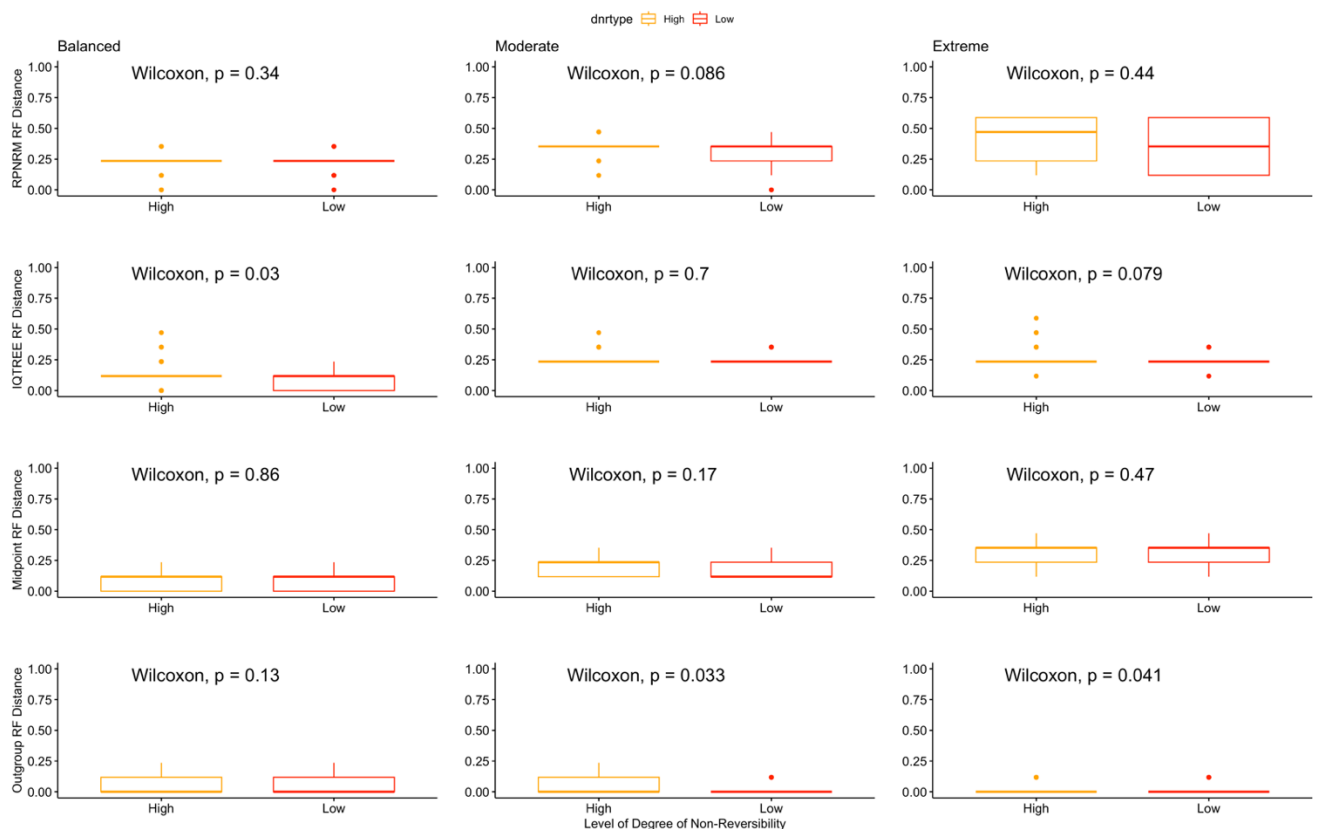


Figure 11 Comparison of normalized RF distances between the true and inferred roots under different rooting methods for high DNR inferred trees versus the low DNR inferred trees. Each row represents results for each rooting method while each column represents results for each tree balance level. Because the size of the tree affects root estimation positively (Figure 10), this comparison assessment was done using the results for the simulation of 4000 sites 10 taxa. The p. values indicate the results of the Wilcoxon test to test for significant differences between normalized RF distances for trees inferred under the assumption of low DNR versus high DNR.

This general decrease in the accuracy of these root node placement methods likely relates to the inherent difficulty of accurately determining both the topologies and branch lengths of trees when different taxon lineages included in the trees are evolving at different rates (Rohlf, Chang, Sokal, & Kim, 1990), (Duchêne, Duchêne, & Ho, 2015). This is particularly pertinent when, as is the case with our simulation setup, a lone taxon that has evolved at a lower rate than the remainder of the taxa in the tree splits from the rest of the tree close to the root node (note taxon F1_BR-07_0 in the extremely unbalanced tree in Figure 9; (Mooers & Heard, 1997) (Gregory, 2008)). In our simulations, irrespective of the rooting method used, when the root position was inaccurately identified these lone “outgroup” taxa were incorrectly clustered with other taxa within the tree (in the case of the extremely unbalanced tree this incorrect clustering was almost always with K_CM_96_96).

In assessing the influence of DNR on the accuracy of inferred roots (Figure 11), I expected that an increase in DNR should specifically increase the accuracy of the RpNRM and IQTREE rooting methods since these make use of non-reversible models to find the root. Although this was the case for the IQTREE method for the balanced tree (albeit with only marginal statistical significance; Figure 4), it was not detectably so for either other methods in the moderately and extremely unbalanced trees. It is unclear why the accuracy of the RpNRM and IQTREE root placement methods did not increase with increasing DNR but, presumably, the upper range of the DNR tested (2.193) does not provide substantially more information on the direction of the evolutionary process than the lower range of DNR tested (0.422). Since the range of DNR examined here encompasses that which has been seen in actual virus sequence datasets, it must therefore be concluded that for datasets of the sizes examined in the simulations, using non-reversible models to root virus phylogenetic trees is, in general, likely to be less accurate than rooting the trees using the outgroup rooting method. Further, the use of non-reversible nucleotide substitution models to root virus phylogenetic trees can be, more accurate than rooting the trees using the midpoint rooting method, but only when trees are imbalanced and/or DNRs are at the higher bounds of what is commonly found in virus sequence datasets.

The Midpoint Rooting Method Outperforms the RpNRM Method on Empirical Data

I conducted further tests to assess the accuracy of the RpNRM, IQ-TREE, and midpoint rooting methods on empirical virus sequence datasets. Specifically, these methods were used to root trees constructed from 28 empirical virus sequence datasets. Each dataset consisted of in-group sequences and one appropriate outgroup sequence that was used to infer the “true” roots of the in-group sequence clades within the 28 trees. We downloaded the nucleotide sequences from the National Centre for Biotechnology Information Taxonomy database (<http://www.ncbi.nlm.nih.gov/taxonomy>). Two versions of each of the 28 datasets were analysed: (1) a version including the outgroup sequence and (2) a version without the outgroup sequence. The datasets were aligned using muscle (Edgar, 2004) as implemented in Aliview (Larsson, 2014) and a phylogenetic tree was constructed for

both versions of each dataset using RAxML v8.2 (Stamatakis, 2016) under the GTR+GAMMA model. The trees constructed using alignments containing an outgroup sequence were rooted using the outgroup sequences. Trees constructed using alignments without an outgroup sequence were rooted using the RpNRM, IQ-TREE and midpoint rooting methods. I again used the normalized RF distance measure to determine how far the RpNRM, IQTREE and midpoint inferred root locations differed from those determined using the outgroup rooting method (Table 11 and Figure 12).

It is important to note that it is impossible to assess the true accuracy with which these methods rooted trees generated from empirical virus datasets because the true root locations of these trees are unknown. Given that the outgroup rooting method was the most accurate among those tested, we tested the accuracy of midpoint, RpNRM and IQ-TREE rooting using the location of the root position identified using the outgroup rooting approach as a proxy for the true root location.

Table 11 Normalized Robinson Foulds distances between the outgroup determined root locations and the 28 empirical datasets used in the study.

Virus	RPNRM distance	Midpoint distance	IQTREE distance	DNR	Colless
BBTVM	0.3076923	0.2051282	0.4615385	0.662	78
BBTVN	0.5217391	0.6086957	0.6956522	0.533	29
BBTVR	0.1276596	0.2978723	0.5957447	0.609	78
BBTVS	0.3529412	0.2745098	0.3921569	0.728	108
BEGOMO9	0.5526316	0.7105263	0.6052632	0.312	73
CMV	0.1481481	0.1481481	0.07407407	0.351	36
FAV_A	0.5714286	0.8571429	0.1904762	0.645	23
FAV_D	0.208	0.25	0.3333	0.646	70
FAV_E	0.4	0.2	0.4	0.357	142
HIVM	0.1304348	0.04347826	0.2608696	0.442	67
HPV6	0.7027027	0.7567568	0.8108108	0.451	52
HPV16	0.3448276	0.3448276	0.4827586	0.371	47
HPV18	0.64	0.64	0.56	0.323	33
HPV45	0.4	0.24	0.32	0.285	33
HRVA_A	0.4	0.1647059	0.1411765	0.3	336
HRVA_B	0.9152542	0.8813559	0.9152542	10.89	120
HMAV_D	0	0.24	0.4	0.646	39

FluA	0.5714286	0.5714286	0.4761905	0.27	31
FluB	0.5	0.136363	0.227272	0.311	30
LPV	0.3076923	0	0.1538462	0.42	10
PRVA	0.533333	0.4166667	0.4166667	0.351	6
SARB	0.2666667	0.133333	0.222222	0.301	86
Sars-cov-2	0.633333	0.633333	0.8	1.536	303
SMV_40	0.22222	0.22222	0.35555555	0.567	70
TTIV	0.296875	0.296875	0.265625	0.279	22
TTSV	0.8	0.2	0.4	1.56	26
TTV	0.238	0.143	0.238	0.513	138
Ebola	0.6341463	0.3414634	0.8292683	0.264	165

The root location identified by RpNRM was only closest to the outgroup root location in 6/28 of the datasets. The midpoints of the analysed trees were closer to the outgroup root locations in these trees for 12/28 of the datasets and the IQ-TREE identified roots were closest for 6/28 of the datasets. For the remaining four datasets the RpNRM, IQ-TREE and midpoint locations were equally close to outgroup locations for 1/28 of the datasets and for 3/28 of the datasets the RpNRM and midpoint roots were equally close. This therefore suggests that, even with empirical virus sequence datasets that display moderate amounts of DNR, the midpoint rooting method is, in general, likely to yield more accurate root locations than either of the methods that employ non-reversible nucleotide substitution models.

I further assessed the influence of DNR and tree balance on the accuracy of inferred empirical datasets by investigating whether the accuracy of inferred roots increased as the DNR increased for each respective rooting method and whether the level of tree balance influenced the accuracy with which root locations were inferred. To do this, for each rooting method I conducted a Pearson correlation test using the ggstatsplot package in R (Patil, 2021) to determine whether a significant association existed between (1) DNR and the normalized RF distances from the true tree; and (2) measures of Colless (tree balance) and normalized RF distances from the true tree.

I found that for all three rooting methods (Figure 12), the correlation between the rooting accuracy (adjudged by normalised RF distances from the true root) and the

DNR was positive but negligible with $\hat{r}_{pearson}$ of 0.252 (RPNRM), 0.24 (IQTREE) and 0.1 (Midpoint): none of these correlation coefficients were significantly different from zero (i.e. $p > 0.05$). This suggests that, across the range of DNR of the analysed datasets, DNR has no detectable impact on the accuracy with which the different methods infer root

Similarly, the correlation between the normalized RF distances and the colless measures of tree balance yielded $\hat{r}_{pearson}$ of 0.1 (RPNRM), 0.33 (IQTREE) and -0.0041 (Midpoint) (Figure 12), none of which were significantly different to zero. Again, suggesting that, for the ranges of tree imbalance present within the analysed empirical datasets, degrees of tree imbalance had no detectable impact on the accuracy with which trees were rooted by any of the three methods.

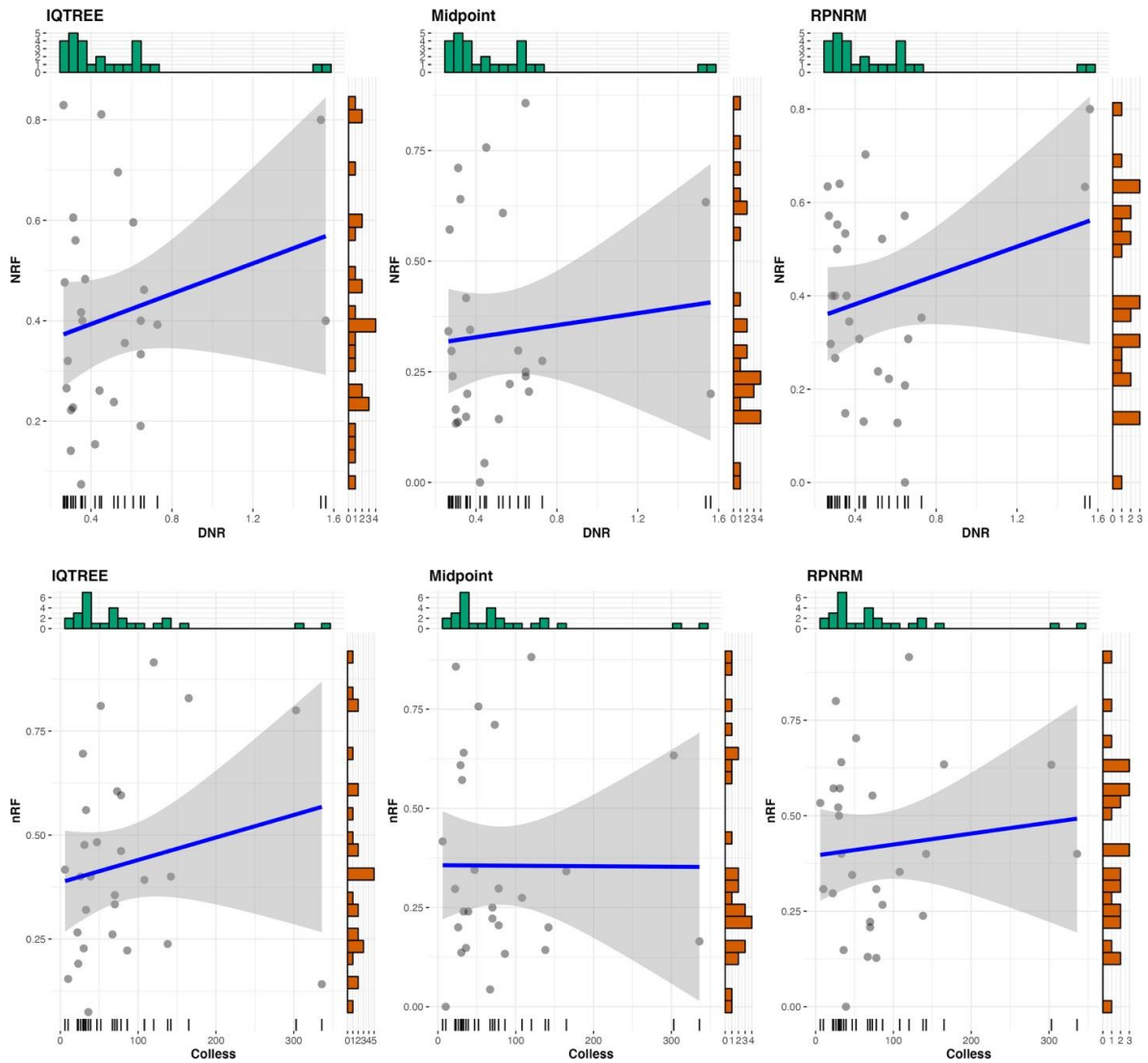


Figure 12 The association between the normalized Robinson-Faulds distance and the degree of non-reversibility, and the association between level of tree balance in inferred trees and the normalized RF distance. The green histogram represents the distribution of the DNR and Colless levels respectively while the orange histogram represents the distribution of the normalized RF distance.

4.4 Conclusion

I have tested whether methods for rooting trees using non-reversible substitution models (as implemented in RPNRM and IQTREE) would provide a good alternative to midpoint tree rooting in the absence of a close enough outgroup sequence that could be used for outgroup rooting: the method that is, justifiably, the current gold-standard for tree rooting. I was especially interested in determining whether methods that use non-reversible substitution models would perform well in instances where (1)

sequence datasets displayed high DNR (such as those for virus genome sequence data) and (2) the phylogenetic trees describing the evolutionary relationships of sequences in such datasets were imbalanced. What I found with both simulated and empirical sequence datasets was that, irrespective of the rooting method used, increasing DNR either reduces the accuracy with which trees are rooted with methods employing non-reversible substitution models only slightly outperforming the midpoint rooting method. Similarly, increasing tree imbalance impacted the accuracy of all rooting methods with the methods employing non-reversible nucleotide substitution models being only marginally less impacted than the midpoint rooting method.

Therefore, even when virus sequence datasets display high DNR and phylogenetic trees are highly imbalanced, using a non-reversible nucleotide substitution model to root phylogenetic trees that are constructed using virus genome sequences is, in most cases, unlikely to yield substantially more accurate root locations than the midpoint rooting method.

This disappointing result does not mean, however, that there is no value in estimating the root locations of trees using methods like those implemented in IQTREE and RpNRM. In the absence of an outgroup sequence, the rooting methods implemented in IQTREE and RpNRM can still be useful if used in conjunction with midpoint rooting to discover the true root location. Given the evidence of high DNR and/or tree imbalance, the rooting methods based on non-reversible substitution models could be very useful with respect to testing whether or not the midpoint of any given tree is, in fact, a likely root location.

Chapter 5: General Conclusion

5.1 General Discussion

I have highlighted how widely used models of nucleotide substitution in phylogenetic software's assume that the evolutionary process, a stochastic Markov Chain process will occur the same way both forward and backward in time even when the arrow of time is inverted (Lio & Goldman, 1998), (Hoff, Orf, Riehm, Darriba, & Stamatakis, 2016), (Tavaré, 1986). Thus Implying that given a relative rate of change from nucleotide j to i and π_i the equilibrium probability of state i , the forward process cannot be distinguished from the backward process hence giving the fundamental definition of time reversibility; $Q_{ji}\pi_i = Q_{ij}\pi_j$ where Q_{ji} is the instantaneous rate of change from j to i (Squartini & Arndt, 2008) (Posada D. , 2003). Mathematical I have shown through the use of the instantaneous rate matrix in equation (13) that time reversibility implies the relative rates i.e., a, b, c, d, e, f for base pairs i and j are equal where i and j can be A, C, G or T and $i \neq j$.

$$Q = \{q_{ij}\} = \begin{pmatrix} - & a\pi_A & b\pi_A & d\pi_A \\ a\pi_C & - & c\pi_C & e\pi_C \\ b\pi_G & c\pi_G & - & f\pi_G \\ d\pi_T & e\pi_T & f\pi_T & - \end{pmatrix}$$

(13)

I have however discussed based on literature ((Sanjuán & Domingo-Calap, 2016), (Cheng, Cahill, Kasai, Nishimura, & Loeb, 1992), (Nguyen, et al., 1992), (Chelico, Pham, Calabrese, & Goodman, 2006), (Sharma, Patnaik, Taggart, & Baysal, 20016)) that the reality of biochemical processes that are involved in causing relative nucleotide mutations are not the same as those that cause the reverse mutations i.e., For example the relative mutation rates of A to G are not the same that cause mutations from G to A and this is true for all other base mutations. But that mutations in viral genomes could be a result of enzyme infidelities during replication, RNA/DNA editing enzymes, ionizing radiation, inorganic oxidizers and chemical mutagens while in many other cases it is due to mutagenic chemical reactions or types of radiation

that, for example, cause G to A or C to U mutations in DNA or RNA, which are not the same as those that cause A to G or U to C mutations.

I have further discussed that Inappropriate nucleotide substitution models can substantially reduce the accuracy with which evolutionary relationships can be inferred (Posada, 2003) and that despite the increase in knowledge on non-reversible models, very little has been done to increase the awareness of the need to use such models in cases of evidence of strand specific mutation bias. Though the lack of use for non-reversible models is attributed to the computational cost involved in modelling with such parameter intensive models (Mai, Sayyari, & Mirarab, 2017), I have also attributed the under-use of non-reversible models to the lack of awareness on how wide non-reversible mutation are experienced by multiple viruses. Thus, my thesis was centred on (1) confirming the non-reversibility of evolution in diverse viral families by testing the goodness of-fit of GTR, NREV6 and NREV12 models to actual viral genome sequence data. (2) assessing how accurately maximum likelihood phylogenies are inferred when using GTR and NREV12 on simulated datasets generated using the NREV12 model. (3) determining the association between degrees of evolutionary non-reversibility and the degree of error in phylogenies that are inferred using GTR and (4) determining the accuracy with which non-reversible models can be used to accurately root Phylogenetic trees.

In chapter two, I compared the goodness-of-fit of the GTR, NREV6 and NREV12 models on 141 viral genomes to verify how wide non-reversibility exists in reality. I found that the 12-rate non-reversible model fits 97% of the datasets better than the general time reversible model (GTR) and the 6-rate non-reversible model (NREV6). I have shown that regardless of the wide use of reversible models in phylogenetic inferencing, there is high evidence of non-reversible mutations in all viral genome types i.e., ssDNA, ssRNA, dsDNA, and dsRNA.

It was of concern that not knowing the effects of the continued use of reversible models on the accuracy at which phylogenetic trees are being reconstructed could undermine the accuracy with which phylogenetic trees i.e., both the arrangement of tree branches and branch lengths are being inferred (Kapli, Flouri, & Telford, 2021). (Naser-Khdour,

Minh, Zhang, Stone, & Lanfear, 2019). I further tested in chapter 3 the impact of non-reversibility on the accuracy of phylogenetic trees under reversible and non-reversible models using simulated datasets that are evolved under the NREV12 model (using a phylogenetic tree which was inferred from an alignment of real sequences (Avian Leukosis virus) with different degrees of non-reversibility and varying average pairwise nucleotide sequence identities. Here I have however shown that using NREV12 a non-reversible nucleotide substitution model against the reversible model to construct a phylogenetic tree from virus genome sequences that display strand-specific nucleotide substitution biases does not guarantee one to yield a more accurate phylogenetic tree. I have also shown that where strand-specific substitution biases yield DNR levels that are higher than ~ 0.5 (such as are found in our SARS-CoV-2, Torque teno sus virus and Banana bunchy top virus datasets) it may be prudent to select a model such as NREV12 (such as is implemented in programs like IQTREE) over GTR as the better of two suboptimal choices.

In chapter 4 I develop an open-source web application (RpNRM) that places a root position on a phylogenetic tree using a non-reversible model NREV12. The goal was to test whether rooting phylogenetic trees using a non-reversible model would provide a good alternative to midpoint tree rooting in the absence of a close enough outgroup sequence that could be used for the outgroup rooting method the current gold-standard for tree rooting. I have shown both with simulated and empirical sequence datasets that an increase in DNR does not affect the good performance of the outgroup rooting method, irrespective of the DNR. I have shown that an increase in DNR does not consistently guarantee a better root position placement by the non-reversible methods than the midpoint method. Further, I have shown that tree imbalance greatly impacts the accuracy of phylogenetic inference with the methods employing non-reversible nucleotide substitution models being only marginally less impacted than the midpoint rooting method while the outgroup method being completely unaffected. Therefore, even when virus sequence datasets display high DNR and phylogenetic trees are highly imbalanced, using a non-reversible nucleotide substitution model to root phylogenetic trees that are constructed using virus genome sequences is, in most cases, unlikely to yield substantially more accurate root locations than the midpoint rooting method.

All together in this thesis, I have focused on showing the practical need for non-reversible models for both tree inference and placement of a root position. In conclusion, regardless of having found that using a non-reversible nucleotide substitution model to construct phylogenetic trees from virus genome sequences that display strand-specific nucleotide substitution biases has no guarantee of yielding more accurate phylogenetic trees when compared to the general time reversible models, there is still a need to consolidate the gap between assumptions made during the modelling of nucleotide substitutions and the reality of actual mutational processes. There is also a need to address the challenges of the lack of availability of data regarding the proportions of viral life cycles during which genomes exist in single and double-stranded states to aid in the prediction of when the use of non-reversible or reversible models might be most justified.

5.2 Limitation of Study

One limitation of the current rooting web-based application RpNRM presented in this thesis currently does not take into account the extent of the degree of non-reversibility of a dataset under study. At this point the application unrealistically assumes that as long as the analysis proceeds beyond the model test stage then it should be rooted accurately by the application. This assumption was however disproved by the simulation study on accuracy assessment where the trees with evidence of non-reversibility did not yield better root inference than other methods.

5.3 Future Works

Despite the current advances in the creation of phylogenetic software's that incorporate realistic nucleotide substitution model like IQTREE, we still lack software that considers the extent of the degree of non-reversibility before model selection. I believe that many of the issues on model misspecification can be addressed by the building of multiple phylogenetic reconstruction and analysis tools that considers more realistic assumptions depicting actual mutational processes but ensures datasets undergoes numerous checks such as checking the irreversibility indices of a given

dataset under study and setting a standard level of DNR that would require non-reversible models for an accurate inference and rooting. In addressing the challenges of the lack of available data regarding the proportions of viral life cycles during which genomes exist in single and double-stranded states there is a need to address this by the creation of a software that can predict the stage during viral life cycle at which a genome was sequenced to aid in model test selection.

References

- Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(934). Retrieved from <https://doi.org/10.1038/s41467-019-08822-w>
- Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G., & Hughes, S. H. (2010). Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology*, *84*(19), 9864-9878.
- Agor, J. K., & Özaltın, O. Y. (2018). Models for predicting the evolution of influenza to inform vaccine strain selection. *Human vaccines & immunotherapeutics*, *14*(3), 678-683.
- Ajbani, S. (2016). HIV vaccine development: current scenario and future prospects. *J AIDS Clin Res*, *7*(11), 626.
- Akand, E. H., & Downard, K. M. (2018). Identification of epistatic mutations and insights into the evolution of the influenza virus using a mass-based protein phylogenetic approach. *Molecular phylogenetics and evolution*, *121*, 132-138.
- Alcami, A., & Koszinowski, U. H. (2000). Viral mechanisms of immune evasion. *Trends in microbiology*, *8*(9), 410-418.
- Allen, J. E., & Whelan, S. (2014). Assessing the state of substitution models describing noncoding RNA evolution. *Genome biology and evolution*, *6*(1), 65-75.
- Anisimova, M., Bielawski, J. P., & Yang, Z. (. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution*, *18*(8), 1585-1592.
- Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in genetics*, *6*, 319.
- Baele, G., Gill, M. S., Lemey, P., & Suchard, M. A. (2019). Markov-modulated continuous-time Markov chains to identify site-and branch-specific evolutionary variation. *arXiv preprint*.
- Baele, G., Van de Peer, Y., & Vansteelandt, S. (2010). Using non-reversible context-dependent evolutionary models to study substitution patterns in primate non-coding sequences. *Journal of molecular evolution*, *17*(1), 34-50.
- Baisnée, P. F., Hampson, S., & Baldi, P. (2002). Why are complementary DNA strands symmetric?. *Bioinformatics*, *18*(8), 1021-1033.
- Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological reviews*, *35*(3), 235-241.
- Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2005). What does structure tell us about virus evolution?. *Current opinion in structural biology*, *15*(6), 655-663.
- Bamford, D., & Zuckerman, M. (2021). Encyclopedia of Virology. *Academic Press*.
- Bettisworth, B., & Stamatakis, A. (2021). Root Digger: a root placement program for phylogenetic trees. *BMC bioinformatics*, *22*(1), 1-20.
- Bottu, G. (2003). *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press.
- Boussau, B., & Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Systematic biology*, *55*(5), 756-768.
- Boykin, L. M., Kubatko, L. S., & Lowrey, T. K. (2010). Comparison of methods for rooting phylogenetic trees: A case study using *Orcuttieae* (Poaceae: Chloridoideae). *Molecular phylogenetics and evolution*, *54*(3), 687-700.

- Brüssow, H. (2021). COVID-19: emergence and mutational diversification of SARS-CoV-2. *Microbial Biotechnology*, 14(3), 756-768.
- Bretscher, M. T., Althaus, C., Müller, V., & Bonhoeffer, S. (2004). Recombination in HIV and the evolution of drug resistance: for better or for worse? *Bioessays*, 26(2), 180-188.
- Brief, T. (2020). Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom. *Epidemiology*, 7.
- Brumme, Z. L., John, M., Carlson, J. M., Brumme, C. J., Chan, D., Brockman, M. A., . . . Mallal, S. (2009). HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS one*, 4(8), e6687.
- Bruslind, L. (2020). *General microbiology*. Retrieved from General Microbiology: <https://open.oregonstate.edu/generalmicrobiology/chapter/the-viruses/>
- Buckley, T. R., & Cunningham, C. W. (2002). The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution*, 19(4), 394-405.
- Calarota, S. A., & Weiner, D. B. (2003). Present status of human HIV vaccine development. *Aids*, 17, S73-S84.
- Cann, A. J. (2008). Replication of viruses. *Encyclopedia of Virology*, 406.
- Castillo Ore, R. M., Caceda, R. E., Huaman, A. A., Williams, M., Hang, J., Juarez, D. E., & Forshey, B. M. (2018). Molecular and antigenic characterization of group C orthobunyaviruses isolated in Peru. *Plos one*, 13(7).
- Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics*, 19 (3 Pt 1), 233.
- Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). Shiny: Web application framework for R. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download>
- Chelico, L., Pham, P., Calabrese, P., & Goodman, M. F. (2006). APOBEC3G DNA deaminase acts processively 3'→ 5' on single-stranded DNA. *Nature structural & molecular biology*, 13(5), 392-399.
- Chen, G. W., Shih, S. R., Hsiao, M. R., Chang, S. C., Lin, S. H., Sun, C. F., & Tsao, K. C. (2007). Multiple genotypes of influenza B viruses cocirculated in Taiwan in 2004 and 2005. *Journal of clinical microbiology*, 45(5), 1515-1522.
- Cheng, K. C., Cahill, D., Kasai, H., Nishimura, S., & Loeb, L. A. (1992). 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes GT and AC substitutions. *Biological Chemistry*, 267(1), 166-172.
- Choi, J. Y., & Smith, D. M. (2021). SARS-CoV-2 variants of concern. *Yonsei medical journal*, 62(11), 961.
- Ciccozzi, M., Lai, A., Zehender, G., Borsetti, A., Cella, E., Ciotti, M., & Angeletti, S. (2019). The phylogenetic approach for viral infectious disease evolution and epidemiology: An updating review. *Journal of Medical Virology*, 91(10), 1707-1724.
- Clark, D. P., & Pazdernik, N. (2012). *Molecular biology*. Elsevier.
- Collins, R. A., Boykin, L. M., Cruickshank, R. H., & Armstrong, K. F. (2012). Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. 3(3), 457-465.
- Combe, M., & Sanjuan, R. (2014). Variation in RNA virus mutation rates across host cells. *PLoS pathogens*, 10(1).
- Conry, M. (2020). Determining the impact of recombination on phylogenetic inference (Doctoral dissertation, The Florida State University).

- Cottam, E. M., Wadsworth, J., Knowles, N. J., & King, D. P. (2009). Full sequencing of viral genomes: practical strategies used for the amplification and characterization of foot-and-mouth disease virus. *In Molecular Epidemiology of Microorganisms*, 217-230 .
- Da Silva, E. V., Da Rosa, A. P., Nunes, M. R., Diniz, J. A., Tesh, R. B., Cruz, A. C., & Vasconcelos, P. F. (2005). Araguari virus, a new member of the family Orthomyxoviridae: serologic, ultrastructural, and molecular characterization. *The American journal of tropical medicine and hygiene*, 73(6), 1050-1058.
- Dance, A. (2021). The incredible diversity of viruses. *Nature*, 595, 22-25.
- Das, T., & Banerjee, A. (2018). Distribution, molecular characterization and diversity of banana bunchy top virus in Tripura, India. *Virusdisease* , 29(2), 157-166.
- Davis, C., Logan, N., Tyson, G., Orton, R., Harvey, W., Haughney, J., & Willett, B. (2021). Reduced neutralisation of the Delta (B. 1.617. 2) SARS-CoV-2 variant of concern following vaccination. *medRxiv*.
- Dearlove, B., Lewitus, E., Bai, H., Li, Y., Reeves, D. B., Joyce, M. G., . . . Rolland, M. (2020). A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proceedings of the National Academy of Sciences*, 117(38), 23652-23662.
- Domingo, E. J., & Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annual review of microbiology*, 51(1), 151-178.
- Domingo, E., García-Crespo, C., Lobo-Vega, R., & Perales, C. (2021). Mutation rates, mutation frequencies, and proofreading-repair activities in RNA virus genetics. *Viruses*, 13(9), 1882.
- Dropulic, L. K., & Cohen, J. I. (2010). Update on new antivirals under development for the treatment of double-stranded DNA virus infections. *Clinical Pharmacology & Therapeutics*, 88(5), 610-619.
- Duchêne, D., Duchêne, S., & Ho, S. Y. (2015). Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. *Molecular Ecology Resources*, 15(4), 785-794.
- Duffy, S., & Holmes, E. C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *Journal of Virology*, 82(2), 957-965.
- Duffy, S., Shackelton, L., & Holmes, E. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 9, 267-276. Retrieved from <https://doi.org/10.1038/nrg2323>
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. . *Nucleic acids research*, 32(5), 1792-1797.
- Edwards, A. W., Cavalli-Sforza, L. L., Heywood, V. H., & McNeill, J. (1964). . Phenetic and phylogenetic classification. *Systematic Association Publication*, 6, 67-76.
- Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. D., Mishra, S., & ... & Sabino, E. C. (2021). Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372(6544), 815-821.
- Fariselli, P., Taccioli, C., Pagani, L., & Maritan, A. (2021). DNA sequence symmetries from randomness: the origin of the Chargaff's second parity rule. *Briefings in bioinformatics*, 22(2), 2172-2181.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6), 368-376.

- Fermin, G. (2018). Virion structure, genome organization, and taxonomy of viruses. *Viruses*, 17.
- Fernandes, J. V., & de Medeiros Fernandes, T. A. (2012). Human papillomavirus: biology and pathogenesis. In Human Papillomavirus and Related Diseases- From Bench to Bedside-A Clinical Perspective. *IntechOpen*.
- Fijalkowska, I. J., Jonczyk, P., Tkaczyk, M. M., Bialoskorska, M., & Schaaper, R. M. (1998). Unequal fidelity of leading and lagging strand DNA replication on the Escherichia coli chromosome. *Proceedings of the National Academy of Science* 95, 17, 10020-10025.
- Fitzsimmons, W. J., Woods, R. J., McCrone, J. T., Woodman, A., Arnold, J. J., Yennawar, M., & Luring, A. S. (2018). A speed–fidelity trade-off determines the mutation rate and virulence of an RNA virus. *PLoS biology*, 16(6), (6).
- Forsdyke, D. R. (2016). Chargaff's First Parity Rule. (Springer, Ed.) *Evolutionary Bioinformatics*, 25-42.
- Forsdyke, D. R., & Mortimer, J. R. (2000). Chargaff's legacy. *Gene*, 261(1), 127-137.
- Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. , . *Proceedings of the National Academy of Sciences*, 117(17), 9241-9243.
- Frank, A. C., & Lobry, J. R. (1991). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, 65-77.
- Furusawa, M. (2012). Implications of fidelity difference between the leading and the lagging strand of DNA for the acceleration of evolution. *Frontiers in oncology* 2, 144.
- Gärtner, K., Wiktorowicz, T., Park, J., Mergia, A., Rethwilm, A., & Scheller, C. (2009). Accuracy estimation of foamy virus genome copying. *Retrovirology*, 6(1), 1-15.
- Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P., & Charnay, P. (1979). Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in E. coli. *Nature*, 281(5733), 646-650.
- Galluzzi, L., Brenner, C., Morselli, E., Touat, Z., & Kroemer, G. (2008). Viral control of mitochondrial apoptosis. *PLoS pathogens*, 4(5).
- Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., & ... & Korber, B. (2002). Diversity considerations in HIV-1 vaccine selection. *Science*,, 296(5577), 2354-2360.
- Gatesy, J., DeSalle, R., & Wahlberg, N. (2007). How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Systematic Biology*, 56(2), 355-363.
- Gergerich, R. C., & Dolja, V. V. (2011). *Introduction to plant viruses, the invisible foe. Plant Health Instr. Available online: <https://www.apsnet.org/edcenter/disandpath/viral/introduction/Pages/PlantViruses.aspx> (accessed on 29 May 2020).*
- Gilbert, M. T., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., & Worobey, M. (2007). The emergence of HIV/AIDS in the Americas and beyond. *Proceedings of the National Academy of Sciences*,, . 104(47), 18566-18570.
- Girard, M. P., Osmanov, S. K., & Kieny, M. P. (2006). A review of vaccine research and development: the human immunodeficiency virus (HIV). *Vaccine*, 24(19), 4062-4081.
- Goraichuk, I. V., Davis, J. F., Kulkarni, A. B., Afonso, C. L., & Suarez, D. L. (2021). A 24-Year-Old Sample Contributes the Complete Genome Sequence of Fowl

- Aviadenovirus D from the United States. *Microbiology Resource Announcements*, 10(1).
- Graham, S. W., Olmstead, R. G., & Barrett, S. C. (2002). Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Molecular biology and evolution*, 19(10), 1769-1781.
- Grard, G., Biek, R., Muyembe, T. J., Fair, J., Wolfe, N., Formenty, P., & Leroy, E. (2011). Emergence of divergent Zaire ebola virus strains in Democratic Republic of the Congo in 2007 and 2008. , 204(suppl_3),. *The Journal of infectious diseases*, 204(suppl_3).
- Gregory, T. R. (2008). Understanding evolutionary trees. *Evolution: Education and Outreach*, 1(2), 121-137.
- Grigoriev, A. (1999). Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus research*, 60(1), 1-19.
- Grubaugh, N. D., Ladner, J. T., Lemey, P., Pybus, O. G., Rambaut, A., Holmes, E. C., & Andersen, K. G. (2019). Tracking virus outbreaks in the twenty-first century. *Nature microbiology*, 4(1), 10-19.
- Guindon, S., Delsuc, F., Dufayard, J. F., & Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *In Bioinformatics for DNA sequence analysis*, 113-137.
- Hampson, A., Barr, I., Cox, N., Donis, R. O., Siddhivinayak, H., Jernigan, D., . . . Zhang, W. (2017). Improving the selection and development of influenza vaccine viruses—Report of a WHO informal consultation on improving influenza vaccine virus selection, Hong Kong SAR, China, 18–20 November 2015. *Vaccine*, 35(8), 1104-1109.
- Hanson, L. (2009). 2 Isolation of Viral DNA from Cultures. *Handbook of Nucleic Acid Purification*. 23.
- Harkins, G., Delpont, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., . . . Varsani, A. (2009). Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology Journal*, 6(1), 1-14.
- Harris, H. M., & Hill, C. (2021). A place for viruses on the tree of life. *Frontiers in Microbiology*, 11, 604048.
- Henderson, P. T., Delaney, J. C., Gu, F., Tannenbaum, S. R., & Essigmann, J. M. (2002). Oxidation of 7, 8-dihydro-8-oxoguanine affords lesions that are potent sources of replication errors in vivo. *Biochemistry*, 41(3), 914-921.
- Hendrix, R. W., Hatfull, G. F., Ford, M. E., Smith, M. C., & Burns, R. N. (2002). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. In *Horizontal gene transfer* . *Academic Press*, 133-VI.
- Hendy, M. D., & Penny, D. (1989). Systematic zoology. *A framework for the quantitative study of evolutionary trees*, 38(4), 297-309.
- Hess, P. N., & De Moraes Russo, C. A. (2007). An empirical test of the midpoint rooting method. *Biological Journal of the Linnean Society*, 92(4), 669-674.
- Hoff, M., Orf, S., Riehm, B., Darriba, D., & Stamatakis, A. (2016). Does the choice of nucleotide substitution models matter topologically? *BMC bioinformatics*,, 17(1), 1-13.
- Holmes, E. C. (2009). The evolutionary genetics of emerging viruses. *Annu. Rev. Ecol. Evol. Syst*, 40, 353-372.
- Holmes, E. C., Rasmussen, A. L., Robertson, D. L., Crits-Christoph, A., Wertheim, J. O., . . . , & Rambaut, A. (2021). The origins of SARS-CoV-2: A critical review. *Cell*, 184(19), 4848-4856.

- Hossain, M. K., Hassanzadeganroudsari, M., & Apostolopoulos, V. (2021). The emergence of new strains of SARS-CoV-2. What does it mean for COVID-19 vaccines? *Expert Review of Vaccines*, 1-4.
- Hovmöller, R., Alexandrov, B., Hardman, J., & Janies, D. (2010). Tracking the geographical spread of avian influenza (H5N1) with multiple phylogenetic trees. *Cladistics*, 26(1), 1-13.
- Huang, Y. W., Dickerman, A. W., Piñeyro, P., Li, L., Fang, L., Kiehne, R. ..., & Meng, X. J. (2013). Origin, evolution, and genotyping of emergent porcine epidemic diarrhea virus strains in the United States. *MBio*, 4(5). doi:e00737-13
- Huelsenbeck, J. P., Bollback, J. P., & Levine, A. M. (2002). Inferring the Root of a Phylogenetic Tree. *Systematic Biology*, Volume 51(1), 32-43. doi:<https://doi.org/10.1080/106351502753475862>
- Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3), 299-314.
- International Committee on Taxonomy of Viruses Executive. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, 5(5), 668.
- Jahn, K., Beerenwinkel, N., & Zhang, L. (2021). The Bourque distances for mutation trees of cancers. *Algorithms for Molecular Biology*, 16(1), 1-15.
- Jermiin, L. S., Jayaswal, V., Ababneh, F. M., & Robinson, J. (2017). Identifying optimal models of evolution. In *Bioinformatics* . 379-420.
- Jiang, N., Gai, X., Yin, D., Zhang, G., Lu, C., Guo, J., & Xia, Z. (2022). Tobacco leaf curl Puer virus: a novel monopartite begomovirus infecting *Nicotiana tabacum* in China. *Archives of Virology*, 167(1), 229-232.
- Joint United Nations Programme on HIV/AIDS., & W. (2008). 2008 report on the global AIDS epidemic. *World Health Organization*.
- Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 21-132.
- Kaczorowska, J., & van der Hoek, L. (2020). Human anelloviruses: diverse, omnipresent and commensal members of the virome. *FEMS microbiology reviews*, 44(3), 305-313.
- Kapli, P., Flouri, T., & Telford, M. J. (2021). Systematic errors in phylogenetic trees. *Current Biology*, 31(2), R59-R64.
- Karki, S., Moniruzzaman, M., & Aylward. (2021). Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae. *Frontiers in microbiology*, 12, 657471. doi: <https://doi.org/10.3389/fmicb.2021.657471>
- Kay, A., & Zoulim, F. (2007). Hepatitis B virus genetic variability and evolution. *Virus research*, 127(2), 164-176.
- Kelchner, S. A., & Thomas, M. A. (2007). Model use in phylogenetics: nine key questions. . *Trends in Ecology & Evolution*, 22(2), 87-94.
- Kelk, S., & Linz, S. (2019). A tight kernel for computing the tree bisection and reconnection distance between two phylogenetic trees. *SIAM Journal on Discrete Mathematics*, 33(3), 1556-1574.
- 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.
- Kinene, T., Wainaina, J., Maina, S., & Boykin, L. M. (2016). Rooting trees, methods for. *Encyclopedia of Evolutionary Biology*. 489.

- Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., & Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology*, 13(5).
- Koonin, E. V., Krupovic, M., & Agol, V. I. (2021). The Baltimore classification of viruses 50 years later: how does it stand in the light of virus evolution?. *Microbiology and Molecular Biology Reviews*, 85(3).
- Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., & Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *science*, 288(5472), 1789-1796.
- Krupovic, M., & Bamford, D. H. (2011). Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. *Current opinion in virology*, 1(2), 118-124.
- Krupovic, M., Ghabrial, S. A., Jiang, D., & Varsani, A. (2016). Genomoviridae: a new family of widespread single-stranded DNA viruses. *Archives of virology*, 161(9), 2633-2643.
- Kuhner, M. K. (2015). Practical performance of tree comparison metrics. *Systematic biology*, 64(2), 205-214.
- Kuhner, M. K., & Yamato, J. (2015). Practical performance of tree comparison metrics. *Systematic Biology*, 64(2), 205-214.
- Kustin, T., & Adi, S. (2021). Biased Mutation and Selection in RNA Viruses. *Molecular Biology and Evolution*, 38(2), 575-588. doi: <https://doi.org/10.1093/molbev/msaa247>
- Kusumoto-Matsuo, R., Kanda, T., & Kukimoto, I. (2011). Rolling circle replication of human papillomavirus type 16 DNA in epithelial cell extracts. *Genes Cells. Genes to Cells*, 23-33. doi:doi: 10.1111/j.1365-2443.2010.01458.x. Epub 2010 Nov 9. PMID: 21059156.
- Lange, C. E., Niama, F. R., Cameron, K., Olson, S., Aime Nina, R., Ondzie, A., & Joly, D. O. (2019). First evidence of a new simian adenovirus clustering with Human mastadenovirus F viruses. *Virology journal*, 16(1), 1-6.
- Lanier, L. L. (2008). Evolutionary struggles between NK cells and viruses. *Nature Reviews Immunology*, 8(4), 259-268.
- Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22), 3276-3278.
- Lazarevic, I., Banko, A., Miljanovic, D., & Cupic, M. (2019). Immune-Escape Hepatitis B Virus Mutations Associated with Viral Reactivation upon Immunosuppression. *Viruses*, 11(9), 778.
- Lefort, V., Longueville, J. E., & Gascuel, O. (2017). SMS: smart model selection in PhyML. *Molecular biology and evolution*, 34(9), 2422-2424.
- Lemey, P., Rambaut, A., & Pybus, O. G. (2006). HIV evolutionary dynamics within and among hosts. *Aids Rev*, 8(3), 125-140.
- Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS computational biology*, 5(9).
- Lemey, P., Salemi, M., & Vandamme, A. M. (2009). The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. *Cambridge University Press*.
- Li, G., Piampongsant, S., Faria, N. R., Voet, A., Pineda-Peña, A. C., Khouri, R., & ... & Theys, K. (2015). An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology*, 12(1), 1-18.
- Lio, P., & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research*, 8(12), 1233-1244.

- Llabrés, M., Rosselló, F., & Valiente, G. (2021). The Generalized Robinson-Foulds Distance for Phylogenetic Trees. *Journal of Computational Biology*.
- Lobry, J. R., & Lobry, C. (1999). Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Molecular biology and evolution*, 16(6), 719-723.
- Louca, S., & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics*, 34(6), 1053-1055.
- Louten, J. (2016). Virus replication. *Essential human virology*, 49.
- Lwoff, A. (1957). The concept of virus. *Microbiology*, 17(2), 239-253.
- Lyons-Weiler, J., Hoelzer, G. A., & Tausch, R. J. (1998). Optimal outgroup analysis. 64(4), 493-511.
- Mai, U., Sayyari, E., & Mirarab, S. (2017). Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLoS One*, 12(8).
- Marchant, W. G., Gautam, S., Hutton, S. F., & Srinivasan, R. (2020). Tomato yellow leaf curl virus-resistant and-susceptible tomato genotypes similarly impact the virus population genetics. *Frontiers in Plant Science*, 11.
- Martínez, A., & Sebastián, J. (2018). Exploring the caudovirales: evaluation of their internal classification and potential relationships with the tectiviridae.
- Mavian, C. P., Marini, S., Magalis, B. R., Vandamme, A. M., Dellicour, S., & Salemi, M. (2020). Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proceedings of the National Academy of Sciences*, 117(23), 12522-12523.
- McCormack, G. P., & Clewley, J. P. (2002). The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Reviews in medical virology*, 12(4), 221-238.
- Mejer, N., Fahnøe, U., Galli, A., Ramirez, S., Weiland, O., Benfield, T., & Bukh, J. (2020). Mutations identified in the hepatitis C virus (HCV) polymerase of patients with chronic HCV treated with ribavirin cause resistance and affect viral replication fidelity. *Antimicrobial agents and chemotherapy*, 64(12).
- Menéndez-Arias, L. (2009). Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. *Viruses*, 1(3), 1137-1165.
- Meng, B., Abdullahi, A., Ferreira, I. A., Goonawardane, N., Saito, A., Kimura, I., & ... & Gupta, R. K. (2022). Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. *Nature*, 603(7902), 706-714.
- Meng, B., Kemp, S. A., Papa, G., Datir, R., Ferreira, I. A., Marelli, S., & ... & Masoli, J. A. (2021). Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B. 1.1. 7. *Cell reports*, 35(13).
- Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., & ... & Ogata, H. (2016). Linking virus genomes with host taxonomy. *Viruses*, 8(3), 66.
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5), 1530-1534.
- Minin, V., Abdo, Z., Joyce, P., & Sullivan, J. (2003). Performance-based selection of likelihood models for phylogeny estimation. *Systematic biology*, 52(5), 674-683.
- Mooers, A. O., & Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *The quarterly review of Biology*, 72 (1), 31-54.

- Morens, D., Folkers, G., & Fauci, A. (2004). The challenge of emerging and re-emerging infectious diseases. *Nature* 430, 242–249. Retrieved from , <https://doi.org/10.1038/nature02759>
- Morris, D. H., Gostic, K. M., Pompei, S., Bedford, T., Łuksza, M., Neher, R. A., . . . McCauley, J. W. (2018). Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends in microbiology*, 26(2), 102-118.
- Moya, A., Holmes, E. C., & González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology*, 2(4), 279-288.
- Mushegian, A. R. (2020). Are there 1031 virus particles on earth, or more, or fewer? *Journal of bacteriology*, 202(9).
- Naqvi, A. A., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., & Hassan, M. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *BBABiochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1866(10).
- Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., & Lanfear, R. (2019). The prevalence and impact of model violations in phylogenetic analysis. *Genome biology and evolution*, 11(12), 3341-3352.
- Naser-Khdour, S., Quang Minh, B., & Lanfear, R. (2022). Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. 71(4),. *Systematic Biology*, 71(4), 959-972.
- Newman, E. N., Holmes, R. K., Craig, H. M., Klein, K. C., Lingappa, J. R., Malim, M. H., & Sheehy, A. M. (2005). Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity. *Current Biology*, 15(2), 166-170.
- Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1), 268-274.
- Nguyen, T., Brunson, D., Crespi, C., Penman, B., Wishnok, J., & Tannenbaum, S. (1992). DNA damage and mutation in human cells exposed to nitric oxide in vitro. *National Academy of Science*, 89(7), 3030-3034.
- O'Carroll, I. P., & Rein, A. (2016). Viral nucleic acids. *Encyclopedia of Cell Biology*, 517.
- Oliveira, S. M., Lordello, C. X., Zardo, L., & Bonvicino, C. M. (2011). Human Papillomavirus in Brazilian women with and without cervical lesions. *Virology journal*, 8(1), 1-6.
- Onwubiko, O., Borst, A., Diaz, A., Passkowski, K., Scheffel, F., Tessmer, I., & Nasheuer, H. (2020, April). SV40 T antigen interactions with ssDNA and replication protein A: a regulatory role of T antigen monomers in lagging strand DNA replication. *National Library of medicine*, 48(7), 3657-3677. doi:10.1093/nar/gkaa138
- Organization, W. H. (2012). Measles virus nomenclature update: 2012. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*. 87(09), 73-80.
- Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., & Ippodrino, R. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of translational medicine*, 18(1), 1-9.
- Park, S. I., Matthijssens, J., Saif, L. J., Kim, H. J., Park, J. G., Alfajaro, M. M., & Cho, K. O. (2011). Reassortment among bovine, porcine and human rotavirus

- strains results in G8P [7] and G6P [7] strains isolated from cattle in South Korea. *Veterinary microbiology*, 152(1-2), 55-66.
- Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, 6(61), 3167.
- Pattengale, N. D., Gottlieb, E. J., & Moret, B. M. (2007). Efficiently computing the Robinson-Foulds metric. *Journal of computational biology*. 14(6), 724-735.
- Patterson, J. L., & Fernandez-Larsson, R. (1990). Molecular mechanisms of action of ribavirin. *Reviews of infectious diseases*. 12(6), 1139-1146.
- Pawlotsky, J. M. (2011). Treatment failure and resistance with direct-acting antiviral drugs against hepatitis C virus. *Hepatology*, 53(5), 1742-1751.
- Peck, K. M., & Luring, A. (2018). Complexities of viral mutation rates. *Journal of virology*, 92(14).
- Peter, S., & Pakorn, A. (2018). Virus classification where do you draw the line? *Archives Virology*(163), 2037–2046. Retrieved from <https://doi.org/10.1007/s00705-018-3938-z>
- Phadungsombat, J., Lin, M. Y., Srimark, N., Yamanaka, A., Nakayama, E. E., Moolasart, V., & Uttayamakul, S. (2018). Emergence of genotype Cosmopolitan of dengue virus type 2 and genotype III of dengue virus type 3 in Thailand. *PloS one*, 13(11).
- PhyML, (. s. (2017). Lefort, V;Longueville, J. E;Gascuel, O. *Molecular biology and evolution*, 34(9), 2422-2424.
- Pipes, L., Wang, H., Huelsenbeck, J. P., & Nielsen, R. (2021). Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *Molecular biology and evolution*, 38(4), 1537-1543.
- Polak, P., & Arndt, P. F. (2008). Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Research*, 18(8), 1216-1223.
- Pond, S. L., Frost, S. D., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5), 676-679.
- Popa, A., Genger, J. W., Nicholson, M. D., Penz, T., Schmid, D., Aberle, S. W., . . . Bergthaler, A. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Science translational medicine*, 12(573). doi:eabe2555
- Posada, D. (2003). Using MODELTEST and PAUP to select a model of nucleotide substitution. (1), pp. 6-5.
- Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5), 793-808.
- Posada, D., & Crandall, K. A. (2001, Posada, David; Keith, A, Crandall). Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Molecular biology and evolution*, 18(6), 897-906.
- Posada, D., & Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic biology*, 50(4), 580-601.
- Posada, D., & Crandall, K. A. (2021). Felsenstein phylogenetic likelihood. *Journal of molecular evolution*, 89(3), 134-145.
- Pybus, O. G., & Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8), 540-550.
- Rambaut, A., Posada, D., Crandall, K. A., & Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 52–61. Retrieved from <https://doi.org/10.1038/nrg1246>

- Ravantti, K. ., & Bamford, D. (2009). Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Bio*, 9(112). doi: doi: 10.1186/1471-2148-9-112
- Rector, A., Lemey, P., Tachezy, R., Mostmans, S., Ghim, S. J., Van Doorslaer, K., & Van Ranst, M. (2007). Ancient papillomavirus-host co-speciation in Felidae. *Genome biology*, 8(4), 1-2.
- Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, 3(2), 217-223.
- Ripplinger, J., & Sullivan, J. (2008). Does choice in model selection affect maximum likelihood analysis? *Systematic biology*, 57(1), 76-85.
- Risso-Ballester, J., & Sanjuán, R. (2019). High fidelity deep sequencing reveals no effect of atm, atr, and DNA-pk cellular DNA damage response pathways on adenovirus mutation rate. *Viruses*. 10(11), 938.
- Ritz, C., & Spiess, A. N. (2008). qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. *Bioinformatics*. 24(13), 1549-1551.
- Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2), 131-147.
- Rochman, N. D., Wolf, Y. I., Faure, G., Mutz, P., Zhang, F., & Koonin, E. V. (2021). Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proceedings of the National Academy of Sciences*, 118(29).
- Rodpothong, P., & Auewarakul, P. (2012). Viral evolution and transmission effectiveness. *World journal of virology*, 1(5), 131.
- Rodriguez, F. J., Oliver, J. L., Marin, A., & Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of theoretical biology*, 142(4), 485-501.
- Roger, A. J., & Hug, L. A. (2006). The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philosophical Transactions of the Royal Society B. Biological Sciences*, 361(1470), 1039-1054.
- Rohlf, F. J., Chang, W. S., Sokal, R. R., & Kim, J. (1990). Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. *Evolution*. 44(6), 1671-1684.
- Roossinck, M. J., & Witzany, G. (2012). Viruses: essential agents of life.
- Rosario, K., Duffy, S., & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology*, 157(10), 1851-1871.
- Ryu, W. S. (2017). Virus life cycle. *Molecular Virology of Human Pathogenic Viruses*, 31.
- Samuel, C. E. (2011). Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. *Virology*, 411(2), 180-193.
- Sanjuán, R. N., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation rates. *Journal of virology*, 84(19), 9733-9748.
- Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and molecular life sciences*. 73(23), 4433-4448.
- Sarairah, H., Bdour, S., & Gharaibeh, W. (2020). The molecular epidemiology and phylogeny of torque teno virus (TTV) in Jordan. *Viruses*, 12(2), 165.
- Sattentau, Q. (2008). Avoiding the void: cell-to-cell spread of human viruses. *Nature Reviews Microbiology*, 6(11), 815-826.

- Schöniger, M., & Von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular phylogenetics and evolution*, 3(3), 240-247.
- Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592-593.
- Schmidt, H. A., Strimmer, K., Vingron, M., & Von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3), 502-504.
- Shafer, R. W., & Schapiro, J. M. (2008). HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS reviews*, 10(2), 67.
- Sharma, S., Patnaik, S. K., Taggart, R. T., & Baysal, B. E. (2016). The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Scientific reports*, 6(1), 1-12.
- Sharp, P. M., & Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1).
- Shearer, P. L., Bonne, N., Clark, P., Sharp, M., & Raidal, S. R. (2008). Beak and feather disease virus infection in cockatiels (*Nymphicus hollandicus*). *Avian pathology*, 37(1), 75-81.
- Shin, S. (2016). Recent update in HIV vaccine development. *Clinical and experimental vaccine research*, 5(1), 6-11.
- Simberloff, D. S., Heck, K. L., McCoy, E. D., & Connor, E. F. (1981). There Have Been No Statistical Tests of Cladistic Biogeographical Hypotheses!.
- Simion, P., Delsuc, F., & Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome?
- Simion, P., Delsuc, F., & Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome?
- Simmonds, P. (2020). Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short-and long-term evolutionary trajectories. *MSphere*, 5(3).
- Simmonds, P., & Aiewsakun, P. (2018). Virus classification—where do you draw the line? *Archives of virology*, 163(8), 2037-2046.
- Simmonds, P., & Ansari, M. (2021). Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. *PLoS Pathog*, 17(6).
- Simmonds, P., & Ansari, M. (2021). Mutation bias implicates RNA editing in a wide range of mammalian RNA viruses. *bioRxiv*, . PLoS Pathog. 2021 Jun 1;17(6):e1009596. doi: 10.1371/journal.pp.
- Simmonds, P., & Ansari, M. A. (2021). Mutation bias implicates RNA editing in a wide range of mammalian RNA viruses. *bioRxiv*.
- Smith, A. B., & Peterson, K. J. (2002). Dating the time of origin of major clades: molecular clocks and the fossil record. *Annual Review of Earth and Planetary Sciences*, 30(1), 65-88.
- Smith, D., & Simmonds, P. (1997). Characteristics of Nucleotide Substitution in the Hepatitis C Virus Genome: Constraints on Sequence Change in Coding Regions at Both Ends of the Genome. *J Mol Evol*, 45, 238-246. doi:<https://doi.org/10.1007/PL00006226>
- Squartini, F., & Arndt, F. P. (2008, December). Quantifying the stationarity and Time Reversibility of the Nucleotide Substitution process. *Molecular Biology and*

- Evolution*, 25(12), 2525-2535. Retrieved from <https://doi.org/10.1093/molbev/msn169>
- Squartini, F., & Arndt, P. F. (2008). Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Molecular biology and evolution*, 25(12), 2525-2535.
- Sridhar, S., Teng, J. L., Chiu, T. H., Lau, S. K., & Woo, P. C. (2017). Hepatitis E virus genotypes and evolution: emergence of camel hepatitis E variants. *International journal of molecular sciences*, 18(4), 869.
- Stamatakis, A. (2016). *The RAxML v8. 2. X Manual*. Retrieved from Heidelberg Institute for Theoretical Studies: <https://cme.h-its.org/exelixis/resource/download/NewManual.pdf>.
- Stern, A., & Andino, R. (2016). Chapter 17 - Viral Evolution: It Is All About Mutations. *Viral Pathogenesis*, 233-240. Retrieved from <https://doi.org/10.1016/B978-0-12-800964-2.00017-3>
- Stern, A., Te Yeh, M., Zinger, T., Smith, M., Wright, C., Ling, G., & Andino, R. (2017). The evolutionary pathway to virulence of an RNA virus. *Cell*, 169(1), 35-46.
- Strimmer, K., von Haeseler, A., & Salemi, A. M. (2003). *Nucleotide substitution models. In The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny*. Cambridge: Cambridge University Press.
- Strimmer, K., von Haeseler, A., & Salemi, M. (2009). Genetic distances and nucleotide substitution models. *The Phylogenetic Handbook*. 111-141.
- Suarez, D. L., & Perdue, M. L. (1998). Multiple alignment comparison of the non-structural genes of influenza A viruses. *Virus research*, 54(1), 59-69.
- Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evo* 40, 318-325. Retrieved from <https://doi.org/10.1007/BF00163236>
- Svarovskaia, E. S., Cheslock, S. R., Zhang, W. H., Hu, W. S., & Pathak, V. K. (2003). Retroviral mutation rates and reverse transcriptase fidelity. *Frontiers in Bioscience-Landmark*, 8(4), 117-134.
- Tabatabaee, Y., Sarker, K., & Warnow, T. (2022). Quintet Rooting: rooting species trees under the multi-species coalescent model. *Bioinformatics*, 38(Supplement_1).
- Tao, K., Tzou, P. L., Nouhin, J., Gupta, R. K., de Oliveira, T., Kosakovsky Pond, S. L., & ... & Shafer, R. W. (2021). The biological and clinical significance of emerging SARS-CoV-2 variants. *Nature Reviews Genetics*, 22(12), 757-773.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*. 17(2), 57-86.
- Tegally, H., Ramuth, M., Amoaka, D., Scheepers, C., Wilkinson, E., Giovanetti, M., & ... & Manraj, S. (2021). A novel and expanding SARS-CoV-2 Variant, B. 1.1. 318, dominates infections in Mauritius.
- Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., & de Oliveira, T. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854), 438-443.
- Telesnitsky, A., & Wolin, S. L. (2016). The host RNAs in retroviral particles. *Viruses*, 8(8), 235.
- Teutsch, K., Schweitzer, F., Knops, E., Kaiser, R., Pfister, H., Verheyen, J., & Di Cristanziano, V. (2015). Early identification of renal transplant recipients with high risk of polyomavirus-associated nephropathy. *Medical microbiology and immunology*, 204(6), 657-664.

- Tian, Y., & Kubatko, L. (2017). Rooting phylogenetic trees under the coalescent model using site pattern probabilities. . *BMC evolutionary biology*, 17(1), 1-11.
- Tria, F. D., Landan, G., & Dagan, T. (2017). Phylogenetic rooting using minimal ancestor deviation. *Nature ecology & evolution*, 1(7), 0193.
- Troiano, E., Bellardi, M. G., & Parrella, G. (2019). *Syringa vulgaris* is a new host for cucumber mosaic virus. . *Phytopathologia Mediterranea*, 58(2).
- Van Der Walt, E., Martin, D. P., Varsani, A., Polston, J. E., & Rybicki, E. P. (2008). Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virology journal*, 5(1), 1-11.
- van Zyl, G., Bale, M., & Kearney, M. (2018). HIV evolution and diversity in ART-treated patients. *Retrovirology*, 115(14). Retrieved from M.F. <https://doi.org/10.1186/s12977-018-0395-4>
- Varsani, A., Lefeuvre, P., Roumagnac, P., & Martin, D. (2018). Notes on recombination and reassortment in multipartite/segmented viruses. *Current opinion in virology*, 33, 156-166.
- Venkatesan, S., Rosenthal, R., Kanu, N., McGranahan, N., Bartek, J., Quezada, S. A., . . . Swanton, C. (2018). Perspective: APOBEC mutagenesis in drug resistance and immune escape in HIV and cancer evolution. *Annals of Oncology*, 29(3), 563-572.
- Viana, R., Moyo, S., Amoako, D. G., Tegally, H., Scheepers, C., Althaus, C. L., & ... & de Oliveira, T. (2022). Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603(7902), 679-686.
- Vignuzzi, M., & López, C. B. (2019). Defective viral genomes are key drivers of the virus–host interaction. *Nature microbiology*, 4(7), 1075-1087.
- Volz, E. M., & Frost, S. D. (2013). Inferring the source of transmission with phylogenetic data. *PLoS computational biology*, 9(12). doi:e1003397.
- Wade, T., Rangel, L. T., Kundu, S., Fournier, G. P., & Bansal, M. (2020). Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families. *PloS one*, 15 (5). doi:e0232950
- Wade, T., Rangel, L. T., Kundu, S., Fournier, G. P., & Bansal, M. (2020). Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families. *PloS one*, 15 (5). doi:e0232950
- Walker, P., Siddell, S., Lefkowitz, E., & al, e. (2019). Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses. *Arch Virol* 164, 2417–2429.
- Walt, v. d., Martin, D., Varsani, A., & al.et. (2008). Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virol J* 5. doi:<https://doi.org/10.1186/1743-422X-5-104>
- Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737-738.
- Webb, B., Rakibuzzaman, A. G., & Ramamoorthy, S. (2020). Torque teno viruses in health and disease. *Virus research*, 285, 198013.
- Wei, S., Shi, M., Chen, X., Sharkey, M., van Achterberg, C., Ye, G., & He, J. (2010). New views on strand asymmetry in insect mitochondrial genomes. *PLoS One*, 5 (9).
- Whidden, C., Zeh, N., & Beiko, R. G. (2014). Supertrees Based on the Subtree Prune-and-Regraft Distance,. *Systematic Biology*, 63(4), 566-581. doi:<https://doi.org/10.1093/sysbio/syu023>
- Whittaker, G. R., Kann, M., & Helenius, A. (2000). Viral entry into the nucleus. *Annual review of cell and developmental biology*, 16(1), 627-651.

- WHO. (2021, July 17). *World Health Organization*. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/hiv-aids>
- Wickner, R. B. (1993). Double-stranded RNA virus replication and packaging. *The Journal of biological chemistry*, 268(6), 3797-3800.
- Williams, T. A., Heaps, S. E., Cherlin, S., Nye, T. M., Boys, R. J., & Embley, T. M. (2015). New substitution models for rooting phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678).
- Worobey, M., & Han, G. Z. (2012). The origins and diversification of HIV. 15-24.
- Worobey, M., & Holmes, E. C. (1999). Evolutionary aspects of recombination in RNA viruses. *Journal of General Virology*, 80(10), 2535-2543.
- Worobey, M., Cox, J., & Gill, D. (2019). The origins of the great pandemic. *Evolution, Medicine, and Public Health*, 2019(1), 18-25.
- Wu, Y. (2009). A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 25(2), 190-196.
- Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J Mol Evol*, 39, 105-111. Retrieved from <https://doi.org/10.1007/BF00178256>
- Yang, Z. (2003). Notes on Calculation of the Transition Probability Matrix $P(t) = \exp(Qt)$. University College London, UK, Tech. Rep
- Yap, V. B., & Speed, T. (2005). Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evolutionary Biology*, 5(1), 1-8.
- Yasaka, R., Nguyen, H. D., Ho, S. Y., Duchêne, S., Korkmaz, S., Katis, N., & Ohshima, K. (2014). The temporal evolution and global spread of Cauliflower mosaic virus, a plant pararetrovirus. *PloS one*, 9(1).
- Youri, P., Newlon, C. S., & KunkelThomas, A. (2002). Yeast origins establish a strand bias for replicational mutagenesis. *Molecular cell* 10, 1, 207-213.
- Yu, Q., König, R., Pillai, S., Chiles, K., Kearney, M., Palmer, S., & Landau, N. R. (2004). Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nature structural & molecular biology*, 11(5), 435-442.
- Zaccaria, G., Lorusso, A., Hierweger, M. M., Malatesta, D., Defourny, S. V., Ruggeri, F., & Marcacci, M. (2020). Detection of Astrovirus in a Cow with Neurological Signs by Nanopore Technology, Italy. *Viruses*, 12(5), 530.
- Zardoya, R. (2021). Quest for the best evolutionary model. *Journal of Molecular Evolution*, 89(3), 146-150.
- Zhang, Q., Feild, T. S., & Antonelli, A. (2015). Assessing the impact of phylogenetic incongruence on taxonomy, floral evolution, biogeographical history, and phylogenetic diversity. *American Journal of Botany*, 102(4), 566-580.
- Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., & Screaton, G. (2021). Evidence of escape of SARS-CoV-2 variant B. 1.351 from natural and vaccine-induced sera. *Cell*, 184(9), 2348-2361.

Appendices

Supplementary files for Chapter 3

Supplementary Table 1 wRF distance summary statistics for the trees inferred using sequences with 75% API.

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Degree of Non-Reversibility: 0			
Mean (SD)	0.566 (0.0666)	0.565 (0.0667)	0.938
Median [Min, Max]	0.565 [0.433, 0.728]	0.562 [0.432, 0.729]	
Degree of Non-Reversibility: 2			
Mean (SD)	0.636 (0.0754)	0.629 (0.0740)	0.429
Median [Min, Max]	0.624 [0.492, 0.812]	0.616 [0.492, 0.801]	
Degree of Non-Reversibility: 4			
Mean (SD)	0.652 (0.0766)	0.639 (0.0747)	0.189
Median [Min, Max]	0.641 [0.514, 0.822]	0.630 [0.507, 0.804]	
Degree of Non-Reversibility: 6			
Mean (SD)	0.659 (0.0793)	0.644 (0.0765)	0.119
Median [Min, Max]	0.658 [0.514, 0.843]	0.642 [0.508, 0.832]	
Degree of Non-Reversibility: 8			
Mean (SD)	0.662 (0.0782)	0.645 (0.0752)	0.0707
Median [Min, Max]	0.655 [0.521, 0.838]	0.642 [0.512, 0.823]	
Degree of Non-Reversibility: 10			
Mean (SD)	0.665 (0.0783)	0.647 (0.0751)	0.064
Median [Min, Max]	0.658 [0.528, 0.852]	0.640 [0.516, 0.835]	
Degree of Non-Reversibility: 12			
Mean (SD)	0.669 (0.0766)	0.650 (0.0739)	0.0553
Median [Min, Max]	0.657 [0.531, 0.849]	0.643 [0.520, 0.833]	
Degree of Non-Reversibility: 14			
Mean (SD)	0.670 (0.0764)	0.651 (0.0737)	0.0559
Median [Min, Max]	0.661 [0.529, 0.844]	0.644 [0.516, 0.832]	
Degree of Non-Reversibility: 16			

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Mean (SD)	0.670 (0.0768)	0.652 (0.0741)	0.0603
Median [Min, Max]	0.663 [0.532, 0.850]	0.643 [0.516, 0.835]	
Degree of Non-Reversibility: 18			
Mean (SD)	0.671 (0.0763)	0.653 (0.0738)	0.0587
Median [Min, Max]	0.665 [0.534, 0.846]	0.645 [0.519, 0.833]	
Degree of Non-Reversibility: 20			
Mean (SD)	0.672 (0.0763)	0.654 (0.0737)	0.0577
Median [Min, Max]	0.669 [0.530, 0.846]	0.649 [0.516, 0.831]	

Supplementary Table 2 wRF distance summary statistics for the trees inferred using sequences with 80% API.

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Degree of Non-Reversibility: 0			
Mean (SD)	0.510 (0.0639)	0.508 (0.0643)	0.876
Median [Min, Max]	0.502 [0.373, 0.665]	0.500 [0.372, 0.665]	
Degree of Non-Reversibility: 2			
Mean (SD)	0.570 (0.0676)	0.563 (0.0662)	0.411
Median [Min, Max]	0.564 [0.438, 0.723]	0.551 [0.437, 0.710]	
Degree of Non-Reversibility: 4			
Mean (SD)	0.581 (0.0702)	0.569 (0.0679)	0.176
Median [Min, Max]	0.572 [0.446, 0.752]	0.556 [0.443, 0.733]	
Degree of Non-Reversibility: 6			
Mean (SD)	0.590 (0.0724)	0.575 (0.0699)	0.111
Median [Min, Max]	0.582 [0.462, 0.775]	0.568 [0.458, 0.755]	
Degree of Non-Reversibility: 8			
Mean (SD)	0.593 (0.0722)	0.576 (0.0692)	0.0744
Median [Min, Max]	0.583 [0.464, 0.760]	0.571 [0.455, 0.738]	
Degree of Non-Reversibility: 10			
Mean (SD)	0.597 (0.0725)	0.579 (0.0694)	0.0652

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Median [Min, Max]	0.587 [0.472, 0.770]	0.569 [0.459, 0.754]	
Degree of Non-Reversibility: 12			
Mean (SD)	0.599 (0.0712)	0.581 (0.0678)	0.056
Median [Min, Max]	0.587 [0.471, 0.781]	0.572 [0.459, 0.759]	
Degree of Non-Reversibility: 14			
Mean (SD)	0.599 (0.0704)	0.582 (0.0671)	0.0575
Median [Min, Max]	0.591 [0.476, 0.775]	0.573 [0.464, 0.753]	
Degree of Non-Reversibility: 16			
Mean (SD)	0.600 (0.0707)	0.582 (0.0678)	0.062
Median [Min, Max]	0.594 [0.474, 0.775]	0.574 [0.461, 0.757]	
Degree of Non-Reversibility: 18			
Mean (SD)	0.600 (0.0703)	0.583 (0.0675)	0.0625
Median [Min, Max]	0.593 [0.478, 0.776]	0.576 [0.466, 0.756]	
Degree of Non-Reversibility: 20			
Mean (SD)	0.601 (0.0701)	0.584 (0.0670)	0.0596
Median [Min, Max]	0.596 [0.486, 0.775]	0.579 [0.473, 0.757]	

Supplementary Table 3 wRF distance summary statistics for the trees inferred using sequences with 85% API.

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Degree of Non-Reversibility: 0			
Mean (SD)	0.460 (0.0575)	0.459 (0.0580)	0.865
Median [Min, Max]	0.455 [0.341, 0.588]	0.451 [0.341, 0.588]	
Degree of Non-Reversibility: 2			
Mean (SD)	0.510 (0.0621)	0.503 (0.0612)	0.396
Median [Min, Max]	0.508 [0.397, 0.663]	0.500 [0.394, 0.651]	
Degree of Non-Reversibility: 4			
Mean (SD)	0.523 (0.0642)	0.511 (0.0620)	0.156
Median [Min, Max]	0.514 [0.412, 0.703]	0.501 [0.401, 0.683]	
Degree of Non-Reversibility: 6			
Mean (SD)	0.529 (0.0652)	0.515 (0.0631)	0.103

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Median [Min, Max]	0.521 [0.424, 0.712]	0.506 [0.412, 0.689]	
Degree of Non-Reversibility: 8			
Mean (SD)	0.532 (0.0659)	0.516 (0.0633)	0.0743
Median [Min, Max]	0.530 [0.413, 0.712]	0.511 [0.406, 0.684]	
Degree of Non-Reversibility: 10			
Mean (SD)	0.535 (0.0653)	0.518 (0.0629)	0.0623
Median [Min, Max]	0.527 [0.418, 0.716]	0.511 [0.410, 0.696]	
Degree of Non-Reversibility: 12			
Mean (SD)	0.537 (0.0652)	0.520 (0.0627)	0.0539
Median [Min, Max]	0.531 [0.427, 0.731]	0.513 [0.416, 0.708]	
Degree of Non-Reversibility: 14			
Mean (SD)	0.538 (0.0651)	0.521 (0.0627)	0.0504
Median [Min, Max]	0.533 [0.431, 0.728]	0.513 [0.421, 0.703]	
Degree of Non-Reversibility: 16			
Mean (SD)	0.539 (0.0645)	0.522 (0.0621)	0.0489
Median [Min, Max]	0.534 [0.431, 0.718]	0.513 [0.422, 0.693]	
Degree of Non-Reversibility: 18			
Mean (SD)	0.540 (0.0646)	0.523 (0.0623)	0.0502
Median [Min, Max]	0.534 [0.434, 0.719]	0.513 [0.423, 0.698]	
Degree of Non-Reversibility: 20			
Mean (SD)	0.541 (0.0642)	0.524 (0.0619)	0.0476
Median [Min, Max]	0.536 [0.438, 0.715]	0.512 [0.428, 0.695]	

Supplementary Table 4 wRF distance summary statistics for the trees inferred using sequences with 90% API.

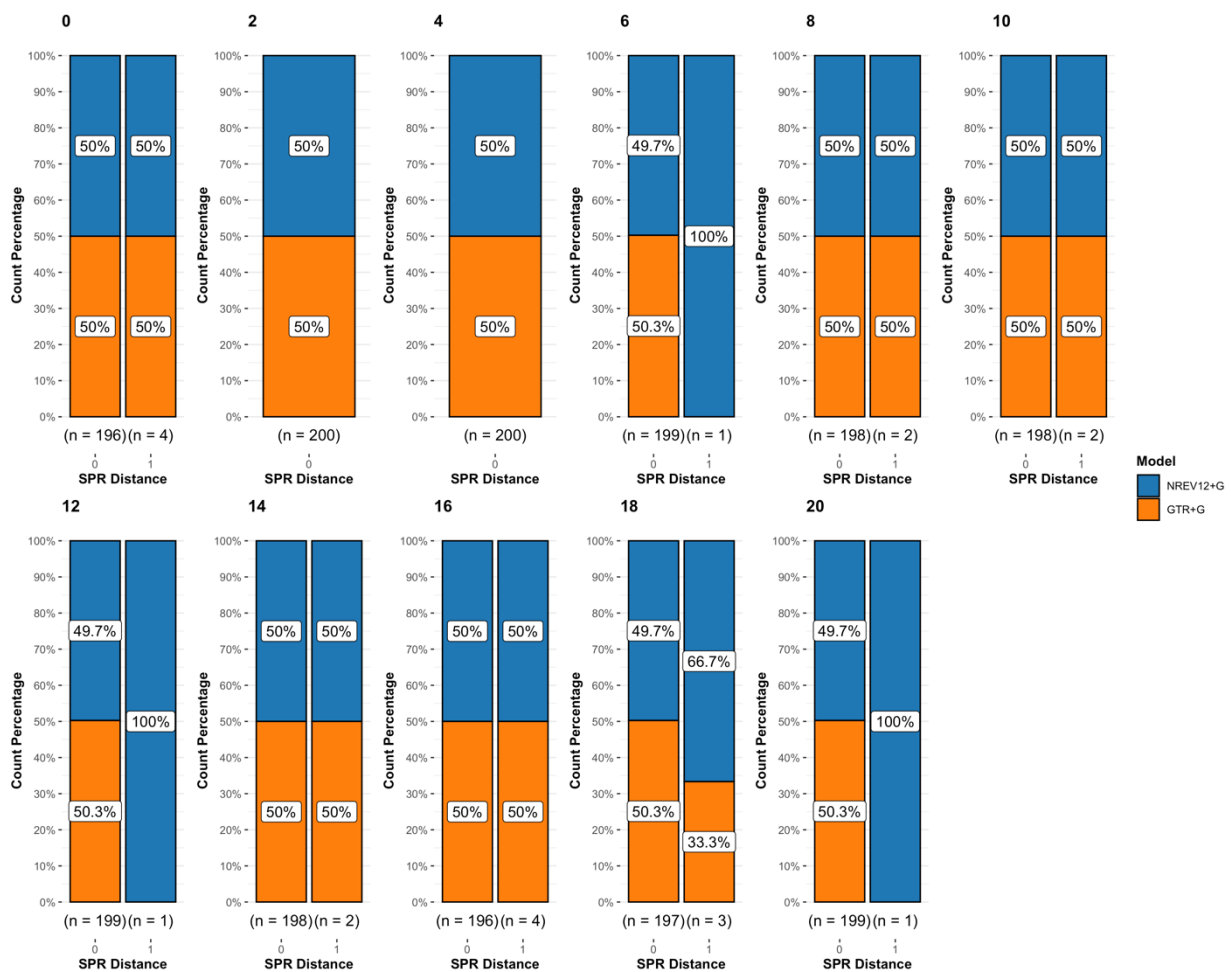
	GTR+G (N=100)	NREV12+G (N=100)	P-value
Degree of Non-Reversibility: 0			
Mean (SD)	0.417 (0.0521)	0.415 (0.0527)	0.834
Median [Min, Max]	0.413 [0.314, 0.541]	0.410 [0.311, 0.541]	

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Degree of Non-Reversibility: 2			
Mean (SD)	0.461 (0.0565)	0.454 (0.0558)	0.338
Median [Min, Max]	0.448 [0.361, 0.599]	0.442 [0.356, 0.587]	
Degree of Non-Reversibility: 4			
Mean (SD)	0.473 (0.0587)	0.461 (0.0568)	0.143
Median [Min, Max]	0.464 [0.375, 0.630]	0.451 [0.366, 0.612]	
Degree of Non-Reversibility: 6			
Mean (SD)	0.478 (0.0609)	0.464 (0.0590)	0.0899
Median [Min, Max]	0.468 [0.384, 0.653]	0.452 [0.374, 0.627]	
Degree of Non-Reversibility: 8			
Mean (SD)	0.480 (0.0614)	0.466 (0.0590)	0.0727
Median [Min, Max]	0.470 [0.380, 0.661]	0.456 [0.367, 0.632]	
Degree of Non-Reversibility: 10			
Mean (SD)	0.482 (0.0606)	0.467 (0.0582)	0.0584
Median [Min, Max]	0.474 [0.377, 0.666]	0.456 [0.363, 0.641]	
Degree of Non-Reversibility: 12			
Mean (SD)	0.484 (0.0608)	0.468 (0.0583)	0.0508
Median [Min, Max]	0.477 [0.382, 0.670]	0.457 [0.369, 0.639]	
Degree of Non-Reversibility: 14			
Mean (SD)	0.485 (0.0597)	0.469 (0.0574)	0.0464
Median [Min, Max]	0.481 [0.382, 0.658]	0.461 [0.369, 0.627]	
Degree of Non-Reversibility: 16			
Mean (SD)	0.486 (0.0592)	0.470 (0.0570)	0.0469
Median [Min, Max]	0.480 [0.388, 0.651]	0.462 [0.374, 0.623]	
Degree of Non-Reversibility: 18			
Mean (SD)	0.486 (0.0585)	0.471 (0.0565)	0.0433
Median [Min, Max]	0.480 [0.387, 0.655]	0.461 [0.373, 0.625]	
Degree of Non-Reversibility: 20			
Mean (SD)	0.487 (0.0587)	0.472 (0.0568)	0.0437
Median [Min, Max]	0.481 [0.389, 0.659]	0.461 [0.374, 0.634]	

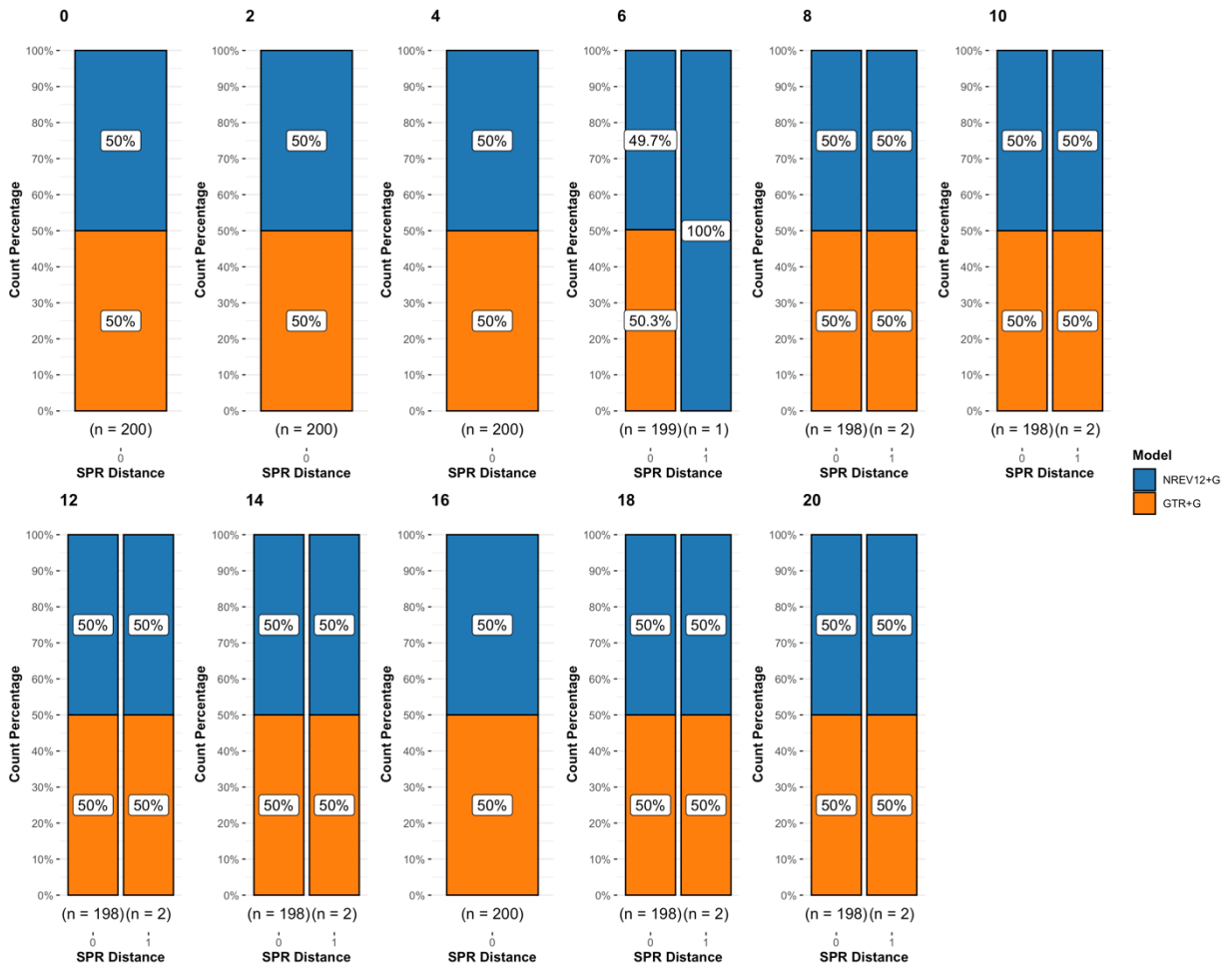
Supplementary Table 5 wRF distance summary statistics for the trees inferred using sequences with 95% API.

	GTR+G (N=100)	NREV12+G (N=100)	P-value
Degree of Non-Reversibility: 0			
Mean (SD)	0.417 (0.0521)	0.415 (0.0527)	0.834
Median [Min, Max]	0.413 [0.314, 0.541]	0.410 [0.311, 0.541]	
Degree of Non-Reversibility: 2			
Mean (SD)	0.461 (0.0565)	0.454 (0.0558)	0.338
Median [Min, Max]	0.448 [0.361, 0.599]	0.442 [0.356, 0.587]	
Degree of Non-Reversibility: 4			
Mean (SD)	0.473 (0.0587)	0.461 (0.0568)	0.143
Median [Min, Max]	0.464 [0.375, 0.630]	0.451 [0.366, 0.612]	
Degree of Non-Reversibility: 6			
Mean (SD)	0.478 (0.0609)	0.464 (0.0590)	0.0899
Median [Min, Max]	0.468 [0.384, 0.653]	0.452 [0.374, 0.627]	
Degree of Non-Reversibility: 8			
Mean (SD)	0.480 (0.0614)	0.466 (0.0590)	0.0727
Median [Min, Max]	0.470 [0.380, 0.661]	0.456 [0.367, 0.632]	
Degree of Non-Reversibility: 10			
Mean (SD)	0.482 (0.0606)	0.467 (0.0582)	0.0584
Median [Min, Max]	0.474 [0.377, 0.666]	0.456 [0.363, 0.641]	
Degree of Non-Reversibility: 12			
Mean (SD)	0.484 (0.0608)	0.468 (0.0583)	0.0508
Median [Min, Max]	0.477 [0.382, 0.670]	0.457 [0.369, 0.639]	
Degree of Non-Reversibility: 14			
Mean (SD)	0.485 (0.0597)	0.469 (0.0574)	0.0464
Median [Min, Max]	0.481 [0.382, 0.658]	0.461 [0.369, 0.627]	
Degree of Non-Reversibility: 16			
Mean (SD)	0.486 (0.0592)	0.470 (0.0570)	0.0469
Median [Min, Max]	0.480 [0.388, 0.651]	0.462 [0.374, 0.623]	
Degree of Non-Reversibility: 18			

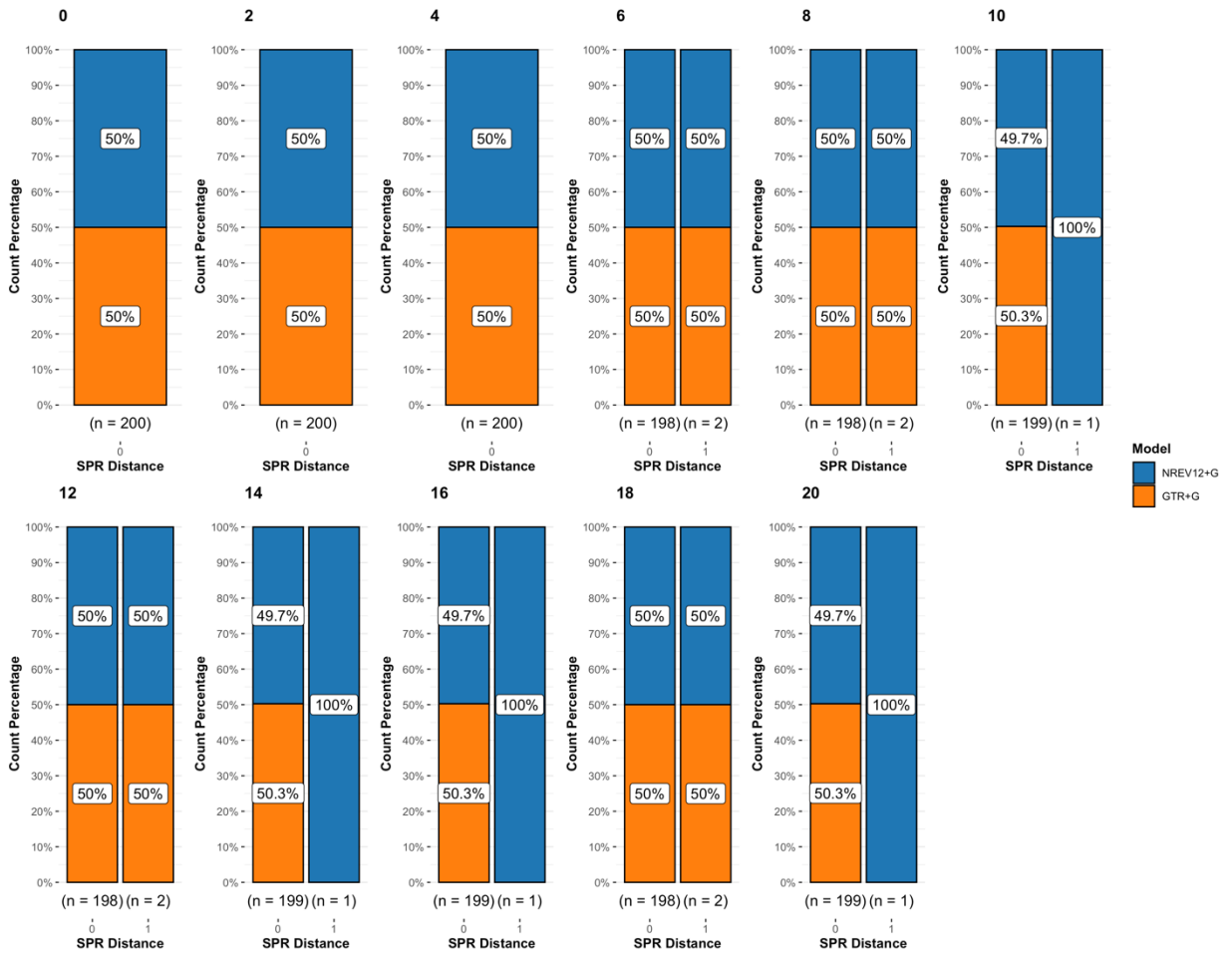
	GTR+G (N=100)	NREV12+G (N=100)	P-value
Mean (SD)	0.486 (0.0585)	0.471 (0.0565)	0.0433
Median [Min, Max]	0.480 [0.387, 0.655]	0.461 [0.373, 0.625]	
Degree of Non-Reversibility: 20			
Mean (SD)	0.487 (0.0587)	0.472 (0.0568)	0.0437
Median [Min, Max]	0.481 [0.389, 0.659]	0.461 [0.374, 0.634]	



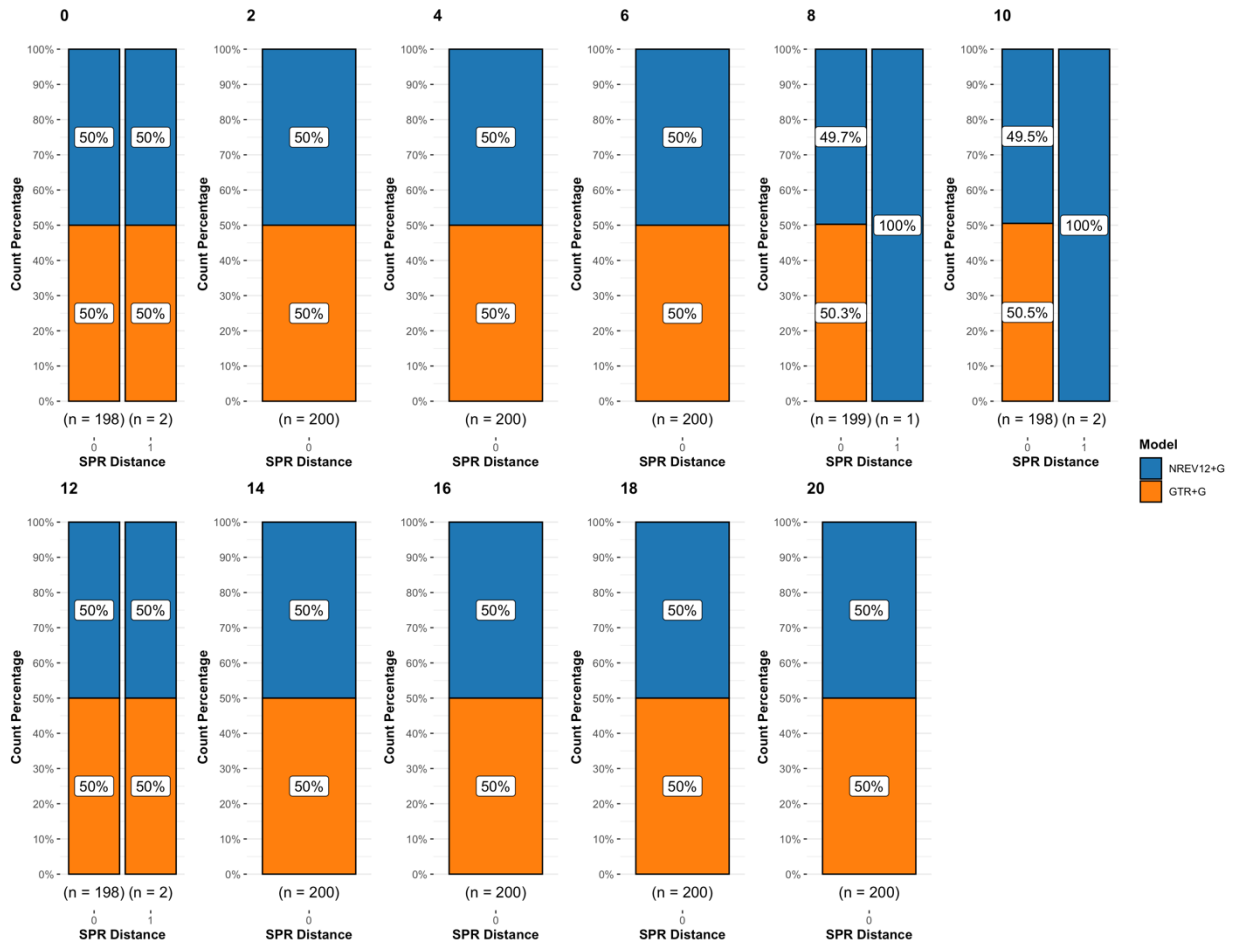
Supplementary Figure 1 SPR distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility for sequences with 75% average pairwise sequence identities.



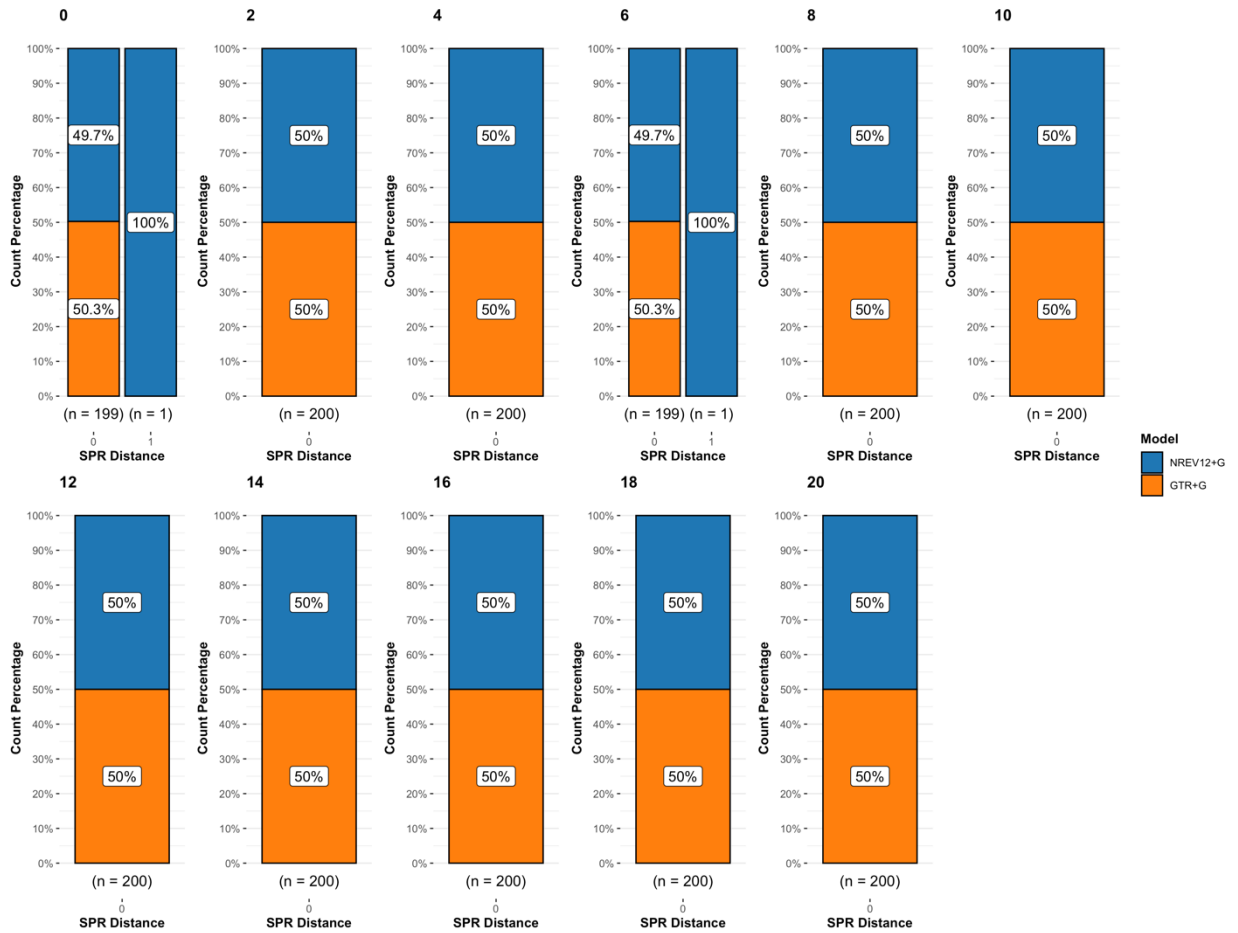
Supplementary Figure 2 SPR distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility for sequences with 80% average pairwise sequence identities.



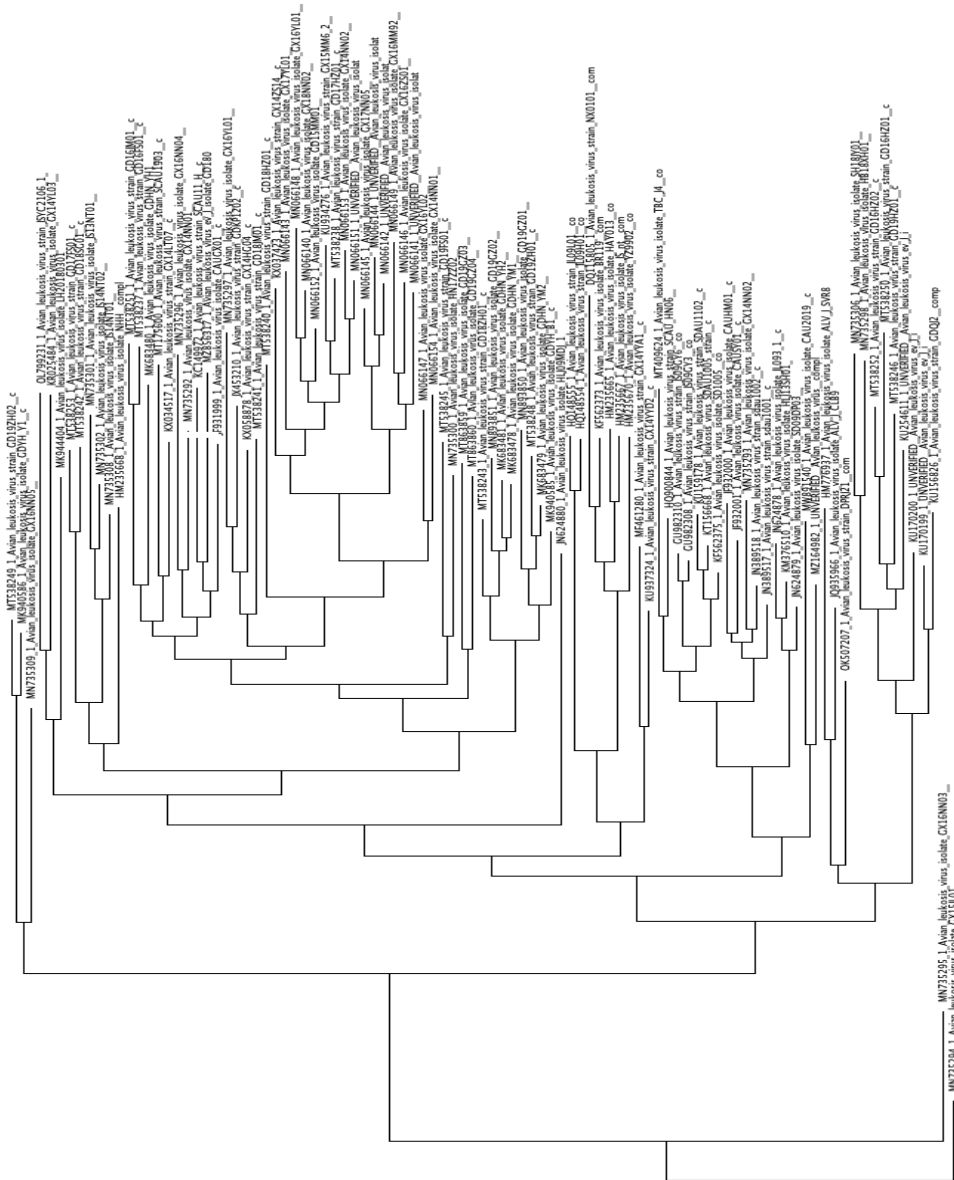
Supplementary Figure 3 SPR distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility for sequences with 85% average pairwise sequence identities.



Supplementary Figure 4 SPR distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility for sequences with 90% average pairwise sequence identities.



Supplementary Figure 5 SPR distances between inferred and true phylogenetic trees for datasets simulated with different degrees of nucleotide substitution non-reversibility for sequences with 95% average pairwise sequence identities.



0.01

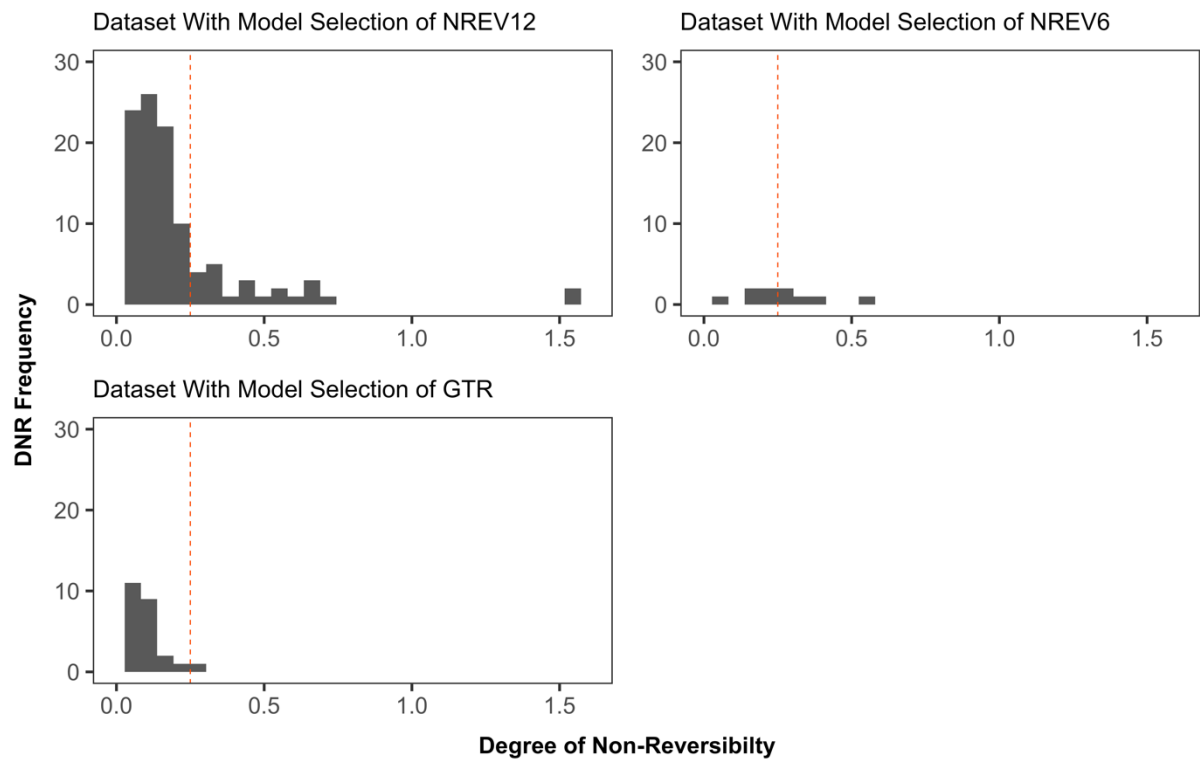


Figure 13 DNR values for data sets under each respective model. The dotted red line indicating 0.25 DNR threshold.