THE UNIVERSITY
of EDINBURGH

The ecology of emerging diseases: virulence and transmissibility of human RNA viruses

Liam Brierley

Thesis submitted for the degree of Doctor of Philosophy
University of Edinburgh
2016

Declaration

All data analyses described herein and the written presentation of this thesis are my own work. Data collection described herein was primarily carried out by myself, with supporting data collection carried out by project students and summer interns (Conor O'Halloran, Claire Taylor, Christopher McCaffery, David McCulloch), in all cases following protocols authored in whole or in part by myself. Otherwise, datasets directly obtained from external sources are clearly attributed. No part of the work herein has been submitted for consideration for any other degree or professional qualification.

Liam Brierley, 2016

Acknowledgements

Firstly, I wish to extend my sincere thanks to my supervisors, Mark Woolhouse and Amy Pedersen, for their expertise and guidance that have culminated in this thesis and helped me to develop as an independent researcher during my PhD, but most of all, for always believing in me.

I must also thank Kevin Olival, Peter Dazsak, Parviez Hosseini, Carlos Zambrana-Torrelio and their many associates at the EcoHealth Alliance, for providing data on viral infections of mammals, and for always approaching our collaborative work with a friendly and supportive demeanour.

I also thank Epigroup and Pedersen lab group members past and present, for their support through assistance in data collection, constructive comments, and helpful discussions. I particularly learned much thanks to Luke McNally, who was always willing to provide extensive technical help and valuable advice. I am very grateful to the NERC and the Wellcome Trust, for the funding that made the work within this thesis possible.

Thank you to Tom G, Christine, Becky, Amy, Doris, Nora, and Eileen, not only for the pleasure of sharing our working life, but for their openness, supportiveness, and solidarity. I would also like to thank Fiona, Jacquie, Kate, and Tasim, without whom this journey would never have been possible. In particular, I am especially thankful to Tom, for always being there for me when no-one else was.

Finally, I would like to dedicate this thesis to the memory of my grandfather, Eric Crowther, the man who taught me the value of hard work. I love you and miss you very much.

Abstract

Emerging infectious diseases continue to represent serious threats to global human health. Novel zoonotic pathogens are continually being recognised, and some ultimately cause significant disease burdens and extensive epidemics. Research and public health initiatives often face emerging pathogens with limited knowledge and resources. Inferences from empirical modelling have begun to uncover the factors determining cross-species transmission and emergence in humans, and subsequently guide risk assessments. However, the dynamics of virulence and transmissibility during the process of emergence are not well understood. Here, I focus on RNA viruses, a priority pathogen type because of their potential for rapid evolution. I use comparative trait-based analyses to investigate how aspects of both host and virus ecology contribute to the risk of virulence and transmissibility within human RNA viruses. To explore these questions, data were collected via systematic literature search protocols.

In the first half of this thesis, I focus on viral determinants of virulence and transmissibility. I ask whether virulence can be predicted by viral traits of tissue tropism, transmission route, transmissibility and taxonomic classification. Using a machine learning approach, the most prominent predictors of severe virulence were breadth of tissue tropism, and nonvector-borne transmission routes. When applied to newly reported viruses as test set, the final model predicted disease severity with 87% accuracy.

Next, I assess support for hypothesised routes of adaptation during emergence using phylogenetic state-switching models. Propensity for adaptation in small 'stepwise' movements versus large 'off-the-shelf' jumps differed between virus taxa, though no single route dominated, suggesting multiple independent trajectories of adaptation to human hosts. In addition, phylogenetic regressions showed vector and respiratory-transmitted viruses to be more likely to progress through early stages of emergence.

In the second half of this thesis, I focus on how dynamics of virulence and transmissibility differ with respect to nonhuman host diversity, identity, and ecology. Using a regression framework, I observe that viruses with a broader mammalian host range exhibited higher risk of severe virulence, but lower risk of transmissibility, which may reflect potential trade-offs of host specificity. Furthermore, viruses with artiodactyl hosts exhibited lower risk of severe virulence and viruses with bat or nonhuman primate hosts exhibited higher risk of transmissibility.

Next, I test hypotheses that mammal species with faster-paced life history may be predisposed to host viruses with greater virulence and transmissibility. Mammal body mass was used as an established proxy for pace of life history. In regression analyses, mammals with faster-paced life history hosted more viruses with severe virulence, though evidence for a relationship with transmissibility was limited.

The broad-scale associations presented in this thesis suggest the evolution of virulence and human-to-human transmissibility during zoonotic emergence is a multifactorial, highly dynamic process influenced by both virus and host ecology. Despite this, general characteristics of high-risk emerging viruses are evident. For example, severe virulence was associated with broad niche diversity of both tissue tropisms at the within-host scale, and host species at the macroecological scale. However, risk factors for virulence and human-to-human transmissibility often did not coincide, which may imply an overarching trade-off between these traits. These analyses can contribute to preparedness and direction within public health strategies by identifying likely candidates for high-impact emergence events among previously known and newly discovered human viruses. The inherent connectivity between RNA viruses, their nonhuman hosts and the resulting implications for human health emphasise the holistic nature of emerging diseases and supports the One Health perspective for infectious disease research.

Lay summary

One of the biggest priorities for global health programmes is to understand when and where new infectious diseases are likely to emerge. The recent emergence of RNA viruses such as SARS-related coronavirus, Ebola virus, and H5N1 influenza virus have all caused significant costs to human health as well as economies. A clear need to predict and prepare for virus emergence is now widely acknowledged. Although models have improved our understanding of the process of virus emergence from animal to human hosts, little is known about what determines an emerging virus' ability to cause severe disease or transmit efficiently between humans. Using an array of statistical models, this thesis aims to identify characteristics of RNA viruses, their hosts, and their wider ecology that increase the risk of severe disease and transmissibility in humans.

Among virus characteristics, the strongest predictors of severe disease were the ability to infect a broad range of cell types, and transmission via respiratory routes, faecal-oral routes, or direct contact. My model correctly predicted severe disease among a group of newly reported viruses with 87% accuracy. I also found that viruses do not adapt to humans exclusively via small evolutionary steps or large evolutionary jumps as previously predicted. Instead, I observed a mixture between these types of adaptation, dependent on taxonomic classification. Finally, viruses with transmission via respiratory routes, or insect or tick vectors were more likely to infect and transmit between humans.

Among host characteristics, viruses infecting a greater diversity of mammal species were more likely to cause severe disease in humans, but less likely to transmit between humans. Ability to cause severe disease or transmit between humans was also dependent on which specific mammal taxonomic orders were infected. Furthermore, I found that mammal species hosted more viruses causing severe disease in humans if they had a 'faster' life strategy with shorter lifespan and rapid reproduction, as indicated by a smaller body mass.

The characteristics I find to be associated with risk of severe disease and transmissibility in humans suggest ways to improve our understanding of how emerging viruses evolve. For example, I often found severe disease and transmissibility to be predicted by different characteristics, suggesting an antagonistic relationship where evolution towards one may inhibit evolution towards the other. This thesis also demonstrates the critical need for future research to consider viruses in the context of their wider ecology. The findings of this thesis can contribute to identifying viruses with high potential risks for public health and ultimately, guide a more preventive global health strategy.

Contents

Declaration ..... **i**
Acknowledgements ..... **ii**
Abstract ..... **iii**
Lay summary ..... **v**
Contents ..... **vii**
List of abbreviations ..... **xi**
List of figures ..... **xii**
List of tables ..... **xiv**
Chapter 1. General introduction **1**
1.1. The global status of emerging RNA viruses 1
1.2. The process of RNA virus emergence 4
1.3. Understanding the determinants of RNA virus emergence 7
1.4. Predicting traits of RNA viruses beyond emergence 10
1.5. Thesis aims & outline 11
Chapter 2. Tropism breadth and transmission ecology predict virulence of emerging human RNA viruses **14**
2.1. Abstract 14
2.2. Introduction 15
2.3. Materials and methods 17
2.3.1. Data collection 17
2.3.2. Classification analysis 21
2.4. Results 23
2.4.1. Virulence of human RNA viruses 23
2.4.2. Classification tree risk factor analysis 24
2.4.3. Ascertainment bias and risk factor temporality 27
2.4.4. Predicted virulence of newly reported viruses 28
2.5. Discussion 30
2.5.1. Ecology and evolution of risk factor traits 31
2.5.2. Ascertainment biases 33
2.5.3. Predictive power for newly reported viruses 34
2.5.4. Analytical limitations 35

2.5.5. Implications for public health....36
2.6. Conclusion....38
Chapter 3. Evolutionary routes of RNA virus emergence and human adaptation 39
3.1. Abstract....39
3.2. Introduction....40
3.3. Materials and methods....46
3.3.1. Pathogen Pyramid level data....46
3.3.2. Virus phylogenies and cladogram....48
3.3.3. State-switching modelling analysis....50
3.3.4. Phylogenetic comparative analysis....53
3.4. Results....54
3.4.1. State-switching models of RNA virus adaptation....54
3.4.2. State-switching models across an RNA virus cladogram....58
3.4.3. Viral trait associations with adaptation....59
3.5. Discussion....61
3.5.1. Sources of human-adapted viruses....64
3.5.2. Evolutionary scope of analysis....66
3.5.3. Analytical limitations....68
3.5.4. Wider implications....70
3.6. Conclusion....71
Chapter 4. Breadth and specificity of mammal host range predict dynamics of RNA virus emergence in humans....72
4.1. Abstract....72
4.2. Introduction....73
4.3. Materials and methods....76
4.3.1. RNA virus trait and mammal host data....76
4.3.2. Host range calculations....78
4.3.3. Statistical analysis....79
4.4. Results....82
4.4.1. Matrices of host-virus relationships....82
4.4.2. Mixed regression analyses of host range....84
4.5. Discussion....91
4.5.1. Analytical limitations....96
4.5.2. Wider implications....97
4.6. Conclusion....98

Chapter 5. Allometry of mammal species predicts ability to host virulent human RNA viruses 99
5.1. Abstract 99
5.2. Introduction 100
5.3. Materials and methods 102
5.3.1. RNA virus and mammal host data 102
5.3.2. Statistical analysis 105
5.4. Results 106
5.5. Discussion 112
5.5.1. Analytical limitations 114
5.5.2. Wider implications 116
5.6. Conclusion 116
Chapter 6. Conclusion 118
6.1. Outline 118
6.2. Insights into virus ecology and evolution 119
6.2.1. Virulence and niche diversity 121
6.2.2. Antagonistic pleiotropy and constraints of human adaptation 122
6.2.3. Evolutionary trade-offs in virulence and transmissibility 124
6.3. Key areas for further study 126
6.3.1. Knowledge and data gaps 126
6.3.2. Study directions and methods 128
6.4. Implications for global health 130
6.5. Concluding remarks 133
References 134
Appendix A. Supplementary material for: Tropism breadth and transmission ecology predict virulence of emerging human RNA viruses 155
A.1. Supplementary methods 155
A.1.1. Bayesian mixed regression model analysis 155
A.2. Supplementary figures 156
A.3. Supplementary tables 163
Appendix B. Supplementary material for: Evolutionary routes of RNA virus emergence and human adaptation 176
B.1. Supplementary figures 176
B.2. Supplementary tables 179

Appendix C. Supplementary material for: Breadth and specificity of mammal host range predict dynamics of RNA virus emergence in humans 186
C.1. Supplementary methods 186
C.1.1. Human specialist sensitivity reanalysis 186
C.2. Supplementary figures 187
C.3. Supplementary tables 192
Appendix D. Supplementary material for: Allometry of mammal species predicts ability to host virulent human RNA viruses 198
D.1. Supplementary methods 198
D.1.1. Virus-centric mixed regression model analysis 198
D.2. Supplementary figures 200
D.3. Supplementary tables 205
Appendix E. Publication: RNA viruses: a case study of the biology of emerging infectious diseases 207
Appendix F. Publication: Assessing the epidemic potential of RNA and DNA viruses 219

List of abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired immune deficiency syndrome
BF	Bayes Factor
CDC	Centers for Disease Control and Prevention
CFR	Case fatality ratio
CI	Confidence interval
CpG	Cytosine-guanine dinucleotide; with cytosine specifically preceding guanine in a nucleotide sequence in the 5' → 3' direction
DNA	Deoxyribonucleic acid; usually in reference to viral genomic material
dsRNA	Double-stranded RNA
EID	Emerging infectious disease
GC	Guanine-cytosine
HFRS	Hantavirus haemorrhagic fever with renal syndrome
HIV	Human immunodeficiency virus
HM	Harmonic mean
HPS	Hantavirus pulmonary syndrome
HTLV	Human T-lymphotropic virus; synonym of Primate T-lymphotropic virus
ICTV	International Committee on Taxonomy of Viruses
LRT	Likelihood ratio test
MCMC	Markov chain Monte Carlo
MERS	Middle East respiratory syndrome
OR	Odds ratio
R_0	Basic reproductive number
RJ-MCMC	Reversible-jump Markov chain Monte Carlo
RNA	Ribonucleic acid; usually in reference to viral genomic material
SARS	Severe acute respiratory syndrome
SIV	Simian immunodeficiency virus
-ssRNA	Single-stranded negative-sense RNA
+ssRNA	Single-stranded positive-sense RNA
ssRNA-RT	Single-stranded reverse-transcribing RNA
TSS	True Skill Statistic; defined as sensitivity + specificity - 1
USAID	United States Agency for International Development
VIF	Variance inflation factor
VIZIONS	Vietnam Initiative on Zoonotic Infections
WHO	World Health Organisation

List of figures

1.1. The 'Pathogen Pyramid' model 5
2.1. Virulence of human RNA viruses, with respect to taxonomy and discovery ... 24
2.2. Final pruned classification tree predicting disease severity for 178 human RNA viruses 25
2.3. Full and jack-knifed tree accuracies across different predictor sets 27
2.4. Informativeness and discovery dynamics with respect to virulence risk factors over time 29
2.5. Application of the final classification tree to 30 newly reported viruses 30
3.1. Potential models of RNA virus adaptation in humans 42
3.2. Example phylogenies under each model of RNA virus adaptation 45
3.3. Taxonomically-structured cladogram of RNA virus species infecting mammals or birds 49
3.4. Consensus RJ-MCMC state-switching models of Pathogen Pyramid levels across viral phylogenies 56
3.5. Consensus RJ-MCMC state-switching model of Pathogen Pyramid levels across taxonomically-structured cladogram 59
4.1. Heatmap of virus species in each viral family known to infect each mammal order 83
4.2. Fitted effects of mammal host range variables in final logistic mixed regression models 85
5.1. Relationship between adult body mass of mammal species and proportion of zoonotic RNA viruses causing 'severe' human disease 108
5.2. Odds ratios describing effect of mammal adult body mass from logistic mixed regression models for data subgroups 109
5.3. Relationship between mammal taxonomic order and proportion of zoonotic RNA viruses capable of human-to-human transmission 111
6.1. Simplified concept map of summarised thesis findings 120
A.1 Classification tree with transmission route variables specified using multiple categories 156
A.2. Classification tree excluding viruses with data quality problems or high data uncertainty 157
A.3. Classification tree excluding viruses only known to infect humans from serological evidence 158

A.4. Full and jack-knifed tree True Skill Statistic values across different predictor sets 159
A.5. Full and jack-knifed tree sensitivities across different predictor sets 160
A.6. Full and jack-knifed tree Negative Predictive Values across different predictor sets 161
A.7. Full and jack-knifed tree relative accuracies across different predictor sets . 162
B.1. Graphical summary of RJ-MCMC state-switching model configurations across virus phylogenies 176
B.2. Graphical summary of RJ-MCMC state-switching model configurations across taxonomically-structured cladogram 177
B.3. Posterior density plots from multinomial phylogenetic mixed regression.... 178
C.1. Histograms of mammal hosts per virus and viruses per mammal host 187
C.2. Heatmap of human-infective virus species in each viral family known to infect each mammal order 188
C.3. Heatmap of human-transmissible virus species in each viral family known to infect each mammal order 189
C.4. Heatmap of sustained human-transmissible virus species in each viral family known to infect each mammal order 190
C.5. Heatmap of virus species causing severe human disease in each viral family known to infect each mammal order 191
D.1. Histograms of zoonotic viruses, viruses causing severe human disease, and human-transmissible viruses per mammal host 200
D.2. Relationship between adult body mass of mammal species and proportion of zoonotic RNA viruses causing 'severe' human disease in a single panel 201
D.3. Relationship between adult body mass of mammal species and proportion of zoonotic RNA viruses causing 'severe' human disease with highlighted influential viruses 202
D.4. Relationship between adult body mass of mammal species and proportion of zoonotic RNA viruses capable of human-to-human transmission, excluding iatrogenic transmission 203
D.5. Relationship between median adult body mass of mammal hosts of zoonotic RNA viruses and risk of causing 'severe' human disease 204

List of tables

2.1. Virus trait data collected for use in classification tree analysis 19

3.1. Posterior support for hypothesised models of virus adaptation across viral phylogenies 55

3.2. Consensus models from RJ-MCMC analyses across viral phylogenies 57

3.3. Posterior support for hypothesised models of virus adaptation across taxonomically-structured cladogram 59

3.4. Consensus models from RJ-MCMC analyses across taxonomically-structured cladogram 60

3.5. Posterior means from MCMC multinomial phylogenetic regression predicting Pathogen Pyramid levels 61

4.1. Outputs from logistic mixed regression predicting human infectivity among mammalian RNA viruses 86

4.2. Outputs from logistic mixed regression predicting human-to-human transmissibility among mammalian RNA viruses 86

4.3. Outputs from logistic mixed regression predicting sustained human-to-human transmissibility among mammalian RNA viruses 88

4.4 Outputs from logistic mixed regression predicting severe disease among human RNA viruses 88

4.5. Fitted random intercepts for each RNA virus family from logistic mixed regressions predicting viral traits 90

5.1. Outputs from logistic mixed regression predicting proportion of zoonotic RNA viruses within mammal species causing severe human disease 107

5.2. Outputs from logistic mixed regression predicting proportion of zoonotic RNA viruses within mammal species capable of human-to-human transmission 110

A.1. Virulence rating data for 180 human RNA virus species 163

A.2. Virulence rating data for 30 newly reported human RNA viruses 171

A.3. Six-rank system of classifying human RNA virus virulence 172

A.4. Diagnostics of final pruned classification trees using alternative definitions of virulence 174

A.5. Posterior means from MCMC mixed logistic regression predicting disease severity 175

B.1. Model configurations with at least 5% frequency in RJ-MCMC analysis across *Picornaviridae* family phylogeny 179
B.2. Model configurations with at least 5% frequency in RJ-MCMC analysis across *Rhabdoviridae* family phylogeny 180
B.3. Model configurations with at least 5% frequency in RJ-MCMC analysis across *Paramyxoviridae* family phylogeny 181
B.4. Model configurations with at least 5% frequency in RJ-MCMC analysis across taxonomically-structured cladogram under branch length assumption set (a) ... 182
B.5. Model configurations with at least 5% frequency in RJ-MCMC analysis across taxonomically-structured cladogram under branch length assumption set (b) ... 183
B.6. Model configurations with at least 5% frequency in RJ-MCMC analysis across taxonomically-structured cladogram under branch length assumption set (c).... 184
B.7. Model configurations with at least 5% frequency in RJ-MCMC analysis across taxonomically-structured cladogram under branch length assumption set (d) ... 185
C.1. Outputs from logistic mixed regressions predicting human infectivity among mammalian RNA viruses using different host range metrics..... 192
C.2. Outputs from logistic mixed regressions predicting human-to-human transmissibility among mammalian RNA viruses using different host range metrics..... 193
C.3. Outputs from logistic mixed regressions predicting sustained human-to-human transmissibility among mammalian RNA viruses using different host range metrics..... 194
C.4. Outputs from logistic mixed regressions predicting severe disease among human RNA viruses 195
C.5. Outputs from likelihood ratio tests when adding random slopes upon host range metrics with respect to viral family..... 196
C.6. Outputs from logistic mixed regression predicting human-to-human transmissibility among mammalian RNA viruses, excluding human-specialist viruses 197
C.7. Outputs from logistic mixed regression predicting sustained human-to-human transmissibility among mammalian RNA viruses, excluding human-specialist viruses..... 197
D.1. Outputs from virus-centric logistic mixed regressions predicting severe disease among zoonotic RNA viruses for complete and restricted datasets..... 205
D.2. Outputs from virus-centric logistic mixed regressions predicting human-to-human transmissibility among zoonotic RNA viruses for complete and restricted datasets..... 206

Chapter 1. General introduction

1.1. The global status of emerging RNA viruses

"It is time to close the book on infectious diseases," was the pronouncement believed to have been declared in the late 1960s by the United States Surgeon General, Dr. William H. Stewart. Although now discredited as a misquote (Spellberg and Taylor-Blake 2013), this statement nevertheless captures an important scientific perspective held by some at the time. However, nearing fifty years later, infectious diseases remain a serious global challenge and a major cause of human morbidity and mortality, with developing nations disproportionately bearing the greatest disease burdens (Mathers et al. 2008; GBD 2013 Mortality and Causes of Death Collaborators 2015). Infectious diseases also cause significant economic burdens (Fonkwo 2008); the total cost of seasonal epidemic influenza in the USA alone is estimated to be $87.1 billion US dollars annually (Molinari et al. 2007).

A primary reason the proverbial "book" has remained open and is likely to do so for the foreseeable future is the continual appearance and impact of emerging infectious diseases (EIDs) and their causative pathogens. To clarify, I use the term "pathogen" to refer to any infectious agent, including all microparasites and macroparasites, regardless of their level of pathogenicity or fitness costs to hosts, if any. In this thesis, I focus on RNA viruses as a key type of pathogen among emerging infectious diseases and the cause of multiple human pandemics with exceptionally high impacts on public health, e.g. SARS coronavirus, HIV, pandemic influenza viruses. The World Health Organisation recently declared a list of seven priority emerging pathogens posing the greatest immediate risk to global health (WHO 2015), all of which are RNA viruses.

The term 'emerging' is often used to refer to a range of epidemiological dynamics, for which there are recent and globally significant examples of human viruses for each. Primarily, emerging viruses are considered to be those not

previously recognised as human pathogens that are newly transmitted from non-human animal sources, referred to as 'zoonotic' transmission (Morse 1995; Jones et al. 2008), e.g. MERS coronavirus, which was first recognised as a human infection in the Middle East in 2012 (Zaki et al. 2012), where outbreaks have caused over 600 fatalities to date (WHO 2016). Additionally, those viruses recurrently appearing after their initial outbreaks have subsided may also be considered emerging, or 're-emerging'. For example, Zaire ebolavirus initially emerged in the 1970s and caused several outbreaks in the 1990s and early 2000s (Groseth et al. 2007), before re-emerging in West Africa in 2013, resulting in the largest and most geographically widespread Ebola virus disease outbreak yet (Spengler et al. 2016). Finally, emerging may also refer to those viruses significantly expanding in geographic range or incidence (Morse 1995), which often involves epidemics within a novel context following release from endemic patterns e.g. the *Aedes* mosquito-borne Zika virus was previously endemic to Sub-Saharan Africa, but has spread to the Yap Islands of the Pacific and to South and Central America, where outbreaks are ongoing and have potentially reached totals of several million cases (Gatherer and Kohl 2015; Weaver et al. 2016). A more mathematical definition of emergence based on the rate of change in incidence has been advocated (Funk et al. 2013). However, incidence data is deficient for many emerging viruses, therefore I use the term 'emerging' as inclusive of all types of dynamics discussed above, except where otherwise stated.

RNA viruses are noted as a pressing priority among all types of emerging pathogen for several reasons. Although some appear to be ancient human pathogens, e.g. rabies virus (Steele and Fernandez 1991), RNA viruses are acknowledged as posing a particularly high risk of emergence, as they are consistently overrepresented amongst emerging and zoonotic pathogens (Cleaveland et al. 2001; Taylor et al. 2001; Woolhouse and Gowtage-Sequeria 2005; Woolhouse and Gaunt 2007). This risk is thought to be partly due to their high evolutionary lability, as RNA viruses lack genetic proofreading mechanisms during replication (Belshaw et al. 2008; Parrish et al. 2008). However, it must be acknowledged that the range of substitution rates

across RNA viruses covers several orders of magnitude and overlaps with some DNA viruses (Holmes and Drummond 2007; Duffy et al. 2008). A continuous rate of human RNA virus emergence has been observed, such that discovery curves estimate approximately 2 new human virus species/year are recognised (among pooled counts of RNA and DNA viruses) (Woolhouse et al. 2008, 2012). The number of viral emergence 'events' has also increased between successive decades (Jones et al. 2008), though a fraction of these reflect novel discoveries of much older, existing human viruses, through increasing sensitivity of viral detection methods (Delwart 2013).

The diversity of human RNA viruses is vast, comprising a great variety of genetic and ecological characteristics. Although all RNA viruses rely on host cell translation machinery to replicate, their genomes vary in structure, which forms one basis of their classification: either single-stranded positive-sense (hereafter '+ssRNA'; also referred to as Group IV), single-stranded negative-sense ('-ssRNA'; Group V), double-stranded ('dsRNA'; Group III), or single-stranded reverse-transcribing ('ssRNA-RT'; Group VI). Within these four genome types, there are 48 recognised taxonomic families of RNA viruses (King et al. 2011). Among these, 19 families are known to infect mammals, with 17 of these 19 containing at least one human-infective virus species (Woolhouse et al. 2013); only the *Arteriviridae* and *Nodaviridae* have not yet been reported to infect humans. The majority of emerging human viruses are ultimately classified under existing genera and families, though the possibility of discovering novel human-infective RNA virus families remains credible (Woolhouse et al. 2008, 2012). Note that throughout the thesis, I follow standardised guidelines for formatting of virus nomenclature (King et al. 2011), which distinguishes a conceptual species (e.g. "*Rabies virus* was classified as a zoonotic species") from a virus in its physical context (e.g. "rabies virus was isolated from a human sample").

1.2. The process of RNA virus emergence

In order to identify their underlying causes and predict their consequences, there is a clear need to understand the origins and processes involved in the emergence of RNA viruses. The majority of human viruses have their immediate origins in non-human animals (Wolfe et al. 2007), and several viruses known only to infect humans are traceable to evolutionary divergence from animal viruses, such as HIV-1 and HIV-2 from primate SIVs (Hahn et al. 2000). Intuitively, it follows that zoonotic status is a major risk factor for pathogens to emerge, particularly among viruses (Taylor et al. 2001). Human viruses appear to be almost exclusively shared with other warm-blooded vertebrate hosts (not considering arthropod vectors as hosts), and predominantly with other mammals over birds (Woolhouse and Gaunt 2007; Woolhouse et al. 2012; Woolhouse and Adair 2013).

The process of emergence from animal to human pathogen (and by extension, host shifts from any one host species to another) has been conceptually modelled in a number of schemata that all demonstrate the same core pathway (Childs et al. 2007; Wolfe et al. 2007; Lloyd-Smith et al. 2009; Woolhouse et al. 2012). Throughout this thesis I make reference to one such system in detail, the Pathogen Pyramid model (Woolhouse and Gaunt 2007; Woolhouse et al. 2012) (Figure 1.1).

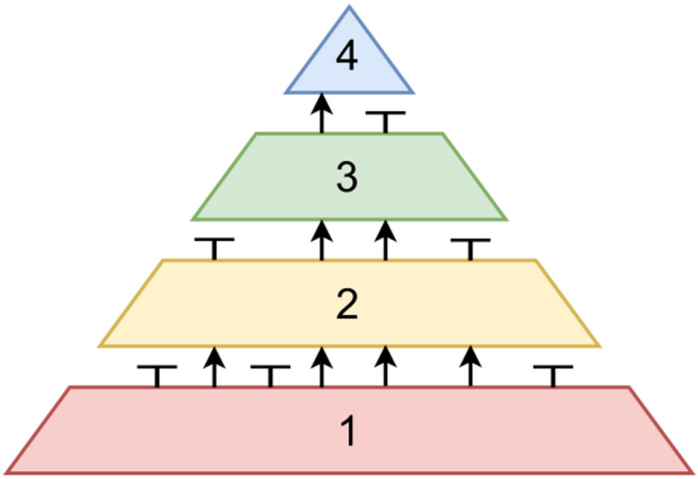


Figure 1.1. The 'Pathogen Pyramid' model, adapted from Woolhouse and Gaunt 2007, showing conceptual levels of pathogens during emergence (see main text for definitions). Arrows illustrate movements between levels, where blocked arrows illustrate presence of a genetic or ecological barrier preventing movement. Note that movements between levels may not necessarily occur in gradual steps and may involve 'jumps' to much higher levels (see Chapter 3). The pyramid shape results from increasingly fewer pathogens fitting the criteria for each additional level.

Within this model, level 1 describes pathogens that are limited to nonhuman hosts only, where no evidence of human infection has been recognised, even if humans are routinely exposed. Level 2 describes pathogens that are capable of infecting humans, but incapable of human-to-human transmission. All human cases of level 2 viruses are therefore the result of zoonotic transmission, e.g. most hantaviruses such as Hantaan, Seoul and Sin Nombre viruses, which transmit from rodents to humans via aerosolised excreta. Level 3 describes human pathogens that are only capable of self-limited human-to-human transmission. This includes both rare single transmission events and short 'stuttering' transmission chains, and can be thought of in terms of $0 < R_0 < 1$, where R_0 denotes the basic reproductive number, i.e. the mean number of secondary cases expected to result from a single primary case, assuming the population is entirely susceptible. An example is Nipah virus, where cases typically result from zoonotic contact with bats or pigs, though short chains of transmission between patients or healthcare workers may occur in clinical

settings. Level 4 describes human pathogens that are capable of long chains of sustained human-to-human transmission, which may include a variety of epidemiological patterns such as epidemic cycling or constant endemic transmission, and can be thought of in terms of $R_0 \geq 1$. An example is the epidemic transmission of seasonal influenza virus subtypes. Some schemata distinguish level 4 pathogens that still infect or are maintained within nonhuman animals (sometimes referred to as level '4a') from those with humans as their only known host species (sometimes referred to as level '4b' or '5') (Wolfe et al. 2007), with the expectation these follow different epidemiological dynamics in humans.

The Pathogen Pyramid model is so shaped to illustrate that the number of pathogens successfully meeting the criteria for each subsequent level decreases (Figure 1.1), as a result of failures to overcome molecular barriers within hosts (e.g. cell receptor availability) or ecological barriers between hosts (e.g. external survivability) (Kuiken et al. 2006). Overcoming such barriers often requires genetic adaptation. The likelihood of adaptation and ultimate emergence depends on both the rate of viral evolution and the adaptive distance to be traversed (Antia et al. 2003; Kuiken et al. 2006; Parrish et al. 2008) (and additionally, but beyond the scope of this schema, any antagonistic coevolution of the host, see Daugherty and Malik 2012). Those necessary traits to overcome barriers of adaptation may be acquired by selection within human hosts or the occurrence of a suitable genotype by chance within animal hosts (Parrish et al. 2008; Pepin et al. 2010). Phenotypically, emergence can therefore appear erratic and unpredictable, where some viruses progress through Pathogen Pyramid levels gradually over long timescales and others seemingly jump straight from level 1 to level 4. The transition between level 2 and level 3 appears subject to particular evolutionary constraints, with little empirically observed evidence this occurs directly (Woolhouse et al. 2016). Instead, it is likely that most level 3 viruses develop from independent introductions from the zoonotic source. However, considering opportunity for mutations to arise and accumulate, level 2 to level 3 progression may be more feasible for viruses causing chronic human

infection, such as the *Retroviridae*. Although they have begun to be formally described in the above manner by quantitative reviews (Woolhouse et al. 2016), these different types of adaptive movements during RNA virus emergence are not well-studied.

1.3. Understanding the determinants of RNA virus emergence

Based on the conceptual models describing how RNA viruses emerge, much research has been devoted to understanding the subsequent question of why they emerge. Many emerging viruses have traditionally been investigated as in-depth case studies to fully map out the factors surrounding their emergence, e.g. Nipah virus emergence in humans was traced to indirect transmission from pigs, and in turn, pig infection was traced back to exposure to wild bats through shared feeding around date palm trees (Chua et al. 2000, 2003). Such highly localised investigations have contributed much to the understanding of specific disease systems.

However, this approach is retrospective, i.e. it can only take place after human infections have already occurred. A more predictive approach for public health has been advocated (Daszak 2009; Morse et al. 2012), involving studies of which types of viruses are most likely to emerge, and under what conditions. Several frameworks have been proposed for this purpose that integrate different scales of study from individuals to global regions (Wilcox and Colwell 2005; Wood et al. 2012; Johnson et al. 2015b). These frameworks are also increasingly broadening in scope to reflect the holistic nature of emerging diseases - the recent 'One Health' perspective suggests that human health can only be fully understood by considering aspects of domestic and wildlife host health, as well as the wider ecological community (Karesh et al. 2012).

As a multifactorial phenomenon, it follows that a predictive approach to viral emergence must be interdisciplinary, where several different methodologies all have a role to play (Wilcox and Colwell 2005; Wood et al. 2012). For example, viral

sequencing efforts are now routinely employed as a clinical detection method and a key tool in the discovery of novel viruses (Delwart 2013; Lipkin 2013). Genomic sequence data also forms the basis of phylogenetic analyses, which can be used to assess evidence for cross-species transmission and reconstruct likely ancestral hosts (Holmes 2009; Kitchen et al. 2011; Drexler et al. 2012). Another example is the disease mapping and biogeographical analyses that have helped understand the environmental determinants of emerging viruses (Guernier et al. 2004; Jones et al. 2008; Dunn et al. 2010). Study at this scale has quantified landscape-level anthropogenic changes, or 'drivers', upon the spatial patterns of disease emergence (Patz et al. 2004; Wilcox and Gubler 2005; Jones et al. 2008).

Within this thesis, I focus on another essential component of predictive frameworks, ecological comparative analyses (or 'trait-based' analyses). These analyses aim to identify associations between phenotypic traits to infer their underlying relationships, typically using species-level data. Comparative analyses are fundamentally broad in their principle, i.e. those traits associated with successful emergence of viruses cannot be understood without also comparatively examining the traits of viruses that have not overcome biological barriers and have failed to emerge (Figure 1.1). The comparative perspective derives from macroecology, though as larger repositories of host-pathogen data are becoming available and statistical modelling methods for sparse and/or highly multidimensional data are improving, this approach is increasingly recognised as applicable to infectious diseases (Nunn 2012).

Several key comparative studies have laid the foundations to understanding characteristics of likely emerging pathogens, including RNA viruses. As previously outlined, the first comparative analyses of human pathogens suggested RNA viruses and zoonotic pathogens were more likely to be considered emerging (Cleaveland et al. 2001; Taylor et al. 2001; Woolhouse and Gowtage-Sequeria 2005). Emerging viruses are also more likely to have a broad range of nonhuman hosts (Cleaveland et al. 2001; Woolhouse and Gowtage-Sequeria 2005), and considering specific host taxa,

the highest numbers of emerging human viruses are known among rodents and ungulates (Woolhouse and Gowtage-Sequeria 2005), although this may partly reflect differences in taxonomic classification of virus families associated with these hosts.

The relationship between humans and nonhuman hosts is increasingly seen as important for zoonotic transmission and emergence. More recent comparative analyses have reported bat species to host a greater diversity of zoonotic viruses than rodent species (Luis et al. 2013). Humans also share more pathogens with nonhuman species that are closely phylogenetically related, though viruses often appear to be less subject to host phylogenetic constraints (Davies and Pedersen 2008; Cooper et al. 2012). A greater proportion of emerging pathogens are also shared with wildlife than domestic animals (Cleaveland et al. 2001), though among domestic species, pathogen sharing is positively correlated with earlier domestication history (Morand et al. 2014), suggesting the importance of human-animal contact in emergence.

Other comparative studies have further expanded on these foundations by examining virological risk factors for emergence. For example, predictions about molecular viral traits likely to increase emergence risk were proposed by Pulliam (2008), and subsequently demonstrated in that viruses replicating in the cytoplasm were more likely to infect humans than those with nuclear replication (Pulliam and Dushoff 2009). However, there are still many open questions surrounding how other virological factors influence emergence, such as transmission route and tissue tropism. Direct contact with blood or tissue has been theorised to present the highest risk of emergence (Pulliam 2008), though comparative analyses have suggested no single transmission route dominates among emerging human pathogens (Woolhouse and Gowtage-Sequeria 2005; Loh et al. 2015). The relationship between transmission route and specific levels of the Pathogen Pyramid is not well characterised (Figure 1.1), though from simple counts, vector-borne viruses appear more likely to infect humans (Woolhouse et al. 2001; Loh et al. 2015), but less likely to exhibit sustained human-to-human transmission (Woolhouse and Adair 2013). The influence of tissue tropism upon viral dynamics during emergence has received little attention, though

one comparative study found substitution rates to be higher among viruses with certain tissue tropisms (Hicks and Duffy 2014), which may in turn represent a predisposition for emergence (Holmes and Drummond 2007; Parrish et al. 2008).

1.4. Predicting traits of RNA viruses beyond emergence

As reviewed in the previous section, the primary focus of previous comparative analyses of RNA viruses has been risk of human infection (in terms of either emergence, zoonotic transmission, or virus sharing). However, despite the often-cited case studies of high-impact emerging viruses such as HIV and SARS coronavirus, not every emerging virus is highly pathogenic. RNA viruses exhibit substantial variation in virulence, which I use to refer to any costs of infection to individual host fitness, in order to broadly cover definitions used by evolutionary biology, ecology and clinical medicine (Day 2002a). Even closely related viruses can demonstrate very different levels of virulence, e.g. Zaire ebolavirus and other ebolaviruses cause extremely debilitating haemorrhagic disease with high mortality (Feldmann and Geisbert 2011), except Reston ebolavirus, where human infections have never presented with clinical disease (Morikawa et al. 2007). Similarly, neither does every zoonotic cross-species transmission event result in efficient human-to-human transmissibility or pandemic spread. Only a small fraction of emerging viruses ultimately reach the top of the Pathogen Pyramid (Figure 1.1) and become established as endemic human pathogens (Woolhouse and Gaunt 2007; Woolhouse and Adair 2013).

Therefore, I identify virulence and human-to-human transmissibility as key phenotypic traits determining the dynamics of emerging viruses and accordingly, representing significant importance to public health. A specific need to incorporate virulence and transmissibility within a wider predictive approach to disease emergence has been emphasised (Morse et al. 2012). Traditionally, the determinants of virulence and transmissibility within human RNA viruses have been studied using theoretical models or experimental animal models, often focusing on a single virus.

In contrast, such determinants have not been well-studied from wider empirical perspectives, though several recent attempts must be highlighted.

Virulence has been reported to positively correlate with external survivability, entry through wounded skin, and a lower infectious dose in comparative analyses pooling all pathogen types (Walther and Ewald 2004; Leggett et al. 2012). Transmissibility has also been associated with pathogen traits. Geoghegan et al. (2016) reported greater risk of human-to-human transmissibility among nonsegmented, nonenveloped, nonvector-borne RNA and DNA viruses, though it must be noted that these traits are highly conserved within viral taxonomic families. Focusing on host traits rather than virus traits, greater risk of human-to-human transmissibility has been reported for viruses with broader ecological diversity of host groups and zoonotic transmission via hunting or meat consumption (Johnson et al. 2015a).

These previous comparative studies on virulence and transmissibility have often been limited in scope to small sets of viruses or pathogens (Walther and Ewald 2004; Leggett et al. 2012), and to my knowledge, have not yet been conducted for a complete inventory of taxonomically-standardised human RNA viruses. Virulence and transmissibility of human viruses have begun to be quantified at this scale (Hay et al. 2013), though there are still many uncertainties and data deficiencies. Several fundamental questions remain regarding virus and host-based determinants of virulence and transmissibility. Many potential predictive traits remain unexplored, and among those examined in previous analyses, the influence of chosen measurements of traits and potential for trait interactions are still largely unknown.

1.5. Thesis aims & outline

In this thesis, I aim to predict key traits underpinning the dynamics of emerging viruses. Specifically, I aim to understand how virulence and transmissibility vary with wider ecological traits of human RNA viruses. Based on previous comparative analyses predicting emergence, I focus on two general types of

predictors – virus traits, e.g. transmission route, tissue tropism, and genetic structure; and nonhuman host traits, e.g. host identity, host diversity, and host life history. Throughout, I use data on traits of viruses at the level of taxonomic species, compiled via systematic literature searches, and I apply a comparative perspective by drawing on a range of predictive statistical modelling techniques.

In the second and third chapters, I focus on virus traits. In the second chapter, I construct classification models via machine learning methods to identify which viral traits may interact to predict virulence. I also assess the potential for ascertainment biases in compiled data by examining virus discovery curves with respect to viral traits. In the third chapter, I assess evidence for alternative hypothesised routes of adaptation towards transmissibility using phylogenetic state-switching models. I then use phylogenetic regression models to test whether increasing levels of transmissibility are associated with ecological or genetic traits.

In the fourth and fifth chapters, I focus on nonhuman host traits. In the fourth chapter, I present an overview of where virulent and human-transmissible viruses are concentrated among mammalian host-virus relationships. I then investigate whether virulence and transmissibility are predicted by host range breadth and infection of specific host taxa using mixed regression methods. In the fifth chapter, I ask whether nonhuman mammal hosts with faster-paced life history are predisposed to host virulent and human-transmissible viruses. I use body mass as a proxy for life history and weighted regression models to account for the wider virus diversity of mammal hosts.

In the sixth and final chapter, I discuss the overall implications of the thesis findings in relation to virus ecology and evolution. I highlight those areas where further work is critically needed and suggest hypotheses for future study. Finally, I provide a perspective on the implications of this thesis for global health programmes and how these findings may contribute to a more effective strategy of preparedness against emerging viruses.

Supplementary material accompanying each of the analytical chapters is provided in separate appendices at the rear of the thesis. Publications describing overviews of the data collected as part of this thesis are also included at the rear, though note that these are not direct adaptations of any of the following thesis chapters.

Chapter 2. Tropism breadth and transmission ecology predict virulence of emerging human RNA viruses

2.1. Abstract

Although studies have begun to identify pathogen traits associated with the emergence of novel viruses, these do not address why viruses exhibit such a wide variation in virulence, a key determinant of risk to population health. I use structured literature searches to review the virulence of each of the 180 known human-infective RNA virus species. I then apply comparative machine learning approaches to determine whether virulence of human viruses can be predicted by their ecological traits, and whether patterns of virus discovery vary with virulence or these traits. Using severity of clinical disease as a measurement of virulence, I determined potential risk factors using predictive classification tree models. Correcting for virus taxonomy, the final classification tree model combining tissue tropism, extent of transmissibility within human populations, and transmission route described disease severity with 88% accuracy. Virus discovery did not vary with virulence alone, but showed complex relationships dependent on both virulence and risk factor traits, though model conclusions were robust to this variation. When applied to 30 newly reported human viruses as a test set, models predicted literature-assigned severity of clinical disease with 87% accuracy compared to a null accuracy of 50%. The risk factors I identify may provide novel perspectives in understanding the evolution of virulence and identifying molecular virulence mechanisms. These risk factors could also improve planning and preparedness in public health strategies as part of a predictive framework for novel human infections.

2.2. Introduction

The emergence of novel infectious diseases continues to represent a threat to global public health. Emerging pathogens have been defined as those newly recognised infections of humans following zoonotic transmission, or those increasing in incidence and/or geographic range (Morse 1995). Recent high-profile examples of emerging pathogens include the discovery of the novel MERS coronavirus from cases of respiratory illness in 2012 (Zaki et al. 2012), and the expansion of the range of chikungunya virus across the Indian Ocean and the Caribbean (Charrel et al. 2014). The emergence of previously unseen viruses means that the set of known human viruses continually increases by around 2 species per year (Woolhouse et al. 2012). Comparative studies have begun to identify trends and ecological risk factors among emerging pathogens. For example, pathogens are more likely to emerge in humans if they have a broad host range, and are RNA viruses (Cleaveland et al. 2001; Taylor et al. 2001; Woolhouse and Gowtage-Sequeria 2005). Here, I focus on understanding the ecological determinants of virulence, using all known human RNA viruses as a study system.

Emerging RNA viruses vary widely in their virulence, with some never having been associated with human disease at all. As an example, Zaire ebolavirus causes severe haemorrhagic fever with outbreaks, including the most recent West African outbreak showing case fatality ratios of ~60% or more (Feldmann and Geisbert 2011; Focosi and Maggi 2015). In comparison, human infections with Reston ebolavirus have never exhibited any evidence of disease symptoms (Morikawa et al. 2007), despite both viruses being members of the genus Ebolavirus. Applying the comparative approach to understand the ecology of virulent viruses could offer valuable synergy with studies of emergence, towards prioritisation and preparedness in the detection of potential new human viruses (Morse et al. 2012).

Few comparative analyses have addressed risk of human pathogen virulence to date (but see Ewald 1983; Walther and Ewald 2004; Leggett et al. 2013), and none have done so exhaustively across the breadth of all currently recognised human

pathogens of a specific type. Several hypotheses regarding how pathogen ecology affects virulence may be derived from theoretical models of evolution. For example, the trade-off hypothesis was developed from ideas that transmission rate between individuals may increase as a function of virulence, but there will be a consequential increase in host mortality (or decrease in host recovery, as the inverse of mortality). As a result, pathogen fitness will be subject to trade-off between virulence and transmissibility over a longer infectious window (Anderson and May 1982; Bremermann and Pickering 1983). The trade-off hypothesis is highly debated, and contested as difficult to empirically characterise and dependent on many other aspects of host-pathogen coevolution (Ebert and Bull 2003; Alizon et al. 2009). However, comparative analysis has been suggested as a method to assess evidence for a virulence-transmission trade-off (Alizon et al. 2009) and based on these core principles, I hypothesise that limited human-to-human transmissibility may act as a predictive risk factor for virulence. It must be noted that evolutionary trade-offs will only apply to coevolved host-virus relationships and that many human viruses result from zoonotic cross-species transmission without onward transmission or adaptation. In these cases, a 'coincidental' non-adapted virulence may result (Levin and Svanborg Edén 1990; Bull 1994), and again, I hypothesise that limited human-to-human transmissibility may predict virulence.

Transmission route may also influence the evolution of virulence. Ewald (1983) suggested that vector-borne pathogens should be less constrained by costs of virulence, i.e. morbidity and immobilisation of the vertebrate host does not impede transmission if it occurs through an arthropod vector. I therefore hypothesise a vector-borne transmission route may predict virulence. Finally, although yet unexplored via theoretical models, it may be an intuitive expectation that systemic infections present with more severe disease than local infections. A broader tissue tropism could therefore also predict virulence.

The currently known set of human RNA virus species is likely an incomplete inventory of the diversity of RNA viruses within humans (Woolhouse et al. 2008,

2012). Virus detectability and reporting is known to depend, among other factors (Chan et al. 2010), on distinctiveness or severity of clinical symptoms (Koopmans 2013). Although these are factors affecting reporting of individual cases, whether there is ascertainment bias towards discovering human virus species with greater virulence has not yet been assessed.

I aim to determine patterns of virulence across the breadth of all known human RNA viruses and use comparative risk factor analysis to ask whether ecological traits of these viruses predict virulence in humans. Specifically, I examine hypotheses that viruses would be more virulent if they: lacked transmissibility within humans; had vector-borne transmission routes; or had greater breadth of tissue tropisms (Table 2.1), independent of any biases in discovery. Finally, I aim to verify the robustness of analyses and assess use of risk factor models as a predictive tool by applying them to newly reported viruses as a test set.

2.3. Materials and methods

2.3.1. Data collection

For each of the 180 recognised human-infective RNA virus species following standardised data compilation efforts and critical assessment protocols (Woolhouse et al. 2013), data on virulence and potential risk factors were collected via a systematic search and review of clinical and epidemiological literature. The following were consulted in turn: clinical virology textbooks (Knipe and Howley 2007; Zuckerman et al. 2009; Richman et al. 2009); references from the dataset described by Woolhouse et al. (2013); literature searches using Google Scholar (search terms: 1) [virus name] AND human, 2) [virus name] AND human AND case, 3) [virus name] AND human AND [fatal* OR death], 4) [virus name] AND human AND [tropi* or isolat*], 5) [virus name] AND human AND transmi*). Searches 3 - 5 were carried out only when fatality, tropism, or transmission data respectively were not already found from previous sources. Data collection and virus name search terms included the full

species name, any synonyms or subspecies (excluding vaccine strains) and the standard virus abbreviation as given by ICTV 9th edition (King et al. 2011).

Although many possible measurements of virulence have been proposed (Day 2002a; Nathanson et al. 2007), even simple metrics like case fatality ratio (CFR) have not been calculated for the majority of human RNA virus species. Therefore, virulence was rated using a simple two-category measure of severity of typical disease in humans. I rated viruses as 'severe' if they firstly had ≥5% CFR where data was available (134/180 viruses including those with zero CFR), otherwise, I rated viruses as 'severe' if they had frequent reports of hospitalisation, were associated with significant morbidity from certain conditions (haemorrhagic fever, seizures/coma, cirrhosis, AIDS, hantavirus pulmonary syndrome, HTLV-associated myelopathy) or were explicitly described as "severe" or "causing severe disease" (Table A.1). I rated viruses as 'nonsevere' if none of these conditions were met. Note that this led to 'nonsevere' ratings for some viruses with clinically severe, but rare syndromes, e.g. dengue virus causes dengue haemorrhagic fever, though this is much rarer than typical acute dengue fever (Knipe and Howley 2007; Zuckerman et al. 2009). To address this, data were also collected on whether the virus has caused fatalities in vulnerable individuals (defined as age 16 and below or 60 and above, immunosuppressed, having co-morbidities, or otherwise cited as being 'at-risk' by sources for specific viruses) and in healthy adults, and whether any 'nonsevere' virus has atypically severe strains (for example, most infections with viruses within the species *Human enterovirus C* are mild; however, poliovirus, which causes severe paralytic disease, is also included as a strain of this species). These were examined both individually and within a composite six-rank system (Table A.3).

Table 2.1. Virus trait data collected for use in classification tree analysis. Data are listed along with motivation for inclusion in analysis, definition of two-level categories, sources used for data collection and pictorial representation in classification tree figures.

Data	Model usage	Specification	Source
Human-to-human transmissibility	Risk factor	Two separate measures of human transmissibility: a) Any ($R_0 > 0$) or none ($R_0 = 0$):  b) Sustained ($R_0 \geq 1$) or nonsustained ($R_0 < 1$): 	Woolhouse et al. (2012) model, Woolhouse et al. (2016) dataset, systematic literature search
Primary transmission route	Risk factor	Vector-borne or nonvector-borne:  Where "nonvector" includes direct, respiratory and faecal-oral routes.	Systematic literature search
Tissue tropism breadth	Risk factor	Single-organ system or multiple organ systems: 	Systematic literature search
Year of discovery	Bias correction	Year of publication of first human infection evidence	Woolhouse et al. (2013) dataset
Virus taxonomy	Bias correction	Genome type: +ssRNA, -ssRNA, dsRNA, RNA-RT) Taxonomic family: n = 18 including "Unassigned"	ICTV 9th edition (King et al. 2011)

Data were collected for three main risk factors: extent of human-to-human transmissibility, transmission route, and tissue tropism breadth. Transmission route data were collected as a multiple-category variable that was subsequently reclassified to a two-category variable (Table 2.1) based on supported groupings from an initial classification tree model (Figure A.1) and previous comparative analyses of viral emergence (Johnson et al. 2015a; Geoghegan et al. 2016). Based on previous conceptual models and a systematic compilation and review of evidence (Woolhouse et al. 2012; Woolhouse et al. 2016), I specified human-to-human transmissibility by constructing two binary variables to denote whether each virus had a) any human-to-human transmissibility, equivalent to $R_0 > 0$ in humans or Pathogen Pyramid level 3 or above (see General Introduction); and b) sustained human-to-human transmissibility, equivalent to $R_0 \geq 1$ or Pathogen Pyramid level 4 (see General Introduction). Transmission route was defined as the primary route the virus is transmitted by, classified as either arthropod vector-borne (excluding mechanical transmission), or nonvector-borne (including direct, faecal-oral and respiratory transmission, which consistently clustered together in preliminary tree analyses (Figure A.1)). Tissue tropism breadth was specified as whether the virus typically infects either single or multiple organ systems. I accepted isolation of the virus, viral proteins or genetic material, or diagnostic symptoms of the virus (such as characteristic histological damage) as evidence of infection within an organ system, but did not accept generalised symptoms such as inflammation.

All virulence and risk factor data pertained to natural or unintentional artificially-acquired human infection only and data from intentional human infection, animal infection, and *in vitro* infection were not considered. Viral taxonomy was corrected for in analyses by including both genome type and taxonomic family as covariates. Additionally, to analyse temporal trends in viral traits and their predictive power for virulence, year of discovery (i.e. first publication of human infection evidence) was obtained for each virus from the dataset described by Woolhouse et al. (2013).

2.3.2. Classification analysis

Comparative risk factor analyses were conducted with classification trees using the R package 'rpart', v4.1-8 (Therneau et al. 2014). Classification trees aim to optimally classify data points into their correct category of outcome variable. Classification tree methods are appropriate for comparative ecological analyses as they easily handle missing predictor data and are capable of fitting complex non-linear interactions. A tree was created by 'recursive partitioning', the repeated splitting of the dataset using every possible permutation of each predictor, and retaining the split that minimises the Gini impurity (De'ath and Fabricius 2000), a measure of how well-separated data points are with respect to different outcome variable categories. To prevent overfitting, all trees were pruned back to the optimal branching size, taken as most common consensus size over 1000 repeats of cross-validation. Tree accuracy was calculated as the overall proportion of viruses correctly classified in outcome variable compared to literature-assigned ratings (assuming these to be 100% accurate as the 'gold standard' or 'ground truth'). Given that my virulence rating only had two categories and many correct classifications may have occurred by chance, I compared accuracy of classification trees to that of the null model, i.e. a model without any predictors that simply assigned the most common virulence rating, 'nonsevere', to all viruses. All modelling was carried out in R, v3.1.1 (R Development Core Team 2015).

To investigate robustness of individual predictors within the classification tree, jack-knifing was carried out for a) the full predictor set, and b) alternative predictor sets, each excluding a different predictor. In all cases, jack-knifing was carried out by sequentially removing each data point before growing and pruning to the optimal tree size, giving 178 jack-knifed trees for each of the predictor sets. Accuracy statistics from these trees were then plotted to visualise estimates of error around the contribution of each predictor, in terms of loss in tree accuracy when this predictor was excluded. Contributions were also measured by visualising accuracy

for c), non-jack-knifed trees built using single predictors only. Additional statistics of interest (True Skill Statistic (Allouche et al. 2006), sensitivity, negative predictive value, and accuracy relative to null) were also obtained from jack-knifed trees (Figure A.4 - A.7). Risk factors were also validated in a traditional logistic regression model framework (Supplementary Methods A.1)

Potential ascertainment biases due to virulence were also explored. Firstly, proportional discovery curves of known human virus species over time relative to total were constructed to compare 'severe' and 'nonsevere' viruses. Secondly, to confirm risk factor robustness over time and independence of any discovery biases, virus data was split into nested subsets of those discovered by the start of each decade, starting with viruses discovered by 1940. For each subset, I calculated the information gain associated with each risk factor variable as a measure of informativeness towards correct virus classification. Information gain is calculated as the difference between the entropy of the dataset before and after splitting via a predictor variable (Kingsford and Salzberg 2008), where entropy is defined as

$$-\sum_{i=1}^{n} p(x_i) * logp(x_i) \qquad (2.1)$$

for outcome variable x with n possible categories and $p(x_i)$ denoting proportion of data points in outcome category i. Information gain was calculated for each risk factor using the attrEval function in the R package 'CORElearn', v0.9.43 (Robnik-Sikonja and Savicky 2014) and was plotted alongside two-way discovery curves split by risk factor-severity rating combinations, to visualise changes in strengths of risk factors over time.

Finally, to test the predictive power of the classification tree, virulence and risk factor data were also collected for newly reported human RNA viruses that have not yet been approved as species, as of ICTV 9th edition Virus Taxonomy (King et al. 2011). These were found via literature searches using Google Scholar (search terms: novel OR new* AND virus, years: 2012-2014) and reference tracing from other

reviewed literature. Virulence and risk factor data collection followed the literature search step of the protocol previously outlined. The final classification tree was then used to generate disease severity predictions and compared in accuracy to the null model, assuming literature-based criteria as a 'ground truth'.

2.4. Results

2.4.1. Virulence of human RNA viruses

Of the 180 human RNA viruses described by Woolhouse et al. (2013), 48 were rated as causing 'severe' clinical disease and 130 as 'nonsevere' (Figure 2.1A, see also Table A.1). Two virus species could not be assigned a disease severity rating and were excluded from all analyses (*Hepatitis delta virus*, which is reliant on hepatitis B virus coinfection; and *Primate T-lymphotropic virus 3*, which may be associated with chronic disease like other T-lymphotropic viruses, but has not been known in humans long enough for cohort observations). Disease severity differed between the 17 viral taxonomic families (Fisher's exact, 1000 simulations, $p < 0.001$), with *Arenaviridae*, *Rhabdoviridae* and *Retroviridae* having the highest fractions of severe-rated viruses (Figure 2.1). Fatalities were reported in healthy adults for 60 viruses and in vulnerable individuals only for an additional 23 viruses, whilst 8 viruses rated 'nonsevere' had severe strains, 6 of which belonged to the family *Picornaviridae*.

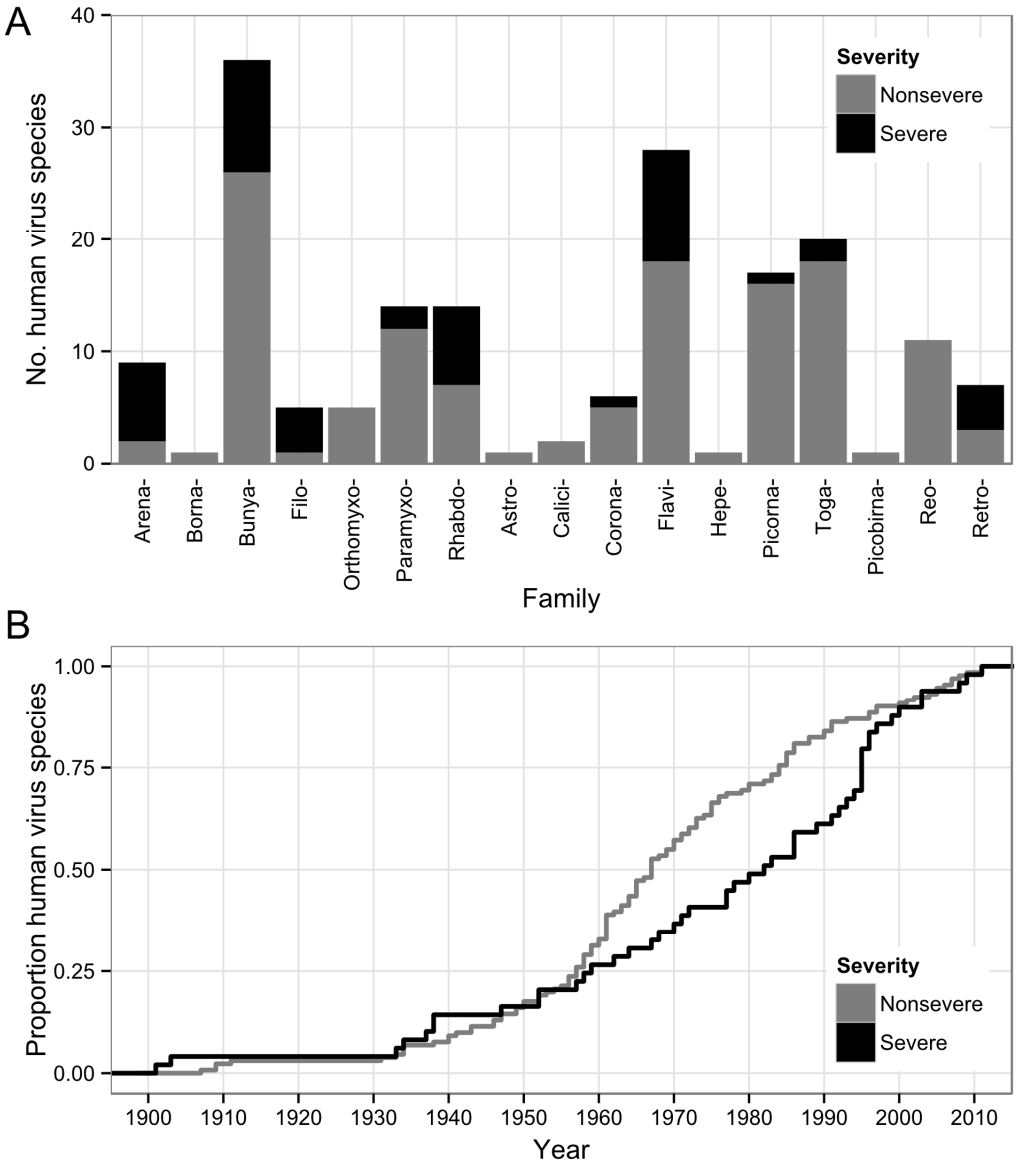


Figure 2.1. Virulence of currently known human RNA viruses, with respect to taxonomy and discovery year. A) Number of human RNA viruses split by taxonomic family, B) curves of virus discovery within humans over time as a proportion of current known total (nonsevere, n = 130; severe, n = 48). Shaded bars/curves denote disease severity rating.

2.4.2. Classification tree risk factor analysis

The final pruned classification tree included all hypothesised risk factors and taxonomic family (Figure 2.2), and classifications matched those from literature-based criteria for 157 of 178 viruses giving a resulting accuracy of 88.2% (95%

confidence interval: 82.5% - 92.6%). When compared to the null model assigning all viruses as nonsevere (accuracy = 73.0%), the classification tree demonstrated significantly greater accuracy (exact binomial one-tailed test, $p < 0.001$). The majority of virus species were classified as nonsevere in a single, large group defined by single-organ tropism and specific taxonomic families. Three separate groups were fitted by the model as associated with severe disease: i) viruses with multi-organ system tropism in the families *Arenaviridae, Bunyaviridae* or *Filoviridae*; ii) viruses with single-organ system tropism and nonvector-borne transmission in the family *Rhabdoviridae*; iii) viruses with single-organ system tropism, nonvector-borne transmission and sustained human-to-human transmissibility in the families *Flaviviridae* and *Retroviridae*.

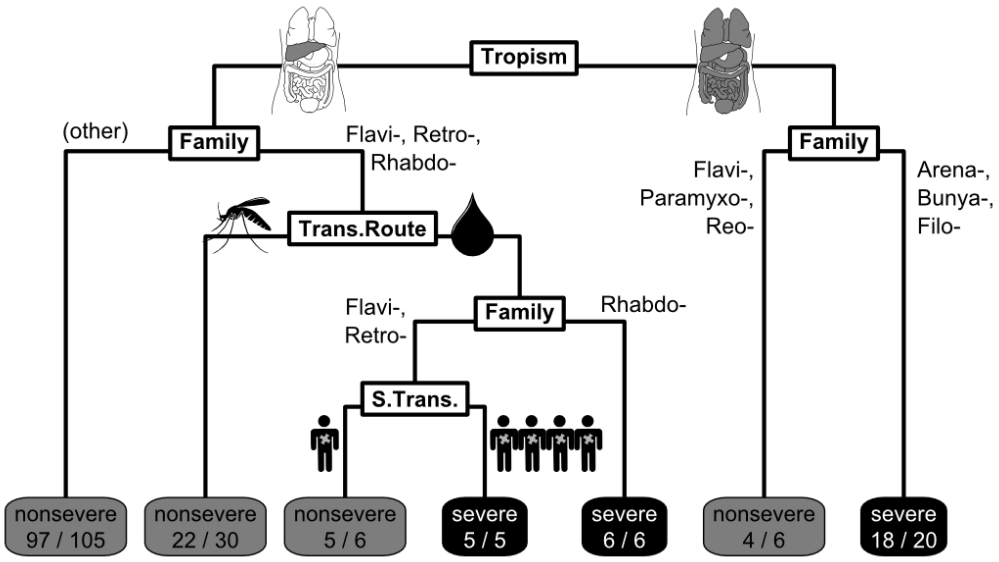


Figure 2.2. Final pruned classification tree predicting disease severity for 178 human RNA viruses. Viruses begin at the top and are classified according to split criteria (white boxes) with risk factors pictorially represented following Table 2.1 until reaching terminal nodes with the model's prediction of disease severity, and the fraction of viruses following that path correctly classified, based on literature-assigned ratings (shaded boxes). 'Trans.Route' denotes primary transmission route, and 'S.Trans.' denotes sustained human-to-human transmissibility.

Of the 21 misclassifications within the tree, 19 were viruses rated as severe through literature review that were misclassified as nonsevere, giving sensitivity of classifying severe disease as 0.604 and specificity as 0.985. These misclassifications occurred primarily in two classification groups (Figure 2.2). The final classification tree structure was robust to removing viruses with low-certainty data (Figure A.2, A.3). Although tree structure differed in some respects, specifying virulence as either a fatality variable alone or combined with severity and severe strains in an ordinal ranking system did not significantly improve accuracy (Table A.3, A.4).

In testing robustness of individual predictors and jack-knifing, trees that excluded each predictor showed substantially reduced accuracy compared to the null model of classifying all viruses as 'nonsevere' (Figure 2.3), except when excluding sustained human transmissibility, which appeared redundant with any human transmissibility. When considered alone in univariate trees, taxonomic family and tropism breadth classified severity with accuracies of 78.1% and 81.2% respectively (Figure 2.3), though still less than the jack-knifed full model, suggesting that no single trait comprehensively explained variation in virulence. Accuracy in univariate trees using human transmissibility and transmission route did not improve upon the null model. Predictor robustness was broadly comparable when examining alternative model performance measures (True Skill Statistic (Allouche et al. 2006), accuracy relative to null, Figure A.4, A.7) and performance measures that prioritised detection of the 'severe' category (sensitivity, Negative Predictive Value, Figure A.5, A.6).

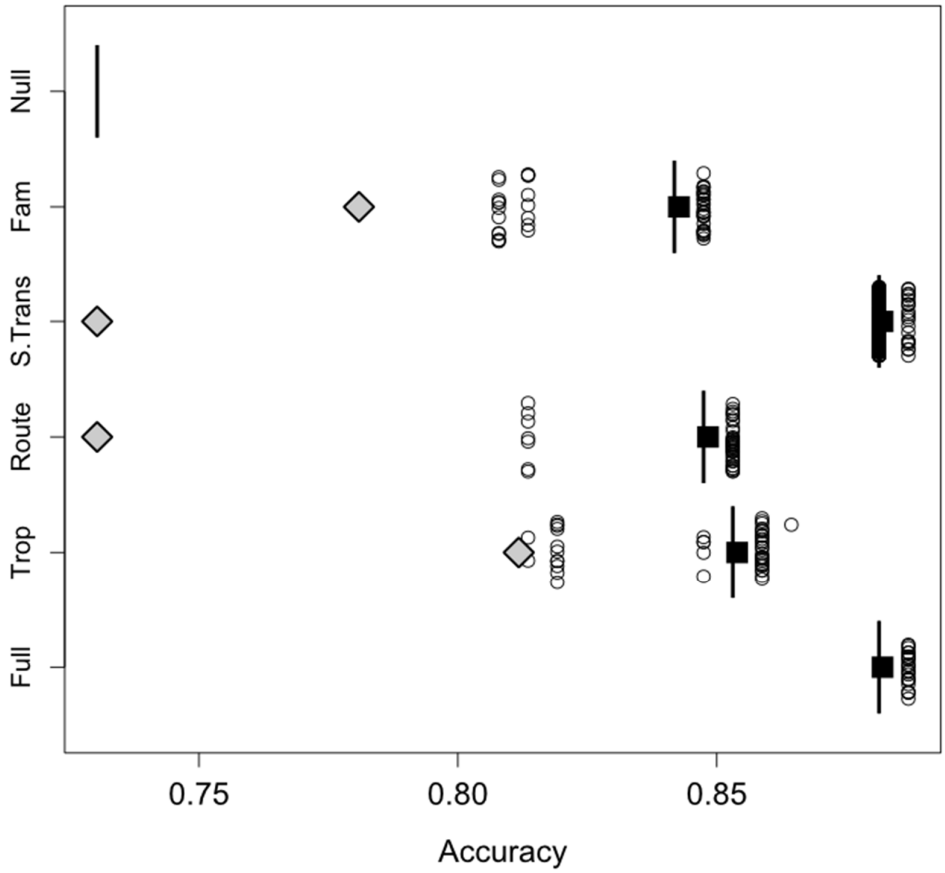


Figure 2.3. Full and jack-knifed tree accuracies across different predictor sets: 'Null' = null model (no predictors), 'Fam' = taxonomic family, 'S.Trans' = sustained human-to-human transmissibility, 'Route' = transmission route, 'Trop' = tropism breadth, 'Full' = full model (all predictors). Solid squares denote accuracy for tree built with full dataset (n=178) and boxes/outlying open circles denote accuracy for 178 trees built with jack-knifed datasets for predictor sets removing predictor given on Y axis (except 'Null' & 'Full'). Grey diamonds denote accuracy for tree built with single predictor given on Y axis only. Jack-knifed trees removing genome type and any human-to-human transmissibility are not depicted as these predictors did not appear in any trees resulting from the full predictor set.

2.4.3. Ascertainment bias and risk factor temporality

Ascertainment bias towards more virulent viruses did not appear present among the full dataset as cumulative proportional discovery curves for severe and

nonsevere-rated viruses showed little difference until 1960, where a relatively higher discovery rate of nonsevere viruses was observed (Figure 2.1B). However, differences in temporal trends were more visible between two-way discovery curves for each severity-risk factor combination (Figure 2.4D - F), for example, almost all currently known severe vector-borne viruses were discovered by the mid-1960s, whilst nonsevere vector-borne and severe nonvector-borne viruses tended to be discovered later (Figure 2.4E). Subsequently, the predictive power of each risk factor also varied over time, based on changes in information gain (Figure 2.4A - C). In earlier, smaller subsets, transmission route and human transmissibility were the most informative risk factors (Figure 2.4B-C). However, as time progressed and sample sizes increased, the informativeness of tissue tropism also increased (Figure 2.4A), ultimately becoming the most informative risk factor and the first branching criteria of the final classification tree (Figure 2.2).

2.4.4. Predicted virulence of newly reported viruses

I identified 30 newly recognised human RNA viruses not yet considered species, 15 of which were rated as causing severe disease using literature-based criteria and 25 of which had complete risk factor predictor data available. Although many were bunyaviruses or hantaviruses, 9 virus families were represented (Table A.2). The final classification tree was applied to these as a test set and correctly predicted the literature-assigned severity rating for 26/30 viruses (86.7% accuracy, 95% confidence interval: 69.2%-96.2%) (Figure 2.5), a significantly greater performance than the null model (null accuracy = 50.0%; exact binomial one-tailed test, $p < 0.001$). In predicting 'severe' viruses, sensitivity was 0.73 and specificity 1.0; all four misclassifications were viruses rated severe by literature-based protocols, but predicted as nonsevere by the tree (*Bunyaviridae: Bhanja virus, Heartland virus, Coronaviridae: MERS coronavirus,* and *Paramyxoviridae: Sosuga virus*).

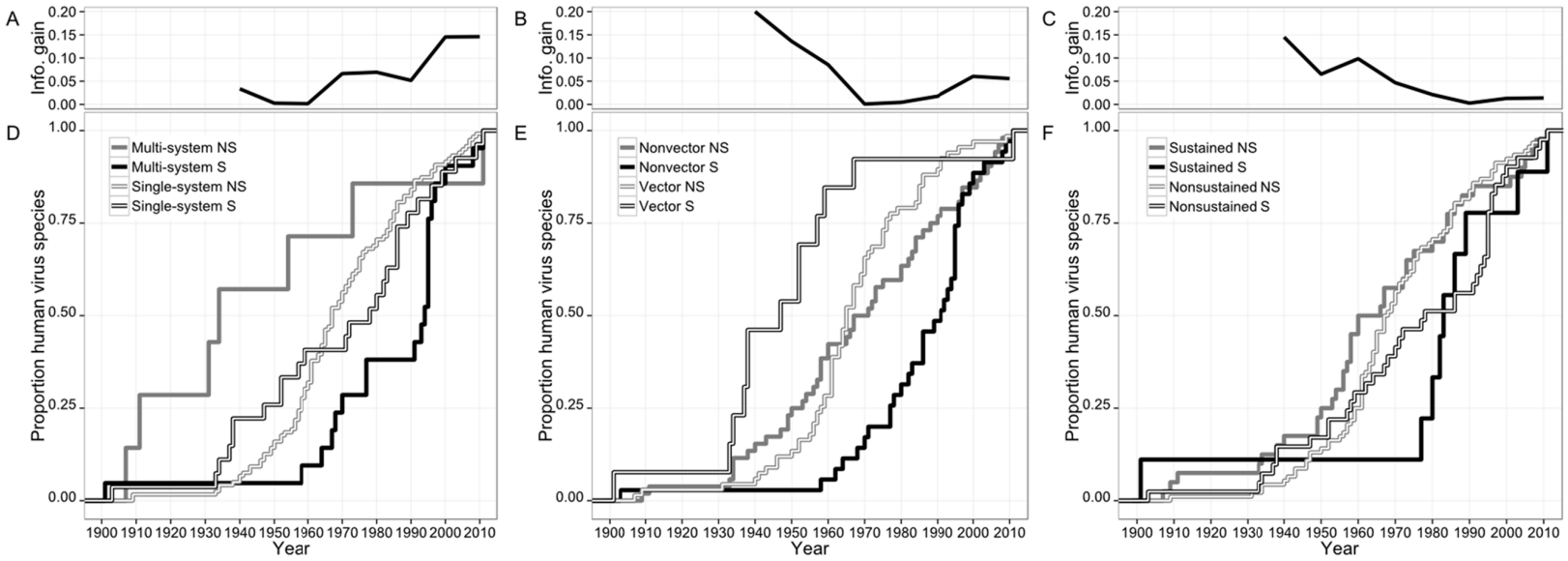


Figure 2.4. Informativeness and discovery dynamics with respect to virulence risk factors over time. A - C) Information gain parameter associated with each risk factor for virulence featured in classification tree (tropism breadth, transmission route, human-to-human transmissibility) over time calculated for subsets of discovered viruses at decade intervals and D - F) corresponding bivariate curves of virus discovery within humans as a proportion of current known total, separated by both disease severity (shaded curves, NS = nonsevere, S = severe) and risk factors (solid/open curves, D) multi-/single-organ system tropism, E) nonvector/vector-borne transmission, F) sustained/nonsustained transmissibility). After excluding viruses with missing risk factor/virulence data, D) n = 174, E) n = 163, F) n = 178.

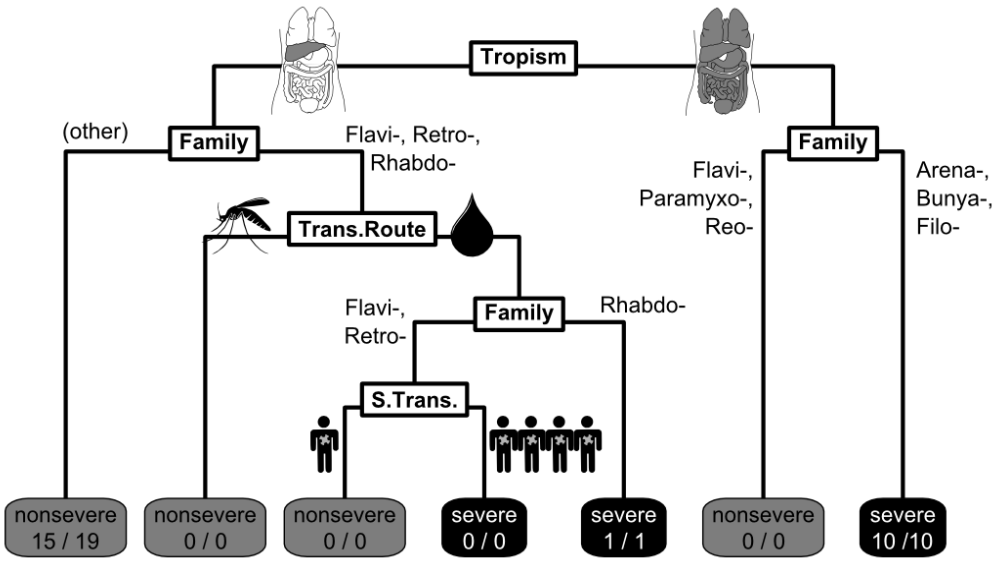


Figure 2.5. Application of the final classification tree to predicting disease severity for an independent dataset of 30 newly reported viruses not yet ratified as species. Viruses are classified according to criteria as in Figure 2.2. 'Trans.Route' denotes primary transmission route, and 'S.Trans.' denotes sustained human-to-human transmissibility.

2.5. Discussion

I present the first comparative analysis of virulence across all known human RNA virus species to my knowledge. I find that disease severity is non-randomly distributed across virus families and that severe disease is predicted by risk factors of tissue tropism breadth, and to a lesser extent, transmission route and human-to-human transmissibility. Specifically, viruses were expected to cause severe disease if they were arenaviruses, bunyaviruses or filoviruses infecting multiple organ systems; or if they infected a single organ system, were nonvector-borne, and were flaviviruses or retroviruses with sustained transmissibility, or rhabdoviruses (regardless of transmissibility). These risk factors were robust to alternative modelling methods, alternative definitions of virulence, and exclusions of poor quality data.

2.5.1. Ecology and evolution of risk factor traits

Tissue tropism breadth was the most informative risk factor (Figure 2.4A) and the first split criteria in the classification tree (Figure 2.2). Few evolutionary studies have predicted how tissue tropism should influence virulence, although it has been noted infection of non-target tissues that do not contribute to transmission may result in an excessive, non-adapted virulence (Levin and Bull 1994). Relationships between virulence and tropism have instead largely been examined from experimental perspectives, which show consistency with my findings, e.g. virulent strains of Newcastle disease virus are detectable in more organ systems of avian embryos than non-virulent strains (Al-Garib et al. 2003). Tropism for multiple organ systems could result in virulence as a function of pathology occurring in multiple bodily areas, increasing intensity of clinical disease. However, the underlying evolutionary dynamics of generalism in tissue tropisms are unknown, and an area of scope for future theoretical models of virulence. In comparison, studies of pathogen generalism in the context of host range have predicted that generalism may increase virulence, though only under specific conditions (Leggett et al. 2013), e.g. within-host competition with specialists. The connection between generalism in tissue tropism and host range is not clear (Taber and Pease 1990), and further evolutionary study will be necessary to resolve the virulence association I observe.

I also found that viruses primarily transmitted by routes other than arthropod vectors were more likely to be virulent, though only among three virus families. Contrastingly, Ewald (1983) previously reported a positive association between virulence and vector-borne transmission in comparative analyses pooling several microparasite types, and suggested virulence has fewer evolutionary costs if vector transmission is independent of host health and mobility. My findings may imply that even if transmission occurs independently, virulence may bring ultimate costs in terms of host mortality before a vector encounter can occur. This previous analysis was also limited to a much smaller range of viruses and may reflect ascertainment in discovery or data availability; I observed that most currently known severe vector-

borne viruses were recognised much earlier than nonsevere vector-borne viruses, or severe viruses transmitted via other routes (Figure 2.4E). Day (2002b) found vector-borne transmission to result in higher virulence in theoretical models only under certain parameter conditions and proposed that arthropod vectoring may increase parasite virulence if this involves a comparatively larger inoculum. However, experimental studies of *Plasmodium chabaudi* show vector-transmitted infections to be less virulent compared to direct inoculation, independent of inoculum size (Spence et al. 2013). The dynamics of virulence in vector-borne systems warrant further attention.

The relationship between virulence and transmissibility appeared more complex. Based on hypothesised virulence-transmissibility trade-offs (Anderson and May 1982; Bremermann and Pickering 1983; Alizon et al. 2009) and the potential for coincidental non-adapted virulence in 'dead-end' zoonotic infections (Levin and Svanborg Edén 1990; Bull 1994), I expected viruses with inefficient or no human-to-human transmission to be more virulent. For the most part I found no association between transmissibility and virulence, except within a small subset of nonvector-borne flaviviruses and retroviruses, where the relationship was counter to expectation. Five virus species of this subset were rated to cause severe disease despite having efficient human-to-human transmissibility (*HIV 1* and *2*, *Primate T-lymphotropic virus 1* and *2*, and *Hepatitis C virus*). These viruses are all typically associated with chronic conditions (specifically AIDS, HTLV-associated myelopathy, and cirrhosis), which may explain why this group does not support evolutionary theory – costs to pathogens of host mortality may be limited if disease occurs after the infectious window has already passed, essentially 'decoupling' virulence and transmission (Bull 1994). However, human-transmissible viruses associated with chronic disease were not rated 'severe' by literature-based criteria in all cases, e.g. several human enteroviruses. The lack of support for a hypothesised negative relationship between virulence and transmissibility may stem from the difficulty in accurately measuring human-to-human transmissibility. In the absence of

standardised metrics, both individual data (e.g. single observations) and population-level data (e.g. R_0 values) were used to categorise transmissibility despite representing different components of the trade-off hypothesis. Improved quantification has been noted as a crucial step towards fully testing this relationship (Alizon et al. 2009).

2.5.2. Ascertainment biases

Although virus detectability is known to depend on severity of cases, I did not observe ascertainment bias with respect to overall disease severity. However, when viruses were subset to only those known in decade intervals, predictive power of risk factors changed over time, likely as a result of shifts in viral discovery focuses. For example, I observed a sharp increase from the 1960s onwards in nonsevere vector-borne viruses (Figure 2.4E), also observed elsewhere and attributed to the efforts of a specific arbovirus discovery program (Rosenberg et al. 2013). Whether the known set of human viruses and virulence risk factors among them are ultimately biased depends on the true picture of human virus diversity, about which it is difficult to speculate given how little of the human and wider global virome has been sampled (Anthony et al. 2013; Delwart 2013).

I also observed some directional bias in the classification tree model – most misclassifications were viruses assigned as severe by literature-based criteria and assigned as nonsevere by the tree, a large fraction of which were vector-borne viruses. This may be partly due to the inflated discovery and designation of vector-borne virus species outlined above, leading the tree to typify taxonomic families dominated by vector-borne viruses (e.g. *Bunyaviridae, Togaviridae*) and vector-borne flaviviruses and rhabdoviruses as broadly nonsevere in the first two prediction groups, respectively (Figure 2.2). Misclassified viruses tended to be rare exceptions among groups of viruses with similar ecology, for example, the equine encephalitis viruses (Eastern, Western) are unusually severe compared to other mosquito-borne alphaviruses. Many of these viruses, including the equine encephalitis viruses, were associated with neural syndromes. This may suggest that beyond broad tissue

tropism, a more specialist neurological tropism could be an alternative evolutionary route to virulence, assuming neural pathology is a true reflection of neural tropism. No other traits considered distinguished between these nonsevere and severe vector-borne viruses, and the acquisition of further ecological traits such as biogeography, or molecular and genetic traits may improve separation of neurological vector-borne viruses.

2.5.3. Predictive power for newly reported viruses

Predictive studies can provide valuable input to risk assessments of novel emerging diseases (Morse et al. 2012). In testing as a predictive tool for virulence risk, the classification tree model correctly predicted the literature-assigned disease severity of 26 of 30 newly reported human viruses not yet ratified as species. However, the predictive potential of this model is subject to the accuracy of both virulence and risk factor data. It must be acknowledged that the literature-based criteria used to calculate accuracy (in the absence of any standardised 'ground truth' for virulence) remains prone to error. Three newly-reported viruses, *Bhanja virus*, *Heartland virus* and *Sosuga virus*, were predicted by the classification tree to be nonsevere, but were assigned a 'severe' rating from the literature protocol as the few known cases resulted in hospitalisation. However, there remains genuine uncertainty as to the true level of virulence for these viruses and the literature-assigned rating may reflect biased detection towards highly symptomatic cases (Koopmans 2013).

I also acknowledge that data on virulence itself may be more accessible information during viral emergence than several predictors used as inputs in my model, particularly tissue tropism, which is not likely to be known with confidence before the first estimate of clinical pathology. To illustrate, MERS coronavirus is a respiratory virus, though its syndrome often exhibits kidney involvement. However, there is currently no diagnostic evidence the virus infects the renal system, and I therefore assigned *MERS coronavirus* as infecting a single organ system only, which may have led to the discrepancy between tree prediction and literature-based rating.

One way to circumvent this and develop timely virulence predictions may be to substitute tissue tropism information using data regarding nonhuman animal infections. For MERS coronavirus, primate and other animal models collectively demonstrate multi-organ tropism with involvement of respiratory, neuronal, gastrointestinal, and several other tissues (van Doremalen and Munster 2015). However, comprehensive laboratory experiments are also unlikely to be accessible in the early stages of emergence. Wild surveillance studies could be consulted, although tissue origins of positive samples may often be left unreported (Young and Olival 2016). The most immediate method of substituting predictor data would be imputation from the nearest phylogenetically related virus, particularly given tissue tropism appears to be a highly conserved trait (Taber and Pease 1990). The above concerns highlight the challenge presented by paucity of data during viral emergence. As genomic methods improve and viral sequence information becomes increasingly easy to obtain, an ultimate target will be the advancement of knowledge such that tissue tropism and receptor usage may be directly inferred via genetic markers from sequences alone.

2.5.4. Analytical limitations

I acknowledge several limitations to the quality of data used, as with any broad comparative analysis. Risk factor data was problematic or missing for certain viruses, e.g. natural transmission route for viruses only known to infect humans by accidental occupational exposure, and breadth of tissue tropism for viruses only known from serological evidence. However, the consistency of findings between alternative, stricter definitions of virulence and data subsets removing viruses with suspected data quality issues suggests scarcity of data does not bias these analyses. I also acknowledge that the chosen model methodology of classification trees can be fragile, with specific tree structures often not being robust to small changes to datasets. Trees are presented here in a predictive context that provides valuable direct interpretability rather than the context of statistical inference. However, the

consistent losses in accuracy across jack-knifed trees when risk factors featuring in the final tree are removed and the finding that single risk factors alone cannot explain full accuracy observed within the final tree (Figure 2.3) provide rudimentary statistical support for the tree model. Virulence predictors were also validated using a logistic regression model, which supported the risk factors and directionalities observed within the classification tree (Table A.5).

There is also potential for confounding within the virulence risk factors I report, as virulence varies with many traits not included in this analysis. For example, although severity of Lassa virus disease superficially varies between infection routes, these differences correspond to variation in geography, which may be due to spatial variation in genotype (Howard 2009). As well as correcting for any broad phylogenetic signal of virulence, the contribution of taxonomic family to the classification tree is likely to have also acted as a proxy for some unmodelled viral traits, particularly molecular characteristics. Many potential confounders explain finer variation of virulence and were not testable at this ecological scale. Virulence can vary substantially between strains of the same virus species, and inference at this resolution would benefit from sequence-based phylogenetic analyses, with the potential to additionally base predictions on genetic markers of virulence. Within individual hosts, clinical symptoms often depend on host traits such as immunocompetence, age, or microbiome (Franco et al. 2003; Mackinnon et al. 2008). My risk factor analysis brings a novel, top-down perspective on virulence at the broadest level, though caution must be exerted in extrapolating the risk factors I find to dynamics of individual infections.

2.5.5. Implications for public health

Care should be also taken in interpreting my predictive model as a measure of population risk. I defined virulence for typical individual infections, independent of prevalence, incidence and population context. For example, I defined case fatality ratios (CFR) of ≥5% as 'severe', though for influenza the most severe category is

defined in recent pandemic severity indices as having CFR >1-2% (Reed et al. 2013). It is therefore important to distinguish clinical virulence from impact; 1-2% CFR would still cause at least 0.9-1.8 million deaths within the USA based on influenza incidence rates (CDC 2007) and would be more of a public health priority than a 'severe' virus with CFR ≥5% that infects many fewer people each year. However, this does not necessarily detract from the value of my model as a predictive tool for newly reported viruses, where ultimate incidence rates are difficult to predict in early emergence.

This work suggests directions for future public health initiatives. Firstly, I have demonstrated how a comparative approach could be used to predict virulence of newly emerging pathogens. This work brings a novel focus that complements comparative models surrounding other aspects of the emergence process, such as those predicting zoonotic transmission from certain hosts (Cleaveland et al. 2001; Pedersen and Davies 2009) or within certain geographic hotspots (Jones et al. 2008). The value of empirical modelling as an inexpensive and rapid tool during early emergence is increasingly being recognised, and similar functional models have been constructed to identify causative pathogens from early outbreak data (Bogich et al. 2013). Secondly, there are growing calls for predictive ecological studies to shape surveillance or intervention strategy of candidate emerging zoonoses (Pulliam 2008; Daszak 2009; Morse et al. 2012). Current surveillance initiatives have just begun to target specific non-human hosts and locations based on empirical studies (Morse et al. 2012). Virulence has already been suggested as a factor that could direct surveillance strategies for viruses, though this has only been explored with respect to virulence in non-human hosts (Levinson et al. 2013). The virulence risk factors I find may suggest that preferentially targeting the nonvector-borne viruses of ecological systems and/or tailoring detection assays towards certain virus families (e.g. *Arenaviridae* or *Rhabdoviridae*) could contribute to a viable strategy to detect future virulent zoonoses.

2.6. Conclusion

This work joins a series of comparative and predictive modelling surrounding emerging infectious diseases. Here, I contribute a novel focus in ecological predictors of virulence of human RNA viruses, which can be combined in holistic frameworks with other models such as those predicting emergence dynamics. As a predictive model, the featured classification tree can offer valuable inference into virulence of newly emerging infections. I propose that future predictive studies and preparedness initiatives with respect to emerging diseases should carefully consider potential for human virulence.

Chapter 3. Evolutionary routes of RNA virus emergence and human adaptation

3.1. Abstract

The majority of human RNA viruses are zoonotic in origin, although the mechanisms of adaptation involved in the emergence of zoonotic pathogens are not well understood. Conceptual models often assume a 'stepwise' route where infectivity and transmissibility develop in distinct stages. However, empirical observations suggest some viruses are capable of 'off-the-shelf' jumps, being sufficiently adapted for epidemic transmissibility immediately after zoonotic transmission. Here, I aim to further understand and characterise these viral routes to human adaptation and their determinants. Firstly, I use state-switching models to assess support for both stepwise and off-the-shelf routes to human adaptation, measured using the Pathogen Pyramid level schema. State-switching models were fitted to existing RNA virus phylogenies where available, and additionally, a cladogram based on taxonomic structure. Secondly, I use phylogenetic comparative analyses to identify whether transmission route and genomic guanine-cytosine content can predict human adaptation. State-switching models variously supported the stepwise or off-the-shelf routes for different RNA virus phylogenies, though for several taxa, neither route was well-supported exclusively, instead resembling a mixture of both stepwise and off-the-shelf movements. When analysed across a broad taxonomic cladogram, mixed dynamics were also observed. Correcting for phylogenetic signal, respiratory and vector-transmitted viruses were more likely to reach greater levels of human adaptation. This suggests that viruses follow a variety of adaptive trajectories, though future public health threats may be broadly predictable regarding certain groups, e.g. nonhuman paramyxoviruses, transmissible human alphaviruses or flaviviruses.

Improved sequence data and higher-resolution phylogenetic inference will be able to build on this approach and further elucidate patterns of viral adaptation.

3.2. Introduction

RNA viruses are consistently overrepresented among emerging human pathogens (Taylor et al. 2001; Woolhouse and Gowtage-Sequeria 2005). As a result of their high rates of replication and lack of proofreading mechanisms, RNA viruses are generally established to have high rates of mutation (Belshaw et al. 2008; Parrish et al. 2008), implying potential for both wide genetic diversity within existing hosts and rapid adaptation to novel hosts. It follows that the majority of human RNA viruses originate from zoonotic spillover or host shifts (Woolhouse and Gowtage-Sequeria 2005). Although many well-studied examples of host shifts have occurred comparatively recently, potential ancient host shifts have been identified through comparative evolutionary methods (Kitchen et al. 2011; Drexler et al. 2012; Longdon et al. 2015b). However, precisely how viral adaptation to novel hosts occurs during host shifts remains poorly characterised, as does whether viruses with different traits follow different adaptive trajectories.

Host shifts have been conceptually modelled as following distinct stages from original host to novel host in a variety of schemata (Childs et al. 2007; Wolfe et al. 2007; Lloyd-Smith et al. 2009). Here, I focus on one such schema, the Pathogen Pyramid model (Woolhouse and Gaunt 2007), and use its criteria to represent RNA virus adaptation within humans. The Pathogen Pyramid is based on transmissibility, and I follow previous suggestions that "human-adapted" might describe a virus capable of sufficient human-to-human transmission to persist in human populations in the absence of any potential nonhuman hosts (Woolhouse et al. 2013).

Viruses present only in animal hosts and not infecting humans are considered "level 1" (Figure 3.1A), whilst those infecting humans from a zoonotic source that have no onward human-to-human transmissibility are considered "level 2". Viruses with human-to-human transmissibility are considered "level 3" if transmission is

self-limiting, usually observed as single events or short chains, or "level 4" if transmission is self-sustaining, which includes fluctuating epidemics or constant endemic transmission. Levels 2, 3, and 4 can be considered as equivalent to values of $R_0 = 0$, $0 < R_0 < 1$, and $R_0 \geq 1$ respectively, where R_0 denotes the basic reproductive number, i.e. the number of secondary cases expected to result from a single primary case among an entirely susceptible population. However, in practice R_0 can be difficult to estimate and is not precisely calculated for the majority of human RNA viruses (Hay et al. 2013, Woolhouse et al. 2016). The pyramid shape reflects the presence of biological "barriers" that viruses with insufficient levels of adaptation may not be able to surpass, for example, host immune response, cell receptor availability, or potential to exit the host (Kuiken et al. 2006). Therefore, a fraction of viruses is prevented from moving upwards at each level, resulting in decreasing numbers of viruses between successive stages of adaptation; only 47 of the 180 human-infective RNA viruses are considered to have reached level 4 (Woolhouse et al. 2013).

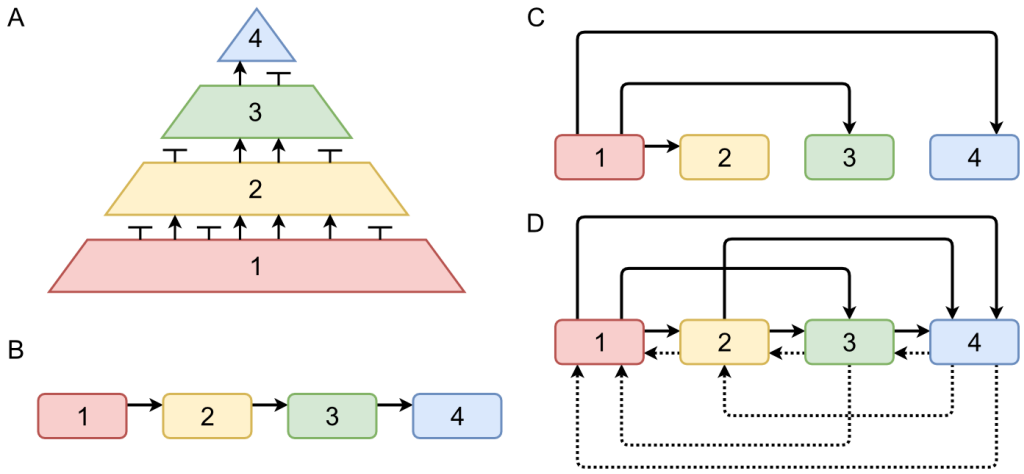


Figure 3.1. Potential models of RNA virus emergence and adaptation in humans. A) The 'Pathogen Pyramid' model, adapted from (Woolhouse and Gaunt 2007), where viruses transition between levels of adaptation (1 – animal only, no human infection; 2 – human infection, no human-to-human transmission; 3 – self-limited human-to-human transmission; 4 – sustained human-to-human transmission) with barriers (blocked arrows) meaning increasingly fewer viruses transition to increasingly higher levels. B) and C) Using Pathogen Pyramid level definitions, hypothesised models of adaptation based on evolutionary routes suggested by theoretical/empirical literature, where B) is the "stepwise" model, and C) is the "off-the-shelf" model. D) Complete RJ-MCMC model space with all possible transitions, including backward transitions (dashed lines).

A key question for disease ecology research as well as for public health is how viruses traverse adaptive landscapes to adapt to novel hosts and become established (i.e. ascend the levels of the Pathogen Pyramid). Several routes to adaptation have been proposed. Generally, schemata of disease emergence assume that adaptation occurs in a stepwise fashion (Figure 3.1B), with selection pressures acting on viruses to induce genetic change successively at each level within humans as the novel host (Pepin et al. 2010). Although the distance between steps is often represented as equivalent, the magnitude of genetic change required for each level of adaptation may vary greatly.

Empirically, a different pattern is often observed among emerging diseases (Woolhouse et al. 2016), where viruses seemingly "jump" straight from the animal

host at level 1 to higher levels of human adaptation (Figure 3.1C), with many high profile examples including SARS coronavirus (Holmes 2005), MERS coronavirus (Breban et al. 2013) and 2009 A-H1N1 influenza virus (Garten et al. 2009). This movement has been previously referred to as an "off-the-shelf" dynamic (Woolhouse and Antia 2007). These off-the-shelf jumps suggest that genetic variation within an animal host can generate a virus that transmits to humans already having sufficient genetic capability for human-to-human transmission (Pepin et al. 2010). Genetic evidence for such pre-adapted genotypes has been identified in sequences of rabies virus during host shifts from bats to carnivores (Kuzmin et al. 2012), though no broad-scale comparative studies have addressed off-the-shelf adaptation to my knowledge (but see Pepin et al. 2010 for a critical assessment for several viruses).

Phylogenetic state-switching models have been used in pathogen evolution studies to demonstrate switching from one host to another over viral phylogenies (Kitchen et al. 2011), but can also be used to test hypotheses regarding routes of movement between states in conceptual multi-state models (Pagel et al. 2004). To my knowledge, these methods have not yet been applied to phenotypes describing human adaptation, and here I aim to use this approach to determine relative support for stepwise or off-the-shelf models, or find the otherwise most likely route of adaptation.

I focus on evidence for these routes of adaptation among RNA viruses at a macroevolutionary resolution spanning taxonomic families and genera, with Pathogen Pyramid levels assigned to individual virus species. Although many viral host shifts involve adaptation at the finer resolution of genetic lineages or strains, traits such as human infectivity, primary host type and vector specificity show detectable phylogenetic signal between virus species (Grard et al. 2010; Kitchen et al. 2011; Longdon et al. 2015). This suggests that despite being prone to changes at the microevolutionary scale, the genetic capability for such traits remains heritable at the macroevolutionary scale. The specific macroevolutionary dynamics surrounding

development of human-to-human transmissibility have not been addressed in depth, although several empirical observations may be used to illustrate.

The alternative hypothesised routes of adaptation described would lead to different macroevolutionary patterns of Pathogen Pyramid level across viral phylogenies, subject to evolutionary rate. Specifically, the stepwise model would imply a highly-structured phylogenetic pattern, whereby common ancestors between viral species gradually progress through Pathogen Pyramid levels, resulting in a detectable gradient of levels along the phylogeny (Figure 3.2A). Although there are no fully-documented observations of a stepwise emergence owing to a rarity of visible level 2 to 3 movements (Woolhouse et al. in review), an appropriate case study might be HIV-1 and 2. These viruses diverged from zoonotic primate SIVs, which are a subset of a larger radiation of primate SIVs that have not been observed to infect humans to date (Hahn et al. 2000; Woolhouse et al. 2016). In contrast, the off-the-shelf model would imply a much more erratic phylogenetic distribution, where virus species with higher Pathogen Pyramid levels arise amidst clusters of level 1 viruses (Figure 3.2B). A typical example is SARS coronavirus, which rapidly adapted to humans following exposure to intermediate hosts (Holmes 2005; Song et al. 2005), yet its closest phylogenetic relatives are numerous ostensibly bat-specific level 1 coronaviruses (Tang et al. 2006).

Aside from the route taken, the ultimate extent of human adaptation that viruses can reach is thought to depend on several types of viral traits (Woolhouse et al.; Pulliam 2008). These include ecological traits, such as host range and transmission route (Woolhouse 2002; Woolhouse and Adair 2013) and molecular or genetic factors, such as cellular replication site, genome segmentation, and nucleotide biases (Pulliam 2008; Pulliam and Dushoff 2009). In this chapter, I focus on one example of each (transmission route, and nucleotide biases, respectively).

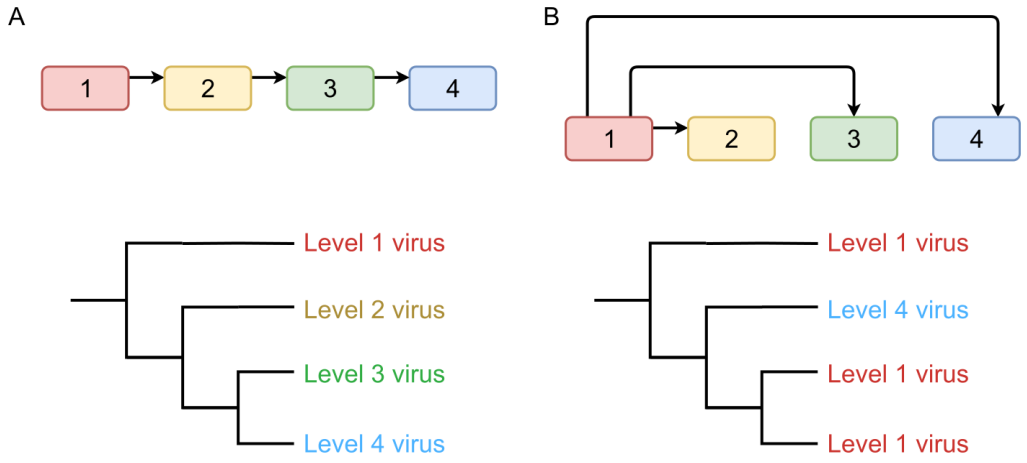


Figure 3.2. Hypothesised models of RNA virus adaptation in humans with corresponding example phylogenies underneath. Phylogenies illustrate potential evolutionary patterns of phenotypic Pathogen Pyramid level between virus species conforming to routes described by A) the stepwise model, B) the off-the-shelf model.

For transmission route, transmission via direct contact has been hypothesised to present the greatest risks of host shifts, involving transmission potential via a range of bodily tissues or fluids with little need for external survivability (Pulliam 2008). However, vector-borne viruses have been shown to be more likely to be zoonotic and/or emerging (Taylor et al. 2001; Woolhouse 2002; Loh et al. 2015). Specifically considering human-to-human transmissibility (levels 3 and 4), preliminary studies suggest vector-borne RNA viruses to be less likely to develop human-to-human transmissibility than those using other routes (Woolhouse and Adair 2013; Woolhouse et al. 2013; Geoghegan et al. 2016). Other comparative work has shown no conclusive single transmission route to act as a risk factor among disease emergence events (Loh et al. 2015), however to my knowledge, no such analyses have corrected for phylogenetic signal in how transmission route might contribute to different phenotypic stages of human adaptation.

Viral nucleotide biases may also play a role in host adaptation. In comparative studies, RNA viruses appear to have similarity in frequency of guanine-cytosine (GC) content to that of their specific host genomes (Bahir et al. 2009). Furthermore,

viruses may adapt to "match" host GC content during host shifts, e.g. human influenza viruses appear to exhibit directional evolution to reduce GC content and frequency of CpG motifs compared to ancestral avian influenza viruses, resembling a level closer to that of the human genome and potentially evading innate immune recognition by mimicry (Rabadan et al. 2006; Greenbaum et al. 2008). This suggests greater risk of human adaptation for viruses with similar GC content to humans, which is notably lower than that of many other mammals (Romiguier et al. 2010).

Based on the above, I hypothesise that viruses transmitted via direct contact and viruses having lower GC content will have increased risk of human adaptation. I also hypothesise that viruses transmitted via vectors will specifically exhibit increased risk of human infection, but decreased risk of human transmissibility.

I aim to test hypotheses surrounding evolutionary routes of RNA virus adaptation to humans across different viral taxa, taking Pathogen Pyramid level as a measure of adaptation. To do this, I source existing phylogenies of virus families or genera where available. I then estimate rates and find consensus models of switching between Pathogen Pyramid levels by adopting a phylogenetic comparative approach. I also apply state-switching models to a wider, taxonomically-structured cladogram as a basic first estimate of routes of adaptation across a complete mammalian and avian RNA virus phylogeny. Finally, I use phylogenetic regression techniques based on this cladogram to understand whether ecological or genetic traits of virus species may predict the extent of their human adaptation, correcting for signal from taxonomic relatedness.

3.3. Materials and methods

3.3.1. Pathogen Pyramid level data

For each of the 180 human-infective RNA virus species identified from previous standardised literature search efforts (Woolhouse et al. 2013), I critically assessed evidence for human transmissibility. I assigned Pathogen Pyramid levels through extensive review of literature, including structured searches and reference

tracing, as described in Woolhouse et al. (2016). Sources were identified through search terms: [virus name term, including all synonyms sourced from ICTV 9th Edition (King et al. 2011)] AND ["human to human" OR "person to person" OR interhuman OR communicab* OR transmi* OR outbreak OR cluster OR household OR iatrogenic OR urban cycle]; indexes used: Google Scholar, ProMED-mail (Yu and Madoff 2004); as well as original references from Woolhouse et al. (2013). R_0 estimates compiled by Hay et al. (2013) were also consulted. In assigning Pathogen Pyramid levels 3 and 4, evidence was accepted as sufficient where literature sources suggested that human-to-human transmissibility is suspected, though not confirmed (e.g. spatial or contact-based clusters of cases without diagnostic confirmation); or possible, though not directly observed (e.g. very large outbreaks in urban communities with few potential animal hosts).

I also identified those level 1 viruses that are known to infect mammals or birds for inclusion in analyses as the most likely sources of zoonotic viruses (Woolhouse and Adair 2013). To represent these viruses, data was supplemented with a list of mammal-infective RNA viruses supplied by the EcoHealth Alliance (Olival et al. in review), collected via structured literature searches (search terms: [virus name including synonyms]; sources used: Web of Knowledge, Google Scholar, Wildlife Disease Association meeting abstracts, Global Mammal Parasite Database (Nunn and Altizer 2005); protocol further described in Levinson et al. 2013). I then supplemented data with level 1 bird-infective viruses obtained via structured literature searches (search terms: [virus name including synonyms] AND ["bird" OR "avian" OR "host range" OR "animal" OR "reservoir"]); sources used: Web of Knowledge, Google Scholar, Scopus). All viruses were standardised to species using ICTV 9th Edition Virus Taxonomy (King et al. 2011), resulting in 374 viruses across 65 genera and 19 families (excluding "unassigned" as a genus or family).

3.3.2. Virus phylogenies and cladogram

To investigate how Pathogen Pyramid level switching and routes of adaptation differ between virus taxa, I sourced existing phylogenies for several families and genera where available, and selected those including species representing at least three different Pathogen Pyramid levels: families *Picornaviridae* (Lu unpublished), *Rhabdoviridae* (Longdon et al. 2015b), and *Paramyxoviridae* (Kitchen et al. 2011), and genera *Flavivirus* and *Alphavirus* (Kitchen et al. 2011), which represent the vast majority of taxonomic richness in their respective families. Sequence tips in these phylogenies were matched to ICTV 9th Edition virus species present in the mammal and bird virus dataset and dropped from the tree if no match could be made. If more than one sequence matched to the same species, a single sequence was randomly chosen to represent that species and all other sequences were dropped. All phylogenies were rooted as per the methods in their respective sources.

Although possible for other types of pathogen (McNally et al. 2014), construction of a complete RNA virus phylogeny is precluded by difficulty in accurately assigning deeper evolutionary branches, as a result of the extreme genetic divergence between families (Holmes 2003b). Following Kümmerli et al. (2014), I therefore constructed a viral cladogram as a basic proxy in lieu of a full mammalian and avian RNA virus phylogeny. The cladogram was assembled based on consecutive divisions of: genome type, taxonomic family, taxonomic genus and taxonomic species, where all sister clades were left as unresolved polytomies (Figure 3.3). I assumed each RNA virus genome type is monophyletic. Although viral monophyly is notoriously hard to determine given the difficulties in estimating ancient viral divergence (Holmes 2003b, 2011), conservation of gene homologues and replication strategies implies monophyly within certain viral genome types (e.g. single-stranded positive-sense RNA viruses) or collections of taxonomic families (Koonin et al. 2006, 2015).

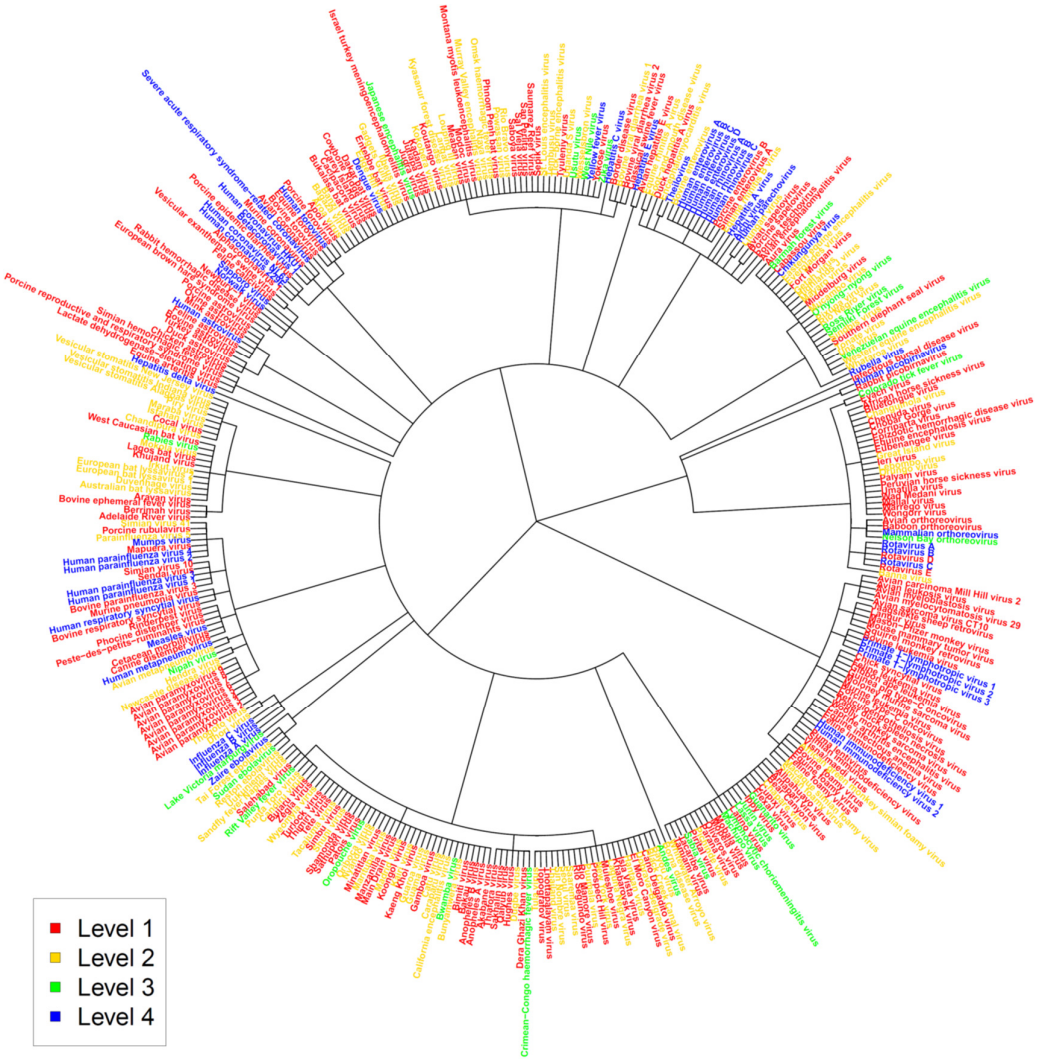


Figure 3.3. Taxonomically-structured cladogram of RNA virus species infecting mammals or birds, using basic branch length assumption set (a), where branch lengths are arbitrarily specified such that deep taxonomic branches (root to genome type, genome type to family) are ten times the length of shallow taxonomic branches (family to genus, genus to species). Colour of species tips represent Pathogen Pyramid level.

Branches were assigned arbitrary lengths where deeper taxonomic branches (those from root to genome type and genome type to family) were ten-fold the length of shallower taxonomic branches (those from family to genus and genus to species) (Figure 3.3). These orders of magnitude between lineage diversification times are consistent with estimates for several RNA virus taxa (Fargette et al. 2008; Li et al.

2015). However, estimation of virus divergence times is often contentious, and there are recent suggestions that within-family diversification (i.e. my shallow taxonomic branches) may be much more ancient among some vertebrate RNA virus families (Taylor et al. 2014). Therefore, in addition to this basic cladogram (a), I tested robustness of state-switching models to branch length assumptions by applying further cladograms with alternative branch length assumption sets where: b) deep taxonomic branches were equal lengths to shallow taxonomic branches; c) deep taxonomic branches were a hundred-fold the length of shallow taxonomic branches; d) all branch lengths were scaled using Grafen's (1989) method for unknown phylogenetic distances, where ages of branching times were proportional to the number of ultimate descendent nodes. All cladograms were rooted at the ancestral node between genome types All phylogenetic construction and manipulation was carried out using package 'ape', v3.2 (Paradis et al. 2004) in R, v3.1.3 (R Development Core Team 2015).

3.3.3. State-switching modelling analysis

State-switching models were implemented using the Multistate function of BayesTraits, v2.0 (Pagel et al. 2004). BayesTraits uses Markov chain Monte Carlo (MCMC) methods to estimate parameters describing rates of transitions between discrete states (in this case, Pathogen Pyramid level) over a phylogenetic tree. As a basic comparison between hypothesised models, posterior likelihoods in the form of estimated harmonic mean were compared between the stepwise and off-the-shelf models (Figure 3.1B, 3.1C), which were specified by restricting the relevant transition parameters to zero (stepwise model: levels 1 to 3, 1 to 4, 2 to 4; off-the-shelf model: levels 2 to 3, 2 to 4, 3 to 4).

To find the most parsimonious model from the entire potential model space (Figure 3.1D), I then used the reversible jump MCMC method (RJ-MCMC) (Pagel and Meade 2006). RJ-MCMC methods flexibly move around the entire model space between iterations by jumping between different model dimensionalities according

to a Markov chain process. Subsequently, RJ-MCMC methods collapse model dimensionality by reducing number of parameters used to an optimum, through fixing multiple transition rates as having the same parameter, or to be zero. For all RJ-MCMC analyses, parameter configurations with at least 5% frequency among all iterations in the RJ-MCMC chain were examined and from these, the consensus model was accepted as the most common parameter configuration. No consensus was thus identified among models over the *Alphavirus* and *Flavivirus* phylogenies (highest model frequency 2.13% and 1.87%, respectively). Although a 5% frequency across all RJ-MCMC iterations may appear low, this still represents substantial support as the potential model space is very large; the number of potential models featuring up to thirteen unique rate parameters (considering fixation at zero as a possible rate parameter) can be considered a Bell number, which describes the number of possible partitions of a set (Wilf 1990) and is calculable recursively as:

$$B_{13} = \sum_{k=0}^{12} \binom{12}{k} B_k = 27644437 \qquad (3.1)$$

Assuming all possible models are equally likely and sampled iterations are independent, the probability of observing a specific parameter configuration at a given iteration is therefore $p = \frac{1}{B_{13}} = 3.62 \times 10^{-8}$, and the probability of observing any parameter configuration at a frequency of at least 5% among n sampled RJ-MCMC iterations follows a cumulative binomial distribution:

$$P(x \geq 0.05n) = \sum_{k=0.05n}^{n} \binom{n}{k} p^k (1-p)^{n-k} \qquad (3.2)$$

As 9950 iterations were sampled in all RJ-MCMC analyses, $n \ll B_{13}$, and the expected number of times any specific parameter configuration would be seen is zero. The probability of observing a model at 5% frequency or higher under this null distribution becomes infinitesimally small for such a small p and large n.

Unless fixed at zero, rate parameters in all state-switching models were assigned an exponential prior with parameter sourced from a uniform hyperprior between 0 and 2. All RJ-MCMC chains were run for 10^7 iterations, discarding the first 5×10^4 as burn-in and sampling every 1000^{th} iteration. Parameter traces were visually inspected to confirm convergence. In all cases, five separate RJ-MCMC runs were conducted to confirm replicability and the run with highest marginal likelihood (as measured by estimated harmonic mean) was retained for further analysis.

Finally, to further assess support for hypotheses regarding routes of human adaptation (Figure 3.1), I used Bayes Factors, which were calculated as the ratio of posterior odds to the prior odds. In this context, the prior odds are the odds of observing the specific hypothesised model out of all possible models, which I calculated as a binomial process, considering each of the twelve model rates as being either zero or nonzero. The stepwise model (Figure 3.1B) was defined as transition rates from Pathogen Pyramid levels 1 to 2, 2 to 3, and 3 to 4 being nonzero and 1 to 3, 1 to 4, and 2 to 4 being zero. Inversely, the off-the-shelf model (Figure 3.1C) was defined as rates from level 1 to 2, 1 to 3, and 1 to 4 being nonzero, and rates from 2 to 3, 2 to 4, and 3 to 4 being zero. In preliminary runs, loss of function was commonly seen in state-switching models featured in RJ-MCMC chains. Therefore, I made no assumptions on loss of function in assessing support for hypothesised models and calculate Bayes Factors based on these forward transition parameters only. The probability of these six rate parameters following either exact hypothesised model are $\left(\frac{1}{2}\right)^6$, giving prior odds of $\frac{\left(\frac{1}{2}\right)^6}{1-\left(\frac{1}{2}\right)^6}$. Prior odds calculations were adjusted accordingly in cases where virus phylogenies used did not contain all four Pathogen Pyramid levels. The posterior odds are the odds of observing the exact hypothesised model within the posterior models featured throughout the sampled RJ-MCMC iterations. Strength of support provided by Bayes Factors is generally accepted as negligible for values <1, minor for values from 1 - 3, moderate for values from 3 - 10, strong for values from 10 - 100, and very strong or decisive for values > 100 (Kass and Raftery 1995).

3.3.4. Phylogenetic comparative analysis

To determine whether ecological or genetic traits may be associated with Pathogen Pyramid levels, data were sourced on transmission route and GC content for each virus species. These traits were selected based on data availability and demonstration of substantial within-taxonomic family variation, as traits that are heavily nested within a phylogenetic structure are difficult to estimate during phylogenetic comparative analysis. Transmission route was defined as the route that viral transmission primarily occurs by, regardless of host type, and was classified as either 'direct contact' (including aerosol material that is directly inoculated), 'respiratory', 'faecal-oral', or 'vector-borne' (excluding mechanical transmission). Transmission route data was collected via consultation of clinical virology sources (Knipe and Howley 2007; Richman et al. 2009; Zuckerman et al. 2009) and structured literature searches (search terms: [virus name including synonyms] AND [transmi* OR *borne OR vector]; source used: Google Scholar). Viral GC content was calculated using sequences obtained from GenBank (Benson et al. 2005) for the type isolates of each virus species, as designated in the ICTV 9th Edition (King et al. 2011). Virus species with missing transmission route or GC content data were excluded from phylogenetic regression analyses, leaving n = 221.

To correct for potential phylogenetic signal in testing for associations between virus traits and Pathogen Pyramid level, I constructed Bayesian phylogenetic mixed regression models with MCMC implementation using the 'MCMCglmm' R package, v2.21 (Hadfield 2010). Phylogenetic regression models were specified with a multinomial error structure, separately comparing likelihood of reaching Pathogen Pyramid levels 2, 3, and 4 compared to a baseline of level 1. Phylogenetic covariance in Pathogen Pyramid level and virus trait predictors was corrected for by specifying the mammalian and avian RNA virus cladogram with assumption set (a) as a random effect in lieu of a complete phylogeny. Viruses excluded due to missing data were

dropped from the cladogram. GC content (%) was modelled under a log-transformation.

Phylogenetic regressions were run with inverse Wishart priors with V = 1 on the random term (RNA virus cladogram) and flattened Gelman priors on the fixed terms (multinomial intercepts, transmission route, log(% GC content)) with a scale of $2 + \pi^{\frac{2}{3}}$, using the 'gelman.prior' function in package 'MCMCglmm', following Gelman et al. (2008). Residual variance was fixed at 1. Phylogenetic regressions were run for 5 x 10^6 iterations, retaining every 1000th iteration, discarding the first 1.25 x 10^6 as burn-in. Convergence was confirmed by inspecting trace output, and suitability of priors was confirmed by inspecting posterior estimates of phylogenetic covariance and confirming values of latent variables were consistent with logit function boundaries. All phylogenetic regression was carried out in R, v3.1.3 (R Development Core Team 2015).

3.4. Results

3.4.1. State-switching models of RNA virus adaptation

I assessed whether the stepwise or off-the-shelf model (Figure 3.1B, 3.1C) best described evolutionary routes of human adaptation over phylogenies of different RNA virus families and genera. The stepwise and off-the-shelf models were both moderately supported for *Rhabdoviridae* adaptation, whereas the stepwise model was much more highly supported for *Alphavirus* and *Flavivirus* adaptation (Table 3.1). Neither model was well-supported for *Picornaviridae* and *Paramyxoviridae* adaptation, with both models receiving zero Bayes Factor support for the *Paramyxoviridae* family.

Table 3.1. Likelihoods and posterior support for hypothesised models of virus adaptation over existing family or genus-level phylogenies. Runs presented are those with the highest marginal likelihood. 'lnHM' = log(harmonic mean), 'BF' = Bayes Factor calculated as odds of observing given model in RJ-MCMC chain compared to naive odds of observing given model from all possible models.

	State-switching model		
Virus phylogeny	**Stepwise**	**Off-the-shelf**	**RJ-MCMC**
Picornaviridae	lnHM = -20.099 BF = 0.33	lnHM = -19.554 BF = 0.75	lnHM = -14.058
Rhabdoviridae	lnHM = -35.734 BF = 5.04	lnHM = -36.917 BF = 5.25	lnHM = -24.499
Paramyxoviridae	lnHM = -85.655 BF = 0	lnHM = -77.829 BF = 0	lnHM = -38.372
Alphavirus genus	lnHM = -36.402 BF = 12.32	lnHM = -48.182 BF = 0.48	lnHM = -30.238
Flavivirus genus	lnHM = -46.471 BF = 10.93	lnHM = -58.052 BF = 2.23	lnHM = -37.775

Examination of the models fitted in RJ-MCMC analyses revealed greater detail in routes of adaptation between virus taxa. The most frequent models accepted as consensus for the *Picornaviridae* and *Paramyxoviridae* families featured off-the-shelf transitions from Pathogen Pyramid level 1 to 4, as well as 2 to 4 (Figure 3.4, Table 3.2), as did all models with at least 5% frequency within RJ-MCMC iterations (Table B.1, B.3). Consistent stepwise gains of function in levels 1 through 4 were generally not seen in examined RJ-MCMC models, although models with least 5% frequency for the *Rhabdoviridae* family often featured a stepwise path from 1 to 3 (Figure 3.4, Table B.2). Although no single model had at least 5% frequency for the *Alphavirus* and *Flavivirus* genera, consistent stepwise progression from level 1 to 4 was much better supported than off-the-shelf movements when summarising over all

model iterations (Figure B.1). Notably, losses of function (i.e. backward transitions) were commonly fitted throughout all RJ-MCMC models. All models featured with at least 5% frequency were fitted with a single rate parameter (Table B.1 – B.3), though posterior mean values of this parameter showed substantial heterogeneity between viral families (Table 3.2). This may suggest different evolutionary rates of host adaptation between different viral taxa, though posterior credible intervals were wide for the *Picornaviridae* and *Rhabdoviridae* parameters (Table 3.2).

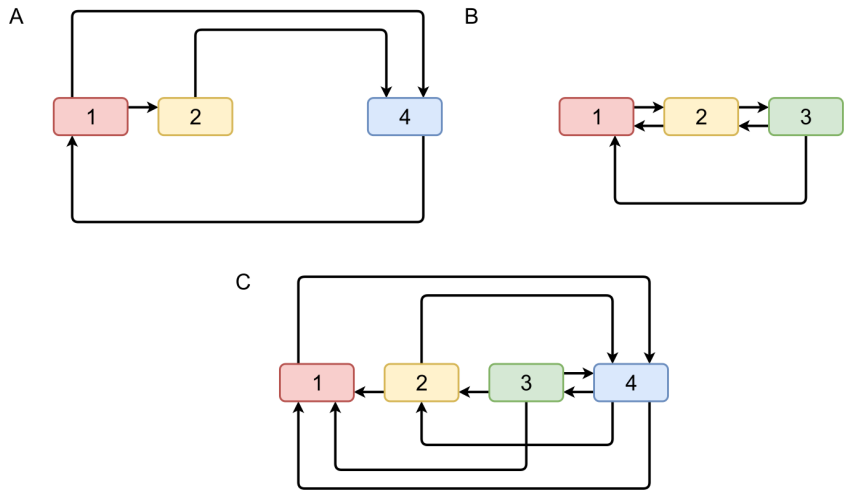


Figure 3.4. Consensus RJ-MCMC state-switching models of Pathogen Pyramid levels across those viral phylogenies having accepted consensus models with ≥5% frequency: A) family *Picornaviridae*, B) family *Rhabdoviridae*, C) family *Paramyxoviridae*.

Table 3.2. Consensus models of viral adaptation from RJ-MCMC analyses over existing family-level phylogenies, for those with accepted consensus of ≥5% frequency. Runs presented are those with the highest marginal likelihood. 'Freq.' = frequency of consensus model, 'No. param.' = number of parameters in consensus model, 'Parameter estimates' = individual posterior mean estimates and standard deviations for pyramid level switching rates in consensus model, with "-" denoting fixed at zero by the RJ-MCMC chain, and "X" denoting those absent due to the phylogeny containing no viruses with certain pyramid levels.

			Parameter estimates											
Virus phylogeny	**Freq.**	**No. param.**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
Picornaviridae	24.3%	1	1.673 ± 1.616	X	1.673 ± 1.616	-	X	1.673 ± 1.616	X	X	X	1.673 ± 1.616	-	X
Rhabdoviridae	29.3%	1	0.332 ± 0.002	-	X	0.332 ± 0.002	0.332 ± 0.002	X	0.332 ± 0.002	0.332 ± 0.002	X	X	X	X
Paramyxoviridae	13.9%	1	-	-	5.079 ± 2.423	5.079 ± 2.423	-	5.079 ± 2.423	5.079 ± 2.423	5.079 ± 2.423	5.079 ± 2.423	5.079 ± 2.423	5.079 ± 2.423	5.079 ± 2.423

3.4.2. State-switching models across an RNA virus cladogram

In testing which model best described adaptation across a preliminary, taxonomically-structured mammalian and avian RNA virus cladogram, both performed poorly with very little or no support (Table 3.3). Instead, the most frequent model accepted as consensus from the RJ-MCMC analysis suggested a mixture of dynamics, with stepwise-like transitions between Pathogen Pyramid levels 1 through 3, and off-the-shelf-like jumping from level 1 to 4 (Figure 3.5). This model was fitted with two parameters, the higher of which was assigned to transitions involving complete loss of human infectivity (levels 2 to 1, 3 to 1) (Table 3.4). Other models with at least 5% frequency were broadly comparable, differing only in a level 3 to 2 parameterisation (Table B.4), and the same mixture of stepwise and off-the-shelf movements was observed when summarising over all iterations (Figure B.2).

Repeated analyses specifying alternative sets generally supported the robustness of the RJ-MCMC analysis to assumptions surrounding cladogram branch lengths. Compared to the basic branch length assumption set (a), the same most frequent consensus model was obtained when using assumption sets (b), where deep taxonomic branches were equal lengths to shallow taxonomic branches, and (c), where deep taxonomic branches were a hundred-fold the lengths of shallow taxonomic branches (Table 3.4), with near-identical rate parameters. For these assumption sets, all models above with at least 5% frequency were also closely comparable (Table B.5, B.6). However, branch length assumption set (d) following Grafen's method gave slightly different models (Table 3.4, B.7), fitting an additional off-the-shelf transition (level 1 to 3) and an additional stepwise transition (level 3 to 4), suggesting higher variability in adaptive routes under this assumption set. The consensus model for this assumption set also featured substantially higher rate parameters (Table 3.4), reflecting the much shorter absolute lengths of shallow taxonomic branches.

Table 3.3. Likelihoods and posterior support for hypothesised models of virus adaptation over taxonomically-structured cladogram under branch length assumption set (a). Runs presented are those with the highest marginal likelihood. 'lnHM' = log(harmonic mean), 'BF' = Bayes Factor calculated as odds of observing given model in RJ-MCMC chain compared to naïve odds of observing given model from all possible models.

State-switching model	lnHM	Bayes Factor
Stepwise model	-528.645	0.013
Off-the-shelf model	-541.742	0
RJ-MCMC model	-393.399	NA

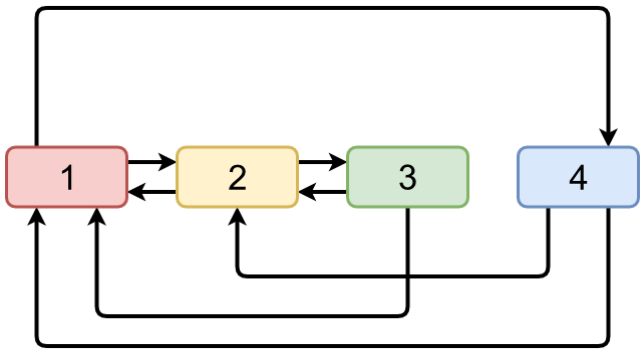


Figure 3.5. Consensus RJ-MCMC state-switching model of Pathogen Pyramid levels across taxonomically-structured cladogram, under branch length assumption set (a).

3.4.3. Viral trait associations with adaptation

Preliminarily correcting for phylogenetic covariance using the mammalian and avian RNA virus cladogram with assumption set (a), phylogenetic mixed regression suggested that Pathogen Pyramid level is strongly associated with transmission route. Specifically, respiratory and vector-transmitted viruses were more likely to reach levels 2 and 3 than directly transmitted viruses (Table 3.5, Figure B.3), although this effect did not extend to level 4. Additionally, there was limited evidence that viruses with greater GC content were more likely to reach levels 2 and 3, where 95% credible intervals marginally included zero (Table 3.5, Figure B.3).

Table 3.4. Consensus models of viral adaptation from RJ-MCMC analyses over taxonomically-structured cladogram under different branch length assumption sets. Runs presented are those with the highest marginal likelihood. 'Freq.' = frequency of consensus model, 'No. param.' = number of parameters in consensus model, 'Parameter estimates' = individual posterior mean estimates and standard deviations for pyramid level switching rates in consensus model, with "-" denoting fixed at zero by the RJ-MCMC chain

			Parameter estimates											
Assumption set	**Freq.**	**No. param.**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
a) Ten-fold deep taxonomic branch lengths	38.5%	2	0.203 ± 0.026	-	0.203 ± 0.026	0.673 ± 0.076	0.203 ± 0.026	-	0.673 ± 0.076	0.203 ± 0.026	-	0.203 ± 0.026	0.203 ± 0.026	-
b) Equal branch lengths	27.2%	2	0.203 ± 0.026	-	0.203 ± 0.026	0.675 ± 0.074	0.203 ± 0.026	-	0.675 ± 0.074	0.203 ± 0.026	-	0.203 ± 0.026	0.203 ± 0.026	-
c) Hundred-fold deep taxonomic branch lengths	35.1%	2	0.202 ± 0.026	-	0.202 ± 0.026	0.669 ± 0.076	0.202 ± 0.026	-	0.669 ± 0.076	0.202 ± 0.026	-	0.203 ± 0.026	0.202 ± 0.026	-
d) Grafen branch lengths	27.8%	2	4.066 ± 0.678	4.066 ± 0.678	4.066 ± 0.678	12.83 ± 1.865	4.066 ± 0.678	-	12.83 ± 1.865	12.83 ± 1.865	12.83 ± 1.865	12.83 ± 1.865	12.83 ± 1.865	-

Table 3.5. Posterior mean coefficients from multinomial phylogenetic MCMC regression for viruses with complete data (n = 221), with 95% credible intervals in brackets. '(intercept)' denotes log odds ratio of a virus having the specified Pathogen Pyramid level compared to level 1 (no human infection), 'TR:' denotes additional log odds of a virus having the specified Pathogen Pyramid level given the specified transmission route compared to a baseline of direct-contact transmission. 95% credible intervals around virus trait predictors that exclude zero are highlighted in bold.

	Pathogen Pyramid level		
Model component	**Level 2**	**Level 3**	**Level 4**
(intercept)	-28.64 (-67.54, 3.89)	-37.30 (-81.87, 0.83)	-2.69 (-27.40, 43.38)
TR: faecal-oral	-4.65 (-11.39, 1.08)	-3.36 (-10.79, 2.75)	-1.39 (-5.83, 1.22)
TR: respiratory	**3.18** **(0.69, 6.04)**	**4.45** **(1.65, 7.69)**	0.08 (-3.99, 2.95)
TR: vector-borne	**3.28** **(1.05, 5.93)**	**3.85** **(1.19, 7.13)**	-2.49 (-5.69, 0.66)
log(% GC content)	6.76 (-1.46, 16.42)	8.50 (-0.84, 20.21)	-0.93 (-11.73, 6.96)

3.5. Discussion

Using existing viral phylogenies, state-switching models show that host switching and emergence in humans variously follows a stepwise or off-the-shelf model for different RNA virus families and genera, although several virus taxa were not well-described by either model exclusively. The most prominent model fits as determined by Bayes Factors were the stepwise model to the genera *Alphavirus* and *Flavivirus*. This is consistent with the emergence dynamics of those human-adapted viruses in these genera such as dengue virus, yellow fever virus, and more recently, chikungyuna virus and Zika virus. Human infections with these viruses have been historically restricted to localised, sylvatic cycles before their progression to large-

scale epidemics and range expansion (Gubler 2004; Powers and Logue 2007; Weaver et al. 2016).

Using a taxonomic cladogram as a tentative substitution for phylogenetic relationships between all RNA viruses of mammals and birds, adaptation did not strictly follow a stepwise or off-the-shelf model and instead resembled a mixture between the two. This pattern was broadly robust to assumptions surrounding relative orders of magnitude between cladogram branch lengths (Table 3.4). The observed intermediate pattern between the stepwise and off-the-shelf models may have resulted from heterogeneity in routes across all virus families. It must be noted that many families featured in the cladogram were not included in the analyses of existing virus phylogenies, as they only contained species limited to specific Pathogen Pyramid levels. For example, the *Astroviridae, Caliciviridae, Coronaviridae, Hepeviridae* and *Picobirnaviridae* families were all limited to levels 1 and 4, with the latter two only containing two mammalian or avian-infective species (King et al. 2011).

The consistency in parameters between accepted consensus models under different branch length assumption sets (Table 3.4) implies that my state-switching models were independent of divergence time between genome types and taxonomic families, and that the majority of transitions between Pathogen Pyramid levels were fitted to occur within virus families. Parameters only varied under branch length assumptions using Grafen's (1989) method, where absolute lengths of within-family and within-genus branches were much shorter.

Correcting for phylogenetic signal, I found that viruses primarily having respiratory and vector-borne transmission to be more likely to infect and exhibit self-limited transmissibility between humans (Pathogen Pyramid levels 2, 3) compared to being restricted to non-human hosts (Pathogen Pyramid level 1). Vector-borne pathogens have been shown elsewhere to be more likely to be zoonotic (Woolhouse 2002; Loh et al. 2015) and additionally, to have a more generalist range of non-human hosts (Johnson et al. 2015a). This may imply vector mobility creates

transmission interfaces between individuals or populations that would not otherwise exist due to behavioural or geographic barriers (Rosenberg and Beard 2011). Any selection pressure upon a virus (or its vector) to maintain broad host exposure may preclude adaptation towards host specificity, which may explain the lack of association with reaching level 4 (see also Chapter 4). Respiratory viruses may also have a broad host exposure as they do not rely on close contact, and aerosolised material can persist outside of the host environment. However, it must be noted that the majority of viruses transmitted to humans by a true respiratory route (i.e. inhaled aerosol material) are arenaviruses and hantaviruses, though reservoir rodent hosts, transmission occurs via a different route - direct contact (Mills 2005), creating complexity in understanding routes to adaptation for these viruses. Although I addressed broad-level patterns, I also note that the influence of transmission route upon adaptation is likely to be highly contextual, depending on frequency of exposures and other viral traits such as environmental survivability (Pulliam 2008).

I also observed tentative evidence suggesting viruses with increasing GC content may be more likely to be human-infective and human-transmissible to a limited extent (Pathogen Pyramid levels 2, 3). This seems to contrast with previous observations that human viruses have decreased GC content and suppressed CpG motifs (Greenbaum et al. 2008; Bahir et al. 2009). However, GC content has also been associated with codon optimisation during viral replication in human cells (Auewarakul 2005), suggesting a more sophisticated structure of genetic biases are involved in human adaptation than simple nucleotide frequency. The evidence for an association between human adaptation and GC content may have been weak as differences in nucleotide biases are much less pronounced within vertebrate viruses than between vertebrate and invertebrate or plant viruses (Greenbaum et al. 2008; Bahir et al. 2009; Kapoor et al. 2010).

3.5.1. Sources of human-adapted viruses

A priority for current public health initiatives is the ability to predict and prepare for emerging viruses based on knowledge of which types of candidate virus are most likely to reach Pathogen Pyramid level 4, i.e. develop sustained or pandemic human-to-human transmissibility (Daszak 2009; Morse et al. 2012; Woolhouse et al. 2015). My analyses suggest that, generally, level 4 viruses may be sourced either via stepwise adaptation from existing human-infective viruses or directly off-the-shelf from zoonotic sources, and appear capable of following a wide variety of adaptive trajectories in between. Although specific routes towards human adaptation are therefore seemingly difficult to predict, it may be possible to predict likely adaptive movements by considering different virus taxa in detail.

Several families were observed to contain only level 1 and level 4 viruses, which may be indicative of off-the-shelf jumps, although the precise adaptive route taken in such cases cannot be inferred. However, evidence for the off-the-shelf route may be supported by wider genetic and epidemiological studies. Within the *Coronaviridae*, both SARS and MERS coronaviruses exhibited human-to-human transmission rapidly during emergence, and civet and human SARS coronavirus sequences appear to have only very few differences, concentrated in receptor-binding sites (Song et al. 2005). This implies that for SARS coronavirus, only a short adaptive distance needed to be traversed to enable emergence in humans.

Level 4 viruses were also consistently sourced by off-the-shelf jumps in models supported with at least 5% frequency for the *Picornaviridae* and *Paramyxoviridae* families (Figure 3.1, Table B.1, B.3). This observation may reflect the presence of highly host-specific, closely related human and nonhuman viruses within these families (e.g. human versus simian enteroviruses, human versus bovine parainfluenzaviruses). However, the family *Paramyxoviridae* also contains Nipah virus, which demonstrated human-to-human transmission immediately following emergence in Bangladesh (Epstein et al. 2006). Therefore, level 1 henipaviruses may represent strong candidates for off-the-shelf adaptation and are also a particular

concern for public health given their severe virulence. Contrastingly, for the *Alphavirus* and *Flavivirus* genera, visualisations over all RJ-MCMC iterations suggested level 4 can be reached by a continuously stepwise route of adaptation (Figure B.1). This suggests current level 3 alphaviruses or flaviviruses are also likely candidates for human adaptation, and it may be most pertinent to focus on those reported to have caused larger outbreaks, e.g. O'nyong-nyong virus (Lanciotti et al. 1998), Venezuelan equine encephalitis virus (Weaver et al. 1996).

When specifying a preliminary mammalian and avian RNA virus cladogram, off-the-shelf jumps were featured as the sole route to reaching level 4 among all models with a supporting frequency of at least 5% under all branch length assumption sets (Table B.4 – B.7), except a single model under assumption set (c) and all models under assumption set (d). In these cases, level 4 was attainable via both off-the-shelf jumps from level 1 and stepwise movements from level 3 (Table B.6, B.7). Over all examined models and branch length assumptions sets, movement from level 2 to 4 was never fitted (Figure B.2, Table B.4 – B.7), which supports observations that level 2 viruses rarely develop sustained human-to-human transmissibility without a considerable extent of adaptation (Woolhouse et al. 2016). Instead, sporadic human-to-human transmissibility via iatrogenic or otherwise anthropogenic means may be a more salient risk for level 2 viruses. To illustrate, the only observed human-to-human transmission of West Nile virus and rabies virus has been through blood products or organ transplants (CDC 2002; Srinivasan et al. 2005), which need not involve substantial genetic adaptation. However, RJ-MCMC models over several virus phylogenies did prominently feature level 2 to 4 transitions (Figure B.1, Table 3.2), further highlighting the importance of heterogeneity between taxa.

No viral traits investigated acted as risk factors for reaching level 4 in human adaptation, although vector-borne and respiratory transmission routes represented risk factors for reaching levels 2 and 3. Combined with the stronger support for the stepwise model in those vector-borne genera of *Alphavirus* and *Flavivirus* (Table 3.1),

this may suggest vector-borne viruses are more likely to adapt to humans in a stepwise manner, particularly if their vectors are anthropophilic (Woolhouse et al. 2016). This is also consistent with comparative genetic studies, which have demonstrated that surface proteins of vector-borne RNA viruses undergo adaptive evolution at a slower rate than those of nonvector-borne viruses (Woelk and Holmes 2002).

3.5.2. Evolutionary scope of analysis

As the methods I used follow an evolutionary perspective, inference of state-switching in my analyses is focused on the timescale of divergence between virus species, and assumes that Pathogen Pyramid level is a genetically-determined trait that does not vary within species. However, Pathogen Pyramid levels were classified based on ecological data (Woolhouse et al. 2016), as current knowledge of genetic markers of adaptation is limited. For many viruses, it is likely that the assigned levels do not reflect their genetic capability for human infectivity or transmissibility, but rather, limitations of their current ecological context. For example, changes to host or vector population dynamics can have major implications upon viral epidemiology, and may cause increases in phenotypic Pathogen Pyramid level without any genetic change (Pepin et al. 2010). Levels 3 and 4 are particularly difficult to distinguish, as viruses having an $R_0 > 1$ can exhibit varying sizes of transmission chains, subject to stochasticity and population heterogeneity (Woolhouse et al. 2016). As these contextual limits are not addressed by the scale of my analysis, this may have introduced error in fitted transition likelihoods.

A further challenge in inference of evolutionary dynamics from ecological data is that losses of function are particularly difficult to validate. Backward transitions frequently featured in consensus models, predicting that losses or switches of function should be relatively common (Figure 3.4, 3.5), i.e. some previously human-adapted viruses should have diverged into nonhuman viruses. These losses of function are difficult to confirm without sufficient understanding of

the genetic determinants of human infectivity, or reliable experimental models. To illustrate, SARS coronavirus is classified as Pathogen Pyramid level 4 though no human cases have been observed since 2004 (WHO 2004), and human-infective strains represented a very limited subset of the genetic diversity of SARS-like coronaviruses in wild hosts (Ren et al. 2006; Lau et al. 2005). The capability for human infectivity and transmissibility among those extant SARS-like coronavirus lineages in animal hosts remains unknown.

My analyses were focused at a macroevolutionary scale, though it must be acknowledged that many viral phenotypes can exhibit rapid changes at the microevolutionary scale, including human infectivity and transmissibility. Although the relationship between different scales of evolutionary dynamics of RNA viruses is largely unclear (but see Antonovics et al. 2015), the adaptive movements identified herein still contribute valuable understanding. These analyses may offer inference towards likely pyramid level transitions at other phylogenetic resolutions and predictions of observable viral emergence events for several reasons.

Firstly, although my scope of analysis covered RNA virus-wide cladograms and family-level phylogenies, the methodology used included some microevolutionary coverage. BayesTraits applies continuous rates of state switching along all branches of inputted trees, in this case, including those branches from genus nodes to species tips. Resulting state-switching models should therefore have at least in part reflected transition likelihoods of pyramid level switches within species. Secondly, what may be considered 'macroevolutionary' for RNA viruses may still have immediate public health relevance. RNA virus evolution is rapid, owing to high mutation rates and lack of proofreading mechanisms (Belshaw et al. 2008; Parrish et al. 2008), and it follows that divergence between viral species or even genera may potentially occur over relatively short timescales (Holmes 2003b). Finally, although dominated by host-virus codivergence, previous phylogenetic analyses have suggested that host shifts, particularly zoonotic host shifts, can be associated with macroevolutionary divergence and viral speciation (Kitchen et al. 2011). Even though

viral species are a somewhat artificial concept, this implies that macroevolutionary patterns have direct implications for the prediction of viral emergence in humans.

A critical next step will be to investigate routes of adaptation at a microevolutionary scale. Pathogen Pyramid level is already known to vary between intraspecific strains of some viral species (e.g. *Betacoronavirus 1* contains strains with host specificity for dogs, cats and swine, as well as humans), and for some viruses, very few genetic changes appear necessary to increase adaptation to humans (Holmes 2005; Li et al. 2005). This would require identification of infective and transmissible phenotypes of specific genetic lineages, which may only be possible through systematic sampling and sequencing efforts to strengthen quality of genetic evidence and markers of adaptation (Pepin et al. 2010). Such sequence-level analyses will be necessary to formally assess support for the routes of adaptation I prospectively identify and ultimately determine how macroevolutionary patterns relate to the likelihood of microevolutionary changes.

3.5.3. Analytical limitations

The analyses presented are intended as a first estimate of the comparative viral dynamics of human adaptation, and several limitations must be acknowledged. Firstly, the scope of available viral phylogenies and assumptions made to construct the wider virus cladogram restricted the phylogenetic accuracy of my analyses. The recent construction of a resolved phylogeny for all RNA viruses with single-stranded negative-sense genomes (Li et al. 2015) suggests that inference at a broader resolution will be feasible for future phylogenetic studies. Furthermore, I examined routes of adaptation within mammalian and avian RNA viruses only, as viral host shifts are overwhelmingly predicted from these host groups over other vertebrates or invertebrates (Woolhouse and Adair 2013; Woolhouse et al. 2013). However, mammalian and avian RNA viruses ultimately diverge from viruses infecting other host types, often more recently than divergence from other mammalian and avian viruses – of the 19 families containing mammalian or avian viruses, 12 also contain

virus species or strains capable of infecting other hosts, excluding arthropod vectors as hosts (King et al. 2011).

Secondly, methodological inaccuracies may have arisen from the use of cladograms featuring unresolved polytomies. Without the bifurcating structure and intermediary nodes of the true phylogeny, many more alternative patterns of transitions could have explained the distribution of Pathogen Pyramid levels at cladogram tips, which would have reduced precision around the fitted transitions. Furthermore, the use of polytomies is likely to have biased methods against detecting the stepwise route of adaptation. To illustrate, a phylogeny conforming to the off-the-shelf model would feature many transitions from level 1 to all other levels (Figure 3.2B). The ability of state-switching analyses to detect such transitions would not be impeded by collapsing this phylogeny to a polytomy, assuming the appropriate ancestral nodes were still assigned as level 1. However, a phylogeny conforming to the stepwise model would feature highly-structured transitions from level 1 through to level 4 (Figure 3.2A). Such structure would be lost when collapsing to a polytomy, resulting in reduced ability to detect all true transitions. Analyses conducted using the RNA virus cladograms therefore represented a more conservative test of the stepwise route of adaptation. Several stepwise movements were consistently observed in cladogram consensus models (Table 3.4), although a level 3 to 4 transition was typically absent and subsequently, little evidence was observed for this exact hypothesised route (Table 3.3).

As with any comparative dataset, there may also be inevitable biases or errors resulting from the data I use. For example, these analyses are based on known virus species, which are subject to ascertainment biases. Specifically, it is likely that many Pathogen Pyramid level 1 viruses are missing from the phylogenies and cladograms used, as very little of the viral diversity in wild mammal and bird hosts is currently known (Anthony et al. 2013; Cooper and Nunn 2013). Undersampling of level 1 viruses would result in fewer lineages with level 1 tips being present, and in terms of transition rate parameters, inflation in the estimated rates of transitions from level 1

to higher levels. Likewise, the estimated rates of backward transitions from higher levels to level 1 would also be inflated, which may explain the comparatively large rate parameters assigned to backward transitions in analyses over viral cladograms (Table 3.4). However, the above mechanism of bias assumes the ancestral states assigned by state-switching models to be fixed. In actuality, completeness of data regarding level 1 lineages is likely to also intricately affect estimation of ancestral states and subsequent transition likelihoods. Therefore, the effects of ascertainment bias are difficult to quantify, and the potential resulting biases either towards or away from the stepwise or the off-the-shelf models are not intuitive. The routes of adaptation I identify should therefore be continuously validated as knowledge of both the human and nonhuman virome improves.

3.5.4. Wider implications

My analyses describe adaptive routes in terms of phenotypes and can synergise with studies aiming to identify the specific underlying genetic changes associated with adaptation. However, current understanding of genetic changes involved in adaptation and at which loci is often limited, making stepwise and off-the-shelf routes difficult to distinguish in practice (Pepin et al. 2010). Additionally, genetic changes associated with adaptation can be highly variable, with many different potential protein or sequence alterations resulting in similar shifts in function (Streicker et al. 2012a). Even if changes can be identified, it is often difficult to determine whether these are essential components of adaptation or whether they adjust an existing phenotype (Pepin et al. 2010). In addition to informing genomic surveillance and further phylogenetic study as previously outlined, the results presented may also set up hypotheses to be tested in experimental studies - if human-adapted viral genotypes replicate sufficiently and persist in suspected reservoir hosts, this would suggest off-the-shelf jumps are feasible, whereas a lack of replication would suggest stepwise adaptation within human hosts and a consequent loss of fitness in other hosts.

3.6. Conclusion

Evolutionarily, RNA viruses may have the genetic capability to transmit from zoonotic sources to humans already having or rapidly developing epidemic potential, although the specific routes towards human adaptation are highly variable. However, dynamics of human adaptation are predictable in a broad, comparative perspective, considering virus taxonomy and transmission route. Future work will develop understanding of adaptive dynamics at higher phylogenetic resolution and further isolate the genetic changes involved in viral adaptation.

Chapter 4. Breadth and specificity of mammal host range predict dynamics of RNA virus emergence in humans

4.1. Abstract

RNA viruses show substantial variation in their host range, where closely related viruses may infect different host taxa. It is well-established that RNA viruses with a broad host range are more likely to infect humans, though less empirical work has addressed other aspects of viral emergence in relation to host range. Following suggestions from preliminary studies, I investigate whether specific RNA virus traits of human infectivity, transmissibility and virulence are associated with the taxonomic or phylogenetic breadth of their nonhuman mammalian host ranges, as well as infection of specific host taxonomic orders. I source data from externally compiled literature searches on known mammal host-virus relationships. I then use logistic mixed regressions to test whether several host range metrics correlate with viral traits whilst correcting for study effort and virus taxonomy. I find that viruses with a broad host range are less likely to exhibit human-to-human transmissibility, but more likely to be highly virulent. I also find that the ability to infect humans correlates with having mammal hosts that are phylogenetically close to humans, rather than host breadth outright. Specific host taxonomic orders also predicted virus traits, e.g. viruses with nonhuman primate hosts were more likely to be human-transmissible, and viruses with artiodactyl hosts were less likely to be highly virulent. These analyses shed further light on the complex relationships between host range and the evolution of emerging viruses. The influence of host range upon viral traits suggests that understanding viruses within the ecological contexts of their nonhuman hosts is critical for surveillance and control of emerging diseases.

4.2. Introduction

The majority of pathogens are known to infect multiple hosts (Cleaveland et al. 2001; Haydon et al. 2002; Fenton and Pedersen 2005). RNA viruses are no exception and are often cited to be the most likely pathogen type to infect multiple hosts, as a result of their affinity for host shifts through high mutation rates and/or lack of replication proofreading (Woolhouse and Gowtage-Sequeria 2005; Parrish et al. 2008). However, host ranges of pathogens exhibit substantial variation in their taxonomic diversity, specificity, and breadth, such that the observed host range of even superficially similar pathogens can be very different. Therefore, the study of pathogens within the context of their host ranges is critical to understanding dynamics of disease, including emergence in novel hosts (Karesh et al. 2012).

Host range is not simply just a product of pathogen characteristics, but instead, a dynamic trait thought to influence the evolutionary trajectories of pathogens. Selection pressures acting on pathogens will be differ between different host environments and may often be antagonistic (Elena et al. 2009). For example, a pathogen selected to efficiently bind to the cell receptors of one host may be less efficient at binding to more divergent receptors among other hosts (Crill et al. 2000). Host range has been observed to correlate with pathogen traits at varying scales, e.g. within hosts, molecular characteristics such as genome size (McNally et al. 2014); and between hosts, transmission rates and persistence dynamics (Dobson 2004). Here, I focus on whether host range correlates with traits of RNA virus species, using comparative analyses.

As the majority of human RNA viruses emerge from a mammal origin (Woolhouse and Adair 2013), I specifically focus on nonhuman mammalian host ranges of RNA viruses, and how they relate to traits of emergence. Previous empirical catalogues of human pathogens have shown viruses with a broad host range to be more likely to be emerging, or to infect humans via zoonotic transmission (Cleaveland et al. 2001; Woolhouse and Gowtage-Sequeria 2005). However, emergence in a novel host species is a multi-step process, involving several

mechanisms and viral traits (Wolfe et al. 2007; Woolhouse and Gaunt 2007; Lloyd-Smith et al. 2009). In addition to human infectivity (the ability to establish an infection within an individual), I use these multi-step schemata to identify human-to-human transmissibility and virulence as key features that both determine the success of RNA virus emergence and represent exceptional importance to public health. Zoonotic viruses show substantial variation in their ability to transmit between humans and cause human disease, and here I ask whether mammalian host range could explain this variation.

Although a broad host range is known to correlate with human infectivity, other measurable aspects of host range may also predispose RNA viruses to infect humans. Viral host shifts are likely to be increasingly successful with decreasing phylogenetic distance (and therefore, increasing biological similarity) between hosts (Wolfe et al. 2000), which has been demonstrated between nonhuman primates and humans (Pedersen and Davies 2009). Human infectivity may also correlate with infection of certain taxa noted to have high richness of zoonotic viruses (Luis et al. 2013). Therefore, I hypothesise that a nonhuman host range with broader taxonomic span, closer phylogenetic distance to humans, and infection of key host types will positively correlate with human infectivity.

To my knowledge, fewer studies have attempted to address how host range influences human transmissibility. However, a recent comparative analysis of 95 zoonotic RNA and DNA viruses reported that viruses infecting a broader range of host groups (combining taxonomic orders and ecology-based groups) were more likely to exhibit human-to-human transmissibility, and it was suggested that evolutionary potential to infect many different hosts also coincides with potential to develop transmissibility between them (Johnson et al. 2015a). Here, I aim to further investigate this finding within an exhaustive dataset comprising all known mammal-RNA virus relationships, by testing multiple standardised metrics of host range. I also distinguish self-limiting from sustained human-to-human transmissibility, as these have very different implications upon emergence. Following this reported finding, I

hypothesise that a broader nonhuman host range will positively correlate with both human-to-human transmissibility measures.

Several evolutionary theories have related host range to virulence. Assuming that an evolutionary trade-off exists coupling virulence and transmission rate (although this idea is heavily debated, see Alizon et al. 2009), higher virulence may be exhibited by pathogens with narrower host ranges as a consequence of efficient selection to increase transmission rate within the specialist host(s) (Leggett et al. 2013). Virulence can also be exhibited as a non-adapted coincident phenotype within 'dead-end' hosts that can be infected, but do not contribute to transmission (and therefore, present no opportunity for host-virus coevolution) (Levin and Svanborg Edén 1990; Bull 1994). Non-adapted virulence would be more likely for infections with generalist viruses, as a broad host range is more likely to include dead-end hosts (Leggett et al. 2013). However, no comparative analyses to my knowledge have investigated how host range correlates with virulence across virus species. In addition to host breadth, the relationship between phylogenetic distance between hosts and virulence has also not been fully addressed. Viruses may more easily exploit and cause disease within novel hosts that are more similar to those they have coevolved with, following the same logic as for infectivity outlined earlier. Alternatively, novel hosts may clear viruses that originate from related hosts more easily if immune strategies are phylogenetically conserved (Holmes and Drummond 2007). These uncertainties make precise hypotheses about nonhuman host range and virulence difficult to specify.

I aim to further understand the relationship between breadth and specificity of nonhuman mammalian host ranges of RNA viruses and their traits surrounding emergence in humans, specifically infectivity, transmissibility, and virulence. To quantify nonhuman mammalian host ranges and RNA virus traits, I use literature-sourced data on all known viral infections of mammal species and viral dynamics in human hosts. Firstly, I use permutation tests and matrix visualisations to investigate how these RNA virus traits are distributed across both virus and mammalian host

taxonomies. Secondly, I use mixed regression models to test whether RNA virus traits can be explained by taxonomic breadth of hosts, phylogenetic breadth of hosts, and infection of specific key host types. These comparative models can contribute to better understanding the biological process of emergence and predicting novel RNA virus dynamics.

4.3. Materials and methods

4.3.1. RNA virus trait and mammal host data

Data on mammal-infective RNA viruses and their known nonhuman host species were sourced from the EcoHealth Alliance (Olival et al. in review), and collected via structured literature searches (search terms: [virus name including synonyms]; sources used: Web of Knowledge, Google Scholar, Wildlife Disease Association meeting abstracts, Global Mammal Parasite Database (Nunn and Altizer 2005); protocol further described in Levinson et al. 2013). After standardising virus names to species level using ICTV 9th Edition Virus Taxonomy (King et al. 2011), these data spanned 281 viruses.

To classify which viruses are human-infective, I used a previous systematic literature review describing a catalogue of 180 known human RNA virus species (Woolhouse et al. 2013), which provided 50 additional viruses not present in the mammal-infective virus data. These 50 were therefore only known to infect humans among mammals and hereafter referred to as 'human-specialist', though the possibility these have nonhuman hosts yet to be discovered must be noted. These combined data spanned 331 virus species within 718 nonhuman mammal host species, across 1991 unique virus species-host species pairings. To classify human transmissibility, data was sourced from a comprehensive systematic review of RNA virus transmissibility potential (Woolhouse et al. 2016), and I assigned binary variables as to whether each virus has a) any human-to-human transmissibility (equivalent to an R_0 value of > 0), and b) self-sustained human-to-human transmissibility (equivalent to an R_0 value of ≥ 1). To classify human virulence, I

assigned a binary variable as to whether each virus typically causes 'severe' or 'nonsevere' disease in humans, where 'severe' met at least one of the following criteria: ≥5% case fatality ratio, frequent reports of hospitalisation, association with significant morbidity from a pre-specified list of conditions (haemorrhagic fever, seizures/coma, cirrhosis, AIDS, hantavirus pulmonary syndrome, HTLV-associated myelopathy) or were explicitly described as "severe" or "causing severe disease" in literature sources (see Chapter 2). Virulence in nonhuman hosts was not considered due to poor coverage of data on viral pathology in mammal hosts (Levinson et al. 2013).

I focused on virus host range dynamics among nonhuman mammals as no comparably thorough data currently exists describing bird-RNA virus relationships. Therefore, I excluded those RNA viruses exclusively using birds as reservoirs (or otherwise originating in birds) that simply 'spill over' into humans or nonhuman mammals without onward transmission. Viruses with bird reservoirs were assessed via structured literature searches (search terms: [virus name including synonyms] AND ["bird" OR "avian" OR "host range" OR "animal" OR "reservoir"]); sources used: Web of Knowledge, Google Scholar, Scopus) and consultation of previous efforts to assign RNA virus reservoirs (Kitchen et al. 2011). Viruses identified as having exclusive bird reservoirs (n = 21) or entirely unknown reservoirs (n = 6) were excluded from analyses.

To standardise data, virus families with less than 5 species present in the combined data were also excluded from analyses (*Arteriviridae, Bornaviridae, Hepeviridae, Picobirnaviridae*, and additionally, viruses unassigned to a family). Mammal host orders with less than 5 species present in the combined data were also excluded from analyses and host breadth calculations (Cingulata, Erinaceomorpha, Peramelemorphia, Proboscidea, Scandentia, Soricomorpha). Phylogenetic host range measures were not calculable for viruses that were only known to infect nonhuman mammal species which did not correspond to any Latin binomial name present in the mammal phylogeny used (Bininda-Emonds et al. 2007). To standardise

comparison between taxonomic and phylogenetic host range measures, these viruses (n = 15) were also excluded. These exclusions left 280 viruses within 683 nonhuman mammal hosts as the primary dataset, across 1702 unique virus species-host species pairings.

4.3.2. Host range calculations

I aimed to investigate whether viral traits of emergence could be explained by different aspects of mammalian host diversity by calculating the following metrics of host range per virus: a) total number of known nonhuman mammal host species, b) total number of known nonhuman mammal host orders, c) phylogenetic breadth of known nonhuman mammal host species, and d) minimum phylogenetic distance to humans among known nonhuman mammal host species. As the viral traits investigated referred to phenotypes in humans, I did not consider humans equivalent to other mammalian host species and therefore did not include humans in calculations of any host range metric. For host phylogenetic measures (c) and (d), I used the mammal supertree phylogeny of Bininda-Emonds et al. (2008). Phylogenetic breadth was calculated as the sum of branch lengths required to connect all nonhuman host species tips, i.e. how far viral infection is known to span across the mammal phylogeny. Phylogenetic breadth was defined as zero for human-specialist viruses. Minimum phylogenetic distance to humans was calculated as the minimum value among all pairwise patristic distances from nonhuman mammal host species to humans. As this measure was not intuitively definable for those human-specialist viruses, minimum phylogenetic distance to humans was analysed separately with a data subset excluding human specialists (n = 239). All phylogenetic manipulation and calculation was carried out using package 'ape', v3.2 (Paradis et al. 2004) in R, v3.1.3 (R Development Core Team 2015).

4.3.3. Statistical analysis

To investigate the distribution of viruses and viral traits across host-virus relationships, I constructed a matrix with rows denoting the 11 mammal host taxonomic orders and columns denoting the 14 virus taxonomic families within the primary dataset. Matrix values were assigned as the number of unique viruses in each family known to infect each mammal order. To test whether the observed host-virus relationships were non-random with respect to taxonomy, I created 999 matrix permutations using the function permatfull() in package 'vegan', v2.3-2 (Oksanen et al. 2015) in R, v3.1.3. Permutations were restricted to maintain fixed row totals to account for column-wise exclusivity, i.e. individual viruses can infect multiple host orders but cannot belong to multiple virus families. The Chi-squared statistic was calculated for the observed matrix and compared to the distribution of Chi-squared values across permutations to create a permutation test with p value calculated as $p = \frac{(N_{crit} + 1)}{(N_{total} + 1)}$ where N_{total} denotes total number of permutations and N_{crit} denotes number of permutations producing a Chi-squared statistic greater than or equal to the value for the observed matrix. Additionally, this matrix was used to calculate further matrices of proportions of viruses exhibiting each viral trait of interest to visualise where human-infective, human-transmissible and viruses causing severe disease clustered within host-virus relationships.

To test whether viral traits were associated with nonhuman mammalian host ranges, I constructed logistic mixed regression models with binomial error structures using function glmmadmb() in package 'glmmADMB' v0.8.3.2 (Skaug et al. 2015) in R, v3.1.3. All viral traits were modelled using four independent model building paths, each containing one of host range metrics (a) – (d). Models predicting transmissibility and sustained transmissibility used the complete primary dataset for host range metrics (a) – (c) (n = 280), and the secondary subset excluding human-specialist viruses for host range metric (d), minimum phylogenetic distance to humans (n = 239). Although the primary dataset included human-specialist viruses that did not infect nonhuman mammals, it did not include analogous viruses neither

infecting humans nor nonhuman mammals (i.e. it was not known how many viruses with a mammal host range of zero have failed to infect humans). To avoid bias from this, models predicting human infectivity used the secondary subset excluding human-specialist viruses for all host range metrics (a) – (d) (n = 239). As virulence was rated in humans only, models predicting virulence were limited to further data subsets featuring only human-infective viruses with complete virulence information (n = 148 for host range metrics (a) – (c); n = 107 for host range metric (d)).

In addition to host range, I also investigated whether specific types of mammal host could predict viral traits by including binary variables in model building paths representing whether viral infection was known in each of the five most common mammal host taxonomic orders in the dataset (Artiodactyla, Carnivora, Chiroptera, Primates, Rodentia). All models were fixed to include two corrective terms: number of literature citations in PubMed search results for virus species names, to correct for ascertainment biases in mammal sampling; and virus taxonomic family as a random effect, to correct for taxonomic signal in viral traits as well as potential unmeasured traits that may covary with taxonomy (for example, substitution rate).

Host range metric (c), phylogenetic breadth, exhibited zero inflation as a result of large numbers of viruses only having a single known host species, or being human specialists. Therefore, phylogenetic breadth was specified in regression models as two simultaneous predictor terms: a binary variable denoting whether phylogenetic breadth was zero or nonzero, and a continuous variable denoting phylogenetic breadth value. All continuous predictors were modelled under a log(covariate + 1) transformation to normalise, except minimum phylogenetic distance to humans, which showed appreciable normality.

Model building paths were constructed using each host range metric (a) – (d), for each of the four viral traits (16 paths in total). Starting with the minimal model of corrective terms only, predictors were sequentially added using a stepwise algorithm, retaining predictors that improved the model fit based on likelihood ratio tests

(LRTs) until no further predictors were retained. Predictors were then confirmed by dropping each term and comparing model fit using LRTs. This process resulted in four models for each trait, each built using a different host range metric path (a) – (d), hereafter referred to as the "best" models. For each best model that featured a host range metric, I tested whether adding random slopes on the host range term with respect to viral family improved model fits via LRTs. For host range metric (c), random slopes were only applied to the continuous component of phylogenetic breadth. Between the best models predicting each trait, the model with the lowest AIC score (Akaike 1974) was accepted and presented as the "final" model. As models testing host range metric (d), minimum phylogenetic distance to humans, used the secondary data subset excluding human specialists, these models were excluded from AIC comparisons with metrics (a) – (c), except for models predicting human infectivity where the secondary data subset was used for all metrics (a) – (d) as previously described.

As some degree of correlation was expected between regression model predictors (e.g. greater known host breadth may result from greater virus sampling effort), I calculated Variance Inflation Factors (VIFs) for each predictor in final regression models to determine whether collinearity had inflated variance estimates around fitted coefficients. VIFs were calculated using the function vif() in package 'car', v2.1-0 (Fox and Weisberg 2011) in R, v3.1.3. Regression assumptions were verified by inspecting independence of final model residuals against fitted values. Final models were also validated against viruses suspected to be influential by removing these and reconstructing the selected model building path. For example, anthroponotic viruses might skew observed patterns in that their human origin implies these are more likely to have certain traits (e.g. human-to-human transmissibility) and also may be more likely to infect specific nonhuman taxa that are closely related to humans (i.e. primates) or domesticated (i.e. some artiodactyls and carnivores). Anthroponotic viruses (n = 4) were identified using a recently

published list of anthroponoses derived from structured literature review (Messenger et al. 2014).

4.4. Results

4.4.1. Matrices of host-virus relationships

Among the complete primary dataset of 280 virus species and 683 nonhuman mammal host species, host-virus relationships showed highly skewed distributions; most viruses were only known to infect zero or a few mammal hosts, and most nonhuman mammal species were only known to host a single virus (Figure C.1). Visualising the taxonomic patterns behind these relationships showed clear associations between host and virus type, with high clustering in certain areas of the host order-virus family matrix, which was otherwise sparse (Figure 4.1). For example, arenaviruses and bunyaviruses were most strongly associated with rodent hosts, whilst paramyxoviruses and retroviruses were most strongly associated with nonhuman primate hosts. This observed clustering pattern showed significant nonindependence between host and virus taxonomy (restricted permutation test, excluding human row, $p = 0.013$). When viral traits of interest were mapped onto the host order-virus family matrix, heterogeneity was again visible (Figure C.2 - C.5). Increasing sparsity was observed with increasing degrees of human-to-human transmissibility (Figure C.2 - C.4), with primate-infective viruses appearing the most consistently human-transmissible. Bat- and rodent-infective viruses appeared the most likely to cause severe human disease (Figure C.5).

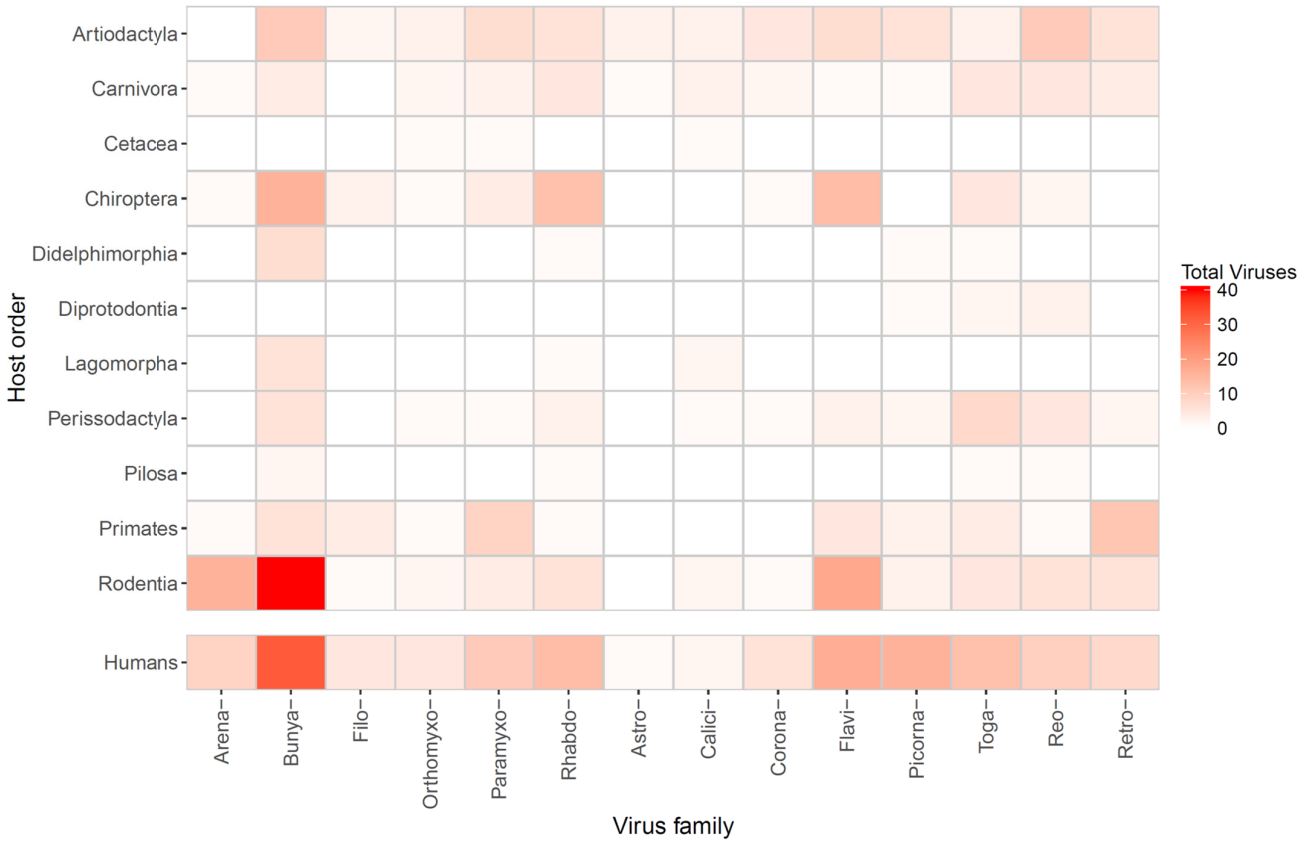


Figure 4.1. Heatmap of nonhuman mammal host taxonomic order versus virus taxonomic family, with colour of cells denoting total number of virus species in each viral family known to infect each mammal order. A separate row indicates the number of virus species in each viral family known to infect humans. Viral families are arranged by genome type (-ssRNA: Arena- to Rhabdo-; +ssRNA: Astro- to Toga; dsRNA: Reo-, ssRNA-RT: Retro-).

4.4.2. Mixed regression analyses of host range

Correcting for sampling biases and virus taxonomy, different aspects of nonhuman mammalian host range were associated with viral traits of emergence in mixed logistic regression models. For human infectivity, the model building path featuring host range metric (d), minimum phylogenetic distance to humans among mammal hosts, was selected as the final model (ΔAIC to next best model = 4.09). Viruses were more likely to be human-infective with a closer minimum phylogenetic distance to humans among mammal hosts (Table 4.1, Figure 4.2A). Bat- and rodent-infective viruses were also more likely to be human-infective (Table 4.1). No other host range metrics predicted human infectivity as these did not appear in the respective best models for each model building path (Table C.1). Three reoviruses that were not human-infective and were only known to infect marsupials among nonhuman mammals were observed as outliers (*Orbivirus: Eubenanagee virus, Wallal virus, Warrego virus*), their hosts having the largest phylogenetic distances to humans (Figure 4.2A). When these viruses were removed, the same final model and effect of phylogenetic distance were retained (LRT = 19.60, $p < 0.001$, odds ratio = 0.984).

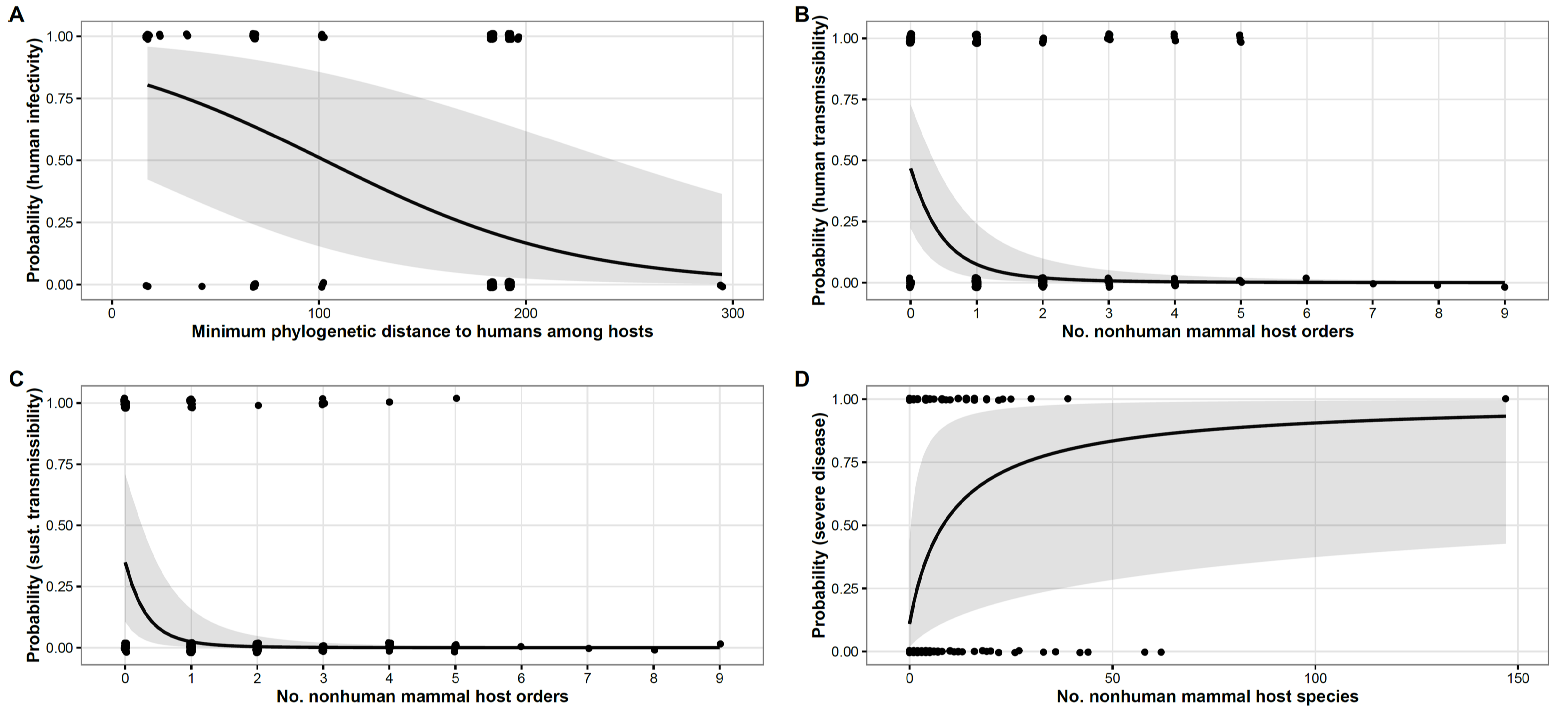


Figure 4.2. Fitted effects of nonhuman mammal host range variables in final logistic mixed regression models predicting viral traits of A) human infectivity (n = 239), B) human-to-human transmissibility (n = 280), C) sustained human-to-human transmissibility (n = 280), and D) severe disease (n = 148). Note that different host range metrics were selected for different final models: A) minimum phylogenetic distance to humans among hosts, B) and C) number of nonhuman host taxonomic orders, and D) number of nonhuman host species. Lines denote model fit with shaded areas denoting 95% confidence interval. Filled circles denote data points and are jittered for visibility. Model fits are calculated holding all other model predictors at constant values (virus citations = median value; Artiodactyl, Carnivora, Chiroptera, Primate, Rodent infection all = 0).

Table 4.1. Outputs from final logistic mixed regression model predicting human infectivity among mammalian RNA viruses, excluding human-specialist viruses (n = 239). Model building path selected based on AIC was model featuring host range metric d) minimum phylogenetic distance to humans (best models from each model building path are presented in Table C.1). LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	1.874 (0.376, 9.348)	-	-
log(Virus citations)	1.388 (1.161, 1.659)	15.05	< 0.001
Min. phylogenetic distance to humans among hosts	0.984 (0.976, 0.992)	20.75	< 0.001
Chiroptera infection	4.257 (1.844, 9.829)	12.14	< 0.001
Rodentia infection	2.630 (1.274, 5.430)	7.00	0.008

Table 4.2. Outputs from final logistic mixed regression model predicting human-to-human transmissibility among mammalian RNA viruses (n = 280). Model building path selected based on AIC was model featuring host range metric b) number of nonhuman mammal host orders (best models from each model building path are presented in Table C.2). LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	0.156 (0.053, 0.457)	-	-
log(Virus citations)	1.685 (1.386, 2.048)	35.99	< 0.001
log(No. mammal host orders)	0.032 (0.009, 0.108)	38.08	< 0.001
Carnivora infection	4.942 (1.487, 16.429)	6.71	0.010
Chiroptera infection	9.514 (2.998, 30.194)	16.09	< 0.001
Primate infection	5.731 (1.965, 16.709)	10.66	0.001

For human-to-human transmissibility, the final model selected featured host range metric (b), number of nonhuman mammal host orders (ΔAIC to next best model = 13.00). Viruses with a narrower range of host orders were more likely to be human-transmissible (Table 4.2, Figure 4.2B). Bat-, carnivore- and primate-infective viruses were also more likely to be human-transmissible (Table 4.2). When considering only sustained human-to-human transmissibility, the final model selected again featured host range metric (b), number of nonhuman mammal host orders (ΔAIC to next best model = 23.15), with a similar resulting model (Table 4.3). However, a sharper decrease in risk of sustained transmissibility with increasing number of host orders was observed (Figure 4.2C). Furthermore, carnivore infection did not feature in this model, and nonhuman primate infection showed a much stronger association with sustained transmissibility (Table 4.3). For both transmissibility and sustained transmissibility, the same final models and effect of nonhuman primate infection were retained when anthroponoses were removed (LRT = 8.98, $p = 0.003$, odds ratio = 5.115; and LRT = 17.33, $p < 0.001$, odds ratio = 17.379 respectively). A closer minimum phylogenetic distance to humans among mammal hosts also positively predicted both measures of human-to-human transmissibility (Table C.2, C.3), though model building path (d) used a secondary data subset and was not directly comparable to the final models selected (Table 4.2, 4.3).

Table 4.3. Outputs from final logistic mixed regression model predicting sustained human-to-human transmissibility among mammalian RNA viruses (n = 280). Model building path selected based on AIC was model featuring host range metric b) number of nonhuman mammal host orders (best models from each model building path are presented in Table C.3). LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	0.106 (0.023, 0.490)	-	-
log(Virus citations)	1.631 (1.288, 2.065)	20.29	< 0.001
log(No. mammal host orders)	0.012 (0.002, 0.057)	49.74	< 0.001
Chiroptera infection	9.101 (1.670, 49.586)	6.68	0.010
Primate infection	21.424 (4.955, 92.634)	20.82	< 0.001

Table 4.4. Outputs from final logistic mixed regression model predicting severe disease among human RNA viruses (n = 148). Model building path selected based on AIC was model featuring host range metric a) number of nonhuman mammal host species (best models from each model building path are presented in Table C.4). LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	0.120 (0.023, 0.620)	-	-
log(Virus citations)	1.004 (0.796, 1.267)	< 0.01	0.972
log(No. mammal host species)	2.574 (1.584, 4.183)	17.53	< 0.001
Artiodactyla infection	0.139 (0.030, 0.643)	7.65	0.006
Primate infection	0.250 (0.061, 1.013)	4.33	0.037

For virulence in humans, the final model selected featured host range metric (a), number of nonhuman mammal host species (ΔAIC to next best model = 3.66), although the model featuring (c) phylogenetic breadth was closely comparable (Table C.4). Viruses with a broader range of mammal host species were more likely to cause severe disease (Table 4.4, Figure 4.2D). Artiodactyl- and primate-infective viruses were less likely to cause severe disease (Table 4.4). In contrast to the other viral traits examined, number of virus citations was not a significant predictor of virulence (Table 4.4), consistent with discovery curves suggesting viruses causing severe disease do not have preferential study effort (see Chapter 2). *Rabies virus* was observed as an outlier, causing severe disease and having a much larger range of host species than other viruses (Figure 4.2D). When removed, the same final model and effect of nonhuman host species breadth were retained (LRT = 14.45, $p < 0.001$, odds ratio = 2.244). Similarly, the association with artiodactyl infection was robust to removal of anthroponoses (LRT = 7.73, $p = 0.005$, odds ratio = 0.146), though the association with nonhuman primate infection was noticeably weaker and was not retained in the final model when removing anthroponoses (LRT = 3.45, $p = 0.063$).

Between each final mixed regression model predicting viral traits, virus taxonomic families showed substantial variation in their fitted random intercepts (Table 4.5). Those viral families in the top quartiles for risk of human infectivity or transmissibility generally did not overlap with those in the top quartiles for risk of severe human disease and instead were often in the bottom quartile for the opposing trait. The sole exception to this was the *Arenaviridae*, being in the top quartile for both transmissibility and severe disease (but the lowest quartile for sustained transmissibility) (Table 4.5). Under no circumstances did the addition of random slopes upon host range metrics improve the fits of best models (Table C.5), except for the best model predicting sustained transmissibility featuring (d) minimum phylogenetic distance to humans. Adding random slopes with respect to virus family

gave a marginally improved fit for this model (LRT = 4.48, p = 0.034), though based on AIC values this was not selected as a final model for further analysis.

Table 4.5. Fitted random intercepts (odds ratios) associated with each RNA virus family from final logistic mixed regression models predicting viral traits. Families within the first quartile (viruses least likely to have the focal trait) are highlighted with italics, whilst families within the last quartile (viruses most likely to have the focal trait) are highlighted in bold.

	Modelled viral trait			
Virus family	**Infectivity**	**Transmissibility**	**Sustained transmissibility**	**Severe disease**
Arenaviridae	0.90	**2.29**	*0.26*	**6.43**
Astroviridae	0.66	0.99	1.71	0.75
Bunyaviridae	1.37	0.66	*0.18*	1.16
Caliciviridae	*0.51*	1.24	**4.75**	0.63
Coronaviridae	0.95	**2.75**	**10.03**	0.94
Filoviridae	1.52	1.98	0.84	**24.92**
Flaviviridae	0.79	*0.33*	0.39	1.54
Orthomyxoviridae	**2.83**	1.85	**6.39**	0.39
Paramyxoviridae	*0.26*	0.51	0.76	1.07
Picornaviridae	1.92	1.84	4.25	*0.38*
Reoviridae	1.65	2.12	4.25	*0.17*
Retroviridae	*0.18*	*0.24*	0.37	**7.23**
Rhabdoviridae	**1.99**	*0.23*	*0.20*	2.22
Togaviridae	**2.81**	**2.18**	0.53	*0.12*

4.5. Discussion

Using comparative analytical methods, I found the nonhuman mammalian host range of RNA viruses to be associated with viral traits integral to human emergence. It is already established that certain taxa of viruses are more likely to emerge in humans (Pulliam 2008) and further to this, I observed areas in host-virus matrices with comparatively higher risk of emergence-associated viral traits, suggesting that combining data on host range and virus taxonomy might improve predictions of emergence. These visualisations also raise biological questions about combinations with noticeably low risks. For example, sustained human-to-human transmissibility was known for four of the nine nonhuman primate-infective paramyxoviruses, but none of the seven artiodactyl-infective paramyxoviruses. This lack of transmissibility is unlikely to be due to limited opportunity for adaptation, as humans will have had exposure to many of these viruses via contact with domestic artiodactyls, suggesting the presence of an adaptive barrier that primate paramyxoviruses may be more likely to overcome.

Considering mixed regression model analyses, I did not replicate findings that a broad host range of mammals predisposes viruses to infect humans, as no host range breadth metrics featured in best models. Instead, I found risk of human infectivity to increase if viruses also infected hosts with phylogenetic similarity to humans, supporting the concept of a phylogenetic constraint on successful cross-species transmission (Wolfe et al. 2007). Parallel to this, alternative model building paths showed nonhuman primate infection to be a predictor of human infectivity in lieu of phylogenetic distance (Table C.1). This supports observations of increasing risk of pathogen sharing between closely related nonhuman primates and humans (Pedersen and Davies 2009; Cooper et al. 2012), but also across host phylogenies among other host-virus relationships (Streicker et al. 2010; Longdon et al. 2011). Fitted mathematical models of viruses within several mammal orders have suggested a sigmoidal shape to this relationship; infectivity decreased relatively little over short phylogenetic distances, and markedly over longer distances (Cuthill and Charleston

2013), which may explain why the simple measure of nonhuman primate infectivity also strongly predicted human infectivity. The lack of correlation between human infectivity and host range breadth may suggest that infectivity does not depend on virus generalism or specialism outright (Roche et al. 2015), but rather the underlying differences in ecological exposure and evolutionary opportunity that these strategies create.

Bats and rodents were additional host orders that significantly predicted human infectivity, both of which have been demonstrated elsewhere to host a relatively high diversity of zoonotic viruses (Luis et al. 2013). Both these orders contain many species that use peridomestic environments, which may result in relatively high human exposure to their viruses. However, both may also have been preferentially sampled for viruses already known to be human-infective following high-profile emergence events, i.e. hantaviruses for rodents and henipaviruses/coronaviruses for bats. A distinctly sharp increase in bat virus study effort since 2004 has been noted (Olival et al. 2012).

I found that viruses with a broad nonhuman host range were less likely to be human-transmissible. This could be explained in terms of competing evolutionary trajectories – genetic adaptation required for transmissibility among humans (or any specific host species) may antagonistically reduce infectivity and/or transmissibility among alternative hosts. Although inference from the regression models presented is limited to ecological correlation and does not address the underlying evolutionary mechanisms, similar ideas have been supported for pathogen infectivity (Elena et al. 2009). Meta-analyses of experimental infections also suggest that the majority of observed failures to infect non-hosts are driven by pathogen specialisation for existing hosts (as opposed to evolution of defences among non-hosts) (Antonovics et al. 2013). As transmissibility follows on from infectivity in conceptual models of host adaptation and specialism (Wolfe et al. 2007; Lloyd-Smith et al. 2009), transmissibility may feasibly be subject to similar constraints.

The observed negative correlation between breadth of nonhuman host range and transmissibility contradicts recent comparative analyses suggesting that zoonotic viruses with a broader host range are more likely to be human-transmissible (Johnson et al. 2015a). Critically, the scope of analyses presented here was wider, including those human-specialist viruses that are sufficiently adapted to transmit between humans without apparent dependency upon on other mammal hosts, further implying patterns of specialisation drive the relationship observed here. It must be noted that many human-specialist viruses such as HIV-1 and 2 have an ultimate zoonotic origin, and that zoonotic hosts are still important to consider when assessing potential human-to-human transmissibility of emerging viruses, particularly if the zoonotic hosts are those in which human-specialist viruses tend to originate.

Post-thesis submission, a reanalysis of models predicting human-to-human transmissibility whilst removing human-specialist viruses was conducted (Supplementary Methods C.1). Within this reanalysis excluding human specialists, no host range metrics acted as predictors of human transmissibility (Table C.6, C.7). This suggests that the observed model slopes and strengths of association between breadth of host range and human transmissibility (Figure 4.2) were primarily driven by an inflated count of human-specialist viruses; human-specialist viruses would be expected to have both human transmissibility and zero nonhuman hosts by definition (assuming host range data was accurate and complete, though see below). Therefore, given the currently available data, I conclude that specialism outright explains phenotypic patterns in human-to-human transmissibility of RNA viruses rather than relative degree of generalism in host range.

Interestingly, even after removing human-specialist viruses, my analyses still do not replicate previous findings of a positive host breadth-human transmissibility relationship (Johnson et al. 2015). This disparity therefore may arise from the alternative data definitions of transmissibility used by Johnson et al. (2015), where vector-borne transmission was excluded from human transmissibility estimates and

groups of species with shared ecology were considered equivalent to taxonomic orders. I also defined human specialists as those viruses infecting humans with no other known nonhuman mammal host species, though it is likely many of these have alternative hosts not yet identified (or not recorded to species level) as a result of sampling inadequacies, and certainly so for the few "specialists" that were not human-transmissible (n = 14).

Although the reanalysis implies human transmissibility of RNA viruses is independent of mammalian host range, this does not necessarily act as evidence against antagonistic pleiotropy in intraspecific transmissibility between different hosts. Rather, evidence for this is unclear at this analytical resolution, and genomic studies will be necessary for further investigation of adaptation and viral fitness between hosts at the molecular resolution (Pepin et al. 2010). Furthermore, nonhuman host range is still important to consider when assessing potential human-to-human transmissibility of emerging viruses. Many human-specialist viruses have an ancestral zoonotic origin, for example, HIV-1 and 2 and it follows that phenotypic capability of human-to-human transmission must arise before or at least concurrently with human specialisation. Although not addressed by this analysis, it is likely that the risk of development of human specialism will be greater for viruses that have an existing specialist or narrow host range, particularly if those hosts are closely related to humans.

In terms of host orders predicting human-to-human transmissibility, nonhuman primate infection was a strong predictor, particularly for sustained transmissibility, and this was not attributable to human-transmissible viruses being known to infect primates as anthroponoses. This may again reflect the short phylogenetic distance between humans and nonhuman primates, and a shorter phylogenetic distance to humans was also featured as a positive predictor of transmissibility in best models resulting from model building path (d) (Table C.2, C.3). Additionally, bat infection also predicted human-to-human transmissibility, which may suggest influence from other aspects of viral ecology. Previous studies have reported viruses to have greater risk of human-to-human transmissibility if

humans are infected through contact during hunting or associated practices (Johnson et al. 2015a), a common route for viral zoonoses from primates and bats as taxonomic groups widely hunted for bushmeat (Mickleburgh et al. 2009). However, bushmeat hunting also occurs for many other mammal taxa (Taylor et al. 2015), though relative global bushmeat pressures are too poorly-quantified for a direct comparison.

Finally, I observed that viruses with a broad nonhuman host range were more likely to exhibit higher virulence in humans. Combined with the previous result that viruses with a broad nonhuman host range were also less likely to be human-transmissible, if humans are 'dead-end' hosts that represent no potential for onward transmission and viral evolution, this would represent evidence of coincident, non-adaptive virulence (Levin and Svanborg Edén 1990; Bull 1994). However, my previous analysis finds little evidence for a direct association between transmissibility and virulence (see Chapter 2). There may instead be a population biology basis to this observation. In cases where multiple host species jointly contribute to viral maintenance (Haydon et al. 2002; Fenton et al. 2015), there may be lower pathogen costs of host mortality within any particular host species, as alternative hosts may still be sufficient for viral persistence (or vice-versa; Leggett et al. (2013) suggest that if virulence reduces host availability this could reciprocally select for a generalist host range). Any resulting selection for virulence could then indirectly increase virulence in novel hosts, including humans. Although this specific idea has not been directly addressed, mathematical models have shown the evolution of virulence to be highly dynamic in multi-host systems, being subject to heterogeneities in host competency, abundance, and interspecific transmission rates (Woolhouse et al. 2001; Gandon 2004; Rigaud et al. 2010).

The observed relationship between nonhuman host range and virulence could additionally be driven by those human-specialist viruses typically associated with mild disease (for example, rhinoviruses or parainfluenzaviruses). This may suggest that a longer coevolutionary history between humans and these viruses has

led to attenuation of virulence, although it must be noted that several human-specialist viruses still cause severe, chronic disease (e.g. HIV, hepatitis C virus; see also Chapter 2). Primate-infective viruses also exhibited lower virulence in humans, though this was not robust to the removal of anthroponoses, reflecting the potential anthroponotic transmission to nonhuman primates of nonsevere, human-specialist viruses. In contrast, the observed lower virulence among artiodactyl-infective viruses was not dependent on anthroponoses and warrants further empirical study.

I did not observe any association between virulence and phylogenetic distance to humans among mammal hosts. In experimental infections of *Drosophila* sigma viruses, virulence also did not follow a predictable linear relationship with phylogenetic distance between host species, and instead showed nonlinear clusters of high and low virulence across the *Drosophila* host phylogeny (Longdon et al. 2015a). If virulence is a phylogenetically conserved trait of hosts as a general rule, this would suggest the level of virulence in nonhuman primates may be an informative predictor of virulence in humans. Although this approach has been traditionally applied in experimental inoculations of nonhuman primate models (Patterson and Carrion 2005), this will be a significant challenge for comparative study as virulence data for wild nonhuman hosts, including primates, is scarce and often hard to quantify (Levinson et al. 2013).

4.5.1. Analytical limitations

The taxonomic host range data and phylogenetic breadth calculations I use here likely do not represent the full nonhuman host range of each virus, which may have introduced inaccuracies to my analyses. The majority of hosts were identified as only having a few virus species (Figure C.1), although calculations extrapolating from intensive sampling efforts of a single bat species would suggest this is a very small fraction of the true viral diversity within mammals (Anthony et al. 2013). Although sampling effort for viruses was directly controlled for, there may be remaining biases as a result of the unequal sampling effort across host taxa, e.g. host orders containing

several common domestic species (Artiodactyla, Carnivora) may be more thoroughly sampled than those with exclusively wild species (e.g. Chiroptera). It is difficult to precisely identify how this would bias analyses. Sampling inequalities could have several influences, e.g. oversampling of domestic taxa may lead to biased inclusion of mammal-exclusive viruses that do not infect humans in the dataset and negatively bias risks of infectivity. However, preferential sampling of otherwise undersampled wild taxa for viruses associated with human outbreaks may lead to biased inclusion of human-infective viruses and positively bias risks of infectivity. Furthermore, host sampling inequalities are known to be present within orders, subject to factors such as biogeography and life history (Cooper and Nunn 2013).

Uncertainty in the classifications of viral traits used are discussed in previous chapters for both data on virulence (see Chapter 2) and transmissibility (see Chapter 3), and the same potential errors also apply to analyses presented here.

4.5.2. Wider implications

My analyses may set up hypotheses for studies of underlying genetic mechanisms behind the associations between nonhuman host range and viral traits. Host specialism within the orthopoxviruses (a genus of DNA viruses) is associated with genome shortening, with extant specialists thought to descend from a generalist virus with a much longer genome (Hendrickson et al. 2010). The relationship between host range, viral traits and genome length warrants further comparative study. Additionally, future studies could identify specific genetic loci associated with host range breadth or specificity, or deeper phylogenetic patterns among viral traits.

These analyses also have implications for public health. The correlations I find between host range and traits of emergence highlight that nonhuman hosts and the wider ecological context should be considered in public health decisions, in line with the One Health perspective. Those key host types my models identify as being associated with human infectivity, transmissibility and virulence could also contribute towards targeted surveillance and interventions (Daszak 2009). However,

if the observed correlations from the models presented reflect a sensitive underlying evolutionary balance, then removal of nonhuman hosts during certain interventions may alter selection pressures and have unpredictable and/or unintended consequences (Bolzoni and De Leo 2013). Resources should first be invested in understanding how host range affects emergence dynamics at a more local, community-based scale.

4.6. Conclusion

Traits of RNA viruses are intrinsically linked to their host range, not only in simple taxonomic breadth but also in the specific types of hosts infected and their phylogenetic relationship with humans. These patterns suggest that nonhuman hosts are a key factor underpinning the emergence of novel viruses. Ultimately, this work represents a strong admonition that human diseases do not operate in a closed system, and their wider ecological context must be considered to further understand their dynamics.

Chapter 5. Allometry of mammal species predicts ability to host virulent human RNA viruses

5.1. Abstract

The majority of human RNA viruses have a mammalian origin. Comparative studies have begun to identify life history traits among mammal species associated with greater virus diversity and ability to host zoonotic viruses. However, it remains poorly understood whether life history traits predispose certain mammals to hosting viruses with high risks of virulence or transmissibility. Using body mass or 'allometry' as a proxy for overall pace of life, I aim to determine whether life history of mammal species can predict hosting zoonotic RNA viruses capable of a) causing severe disease in humans, and b) human-to-human transmission. I obtain data on mammal body mass, virus traits, and reported mammal-virus relationships by conducting or externally sourcing systematic literature searches. Correcting for taxonomy and study effort, I use mixed logistic regressions to investigate whether proportions of virulent and human-transmissible viruses are predicted by body mass among 524 mammal species. I observed that mammal species with a smaller body mass (and therefore, a faster pace of life history) were more likely to host viruses causing severe human disease. The relationship between human-to-human transmissibility and mammalian host body mass was sensitive to data definitions of transmissibility and subgroup analyses, and largely remains unclear. My analyses suggest that host life history may play a significant role in the evolution of virus traits, and further studies of virulence should consider the context of host ecology. Efficiency of global health strategies may be improved by preferentially targeting surveillance and interventions towards mammals with specific life histories.

5.2. Introduction

Newly recognised viral infections are continuing to emerge at a persistent rate (Woolhouse et al. 2012). Although several definitions of 'emergence' are based solely upon infectivity (Morse 1995), emerging viruses vary greatly in their ultimate public health impact. This is partly determined by virus traits such as virulence and human-to-human transmissibility, with several key virulent and/or human-transmissible viruses having recently been designated a priority for global health (WHO 2015). The majority of emerging and re-emerging RNA viruses are zoonotic (Taylor et al. 2001; Woolhouse and Gowtage-Sequeria 2005), and among those, the majority are shared with nonhuman mammals. Certain risk factors are known to predispose mammals to hosting zoonotic viruses, e.g. phylogenetic relatedness to humans, primarily within hominid apes and wider primates (Pedersen and Davies 2009), and degree of human contact, primarily within domestic mammals (Cleaveland et al. 2001).

Several comparative analyses have shown the diversity and zoonotic potential of viruses hosted by mammals to additionally depend on their ecology and life history. Mammals exhibit a very broad range of ecological strategies and show substantial variation in their patterns of maturity, activity, reproduction, etc. (Lindstedt and Calder 1981; Ricklefs and Wikelski 2002). Greater virus richness has been observed among mammals with greater body mass, longer lifespan, earlier and more frequent reproduction, specific diets, greater range use intensity and greater population genetic structure (Nunn and Dokey 2006; Lindenfors et al. 2007; Turmelle and Olival 2009; Luis et al. 2015). Fewer studies have directly addressed zoonotic transmission in the context of mammal life history, though both ability to host zoonotic viruses and richness of zoonotic viruses hosted appear to follow similar principles, being greater in species with greater lifespan and reproductive potential (Luis et al. 2013; Han et al. 2015b).

However, mammal life history influences more than simply the diversity of viruses hosted. Mammals with differing life history strategies are likely to represent very different host environments and subsequently, selection pressures for viruses.

Consequently, host ecology and life history may shape the trajectory of viral evolution (Streicker et al. 2012b). Therefore, a key question that remains is how life history influences the traits of viruses hosted by mammal species. Correspondingly, there is also a pressing public health need to identify likely host species of future high-impact emerging zoonoses (Daszak 2009), beyond simply predicting virus sharing. Here, I ask whether life history can predict ability to host priority RNA viruses, in terms of virulence and transmissibility within humans.

Although the term 'life history' covers a broad spectrum of traits such as activity, maturity and reproduction, variation in life history is often described using a single axis of "pace of life", with fast-paced species having comparatively shorter lifespans, but faster metabolism and higher fecundity, and *vice versa* for slow-paced species (Promislow and Harvey 1990; Ricklefs and Wikelski 2002). Additionally, many of the traits following such a gradient strongly correlate with overall body mass or 'allometry', such that body mass is a well-established proxy for pace of life (Lindstedt and Calder 1981; West and Brown 2005), and here I focus on allometry to represent pace of life in the context of hosting virulent and human-transmissible viruses.

Evolutionary pressure to increase virulence and transmissibility may be anticipated in mammals with certain life histories for several reasons. Firstly, mammal species with a faster-paced life history might exhibit narrower, more opportunistic windows of viral transmission potential as a result of shorter lifespans and intensive reproduction. This may introduce selection pressure for viruses to increase their intensity of replication and subsequently, virulence and transmissibility (Nidelet et al. 2009).

Secondly, higher investment in processes such as reproduction and metabolism among faster-paced mammal species may restrict investment in immune function via resource trade-offs (Lochmiller and Deerenberg 2000; Tella et al. 2002). Differences are also predicted with life history in where immune investment is allocated – faster-paced species may preferentially invest in innate immunity versus

costlier adaptive immunity (Lee 2006; Previtali et al. 2012). One might expect this to extend to greater potential for host exploitation and evolution of increased virulence and transmissibility.

Using body mass as an allometric proxy, I aim to test whether pace of life history in mammal species predicts their ability to host zoonotic RNA viruses with specific traits of virulence and transmissibility within humans. Based on the above, I hypothesise that mammal species with smaller body mass (and therefore, faster-paced life history) will be more likely to host viruses with greater virulence, and viruses capable of human-to-human transmission.

Previous studies that identified potential mammalian hosts of zoonotic viruses have largely done so on the basis of zoonotic virus richness (Luis et al. 2013, 2015), though this does not account for baseline viral diversity within each mammal species. Instead of richness, I focus on the relative proportions of virulent and human-transmissible viruses out of all hosted zoonotic viruses. I source known mammal host-virus relationships, mammal body mass data, and virus trait data using datasets constructed from systematic literature searches. Correcting for mammal taxonomy and study effort, I used mixed logistic regression models to investigate whether body mass of mammal species predicts their proportions of virulent and human-transmissible viruses. I then conduct subgroup analyses to test whether specific transmission routes or viral families are driving observed relationships.

5.3. Materials and methods

5.3.1. RNA virus and mammal host data

Zoonotic RNA viruses were defined as those known to infect both humans and non-human mammals. Human-infective RNA viruses were defined using a published empirical review of human RNA virus species (Woolhouse et al. 2013). Mammal RNA viruses and their known host species were sourced from the EcoHealth Alliance (Olival et al. in review) and were originally compiled via structured literature searches (search terms: [virus name including synonyms];

sources used: Web of Knowledge, Google Scholar, Wildlife Disease Association meeting abstracts, Global Mammal Parasite Database (Nunn and Altizer 2005); protocol further described in Levinson et al. (2013). Virus detection via molecular methods (e.g. virus isolation, PCR-based methods) and serological methods (e.g. serum neutralisation, antigen detection assays) were both accepted as sufficient evidence of host infection. Data from experimental inoculations, studies of captive individuals (e.g. zoological parks and breeding facilities), or cell culture detections were excluded. Host species were accepted based solely on evidence of infection, via either molecular (e.g. virus isolation, PCR-based methods) or serological (e.g. serum neutralisation, antigen detection assays) diagnostic methods. Diagnostic evidence from experimental infections, captive individuals (zoological parks and breeding facilities), and cell cultures were not considered. Virus names were standardised to species using ICTV 9th Edition Virus Taxonomy (King et al. 2011).

For each zoonotic RNA virus species represented, virulence and transmissibility within humans were classified using previous analyses/sources. Firstly, I defined virulence as a binary classification based on whether viruses typically cause either 'severe' or 'nonsevere' disease in humans, where 'severe' met at least one of the following criteria: ≥5% case fatality ratio, frequent reports of hospitalisation, association with significant morbidity from a prespecified list of conditions (haemorrhagic fever, seizures/coma, cirrhosis, AIDS, hantavirus pulmonary syndrome, HTLV-associated myelopathy) or were explicitly described as "severe" or "causing severe disease" in literature sources (see Chapter 2). Secondly, I defined viruses capable of human-to-human transmission as described in a comprehensive empirical review (Woolhouse et al. 2016), and accepted any observed route of human-to-human transmissibility, including vertical or iatrogenic routes. Following Woolhouse et al. (2016), I accepted viruses as human-transmissible where transmissibility is suspected but unconfirmed (e.g. spatial or contact-based clusters of cases without diagnostic confirmation) or possible but not yet directly observed (e.g. very large outbreaks in urban communities with few potential animal hosts). One

virus was removed due to difficulty in estimating virulence (*Retroviridae*: *Primate T-lymphotropic virus 3*). To understand whether wider viral traits were driving any observed patterns, I also compiled the primary transmission route of each virus where known using the same literature sources outlined above (see Chapter 2), classified as either 'vector-borne' (excluding mechanical transmission) or 'nonvector-borne' (including direct contact, respiratory and faecal-oral transmission). This grouping was chosen as transmission routes occasionally vary between different hosts, but do not usually cross this dichotomy.

For each mammal host species with at least one zoonotic virus, I calculated the proportions of viruses causing severe human disease and being capable of human-to-human transmission out of their total number of known zoonotic viruses. Body mass data for mammal species was sourced from PanTHERIA (Jones et al. 2009), an open-access database of mammal life history traits, and was defined therein as adult body mass in grams, averaged over all data sources consulted. Zoonotic virus richness in mammal species has been consistently demonstrated to correlate with sampling effort (Nunn et al. 2003; Lindenfors et al. 2007; Luis et al. 2013). To correct for any additional bias from preferentially sampling mammal hosts for virulent and human-transmissible viruses, I obtained a count of virus research citations for the Latin binomial name of each mammal species via PubMed, while attempting to minimise inclusion of experimental studies (search terms: [Genus] + [species] + virus NOT experiment* NOT "cell line"). To standardise the mammal taxonomic orders represented, orders with less than 10 species were excluded. When removing mammal species with missing body mass data, erroneous citation counts (one species, *Axis axis*), or belonging to an order with less than 10 species, data was filtered from 627 to 524 mammal species, within six taxonomic orders (Artiodactyla, Carnivora, Chiroptera, Lagomorpha, Primates, Rodentia). Mammal species within this final dataset were host to 125 zoonotic RNA viruses in total.

5.3.2. Statistical analysis

To investigate whether mammal body mass correlated with ability to host virulent and human-transmissible viruses, I constructed logistic mixed regression models where outcomes were specified as proportion of zoonotic viruses causing severe human disease and being capable of human-to-human transmission, respectively. Mammal species were weighted within models by their total zoonotic virus richness to account for their baseline virus diversity. Model covariates were specified as body mass and citation counts, both of which were modelled under a log(covariate+1) transformation. Model covariates were assessed via likelihood ratio tests (LRTs) between the full model and a reduced model removing the covariate. As a basic correction for unmeasured mammal traits and phylogenetic structure in host-virus relationships, mammal taxonomic order was specified as a random effect. Separate models were created featuring a) random intercepts, and b) random intercepts plus random slopes upon the body mass covariate. Models (a) and (b) were compared using LRTs, retaining the best fitting model. All mixed regression models were conducted using function glmer() in package 'lme4', v1.1-11 (Bates et al. 2015) within R, v3.2.2 (R Development Core Team 2015). Confidence intervals were calculated for fitted model plots using 1000 simulations over both fixed and random terms using function predictInterval() in package 'merTools', v0.2.0 (Knowles and Frederick 2011).

As body mass has been demonstrated to correlate with study effort in certain mammal orders (Cooper and Nunn 2013; Brooke et al. 2014; Han et al. 2015a), I used Variance Inflation Factors (VIFs) to assess collinearity between body mass and citation counts within retained models. VIFs were calculated using a function adapted for mixed regression models (Frank 2011) based on the function vif() in package 'rms', v3.2-0 (Harrell 2011).

To understand whether any observed associations were being driven by specific types of viruses, data compilation and analyses were repeated for four data subgroups, limiting those viruses considered to: a) vector-borne viruses (resulting in

n = 258 mammal species analysed), b) nonvector-borne viruses (n = 388 mammal species), and two viral families producing adequate sample sizes, c) *Bunyaviridae* (n = 126 mammal species) and d) *Flaviviridae* (n = 157 mammal species). Model sensitivity was also assessed by repeating analyses removing those virus species suspected to be influential due to their exceptionally broad mammalian host range (*Rhabdoviridae*: *Rabies virus, Picornaviridae*: *Foot-and-mouth disease virus*).

5.4. Results

Of the 125 zoonotic RNA viruses represented within compiled data, 42 were classified as 'severe' in their human virulence, and 46 exhibited evidence of human-to-human transmissibility. Among the 524 mammal host species represented, there was strong overdispersion in the number of known zoonotic viruses hosted, as well as the number of severe and human-transmissible viruses (Figure D.1). The majority of mammal species were only known to host a single zoonotic virus, leading to many values of zero or one in proportions of severe and human-transmissible viruses. Three of the five mammal species hosting the highest number of severe viruses (n = 7 viruses), were fruit bats (*Artibeus lituratus, Artibeus jamaicensis, Rousettus leschenaultii)*, with the remainder being common rodents (*Mus musculus, Peromyscus leucopus)*. The mammal with the highest number of human-transmissible viruses was again a fruit bat, *Rousettus leschenaultii* (n = 11 viruses), followed by species that include domestic variants: house mice (*Mus musculus;* n = 9), cattle (*Bos taurus;* n = 9) and swine (*Sus scrofa;* n = 8).

Considering human virulence, mammals with a smaller body mass hosted a greater proportion of viruses causing severe disease (Table 5.1, Figure 5.1). No relationship was observed with sampling effort (Table 5.1), and no influence of collinearity upon fitted model variances was detected (VIF = 1.007). Fitted model intercepts under a common regression slope suggested carnivores, rabbits and hares (order Lagomorpha), and bats (order Chiroptera) hosted the greatest proportion of viruses causing severe disease, and primate and artiodactyl species hosted the least

(Figure 5.1), though there was general overlap in 95% confidence intervals of fitted model lines between mammal orders (Figure D.2). Introducing random slopes with respect to mammal order did not improve model fit (LRT = 2.83, p = 0.243).

Table 5.1. Outputs from logistic mixed regression model predicting proportion of zoonotic RNA viruses causing 'severe' human disease within 524 mammal host species. Variance in random intercept term of mammal taxonomic order = 0.289. LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	2.297 (1.679, 8.505)	-	-
log(Adult body mass (g))	0.822 (0.757, 0.894)	19.91	< 0.001
log(Mammal species citation counts)	0.961 (0.890, 1.038)	1.02	0.312

Rabies virus and *Foot-and-mouth disease virus* were suspected to be influential in analyses due to their broad host ranges. When mammal species hosting either of these viruses were examined, a large number (n = 100) appeared to host one of these as their only recognised zoonotic virus (particularly artiodactyls, carnivore, and bat species), which could have introduced potential bias to regression slope estimates. However, the negative relationship between proportion of viruses causing severe disease and body mass was robust to their removal (removing *Rabies virus*, odds ratio = 0.739, LRT = 23.06, p < 0.001; removing *Foot-and-mouth disease virus*, odds ratio = 0.823, LRT = 21.33, p < 0.001, Figure D.3). This relationship was also consistently retained when data were limited to subgroups considering viruses with specific transmission routes or taxonomic families (Figure 5.2).

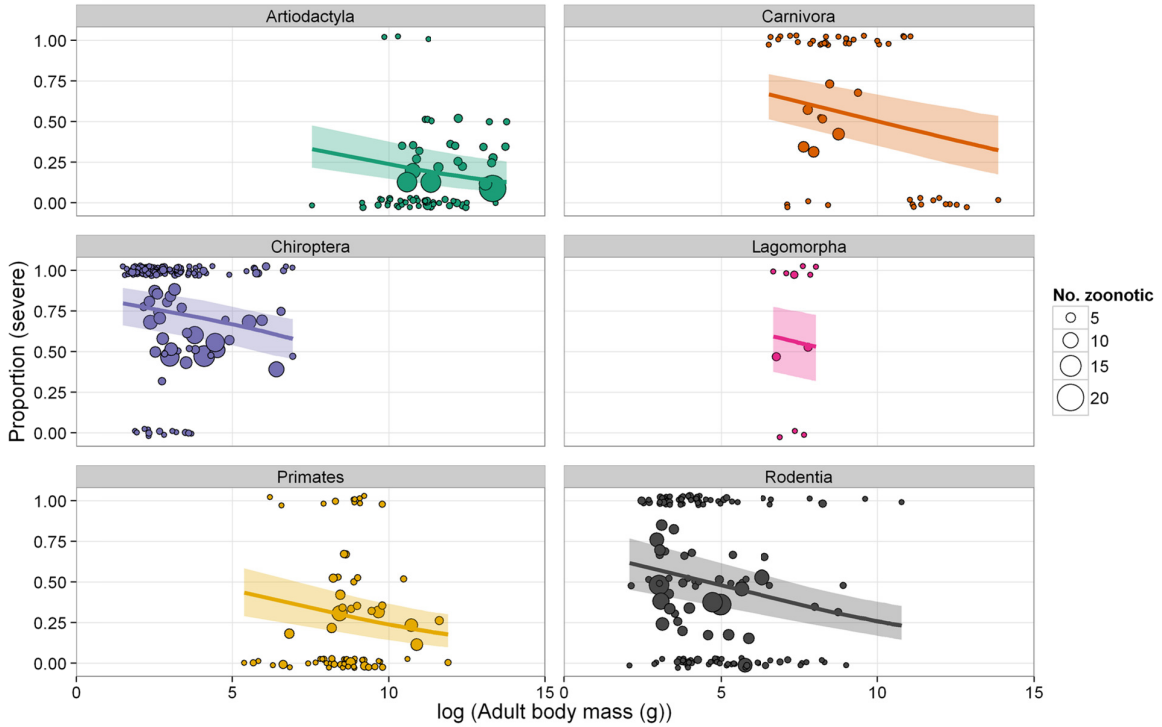


Figure 5.1. Relationship between adult body mass in grams and proportion of zoonotic RNA viruses causing 'severe' human disease among 524 mammal species within six taxonomic orders, illustrated using separate panels and colour schemes. Filled circles denote individual host species where circle size denotes number of zoonotic viruses (i.e. denominator in calculated proportion on the y axis) and model weighting in logistic mixed regression. Filled circles are jittered for visibility. Lines denote fitted effect from logistic mixed regression model with random intercepts for each taxonomic order, holding mammal species citations constant at the median value. Shaded areas denote 95% confidence interval from 1000 simulations accounting for variances of both the regression slope and random intercepts

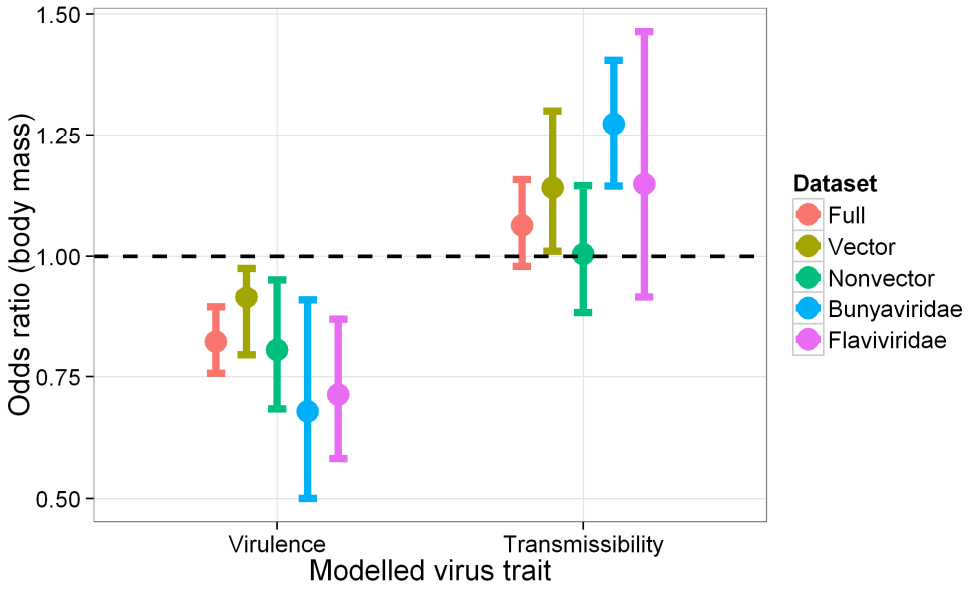


Figure 5.2. Odds ratios with 95% confidence intervals describing effect of mammal adult body mass from logistic mixed regression models predicting proportion of zoonotic RNA viruses causing 'severe' human disease and capable of human-to-human transmission, respectively. Colours denote effects from fitted models for data subgroups, limiting viruses considered to those with specific transmission routes (vector-borne, nonvector-borne) or from specific taxonomic families (*Bunyaviridae*, *Flaviviridae*). Odds ratios from models with full data are presented in red for comparison. Dashed line denotes the null expectation of an odds ratio value of 1.

Considering human-to-human transmissibility, no relationship with mammal body mass was observed, nor with sampling effort (Table 5.2A), and no collinearity was detected (VIF = 1.003). Therefore, to visualise potential differences between mammal taxonomic orders, a subsidiary model was created specifying order as a random intercept and no other terms. As for virulence, carnivore and bat species appeared to host the greatest proportion of viruses capable of human-to-human transmission, though confidence intervals suggested limited distinction between orders (Figure 5.3).

Table 5.2. Outputs from logistic mixed regression models predicting proportion of zoonotic RNA viruses capable of human-to-human transmission within 524 mammal host species. Models presented use different datasets where transmissibility is defined as A) any observed transmissibility following Woolhouse et al. (2016), and B) excluding iatrogenic transmission, which reclassifies six viruses as non-transmissible. Variance in random intercept term of mammal taxonomic order for A) = 0.499, B) = 0.312. LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
A) Standard transmissibility classification			
(intercept)	0.209 (0.154, 1.040)	-	-
log(Adult body mass (g))	1.063 (0.979, 1.158)	2.07	0.150
log(Mammal species citation counts)	0.989 (0.920, 1.062)	0.10	0.756
B) Reclassified transmissibility excluding iatrogenic transmission			
(intercept)	0.066 (0.042, 0.268)	-	-
log(Adult body mass (g))	1.137 (1.040, 1.251)	8.48	0.004
log(Mammal species citation counts)	1.041 (0.963, 1.125)	1.01	0.314

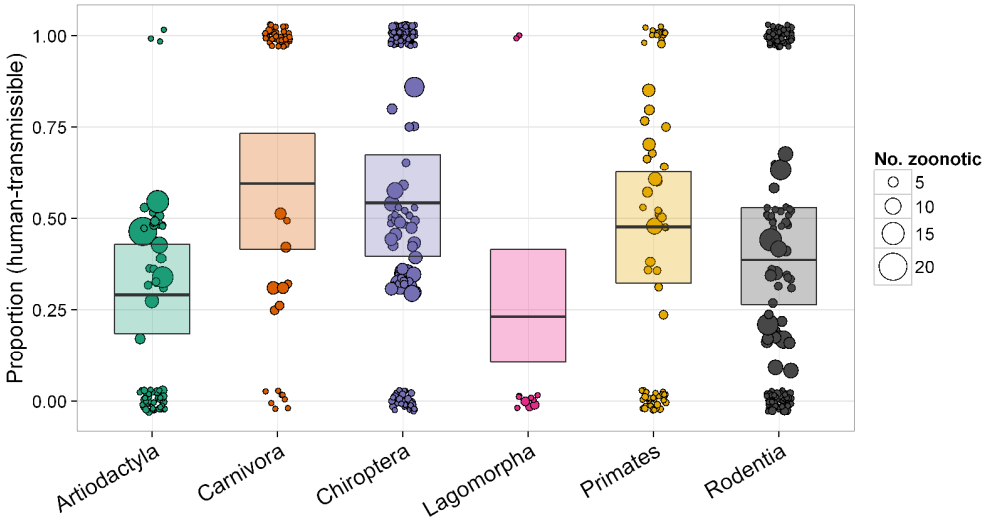


Figure 5.3. Relationship between mammal taxonomic order and proportion of zoonotic RNA viruses capable of human-to-human transmission among 524 mammal species. Filled circles denote individual host species where circle size denotes number of zoonotic viruses (i.e. denominator in calculated proportion on the y axis) and model weighting in logistic mixed regression. Filled circles are jittered along both axes for visibility. Boxes denote predictions and 95% confidence interval from 1000 simulations derived from a subsidiary model containing only random intercepts for each taxonomic order, as neither body mass nor citation counts were significantly associated with proportion transmissible.

As for the full dataset, no relationship was visible between proportion of viruses capable of human-to-human transmission and mammal body mass when limiting data to subgroups considering nonvector-borne viruses or flaviviruses (Figure 5.2). However, a positive relationship was observed when considering bunyaviruses and vector-borne viruses, though this was much more tentative for the latter subgroup. To further explore this model sensitivity, human-to-human transmissibility was redefined to exclude transmission through iatrogenic routes, reclassifying six viruses as non-transmissible (*Arenaviridae*: *Lymphocytic choriomeningitis virus; Flaviviridae*: *Japanese encephalitis virus, Usutu virus, West Nile virus*; *Reoviridae*: *Colorado tick fever virus; Rhabdoviridae*: *Rabies virus*). A positive relationship was also visible within the full dataset under this reclassification

(Table 5.2B), where mammals with a larger body mass hosted a greater proportion of viruses capable of human-to-human transmission (Figure D.4).

Analogous results were obtained when the data structure was reversed to conduct analyses at the level of virus species rather than host species (Supplementary Methods D.1). Specifically, viruses with a smaller average body mass among their hosts were more likely to cause severe disease, but were not more likely to be human-transmissible (Table D.1A, D.2A, Figure D.5A). Similar conclusions were also confirmed when restricting data to host-virus pairs with stringent diagnostic evidence of within-host replication (Table D.1B, D.2B, Figure D.5B).

5.5. Discussion

Using a comparative trait-based approach, I found evidence for a relationship between host body mass and traits of emergence among zoonotic mammalian RNA viruses. Specifically, smaller mammals were more likely to host viruses causing severe disease in humans. This is consistent with my initial hypothesis that, following an allometric proxy, mammal hosts with faster-paced life history may have increased suitability to host virulent viruses. Assuming allometry is an accurate proxy for overall pace of life history, two key elements need to be confirmed to better understand any underlying causality between host life history and virulence in humans.

Firstly, pace of life must influence selection for virulence within the mammal host itself, either directly or indirectly through trade-offs, which may occur via several potential mechanisms. For example, hosts with faster pace of life may have differential investment in immunity. Generalised immune function has been reported to be lower in species with lower body mass and shorter lifespan across 50 bird species (Tella et al. 2002). However, the multitude of pathways within the vertebrate immune system means this is likely not a simple relationship, and instead pace of life may determine where immune investment is allocated. For example, species with faster-paced life history exhibited higher metrics of innate immunity and

lower metrics of adaptive or antibody-mediated immunity in assays of wild rodents (Previtali et al. 2012). Hosts with a faster pace of life may also present limited windows for viral transmission, creating a selection pressure to increase replication or viral load. Fewer studies have addressed this idea, though viral replication rates appear to be higher in cell types with higher turnover (Hicks and Duffy 2014). It remains to be tested whether viral replication mirrors host turnover at the wider scale of host species longevity and life history. Although any evolutionary mechanisms are difficult to precisely identify, there is empirical evidence that the level of virulence observed in animal hosts reflects their life history. Mammals and birds with smaller body mass experienced faster onset of symptoms and mortality in a meta-analysis of microparasites, including two RNA viruses (Cable et al. 2007). Virulence has also been reported to vary with pace of life in other host-parasite systems (Nidelet et al. 2009; Johnson et al. 2012).

Secondly, any evolved level of virulence must then persist over zoonotic cross-species transmission, independent of the equivalent pace of life of humans. This would imply a detectable correlation between virulence within animal and human hosts, though the extent to which animal hosts show pathology varies (Levinson et al. 2013) with several highly virulent human viruses not being known to cause disease in their suspected natural host range (Halpin et al. 2007). Although the level of virulence can be a non-adapted, coincidental phenotype following cross-species transmission to a novel host (Levin and Svanborg Edén 1990; Bull 1994), some predictability of virulence is evident from experimental studies, based on host phylogenies (Longdon et al. 2015a). Evolutionary study of the relative influences of both original and novel host species pace of life history upon virulence during cross-species transmission will be a crucial next step towards understanding risks surrounding emerging viruses.

Mammals with larger body mass appeared more likely to host viruses capable of human-to-human transmission only under select conditions, and the wider relationship between pace of life history and transmissibility remains unclear. An

interaction between pace of life and microparasite transmissibility within nonhuman hosts has been proposed in several mathematical models (De Leo and Dobson 1996; Han et al. 2015a). However, evolutionary potential for human-to-human transmissibility to develop following zoonotic transmission is likely determined by a wider range of factors (see Chapter 3, also Pepin et al. 2010).

Among mammal orders, bats, lagomorphs and carnivores were observed to have the highest proportions of viruses causing severe disease (corrected for body mass), and artiodactyls the lowest. However, species within these orders often hosted viruses suspected as influential in regression modelling (Figure D.3). Sensitivity analyses suggested that although regression slopes over body mass remained consistent, *Rabies virus* visibly inflated the fitted intercept for carnivores (Figure D.3). However, fitted intercepts for artiodactyls and bats were comparable after removing these viruses. My analyses may tentatively suggest support towards bats as reportedly "special" hosts of viruses causing severe disease and mortality in humans (Dobson 2005), and requiring focused study.

5.5.1. Analytical limitations

As with any comparative trait-based modelling, my analyses are prone to error from gaps in data availability. Mammal host-virus data and life history data were sourced from standardised datasets compiled using systematic literature searches. I implicitly assumed this data to be representative across all mammals. However, most mammal species were only known to host a single virus as data coverage of the complete mammal virome remains poor (Anthony et al. 2013; Cooper and Nunn 2013). I adjusted models to place greater confidence upon well-studied species by weighting with respect to the number of known zoonotic viruses and explicitly correcting for citation counts, though ongoing study will be needed to confirm my findings as coverage of mammal-virus data improves. Although life history data is often incomplete, contingent on species rarity and difficulty of sampling (Penone et al. 2014), body mass data coverage among mammal species

included here was satisfactory (93.4%). There are also potential biases within the data collection protocol and classification systems I use for virulence and human-to-human transmissibility, discussed in detail in previous chapters (see Chapter 2, Chapter 3).

Additionally, analysis of data aggregated at the species level for mammal hosts brings several limits to inference. I detected differences in ability to host viruses causing severe human disease between mammal species, though it is possible for greater unmeasured variation to occur within species. Viral communities can vary highly even among geographically close intraspecific populations or individuals (Anthony et al. 2015). Other host traits that might indirectly govern relationships observed here may also vary within species, e.g. individual or seasonal differences in immunity (Lee 2006). Although the large amount of unexplained variation in hosting viruses causing severe disease (Figure 5.1) likely reflects these processes, comparative study between host species remains essential. To my knowledge, the analyses herein represent the first empirical tests of hypotheses relating nonhuman host life history to traits of emerging human viruses. There is also an urgent need among global health programmes to understand the broad-scale trends of viral emergence. To improve inference and precision of risk assessments, my analyses can be followed by trait-based study of mammal-virus relationships at more precise ecological scales (Johnson et al. 2015b).

Finally, my analyses are built on mammal host-virus data collected via literature searches, including studies of cross-sectional sampling. With this scope of data, it is currently not possible to distinguish between mammal species that are true maintenance hosts or 'reservoirs' (and therefore, have significant coevolutionary relationships with their respective viruses) from dead-end host species that do not contribute to onward transmission. This introduces uncertainty regarding any underlying explanations for the associations I observe, as host pace of life cannot shape viral selection without coevolution. To strengthen causality, analyses could be repeated considering only known reservoir species. However, reservoir status is

challenging to establish without in-depth study (Viana et al. 2014), and no reservoir has yet been identified for many zoonotic viruses.

5.5.2. Wider implications

Study of mammalian life history in relation to pathogen traits has significant implications for public health. As the importance of the wider ecology of infectious diseases is increasingly recognised under 'One Health' perspectives, there are calls to target surveillance and control programmes for zoonoses based on insights from empirical models (Daszak 2009). In previous studies of virus richness, correlations with life history have varied between different mammal groups or analytical methods, though generally, larger mammal species with slower-paced life history have been associated with a greater diversity of zoonotic viruses (Lindenfors et al. 2007; Luis et al. 2013, 2015). Based on this, species with a slower pace of life would appear to be priority targets for surveillance and intervention. However, my results suggest that, independent of diversity, those mammals hosting viruses with the greatest potential public health impact are actually smaller species with a faster pace of life. This disparity highlights the wide phenotypic variation among emerging viruses and the essential need for studies of parasite richness to be complemented by trait-based analyses. The most efficient public health strategies concerning mammal hosts may therefore involve a compromise between sampling virus-rich species and species suited to hosting virulent viruses. There is a clear need for further research of mammalian viromes in order to improve and direct risk assessments for emerging zoonoses.

5.6. Conclusion

Analogous to aspects of traditional ecology, virulence and potentially other traits of RNA viruses hosted follows a gradient of life history pace within mammals. Further work is urgently needed to identify the underlying evolutionary pressures that govern these host-virus relationships. The influence of mammal life history

upon zoonotic virus traits supports the growing recognition of nonhuman hosts as a key component of global human health. Risks of future virulent zoonoses may be mitigated by targeting surveillance and control to host species with specific life histories.

Chapter 6. Conclusion

6.1. Outline

This thesis aimed to identify and quantify the influence of ecological virus traits and host traits upon the virulence and human-to-human transmissibility of RNA viruses, with a view to better understanding viral emergence. I have presented, to my knowledge, the first dataset for virulence and transmissibility based on a structured review of literature that covers all known taxonomically-standardised human pathogens of a single type. By compiling such data, I was able to conduct comparative analyses of these characteristics using a comprehensively broad scope. I have demonstrated that while often non-linear and highly holistic, ecological traits of both viruses and hosts have discernible associations with virulence and transmissibility and are likely to play key roles in the ultimate dynamics of emerging RNA viruses.

Specifically, I found that the traits with the highest predictive power for virulence in classification models were tissue tropism breadth and transmission route, where broad tissue tropisms and nonvector-borne transmission predicted severe disease. Transmissibility itself was a weaker predictor of virulence, except within a small subset featuring viruses causing chronic infections, where sustained transmissibility predicted severe disease. Increased risk of virulence was also observed for viruses infecting a broader diversity of mammal host species. Furthermore, mammal host species with a smaller body mass (and by extension, a faster pace of overall life history) were more likely to host a higher proportion of zoonotic viruses associated with human virulence.

Increased risk of human-to-human transmissibility (though not self-sustained transmissibility) was observed for viruses with vector-borne or respiratory transmission routes. Increased risk of all forms of human-to-human transmissibility was observed for viruses infecting a narrower diversity of mammal taxonomic orders, as well as mammals with close phylogenetic similarity to humans. Additionally, adaptation towards transmissibility appeared to follow a variety of potential

evolutionary routes in state switching models, rather than hypothesised routes of stepwise or off-the-shelf movements exclusively. Finally, host and virus taxonomy were additional important influences upon both virulence and transmissibility throughout analyses. These findings are summarised in Figure 6.1.

6.2. Insights into virus ecology and evolution

The various predictive virus traits and host traits I find to be associated with virulence and transmissibility are diverse and vary widely in their scale (ranging from within-host tissue tropism, to known global diversity of host taxa). The connections between emerging virus dynamics and this highly varied range of predictors further confirms disease emergence as a complex, multifactorial phenomenon, and supports the use of frameworks applying a holistic approach to emerging pathogens (Wilcox and Colwell 2005; Wood et al. 2012).

The potential mechanisms, influences, and recommended avenues to aid understanding of each of these specific relationships in Figure 6.1 are discussed in detail in their respective chapters. However, several overarching trends and implications become evident when considering how the findings of this thesis may interrelate.

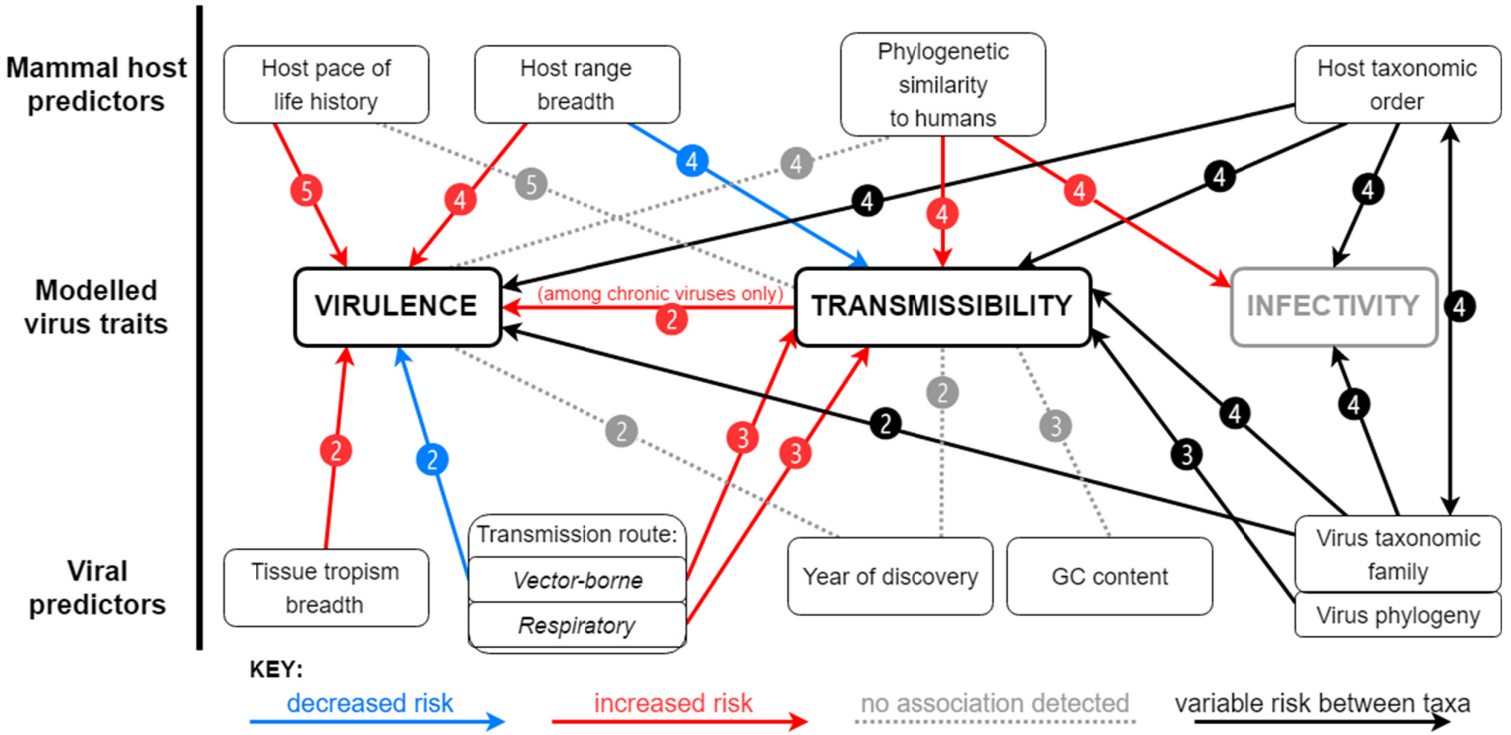


Figure 6.1. Simplified concept map of summarised thesis findings. Viral traits of central focus to the thesis, i.e. virulence and human-to-human transmissibility, are in the central row, along with additional modelled trait of infectivity. Top and bottom rows group potential nonhuman mammal host or virus trait predictors addressed throughout the thesis. Arrows denote relationships observed in analyses, where colours denote direction of association: blue - negative association (and therefore, a decreased risk of the trait in question); red – positive association (and therefore, increased risk); dotted grey – inconclusive or no association detected; and black – variably increased or decreased risk for different taxa. Numbered circles refer to the thesis chapter addressing evidence for each association

6.2.1. Virulence and niche diversity

Higher risk of virulence in humans was associated with broad niche diversity of RNA viruses at different ecological levels: both within-host (breadth of tissue tropism) and in host macroecology (breadth of mammal host species). This provides a phenomenological link between these two scales of disease ecology (Johnson et al. 2015b). An understanding of the mechanisms connecting these scales is needed to truly characterise the relationship between niche generalism and virulence, though it is currently unclear whether breadths of tissue tropism and host range are related. Although a broad tissue tropism and host range coincides for certain virus taxa (e.g. henipaviruses (Zeltina et al. 2016) and among DNA viruses, dependoparvoviruses (Hueffer and Parrish 2003)), there are suggestions these traits may not be strongly coupled, with tissue tropism likely to be more evolutionarily conserved than host range (Taber and Pease 1990). It also remains to be tested whether the association between risk of human virulence and niche diversity holds at intermediate scales, i.e. breadth of genotypes within host species, or breadth of host species within communities.

In a review of evolutionary processes, generalism in host range was predicted to lead to lower virulence than specialism in the majority of scenarios (Leggett et al. 2013), contradicting these findings. To further investigate the underlying causes behind this association, it will be essential to understand how niche diversity arises among emerging viruses. For example, my observations may hint towards a molecular mechanism of niche diversity as a critical contributor to viral virulence, such as usage of multiple receptors, or a single, highly-conserved receptor. As an illustration, henipaviruses cause severe human disease and enter cells via the ephrin-B2 surface protein, which is both expressed in a wide range of tissues and conserved across a wide range of mammals (Zeltina et al. 2016), allowing a diverse range of both tropisms and hosts. Several viruses analysed likely have broad niche generalism as a viral species because they comprise large numbers of strains or subspecies, each of which may individually have specialism for a single host or small group of related

hosts. This pattern has been recognised in the phylodynamics of *Rabies virus* (Streicker et al. 2010), with potential implications for the evolution of rabies virulence (Mollentze et al. 2014). It is not possible to distinguish which of these mechanisms explain niche diversity, and for which viral taxa, at the scale of my analyses. However, further hypotheses regarding generalism and virulence may be tested in future via study beyond the species resolution.

6.2.2. Antagonistic pleiotropy and constraints of human adaptation

I observed that human-to-human transmissibility, particularly sustained transmissibility, which may be thought of as an important phenotypic marker of human adaptation, became less likely with increasing breadth of host range of nonhuman mammals. This may be indicative of antagonistic pleiotropy, whereby genetic adaptation to one host consequently reduces suitability for other hosts. This phenomenon is well-documented in numerous experimental studies that show a negative trade-off between viral fitness within different host types (Crill et al. 2000; Turner and Elena 2000; Elena et al. 2009). Although antagonistic pleiotropy has primarily been explored with respect to intracellular infectivity and replication, my findings may imply identifiable patterns or strategies in host adaptation specifically regarding transmissibility, i.e. generalism, where viruses may exhibit some limited transmissibility in each of a number of hosts; and specialism, where viruses are adapted to efficient transmission within a single or few host(s).

Patterns of host specialisation in infectivity have generally been observed to be driven by viral adaptation rather than host defence (Antonovics et al. 2013), though may follow constraints of host phylogenies (Longdon et al. 2011; Cooper et al. 2012). Similar constraints to transmissibility are suggested by the relationships I observe between human-to-human transmissibility and phylogenetic similarity of nonhuman hosts to humans, as well as infection of nonhuman primates (Chapter 4). These host scenarios likely involve a shorter adaptive distance to be traversed to

develop transmissibility (Davies and Pedersen 2008; Parrish et al. 2008). Although not closely related to humans, bat-infective viruses also appeared more likely to develop sustained human-to-human transmissibility. Suggestions that bats are special or unique amongst zoonotic virus sources are still under close scrutiny (Dobson 2005; Olival et al. 2012).

Assuming distinctive strategies of generalism and specialism, regarding transmissibility, the important exceptions for both scientific study and public health priorities would be viruses that are capable of sustained transmission within a large number of distantly-related hosts. One potential determinant of these dynamics that warrants further exploration is viral genome size, as a larger genome may offer plasticity and scope for sufficient adaptation to multiple hosts (Holmes 2003a), a pattern observed within a genus of DNA viruses, the orthopoxviruses (Hendrickson et al. 2010).

Finally, adaptation may be particularly subject to pleiotropic or other genetic constraints for viruses with vector-borne transmission, as replication must occur in very different host environments of arthropods and vertebrates (Woelk and Holmes 2002; Holmes 2003a). Vector-borne viruses also showed little evidence of off-the-shelf jumps in adaptation (Chapter 3). In addition to evolutionary constraints, if arthropod vectors facilitate wide host exposure then infection of multiple hosts is likely (Woolhouse et al. 2001; Johnson et al. 2015a), but if this exposure is nonselective, vector transmission may simultaneously reduce opportunity for adaptation towards specialism. This is supported by my observations that vector-borne viruses were more likely to infect and transmit between humans in self-limited chains, but were not more likely to transmit between humans in self-sustained chains without need for other hosts (Chapter 3). It has been noted that there are very few human-adapted vector-borne viruses (Woolhouse and Adair 2013; Woolhouse et al. 2016), the only examples being dengue virus, yellow fever virus, chikungunya virus, and additionally reconsidered since these citations, Zika virus. A crucial determinant of host adaptation for vector-borne viruses is the ecological feeding preference of the

arthropod vectors, and each of the human-adapted vector-borne viruses has at least one vector species recognised to be anthropophilic (Gratz 2004; Chouin-Carneiro et al. 2016). It is evident that generally, vector-borne viruses follow different dynamics in their epidemiology and ecology to those of other viruses.

6.2.3. Evolutionary trade-offs in virulence and transmissibility

The trade-off hypothesis of virulence posits that transmission rate between individuals will increase as a function of virulence, by means of pathogen replication or development of symptoms. However, host mortality will consequently also increase as a function of virulence, such that resulting pathogen fitness or R_0 follows a trade-off between virulence and transmission throughout a longer infected host lifespan (Anderson and May 1982; Bremermann and Pickering 1983).

I observed that those ecological risk factors for human virulence predominantly did not coincide with those for human-to-human transmissibility, and in several cases, showed opposite directionality (Figure 6.1). For example, breadth of mammal host range was positively associated with virulence but negatively associated with transmissibility (Chapter 4), and vector-borne transmission generally predicted nonsevere disease (Chapter 2) but also predicted development of self-limited transmissibility (Chapter 3). These correlations would appear to support a broad scale trade-off between virulence and transmissibility within human hosts, at least at the resolution of viral species.

However, inference of a trade-off from these analyses may be blurred by the broad data definitions applied. Use of a standardised transmissibility measure was precluded by the large number of human viruses that are poorly studied. To ensure coverage of a broad diversity of viruses, human-to-human transmissibility was defined using both evidence of individual-scale transmissibility (e.g. case reports of transmission events), which might be considered equivalent to the transmissibility axis within the trade-off hypothesis, and population-scale transmissibility (e.g. R_0

calculations), which might be considered equivalent to viral fitness and the ultimate outcome of the trade-off hypothesis.

Furthermore, the observed patterns are likely to at least partly reflect differences in coevolutionary relationships between viruses and humans as hosts – if humans have not coevolved with a pathogen and are infected via zoonotic transmission without onward human-to-human transmission, virulence in humans will not be shaped by adaptation and can be coincidentally severe (Levin and Svanborg Edén 1990; Ebert and Bull 2008). However, the *a priori* expectations for levels of non-adapted virulence are not well-established. In experimental inoculations of *Drosophila* C virus (Longdon et al. 2015a), both increases and decreases in virulence were observed, broadly following phylogenetic clusters of *Drosophila* hosts that experienced either high or low virulence. If the absence of coevolution were to consistently result in a single direction of change in virulence in humans, this would suggest a detectable relationship between virulence and human-to-human transmissibility.

Potential for such a direct relationship was explored in Chapter 2, however, transmissibility appeared to be a poorer predictor of virulence than other traits. Human-to-human transmissibility only featured in the classification model for a subset of viruses associated with chronic disease, where, counter to expectation, human-transmissible viruses were predicted to be more highly virulent. Viruses associated with chronic disease may cause symptoms well after their transmission window, essentially decoupling the two (Bull 1994), which would again indicate a non-adapted virulence. Contrastingly, among pooled RNA and DNA viruses, human-to-human transmissibility was found to be associated with lower case fatality ratios (Geoghegan et al. 2016). However, this analysis also found greater risk of transmissibility for viruses with chronic infections, nonvector-borne transmission, and specific genomic structures, all of which are also likely to influence the evolution of virulence. Without a more holistic framework controlling for the complex

interplay between these traits, the precise relationship between virulence and transmissibility is difficult to elucidate.

Brought together, the above considerations imply that the nature of any trade-offs or other form of relationship between virulence and transmissibility will be highly dynamic and subject to many other viral traits (Ebert and Bull 2003; Alizon et al. 2009), making it particularly challenging to characterise at this comparative scale of study. Furthermore, virulence and transmissibility are often highly contextual, showing heterogeneity both within viral species and between host taxa. Therefore, developments in testing evolutionary models of virulence may be achieved by comparative or experimental analyses of viral strains and alternative host species (Alizon et al. 2009; Bull and Lauring 2014).

6.3. Key areas for further study

6.3.1. Knowledge and data gaps

Throughout the chapters presented, several core areas are apparent where current knowledge is deficient and may be improved upon with further data collection efforts. I investigated how risks of virulence and transmissibility vary with ecological viral traits, though how these risks are shaped by further traits, such as molecular factors, is not well-understood and may describe unexplained variation within my analyses. One key determinant of emergence not addressed here is receptor usage, as success of cross-species transmission will be influenced by conservation of receptors between hosts (Woolhouse et al. 2012). In addition, genetic traits such as substitution rates and loci associated with virulent or transmissible phenotypes (Pepin et al. 2010) are still largely unknown for human viruses, though proteomics and sequencing efforts have the potential to address this deficiency as they improve in coverage (Delwart 2013; Lipkin 2013).

Regarding data on nonhuman hosts, I primarily used predictors based on mammal species known to be infected, though the extent of known mammal-virus relationships is likely to be a very small fraction of the complete mammal virome

(Anthony et al. 2013), as data is deficient even for well-studied taxa (Cooper and Nunn 2013). Systematic, wide-ranging sampling strategies of animal hosts will therefore be necessary to improve predictive models. Beyond mammal species infected, a more meaningful basis for biological predictions might be those mammal species that viruses originated within or those that act as reservoirs (noting that these may not necessarily be the same species). The value of comprehensively documenting the origins of every human pathogen has been emphasised (Wolfe et al. 2007; Morse et al. 2012) and phylogenetic studies are closing this knowledge gap by assessing evidence for ancestral hosts among mammalian RNA virus families (Drexler et al. 2012; Longdon et al. 2015b). It is also important to identify reservoir hosts, i.e. those hosts viruses can be maintained in and transmit from, though these are difficult to accurately determine, e.g. reservoir potential may be synergistic and subject to presence of other host species (Haydon et al. 2002; Viana et al. 2014). Mathematical frameworks have been proposed to identify reservoirs through transmissibility metrics, based on epidemiological data (Fenton et al. 2015). Data on reservoir status and transmissibility would also allow a test of the presence of antagonistic pleiotropy across mammal-virus systems (see 'Insights into virus ecology and evolution'). Multivariate analyses could then be constructed following Chapter 3, to understand whether gain of transmissibility in one host directly coincides with loss in another.

In addition to transmissibility, I also was unable to directly compare virulence in humans to virulence in nonhuman hosts. Efforts to systematically compile data on nonhuman virulence would also allow verification of the traits I find associated with virulence in additional species beyond humans, whether reservoirs or incidental dead-end hosts. Initial attempts at structured data compilation suggest a wide variation in viral virulence within mammals (Levinson et al. 2013). Problematically, pathology and mortality are usually difficult to quantify in wild observational studies, creating incentives to turn to experimental studies. Experimental inoculations have been classically used to assess virulence, though existing data is likely to be focused towards specific laboratory models rather than species among natural host ranges.

6.3.2. Study directions and methods

Many determinants of cross-species transmission and emergence have been characterised (Taylor et al. 2001; Woolhouse and Gowtage-Sequeria 2005; Wolfe et al. 2007), and this thesis adds new dimensions in characterising determinants of virulence and transmissibility. Assuming adequate data were readily available, inference of emergence risk may be improved by several directions of study that can build upon the analyses presented herein.

Cross-species transmission is well-accepted to be dependent on both phylogenetic similarity and ecological contact (Wolfe et al. 2007), though the potential interdependencies of such risk factors remain ambiguous. One study directly comparing contributions of both phylogeny and contact for West Nile virus among birds concluded that neglecting either can result in wrongful conclusions (Roche et al. 2015). Models of zoonotic pathogen sharing combining phylogeny and contact have to date been mainly limited to primates (Pedersen and Davies 2009), representing a narrow window of the much wider vertebrate phylogeny. Better metrics of human contact are also required, as previous studies have often used the generalised measure of geographic overlap with human density (Pedersen and Davies 2009; Han et al. 2015b). One possibility may be creation of a conceptual gradient based on proximity to humans in both physical habitat and interfaces created by human activity (e.g. domesticity, population management, resource harvesting). Broader analyses of virus sharing that integrate phylogeny and ecology are currently under development (Olival et al. in review).

More broadly, methodologies that use highly-structured statistical frameworks also offer scope to investigate how determinants of disease emergence interact. For example, network-based modelling offers a natural way of representing interactions between hosts, pathogens or both (Godfrey 2013). Network models could allow insight into risks of viral virulence and transmissibility by quantifying the structure of: within-host interactions between viruses (Murall et al. 2012; Griffiths et

al. 2014); similarities between viruses in their ecological or clinical traits (Bogich et al. 2013); or perhaps most significantly, similarities between hosts in their known virome, in order to determine likely sources of emerging zoonoses (Gómez et al. 2013; Morand et al. 2014). Alternatively, structural equation modelling or similar methods could be used to investigate proposed routes of causality by integrating emergence predictors in a pathway-based model (Plowright et al. 2008). With these methods, specific hypotheses surrounding virulence and transmissibility predictors may be tested based on the connections and directionalities I observed throughout the thesis (Figure 6.1).

The next major methodological developments in comparative studies of disease emergence are likely to be transitions from species-based analyses to sequence-based analyses, especially considering sequencing efforts are increasing in efficiency and reducing in cost (Lipkin 2013). Firstly, this would assist in a move towards models that fully account for virus phylogenies – many of these associations I report are likely to have strong phylogenetic determinants yet to be fully characterised. I specified taxonomic family as a basic correction for phylogenetic relatedness, and was only able to specify viral phylogenies in analyses focusing on single families (Chapter 3). Comparatively high substitution rates and short genome lengths among RNA viruses mean that extreme divergence between families has impeded the reconstruction of a single complete phylogeny (Holmes 2003b). However, phylogenies at a wider resolution than taxonomic family have recently begun to be reconstructed from large-scale sampling efforts (Li et al. 2015).

Improvements to sequence data and viral phylogenetics will also synergise with trait-based models. Phylogenetic analyses could infer patterns in determinants of virulence and transmissibility at a finer resolution than viral species, assuming that phenotypic traits may be accurately assigned to sequences. Potential steps towards sequence-level trait data include closer collaboration between clinical diagnostics and genomic sampling efforts, experimental inoculation using animal models, or inference from known genetic signals, e.g. virulence-associated markers or mutations

(Brault et al. 2007). Phylogenetic models of infectivity based on sample host origins of sequences show a promising route to better understanding viral host switching (Streicker et al. 2010; Longdon et al. 2011, 2015a; Drexler et al. 2012). As identifying traits for individual sequences is likely to be resource-intensive, scope for phylogenetic analyses may be most efficient with a small, high-resolution focus. For example, The Vietnam Initiative on Zoonotic Infections (VIZIONS) focuses on sampling and phylogenetic analysis of viruses among a human cohort identified as at-risk for zoonotic transmission, and the animals they are exposed to (Rabaa et al. 2015).

6.4. Implications for global health

RNA viruses represent a major threat to global health, with concerns surrounding their potential for rapid emergence, pandemic spread, and debilitating clinical disease. Although public health infrastructures such as surveillance networks and reporting systems are globally improving (Chan et al. 2010), resources remain limited. A strategy shaped by scientific prediction would use available resources more effectively, and comparative analyses have been advocated as useful tools to aid in public health decisions (Daszak 2009; Morse et al. 2012). More informed and targeted strategies are beginning to be implemented through projects like USAID's Emerging Pandemic Threats programme (USAID 2009) that directly feed into to policy and applied risk assessments. However, there is still a clear shortage of targeted programmes for zoonotic transmission and disease emergence (Morse et al. 2012), and the analyses within this thesis have implications for the development of new public health strategies.

Overall, this thesis highlights that future emerging viruses will originate from a diverse range of zoonotic sources and through a diverse range of routes. At first, a tailored public health strategy may seem impossibly broad. However, emergence dynamics are not entirely unpredictable. I find that human RNA viruses tend to be shared with phylogenetically related species, are most likely to develop human

transmissibility following vector-borne or respiratory routes, and are most likely to cause severe human disease following nonvector-borne routes or when infecting a wide range of hosts and tissue types.

This information can be used to allocate resources in several ways. Firstly, regarding surveillance, sampling methodologies and assays can be prioritised towards those specific virus taxa of concern, e.g. coronaviruses and orthomyxoviruses as the most likely candidates for human transmissibility (Chapter 4), or arenaviruses and filoviruses as the most likely candidates for severe human disease (Chapters 2, 4). Surveillance of specific targets has also been advocated (Morse et al. 2012). Considering likely sources of viruses with epidemic potential (i.e. sustained human-to-human transmissibility) within state-switching models (Chapter 3), surveillance could be tailored to animal hosts for paramyxoviruses, where epidemic potential was most likely following zoonotic transmission; and large human populations for alphaviruses and flaviviruses, where epidemic potential was most likely following existing self-limited human-to-human transmission. Furthermore, my findings support previously proposed human populations to target, such as those in regular domestic or occupational contact with arthropod vectors or key mammal host taxa of bats and primates, e.g. forest workers or bushmeat hunters (Wolfe et al. 2000).

Beyond surveillance, the factors I find associated with virulence and transmissibility offer criteria for quantifying viruses during risk assessments (Mangen et al. 2010). The virulence classification model presented (Chapter 2) may be also applied directly or within a larger framework as an immediate, inexpensive risk assessment tool. Regarding ultimate control initiatives, the priority virus families identified by my analyses may represent key coverage areas for future vaccine development, and those priority transmission routes may represent important considerations in selecting intervention methods.

Inclusion of wild animal populations within surveillance and control strategies is also integral to public health (Wendt et al. 2014; Johnson et al. 2015b), and supported by the associations I observe between virulence, transmissibility, and

mammal host traits. Several existing passive surveillance systems anticipate human outbreaks and issue warnings to regional public health divisions based on reports of infection in certain indicator hosts, e.g. birds for mosquito-borne viruses (Wendt et al. 2014). The same could be implemented for those key host groups I find associated with human-transmissible viruses (Chapter 4).

Although the finding that virulence correlates with a broad range of mammal hosts (Chapter 4) may initially suggest a need for widespread control measures across many mammal taxa, it will first be necessary to quantify the contributions of each nonhuman host to zoonotic transmission (Woolhouse et al. 2001; Fenton et al. 2015). Overall, interventions at the zoonotic interface (such as contact barriers or reduction of bushmeat pressure) may be most productive when targeted to bats, primates, wild carnivores, or mammal species with faster-paced life history (Chapters 4, 5). Unfortunately, the latter may present a significant challenge in that fast-paced mammals are often capable of survival in highly changeable conditions, including peridomestic and urban environments (Mills 2006). A further method of targeted control may be population management or culling, though it must be stressed that any culling must be carried out carefully - sufficient prior understanding of local ecology is critical to minimise risks of unintended consequences. For example, removal of established reservoir hosts may result in replacement by opportunistic, short-lived species that may host more virulent human viruses (Chapter 5) or indirectly alter selection pressures for virulence and transmission (Bolzoni and De Leo 2013).

In the longer-term, recommendations for public health strategies will be refined as knowledge expands from wider data availability and better model inference. Therefore, these suggested strategies should be periodically revisited and evaluated to avoid becoming obsolete or inflexible (Wilcox and Colwell 2005). The value of predictive comparative study for this purpose is clear. Although disease emergence and public health crises are often contingent on single chance events, such as the translocation of a single arthropod vector or transmission of a single rare

genotype, predictive models as I present in this thesis can aid in understanding and quantifying disease emergence by addressing the wider risk of such events with a universal perspective.

6.5. Concluding remarks

The emergence of novel zoonotic viruses in the near future is inevitable, posing an immediate threat to global health regarding risks of increasing disease burdens and pandemic spread. This thesis has demonstrated that the determinants of viral virulence and transmissibility within humans are diverse, but broadly predictable. The ecology and evolution of both viruses and their associated hosts play significant roles in shaping viral phenotypes. The work presented herein offers inference into the likely trait profiles of high-impact emerging viruses and can contribute to a more effective, pre-emptive, and adaptive public health strategy. This thesis highlights the fundamental connectivity between pathogens, humans, and their wider hosts, and amidst a frame that brings together clinicians, veterinarians, virologists and ecologists, fits a further piece of the ever-challenging puzzle of emerging infectious diseases.

References

Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716–23.

Al-Garib SO, Gielkens ALJ, Gruys E, Peeters BPH, Koch G. Tissue tropism in the chicken embryo of non-virulent and virulent Newcastle diseases strains that express green fluorescence protein. Avian Pathol. 2003;32(6):591–6.

Alizon S, Hurford A, Mideo N, Van Baalen M. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. J Evol Biol. 2009;22(2):245–59.

Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol. 2006;43(6):1223–32.

Anderson RM, May RM. Coevolution of hosts and parasites. Parasitology. 1982;85(2):411–26.

Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, Solovyov A, et al. A strategy to estimate unknown viral diversity in mammals. mBio. 2013;4(5):e00598-13.

Anthony SJ, Islam A, Johnson C, Navarrete-Macias I, Liang E, Jain K, et al. Non-random patterns in viral diversity. Nat Commun. 2015;6:8147.

Antia R, Regoes RR, Koella JC, Bergstrom CT. The role of evolution in the emergence of infectious diseases. Nature. 2003;426(6967):658–61.

Antonovics J, Boots M, Ebert D, Koskella B, Poss M, Sadd BM. The origin of specificity by means of natural selection: evolved and nonhost resistance in host-pathogen interactions. Evolution. 2013;67(1):1–9.

Auewarakul P. Composition bias and genome polarity of RNA viruses. Virus Res. 2005;109(1):33–7.

Bahir I, Fromer M, Prat Y, Linial M. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol Syst Biol. 2009;5:311.

Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1–48.

Belshaw R, Gardner A, Rambaut A, Pybus OG. Pacing a small cage: mutation and RNA viruses. Trends Ecol Evol. 2008;23(4):188–93.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2005;33:D34–8.

Bininda-Emonds OR., Cardillo M, Jones KE, MacPhee RD., Beck RM., Grenyer R, et al. The delayed rise of present-day mammals. Nature. 2007;446(7135):507–12.

Bogich TL, Funk S, Malcolm TR, Chhun N, Epstein JH, Chmura AA, et al. Using network theory to identify the causes of disease outbreaks of unknown origin. J R Soc Interface. 2013;10(81):20120904.

Bolzoni L, De Leo GA. Unexpected consequences of culling on the eradication of wildlife diseases: the role of virulence evolution. Am Nat. 2013;181(3):301–13.

Brault AC, Huang CY-H, Langevin SA, Kinney RM, Bowen RA, Ramey WN, et al. A single positively selected West Nile viral mutation confers increased virogenesis in American crows. Nat Genet. 2007;39(9):1162–6.

Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. Lancet. 2013;382(9893):694–9.

Bremermann HJ, Pickering J. A game-theoretical model of parasite virulence. J Theor Biol. 1983;100(3):411–26.

Brooke ZM, Bielby J, Nambiar K, Carbone C. Correlates of research effort in carnivores: body size, range size and diet matter. PLoS ONE. 2014;9(4):e93195.

Bull JJ. Perspective: virulence. Evolution. 1994;48(5):1423–37.

Bull JJ, Lauring AS. Theory and empiricism in virulence evolution. PLoS Pathog. 2014;10(10):e1004387.

Cable JM, Enquist BJ, Moses ME. The allometry of host-pathogen interactions. PLoS ONE. 2007;2(11):e1130.

CDC. Update: Investigations of West Nile virus infections in recipients of organ transplantation and blood transfusion--Michigan, 2002. MMWR Morb Mortal Wkly Rep. 2002;51(39):879.

CDC. Interim pre-pandemic planning guidance; community strategy for pandemic influenza mitigation in the United States : early, targeted, layered use of nonpharmaceutical interventions. Centers for Disease Control and Prevention (U.S.); 2007. Available from: http://stacks.cdc.gov/view/cdc/11425/

Chan EH, Brewer TF, Madoff LC, Pollack MP, Sonricker AL, Keller M, et al. Global capacity for emerging infectious disease detection. Proc Natl Acad Sci USA. 2010;107(50):21701–6.

Charrel RN, Leparc-Goffart I, Gallian P, de Lamballerie X. Globalization of Chikungunya: 10 years to invade the world. Clin Microbiol Infect. 2014;20(7):662–3.

Childs JE, Richt JA, Mackenzie JS. Introduction: conceptualizing and partitioning the emergence process of zoonotic viruses from wildlife to humans. Curr Top Microbiol Immunol. 2007;315:1–31.

Chouin-Carneiro T, Vega-Rua A, Vazeille M, Yebakima A, Girod R, Goindin D, et al. Differential susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika virus. PLoS Negl Trop Dis. 2016;10(3):e0004543.

Chua KB, Bellini WJ, Rota PA. Nipah virus outbreak in Malaysia. J Clin Virol. 2003;26(3):265–75.

Chua KB, Bellini WJ, Rota PA, Harcourt BH, Tamin A, Lam SK, et al. Nipah virus: a recently emergent deadly paramyxovirus. Science. 2000;288(5470):1432–5.

Cleaveland S, Laurenson MK, Taylor LH. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Phil Trans R Soc B. 2001;356(1411):991–9.

Cooper N, Griffin R, Franz M, Omotayo M, Nunn CL. Phylogenetic host specificity and understanding parasite sharing in primates. Ecol Lett. 2012;15(12):1370–7.

Cooper N, Nunn CL. Identifying future zoonotic disease threats. Evol Med Public Health. 2013;2013(1):27–36.

Crill WD, Wichman HA, Bull JJ. Evolutionary reversals during viral adaptation to alternating hosts. Genetics. 2000;154(1):27–37.

Cuthill JH, Charleston MA. A simple model explains the dynamics of preferential host switching among mammal RNA viruses. Evolution. 2013;67(4):980–90.

Daszak P. A call for 'smart surveillance': a lesson learned from H1N1. EcoHealth. 2009;6(1):1–2.

Daugherty MD, Malik HS. Rules of engagement: molecular insights from host-virus arms races. Annu Rev Genet. 2012;46(1):677–700.

Davies TJ, Pedersen AB. Phylogeny and geography predict pathogen community similarity in wild primates and humans. Proc R Soc B. 2008;275(1643):1695–701.

Day T. On the evolution of virulence and the relationship between various measures of mortality. Proc R Soc B. 2002a;269(1498):1317–23.

Day T. The evolution of virulence in vector-borne and directly transmitted parasites. Theor Popul Biol. 2002b;62(2):199–213.

De Leo GA, Dobson AP. Allometry and simple epidemic models for microparasites. Nature. 1996;379(6567):720–2.

De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 2000;81(11):3178–92.

Delwart E. A roadmap to the human virome. PLoS Pathog. 2013;9(2):e1003146.

Dobson A. Population dynamics of pathogens with multiple host species. Am Nat. 2004;164(5):S64–78.

Dobson AP. What links bats to emerging infectious diseases? Science. 2005;310(5748):628–9.

Drexler JF, Corman VM, Müller MA, Maganga GD, Vallo P, Binger T, et al. Bats host major mammalian paramyxoviruses. Nat Commun. 2012;3:796.

Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267–76.

Dunn RR, Davies TJ, Harris NC, Gavin MC. Global drivers of human pathogen richness and prevalence. Proc R Soc B. 2010;277(1694):2587–95.

Ebert D, Bull JJ. Challenging the trade-off model for the evolution of virulence: is virulence management feasible? Trends Microbiol. 2003;11(1):15–20.

Ebert D, Bull JJ. The evolution and expression of virulence. In: Stearns SC, Koella JC, editors. Evolution in health and disease. Oxford University Press; 2008. p. 153–67.

Elena SF, Agudelo-Romero P, Lalić J. The evolution of viruses in multi-host fitness landscapes. Open Virol J. 2009;3:1–6.

Epstein JH, Field HE, Luby S, Pulliam JRC, Daszak P. Nipah virus: Impact, origins, and causes of emergence. Curr Infect Dis Rep. 2006;8(1):59–65.

Ewald PW. Host-parasite relations, vectors, and the evolution of disease severity. Ann Rev Ecol Syst. 1983;14:465–85.

Fargette D, Pinel-Galzi A, Sérémé D, Lacombe S, Hébrard E, Traoré O, et al. Diversification of rice yellow mottle virus and related viruses spans the history of agriculture from the Neolithic to the present. PLoS Pathog. 2008;4(8):e1000125.

Feldmann H, Geisbert TW. Ebola haemorrhagic fever. Lancet. 2011;377(9768):849–62.

Fenton A, Pedersen AB. Community epidemiology framework for classifying disease threats. Emerg Infect Dis. 2005;11(12):1815–21.

Fenton A, Streicker DG, Petchey OL, Pedersen AB. Are all hosts created equal? Partitioning host species contributions to parasite persistence in multihost communities. Am Nat. 2015;186(5):610–22.

Focosi D, Maggi F. Estimates of Ebola virus case-fatality ratio in the 2014 West African outbreak. Clin Infect Dis. 2015;60(5):829.

Fonkwo PN. Pricing infectious disease. EMBO Rep. 2008;9:S13–7.

Fox J, Weisberg S. An R companion to applied regression. Second. Thousand Oaks CA: Sage; 2011.

Franco DJ, Vago AR, Chiari E, Meira FCA, Galvão LMC, Machado CRS. Trypanosoma cruzi: mixture of two populations can modify virulence and tissue tropism in rat. Exp Parasitol. 2003;104(1–2):54–61.

Frank AF. mer-utils. R script. 2011. Available from: https://github.com/aufrank/R-hacks/blob/master/mer-utils.R

Funk S, Bogich TL, Jones KE, Kilpatrick AM, Daszak P. Quantifying trends in disease impact to produce a consistent and reproducible definition of an emerging infectious disease. PLoS ONE. 2013;8(8):e69951.

Gandon S. Evolution of multihost parasites. Evolution. 2004;58(3):455–69.

Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science. 2009;325(5937):197–201.

Gatherer D, Kohl A. Zika virus: a previously slow pandemic spreads rapidly through the Americas. J Gen Virol. 2015;97(2):269–73.

GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;385(9963):117–71.

Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Biol. 2008;2(4):1360–83.

Geoghegan JL, Senior AM, Giallonardo FD, Holmes EC. Virological factors that increase the transmissibility of emerging human viruses. Proc Natl Acad Sci USA. 2016;113(15):4170–5.

Godfrey SS. Networks and the ecology of parasite transmission: A framework for wildlife parasitology. Int J Parasitol Parasites Wildl. 2013;2:235–45.

Gómez JM, Nunn CL, Verdú M. Centrality in primate–parasite networks reveals the potential for the transmission of emerging infectious diseases to humans. Proc Natl Acad Sci USA. 2013;110(19):7738–41.

Grafen A. The phylogenetic regression. Phil Trans R Soc B. 1989;326(1233):119–57.

Grard G, Moureau G, Charrel RN, Holmes EC, Gould EA, de Lamballerie X de. Genomics and evolution of Aedes-borne flaviviruses. J Gen Virol. 2010;91(1):87–94.

Gratz NG. Critical review of the vector status of Aedes albopictus. Med Vet Entomol. 2004;18(3):215–27.

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 2008;4(6):e1000079.

Griffiths EC, Pedersen AB, Fenton A, Petchey OL. Analysis of a summary network of co-infection in humans reveals that parasites interact most via shared resources. Proc R Soc B. 2014;281(1782):20132286.

Groseth A, Feldmann H, Strong JE. The ecology of Ebola virus. Trends Microbiol. 2007;15(9):408–16.

Gubler DJ. The changing epidemiology of yellow fever and dengue, 1900 to 2003: full circle? Comp Immunol Microbiol Infect Dis. 2004;27(5):319–30.

Guernier V, Hochberg ME, Guégan JF. Ecology drives the worldwide distribution of human diseases. PLoS Biol. 2004;2(6):e141.

Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Soft. 2010;33(2):1–22.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science. 2000;287(5453):607–14.

Halpin K, Hyatt AD, Plowright RK, Epstein JH, Daszak P, Field HE, et al. Emerging viruses: coming in on a wrinkled wing and a prayer. Clin Infect Dis. 2007;44(5):711–7.

Han BA, Park AW, Jolles AE, Altizer S. Infectious disease transmission and behavioural allometry in wild mammals. J Anim Ecol. 2015a;84(3):637–46.

Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. Proc Natl Acad Sci USA. 2015b;112(22):7039–44.

Harrell FEJ. rms: regression modeling strategies. R package version 3.2-0. 2011. Available from: http://cran.r-project.org/package=rms

Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, et al. Global mapping of infectious disease. Phil Trans R Soc B. 2013;368(1614):20120250.

Haydon DT, Cleaveland S, Taylor LH, Laurenson MK. Identifying reservoirs of infection: a conceptual and practical challenge. Emerg Infect Dis. 2002;8(12):1468–73.

Hendrickson RC, Wang C, Hatcher EL, Lefkowitz EJ. Orthopoxvirus genome evolution: the role of gene loss. Viruses. 2010;2(9):1933–67.

Hicks AL, Duffy S. Cell tropism predicts long-term nucleotide substitution rates of mammalian RNA viruses. PLoS Pathog. 2014;10(1):e1003838.

Holmes E, Drummond A. The evolutionary genetics of viral emergence. In: Childs JE, Mackenzie JS, Richt JA, editors. Wildlife and emerging zoonotic diseases: the biology, circumstances and consequences of cross-species transmission. Springer Berlin Heidelberg; 2007. p. 51–66.

Holmes EC. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. 2003a;11(12):543–6.

Holmes EC. Molecular clocks and the puzzle of RNA virus origins. J Virol. 2003b;77(7):3893–7.

Holmes EC. The evolution and emergence of RNA viruses. Oxford; New York: Oxford University Press; 2009.

Holmes EC. What does virus evolution tell us about virus origins? J Virol. 2011;85(11):5247–51.

Holmes KV. Adaptation of SARS coronavirus to humans. Science. 2005;309(5742):1822–3.

Howard CR. Arenaviruses. In: Zuckerman AJ, Banatvala JE, Schoub BD, Griffiths PD, Mortimer P, editors. Principles and practice of clinical virology. John Wiley & Sons, Ltd; 2009. p. 733–54.

Hueffer K, Parrish CR. Parvovirus host range, cell tropism and evolution. Curr Opin Microbiol. 2003;6(4):392–8.

Johnson CK, Hitchens PL, Evans TS, Goldstein T, Thomas K, Clements A, et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. Sci Rep. 2015a;5:14830.

Johnson PTJ, Rohr JR, Hoverman JT, Kellermanns E, Bowerman J, Lunde KB. Living fast and dying of infection: host life history drives interspecific variation in infection and disease risk. Ecol Lett. 2012;15(3):235–42.

Johnson PTJ, Roode JC de, Fenton A. Why infectious disease research needs community ecology. Science. 2015b;349(6252):1259504.

Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, et al. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology. 2009;90(9):2648.

Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451(7181):990–3.

Kapoor A, Simmonds P, Lipkin WI, Zaidi S, Delwart E. Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. J Virol. 2010;84(19):10322–8.

Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, et al. Ecology of zoonoses: natural and unnatural histories. Lancet. 2012;380(9857):1936–45.

Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90(430):773–95.

King AM, Lefkowitz E, Adams MJ, Carstens EB. Virus taxonomy: Ninth report of the International Committee on Taxonomy of Viruses. Elsevier; 2011.

Kingsford C, Salzberg SL. What are decision trees? Nat Biotech. 2008;26(9):1011–3.

Kitchen A, Shackelton LA, Holmes EC. Family level phylogenies reveal modes of macroevolution in RNA viruses. Proc Natl Acad Sci USA. 2011;108(1):238–43.

Knipe DM, Howley PM. Fields virology, 5th Edition. Lippincott Williams & Wilkins; 2007.

Knowles JE, Frederick C. merTools: tools for analyzing mixed effect regression models. R package version 0.2.0. 2011. Available from: http://cran.r-project.org/package=merTools

Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology. 2015;479:2–25.

Koonin EV, Senkevich TG, Dolja VV. The ancient virus world and evolution of cells. Biol Direct. 2006;1:29.

Koopmans M. Surveillance strategy for early detection of unusual infectious disease events. Curr Opin Virol. 2013;3(2):185–91.

Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. Host species barriers to influenza virus infections. Science. 2006;312(5772):394–7.

Kümmerli R, Schiessl KT, Waldvogel T, McNeill K, Ackermann M. Habitat structure and the evolution of diffusible siderophores in bacteria. Ecol Lett. 2014;17(12):1536–44.

Kuzmin IV, Shi M, Orciari LA, Yager PA, Velasco-Villa A, Kuzmina NA, et al. Molecular inferences suggest multiple host shifts of rabies viruses from bats to mesocarnivores in arizona during 2001–2009. PLoS Pathog. 2012;8(6):e1002786.

Lanciotti RS, Ludwig ML, Rwaguma EB, Lutwama JJ, Kram TM, Karabatsos N, et al. Emergence of epidemic O'nyong-nyong fever in Uganda after a 35-year absence: genetic characterization of the virus. Virology. 1998;252(1):258–68.

Lau SKP, Woo PCY, Li KSM, Huang Y, Tsoi H-W, Wong BHL, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci USA. 2005;102(39):14040–5.

Lee KA. Linking immune defenses and life history at the levels of the individual and the species. Integr Comp Biol. 2006;46(6):1000–15.

Leggett HC, Buckling A, Long GH, Boots M. Generalism and the evolution of parasite virulence. Trends Ecol Evol. 2013;28(10):592–6.

Leggett HC, Cornwallis CK, West SA. Mechanisms of pathogenesis, infective dose and virulence in human parasites. PLoS Pathog. 2012;8(2):e1002512.

Levin B., Svanborg Edén C. Selection and evolution of virulence in bacteria: an ecumenical excursion and modest suggestion. Parasitology. 1990;100(Supplement S1):S103–S115.

Levin BR, Bull JJ. Short-sighted evolution and the virulence of pathogenic microorganisms. Trends Microbiol. 1994;2(3):76–81.

Levinson J, Bogich TL, Olival KJ, Epstein JH, Johnson CK, Karesh W, et al. Targeting surveillance for zoonotic virus discovery. Emerg Infect Dis. 2013;19(5):743–7.

Li C-X, Shi M, Tian J-H, Lin X-D, Kang Y-J, Chen L-J, et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. eLife. 2015;4:e05378.

Li W, Zhang C, Sui J, Kuhn JH, Moore MJ, Luo S, et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO. 2005;24(8):1634–43.

Lindenfors P, Nunn CL, Jones KE, Cunningham AA, Sechrest W, Gittleman JL. Parasite species richness in carnivores: effects of host body mass, latitude, geographical range and population density. Global Ecol Biogeogr. 2007;16(4):496–509.

Lindstedt SL, Calder WA. Body size, physiological time, and longevity of homeothermic animals. Q Rev Biol. 1981;56(1):1–16.

Lipkin WI. The changing face of pathogen discovery and surveillance. Nat Rev Micro. 2013;11(2):133–41.

Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, Dobson AP, et al. Epidemic dynamics at the human-animal interface. Science. 2009;326(5958):1362–7.

Lochmiller RL, Deerenberg C. Trade-offs in evolutionary immunology: just what is the cost of immunity? Oikos. 2000;88(1):87–98.

Loh EH, Zambrana-Torrelio C, Olival KJ, Bogich TL, Johnson CK, Mazet JAK, et al. Targeting transmission pathways for emerging zoonotic disease surveillance and control. Vector Borne Zoonotic Dis. 2015;15(7):432–7.

Longdon B, Hadfield JD, Day JP, Smith SCL, McGonigle JE, Cogni R, et al. The causes and consequences of changes in virulence following pathogen host shifts. PLoS Pathog. 2015a;11(3):e1004728.

Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. Host phylogeny determines viral persistence and replication in novel hosts. PLoS Pathog. 2011;7(9):e1002260.

Longdon B, Murray GG, Palmer WJ, Day JP, Parker DJ, Welch JJ, et al. The evolution, diversity and host associations of rhabdoviruses. Virus Evol. 2015b;1(1):vev014.

Lu L. Unpublished data.

Luis AD, Hayman DTS, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JRC, et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proc R Soc B. 2013;280(1756):20122753.

Luis AD, O'Shea TJ, Hayman DTS, Wood JLN, Cunningham AA, Gilbert AT, et al. Network analysis of host–virus communities in bats and rodents reveals determinants of cross-species transmission. Ecol Lett. 2015;18(11):1153–62.

Mackinnon MJ, Gandon S, Read AF. Virulence evolution in response to vaccination: The case of malaria. Vaccine. 2008;26(Supplement 3):C42–52.

Mangen M-JJ, Batz MB, Käsbohrer A, Hald T, Morris JG, Taylor M, et al. Integrated approaches for the public health prioritization of foodborne and zoonotic pathogens. Risk Anal. 2010;30(5):782–97.

Mathers C, Fat DM, Organization WH, Boerma JT. The global burden of disease: 2004 update. World Health Organization; 2008.

McNally L, Viana M, Brown SP. Cooperative secretions facilitate host range expansion in bacteria. Nat Commun. 2014;5:5494.

Messenger AM, Barnes AN, Gray GC. Reverse zoonotic disease transmission (zooanthroponosis): a systematic review of seldom-documented human biological threats to animals. PLoS ONE. 2014;9(2):e89055.

Mickleburgh S, Waylen K, Racey P. Bats as bushmeat: a global review. Oryx. 2009;43(2):217–34.

Mills JN. Regulation of rodent-borne viruses in the natural host: implications for human disease. In: Peters CJ, Calisher CH, editors. Infectious diseases from nature: mechanisms of viral emergence and persistence. Springer Vienna; 2005. p. 45–57.

Mills JN. Biodiversity loss and emerging infectious disease: an example from the rodent-borne hemorrhagic fevers. Biodiversity. 2006;7(1):9–17.

Molinari N-AM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. Vaccine. 2007;25(27):5086–96.

Mollentze N, Biek R, Streicker DG. The role of viral evolution in rabies host shifts and emergence. Curr Opin Virol. 2014;8:68–72.

Morand S, McIntyre KM, Baylis M. Domesticated animals and human infectious diseases of zoonotic origins: Domestication time matters. Infect Genet Evol. 2014;24:76–81.

Morikawa S, Saijo M, Kurane I. Current knowledge on lower virulence of Reston Ebola virus. Comp Immunol Microb. 2007;30(5–6):391–8.

Morse SS. Factors in the emergence of infectious diseases. Emerg Infect Dis. 1995;1(1):7–15.

Morse SS, Mazet JA, Woolhouse MEJ, Parrish CR, Carroll D, Karesh WB, et al. Prediction and prevention of the next pandemic zoonosis. Lancet. 2012;380(9857):1956–65.

Murall CL, McCann KS, Bauch CT. Food webs in the human body: linking ecological theory to viral dynamics. PLoS ONE. 2012;7(11):e48812.

Nathanson N, Gonzalez-Scarano F, Nathanson N. Viral virulence. In: Viral Pathogenesis and Immunity. Academic Press; 2007. p. 113–29.

Nidelet T, Koella JC, Kaltz O. Effects of shortened host life span on the evolution of parasite life history and virulence in a microbial host-parasite system. BMC Evol Biol. 2009;9(1):65.

Nunn CL. Primate disease ecology in comparative and theoretical perspective. Am J Primatol. 2012;74(6):497–509.

Nunn CL, Altizer S, Jones KE, Sechrest W. Comparative tests of parasite species richness in primates. Am Nat. 2003;162(5):597–614.

Nunn CL, Altizer SM. The global mammal parasite database: an online resource for infectious disease records in wild primates. Evol Anthropol. 2005;14(1):1–2.

Nunn CL, Dokey AT-W. Ranging patterns and parasitism in primates. Global Ecol Biogeogr. 2006;2(3):351–4.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. vegan: community ecology package. 2015. Available from: http://cran.r-project.org/package=vegan

Olival KJ, Epstein JH, Wang L-F, Field HE. Are bats exceptional viral reservoirs? In: Aguirre AA, Ostfeld RS, Daszak P, editors. New directions in conservation medicine: Applied cases of ecological health. Oxford University Press; 2012. p. 195–212.

Olival KJ, Hosseini PR, Bogich TL, Zambrana-Torrelio CM, Daszak P. Host and viral traits predict zoonotic spillover from mammals. Nature. (in review).

Pagel M, Meade A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat. 2006;167(6):808–25.

Pagel M, Meade A, Barker D. Bayesian estimation of ancestral character states on phylogenies. Syst Biol. 2004;53(5):673–84.

Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.

Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, et al. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev. 2008;72(3):457–70.

Patterson JL, Carrion R. Demand for nonhuman primate resources in the age of biodefense. ILAR J. 2005;46(1):15–22.

Patz JA, Daszak P, Tabor GM, Aguirre AA, Pearl M, Epstein J, et al. Unhealthy landscapes: policy recommendations on land use change and infectious disease emergence. Environ Health Persp. 2004;112(10):1092–8.

Pedersen AB, Davies TJ. Cross-species pathogen transmission and disease emergence in primates. EcoHealth. 2009;6(4):496–508.

Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, et al. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol Evol. 2014;5(9):961–70.

Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO. Identifying genetic markers of adaptation for surveillance of viral host jumps. Nat Rev Micro. 2010;8(11):802–13.

Plowright RK, Sokolow SH, Gorman ME, Daszak P, Foley JE. Causal inference in disease ecology: investigating ecological drivers of disease emergence. Front Ecol Environ. 2008;6(8):420–9.

Plummer M, Best N, Cowles K, Vines K, Sarkar D, Almond R. coda: Output analysis and diagnostics for MCMC. R package version 0.16-1. 2012. Available from: http://cran.r-project.org/package=coda

Powers AM, Logue CH. Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. J Gen Virol. 2007;88(9):2363–77.

Previtali MA, Ostfeld RS, Keesing F, Jolles AE, Hanselmann R, Martin LB. Relationship between pace of life and immune responses in wild rodents. Oikos. 2012;121(9):1483–92.

Promislow DEL, Harvey PH. Living fast and dying young: A comparative analysis of life-history variation among mammals. J Zool. 1990;220(3):417–37.

Pulliam JRC. Viral host jumps: moving toward a predictive framework. EcoHealth. 2008;5(1):80–91.

Pulliam JRC, Dushoff J. Ability to replicate in the cytoplasm predicts zoonotic transmission of livestock viruses. J Infect Dis. 2009;199(4):565–8.

Rabaa MA, Tue NT, Phuc TM, Carrique-Mas J, Saylors K, Cotten M, et al. The Vietnam Initiative on Zoonotic Infections (VIZIONS): a strategic approach to studying emerging zoonotic infectious diseases. EcoHealth. 2015;12(4):726–35.

Rabadan R, Levine AJ, Robins H. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. J Virol. 2006;80(23):11887–91.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www. R-project. org; 2015.

Reed C, Biggerstaff M, Finelli L, Koonin LM, Beauvais D, Uzicanin A, et al. Novel framework for assessing epidemiologic effects of influenza epidemics and pandemics. Emerg Infect Dis. 2013;19(1):85–91.

Ren W, Li W, Yu M, Hao P, Zhang Y, Zhou P, et al. Full-length genome sequences of two SARS-like coronaviruses in horseshoe bats and genetic variation analysis. J Gen Virol. 2006;87(11):3355–9.

Richman DD, Whitley RJ, Hayden FG. Clinical virology. John Wiley & Sons; 2009.

Ricklefs RE, Wikelski M. The physiology/life-history nexus. Trends Ecol Evol. 2002;17(10):462–8.

Rigaud T, Perrot-Minnot M-J, Brown MJF. Parasite and host assemblages: embracing the reality will improve our knowledge of parasite transmission and virulence. Proc R Soc B. 2010;277(1701):3693–702.

Robnik-Sikonja M, Savicky P. CORElearn: Classification, regression, feature evaluation and ordinal evaluation. R package version 0.9.43. 2014; Available from: http://cran.r-project.org/package=CORElearn

Roche B, Morand S, Elguero E, Balenghien T, Guégan J-F, Gaidet N. Does host receptivity or host exposure drives dynamics of infectious diseases? The case of West Nile Virus in wild birds. Infect Genet Evol. 2015;33:11–9.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. Genome Res. 2010;20(8):1001–9.

Rosenberg R, Beard CB. Vector-borne Infections. Emerg Infect Dis. 2011;17(5):769–70.

Rosenberg R, Johansson MA, Powers AM, Miller BR. Search strategy has influenced the discovery rate of human viruses. Proc Natl Acad Sci USA. 2013;110(34):13961–4.

Skaug H, Fournier D, Bolker B, Magnusson A, Nielsen A. Generalized linear mixed models using 'AD Model Builder'. 2015. Available from: http://glmmadmb.r-forge.r-project.org

Song H-D, Tu C-C, Zhang G-W, Wang S-Y, Zheng K, Lei L-C, et al. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci USA. 2005;102(7):2430–5.

Spellberg B, Taylor-Blake B. On the exoneration of Dr. William H. Stewart: debunking an urban legend. Infect Dis Poverty. 2013;2:3.

Spence PJ, Jarra W, Lévy P, Reid AJ, Chappell L, Brugat T, et al. Vector transmission regulates immune control of Plasmodium virulence. Nature. 2013;498(7453):228–31.

Spengler JR, Ervin ED, Towner JS, Rollin PE, Nichol ST. Perspectives on West Africa Ebola virus disease outbreak, 2013–2016. Emerg Infect Dis. 2016;22(6):956-63.

Srinivasan A, Burton EC, Kuehnert MJ, Rupprecht C, Sutker WL, Ksiazek TG, et al. Transmission of rabies virus from an organ donor to four transplant recipients. N Engl J Med. 2005;352(11):1103–11.

Steele JH, Fernandez PJ. History of rabies and global aspects. In: Baer GM, editor. The natural history of rabies. Boca Raton, FL: CRC press; 1991. p. 1–26.

Streicker DG, Altizer SM, Velasco-Villa A, Rupprecht CE. Variable evolutionary routes to host establishment across repeated rabies virus host shifts among bats. Proc Natl Acad Sci USA. 2012a;109(48):19715–20.

Streicker DG, Lemey P, Velasco-Villa A, Rupprecht CE. Rates of viral evolution are linked to host geography in bat rabies. PLoS Pathog. 2012b;8(5):e1002720.

Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science. 2010;329(5992):676–9.

Taber SW, Pease CM. Paramyxovirus phylogeny: tissue tropism evolves slower than host specificity. Evolution. 1990;44(2):435–8.

Tang XC, Zhang JX, Zhang SY, Wang P, Fan XH, Li LF, et al. Prevalence and genetic diversity of coronaviruses in bats from China. J Virol. 2006;80(15):7481–90.

Taylor DJ, Ballinger MJ, Zhan JJ, Hanzly LE, Bruenn JA. Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the Miocene. PeerJ. 2014;2:e556.

Taylor G, Scharlemann JPW, Rowcliffe M, Kümpel N, Harfoot MBJ, Fa JE, et al. Synthesising bushmeat research effort in West and Central Africa: A new regional database. Biol Conserv. 2015;181:199–205.

Taylor LH, Latham SM, Woolhouse MEJ. Risk factors for human disease emergence. Phil Trans R Soc B. 2001;356(1411):983–9.

Tella JL, Scheuerlein A, Ricklefs RE. Is cell–mediated immunity related to the evolution of life-history strategies in birds? Proc R Soc B. 2002;269(1495):1059–66.

Therneau TM, Atkinson B, Ripley B. rpart: Recursive partitioning and regression Trees. R package version 4.1-8. 2014. Available from: http://cran.r-project.org/package=rpart

Turmelle AS, Olival KJ. Correlates of viral richness in bats (order Chiroptera). EcoHealth. 2009;6(4):522–39.

Turner PE, Elena SF. Cost of host radiation in an RNA virus. Genetics. 2000;156(4):1465–70.

USAID. USAID launches Emerging Pandemic Threats program [press release]. United States Agency for International Development. 2009. Available from: https://www.usaid.gov/news-information/press-releases/usaid-launches-emerging-pandemic-threats-program

van Doremalen N, Munster VJ. Animal models of Middle East respiratory syndrome coronavirus infection. Antivir Res. 2015;122:28–38.

Viana M, Mancy R, Biek R, Cleaveland S, Cross PC, Lloyd-Smith JO, et al. Assembling evidence for identifying reservoirs of infection. Trends Ecol Evol. 2014;29(5):270–9.

Walther BA, Ewald PW. Pathogen survival in the external environment and the evolution of virulence. Biol Rev. 2004;79(4):849–69.

Weaver SC, Costa F, Garcia-Blanco MA, Ko AI, Ribeiro GS, Saade G, et al. Zika virus: history, emergence, biology, and prospects for control. Antivir Res. 2016;130:69–80.

Weaver SC, Salas R, Rico-Hesse R, Ludwig GV, Oberste MS, Boshell J, et al. Re-emergence of epidemic Venezuelan equine encephalomyelitis in South America. Lancet. 1996;348(9025):436–40.

Wendt A, Kreienbrock L, Campe A. Zoonotic disease surveillance – inventory of systems integrating human and animal disease information. Zoonoses Public Health. 2014;62(1):61–74.

West GB, Brown JH. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. J Exp Biol. 2005;208(9):1575–92.

WHO. China's latest SARS outbreak has been contained, but biosafety concerns remain – Update 7. 2004. Available from: http://www.who.int/csr/don/2004_05_18a/en/

WHO. Blueprint for R&D preparedness and response to public health emergencies due to highly infectious pathogens. Geneva; 2015. Available from: http://www.who.int/csr/research-and-development/meeting-report-prioritization.pdf

WHO. Middle East respiratory syndrome coronavirus (MERS-CoV) – Qatar. 2016. Available from: http://www.who.int/csr/don/16-may-2016-mers-qatar/en/

Wilcox BA, Colwell RR. Emerging and reemerging infectious diseases: biocomplexity as an interdisciplinary paradigm. EcoHealth. 2005;2(4):244–57.

Wilcox BA, Gubler DJ. Disease ecology and the global emergence of zoonotic pathogens. Environ Health Prev Med. 2005;10(5):263–72.

Wilf HS. Generatingfunctionology. New York, NY: Academic Press; 1990.

Woelk CH, Holmes EC. Reduced positive selection in vector-borne RNA viruses. Mol Biol Evol. 2002;19(12):2333–6.

Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447(7142):279–83.

Wolfe ND, Eitel MN, Gockowski J, Muchaal PK, Nolte C, Prosser AT, et al. Deforestation, hunting and the ecology of microbial emergence. Glob Change & Human Health. 2000;1(1):10–25.

Wood JLN, Leach M, Waldman L, MacGregor H, Fooks AR, Jones KE, et al. A framework for the study of zoonotic disease emergence and its drivers: spillover of bat pathogens as a case study. Phil Trans R Soc B. 2012;367(1604):2881–92.

Woolhouse MEJ. Population biology of emerging and re-emerging pathogens. Trends Microbiol. 2002;10(10):S3–7.

Woolhouse MEJ, Adair K. Ecological and taxonomic variation among human RNA viruses. J Clin Virol. 2013;58(2):344–5.

Woolhouse MEJ, Adair K, Brierley L. RNA viruses: a case study of the biology of emerging infectious diseases. Microbiol Spectrum. 2013;1(1):OH-0001-2012.

Woolhouse MEJ, Antia R. Emergence of new infectious diseases. In: Stearns SC, Koella JC, editors. Evolution in Health and Disease. Oxford University Press; 2007. p. 215–28.

Woolhouse MEJ, Brierley L, McCaffery C, Lycett S. Assessing the epidemic potential of RNA and DNA viruses. Emerg Infect Dis. 2016;22(12):2037–44.

Woolhouse MEJ, Gaunt E. Ecological origins of novel human pathogens. Crit Rev Microbiol. 2007;33(4):231–42.

Woolhouse MEJ, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. Emerg Infect Dis. 2005;11(12):1842–7.

Woolhouse MEJ, Howey R, Gaunt E, Reilly L, Chase-Topping M, Savill N. Temporal trends in the discovery of human viruses. Proc R Soc B. 2008;275(1647):2111–5.

Woolhouse MEJ, Rambaut A, Kellam P. Lessons from Ebola: Improving infectious disease surveillance to inform outbreak management. Sci Transl Med. 2015;7(307):307rv5.

Woolhouse MEJ, Scott F, Hudson Z, Howey R, Chase-Topping M. Human viruses: discovery and emergence. Phil Trans R Soc B. 2012;367(1604):2864–71.

Woolhouse MEJ, Taylor LH, Haydon DT. Population biology of multihost pathogens. Science. 2001;292(5519):1109–12.

Young CCW, Olival KJ. Optimizing viral discovery in bats. PLoS ONE. 2016;11(2):e0149237.

Yu VL, Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis. 2004;39(2):227–32.

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012;367(19):1814–20.

Zeltina A, Bowden TA, Lee B. Emerging paramyxoviruses: receptor tropism and zoonotic potential. PLoS Pathog. 2016;12(2):e1005390.

Zuckerman AJ, Banatvala JE, Griffiths P, Schoub B, Mortimer P. Principles and practice of clinical virology. John Wiley & Sons; 2009.

Appendix A. Supplementary material for: Tropism breadth and transmission ecology predict virulence of emerging human RNA viruses

A.1. Supplementary methods

A.1.1. Bayesian mixed regression model analysis

As a statistical validation of the risk factors highlighted by the final classification tree (Figure 2.2), I also carried out risk factor analysis by constructing a Bayesian mixed logistic regression model with Markov chain Monte Carlo estimation (MCMC), using package 'MCMCglmm', v2.22.1 (Hadfield 2010) in R, v3.1.1 (R Development Core Team 2015). Predictors were specified as in Table 2.1, except that human-to-human transmissibility was only specified as sustained versus nonsustained, and taxonomic family was specified as a random effect as a basic control for phylogenetic relatedness. The model was specified with residual variance fixed at 1, an inverse Wishart prior with V = 1 on the random term of taxonomic family, and flattened Gelman priors on the fixed terms using the 'gelman.prior' function in package 'MCMCglmm', following recommendations by Gelman et al. (2008). Gelman priors were scaled by a factor of $1 + V_{fam} + \pi^{\frac{2}{3}}$, where V_{fam} denotes the first posterior estimate of the residual variance of the taxonomic family random effect. The model was run for 5 x 10^6 iterations, retaining every 1000^{th} iteration, discarding the first 1.25 x 10^6 as burn-in, and inspecting trace output and assessing convergence using the gelman.diag function in package 'coda', v0.18-1 (Plummer et al. 2012). As a result of missing predictor data, 17 virus species were excluded from the model, leaving n = 161. The Bayesian mixed regression model supported the risk factors from the final classification tree. Specifically, viruses with multi-organ system

tropism were more much more likely to cause severe disease (Table A.5). Viruses with vector-borne transmission and viruses with sustained human-to-human transmissibility were less likely to cause severe disease, though these were weaker risk factors (Table A.5).

A.2. Supplementary figures

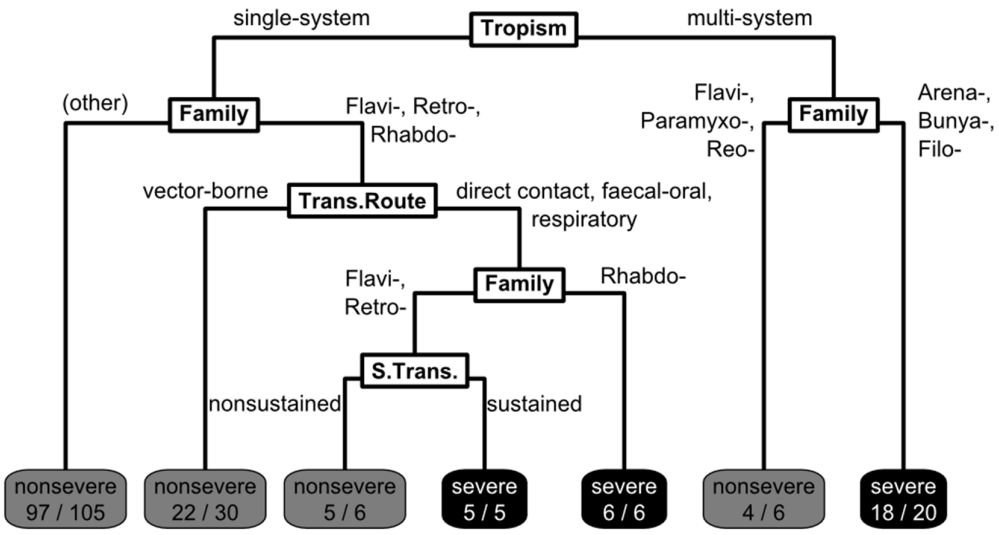


Figure A.1. Pruned classification tree run with same methodology/dataset as Figure 2.2, with transmission route variables specified using multiple categories (direct contact, faecal-oral, respiratory, vector-borne). Resulting tree is identical to that of Figure 2.2.

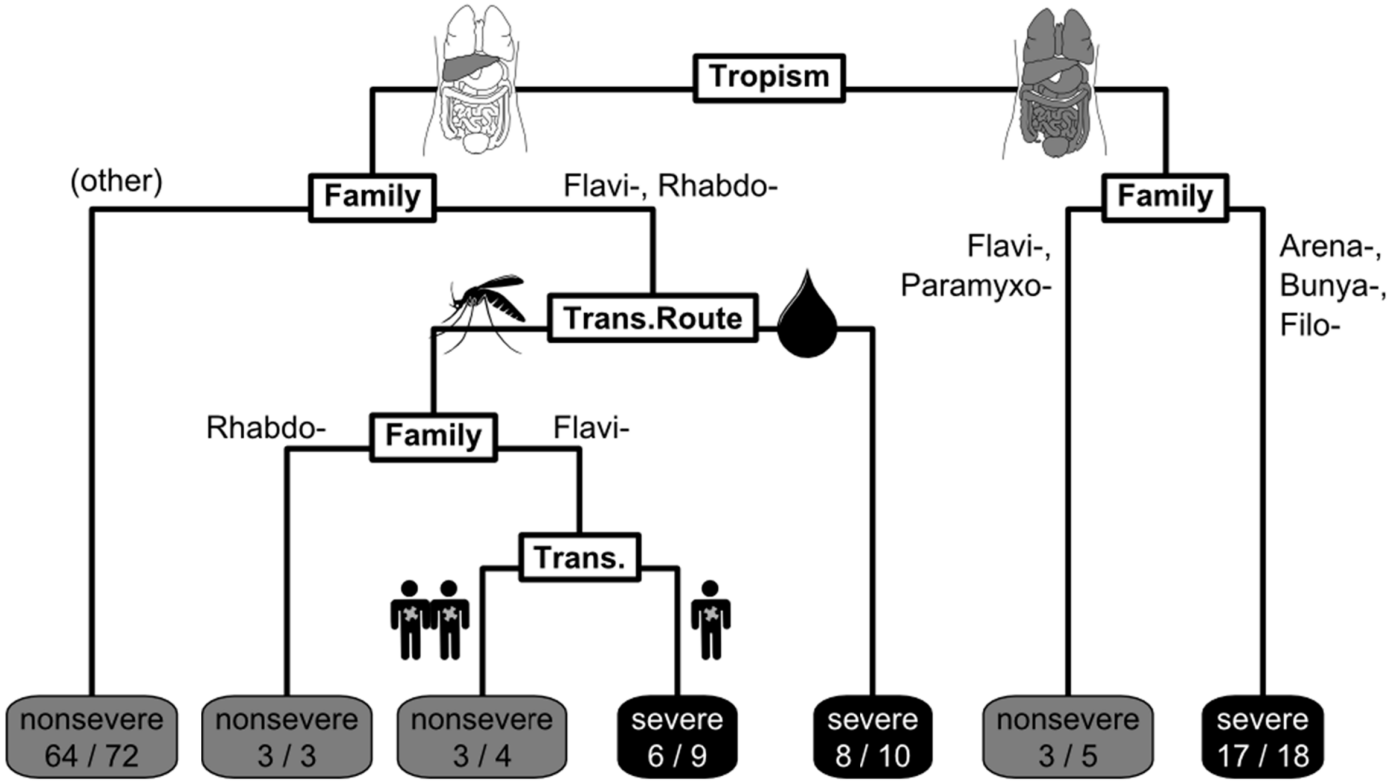


Figure A.2. Classification tree run with same methodology/dataset as Figure 2.2, excluding viruses with data quality problems or high data uncertainty, leaving n = 121. Risk factors are pictorially represented following Table 2.1. Resulting pruned tree retains a similar overall structure to that of Figure 2.2 and 86.0% predictive accuracy. 'Trans.Route' denotes primary transmission route, and 'Trans.' denotes any human-to-human transmissibility.

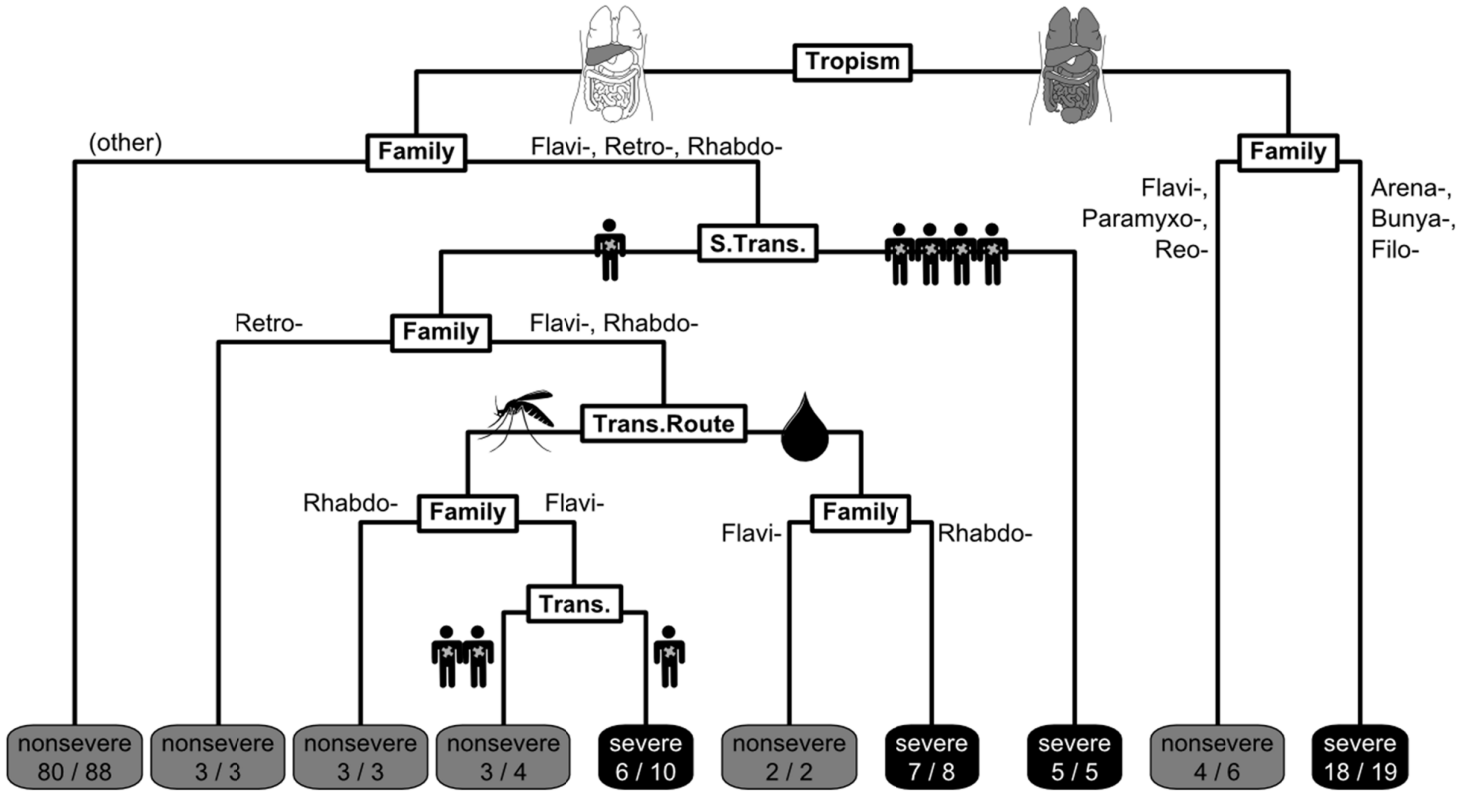


Figure A.3. Classification tree run with same methodology/dataset as Figure 2.2, excluding viruses only known to infect humans from serological evidence, leaving n = 148. Risk factors are pictorially represented following Table 2.1. Resulting pruned tree is larger, but retains as a similar overall structure to that of Figure 2.2 and 88.5% predictive accuracy. 'Trans.Route' denotes primary transmission route, 'Trans.' denotes any human-to-human transmissibility, and 'S.Trans.' denotes sustained human-to-human transmissibility.

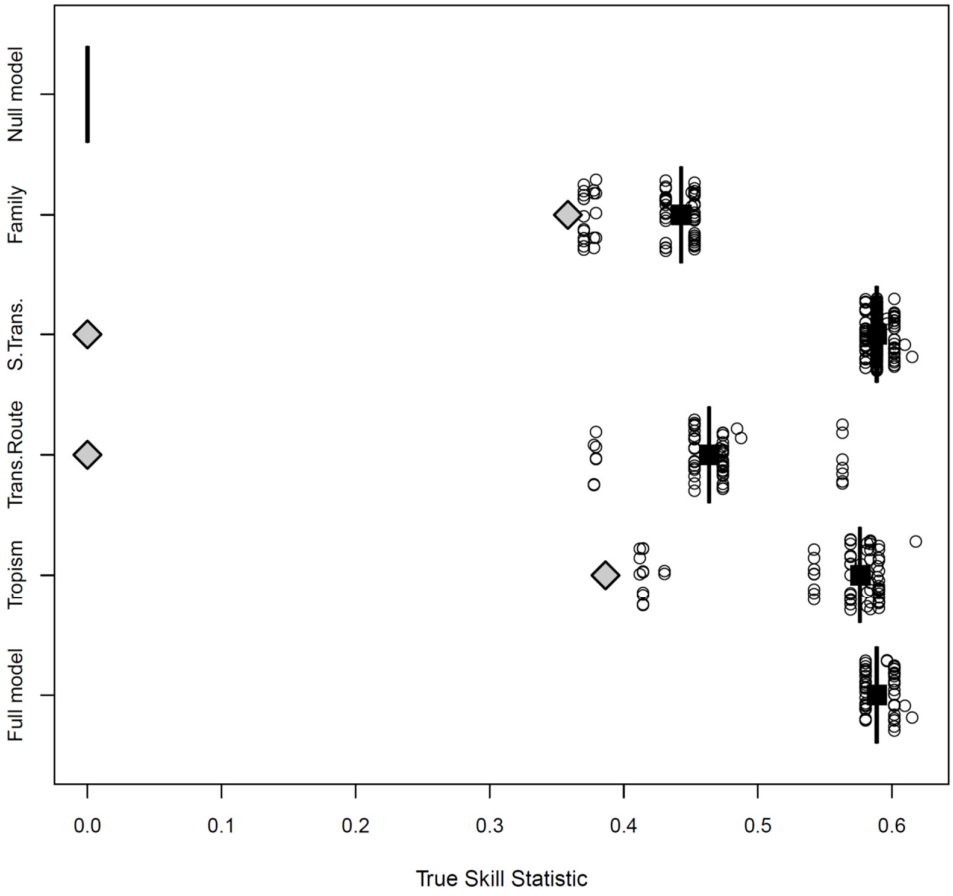


Figure A.4. True Skill Statistic (TSS, calculated as sensitivity + specificity - 1) values for resulting trees when specifying different predictor sets and jack-knifed datasets as in Figure 2.3. Solid squares denote TSS for tree built with full dataset (n=178) and boxes/outlying open circles denote TSS for 178 trees built with jack-knifed datasets for predictor sets removing predictor given on Y axis (except 'Null' & 'Full'). Grey diamonds denote TSS for tree built with single predictor given on Y axis only.

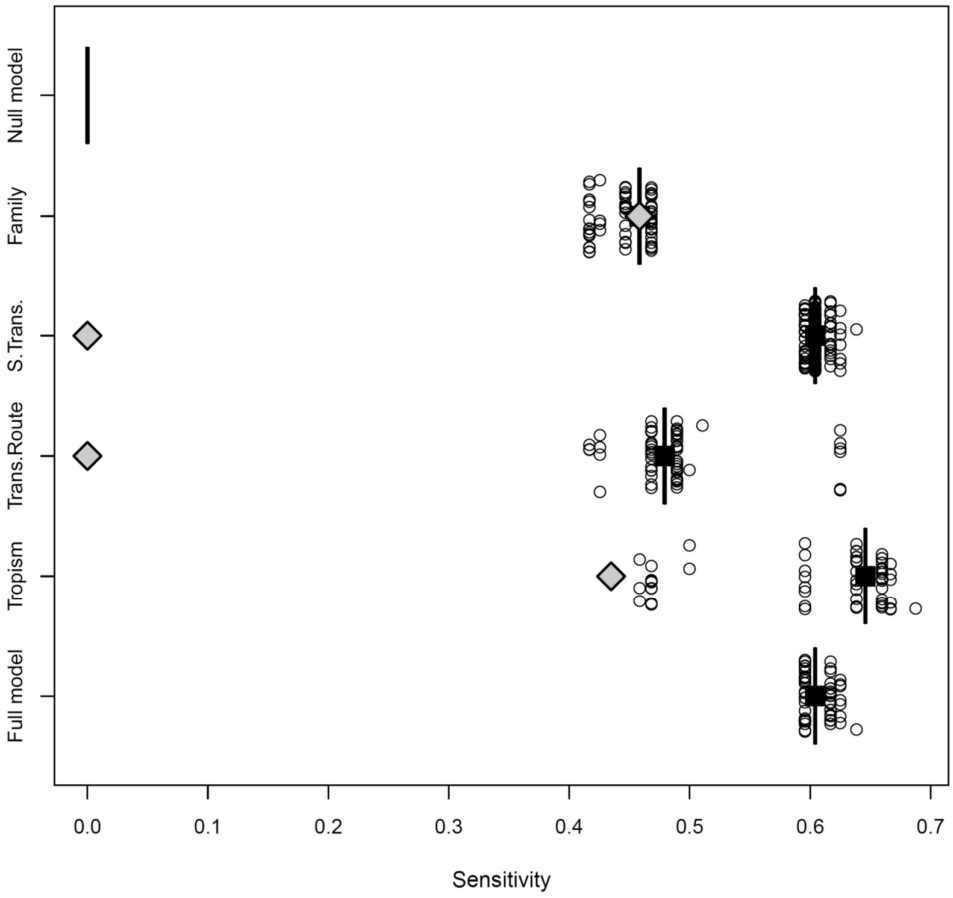


Figure A.5. Sensitivity values for resulting trees when specifying different predictor sets and jack-knifed datasets as in Figure 2.3. Solid squares denote sensitivity for tree built with full dataset (n=178) and boxes/outlying open circles denote sensitivity for 178 trees built with jack-knifed datasets for predictor sets removing predictor given on Y axis (except 'Null' & 'Full'). Grey diamonds denote sensitivity for tree built with single predictor given on Y axis only.

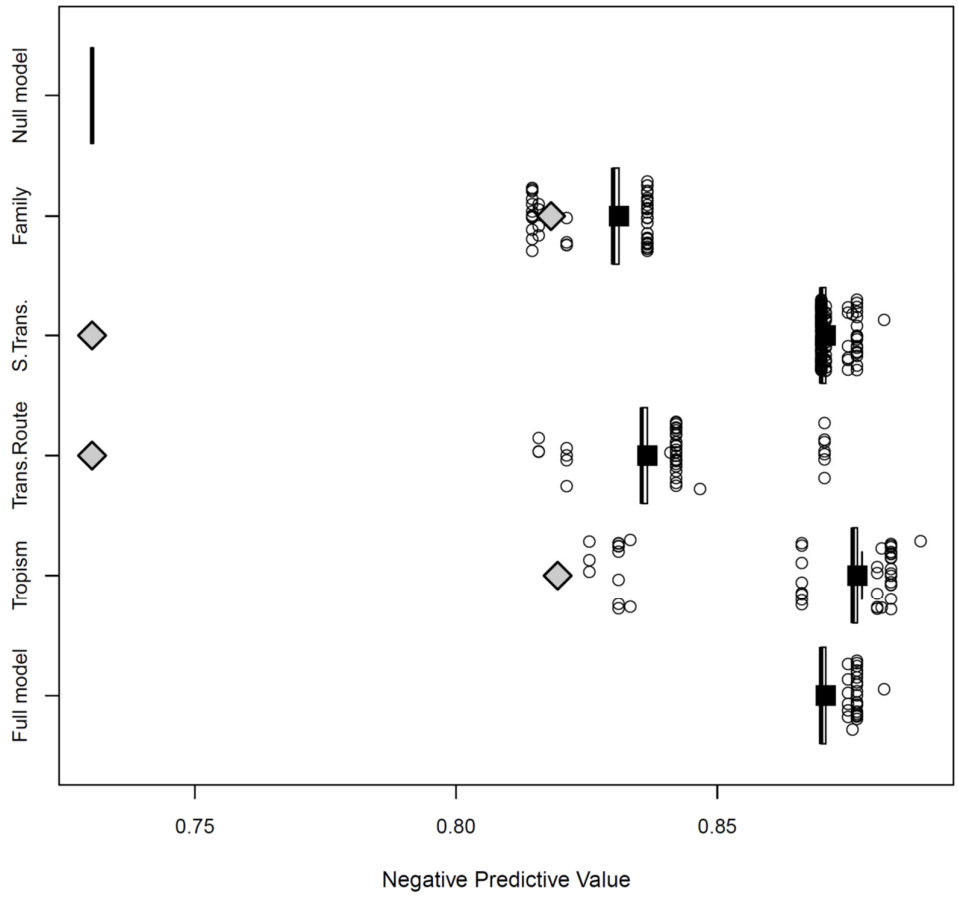


Figure A.6. Negative Predictive Values (NPV, calculated as correctly classified nonseveres/(correctly classified nonseveres + misclassified severes)) for resulting trees when specifying different predictor sets and jack-knifed datasets as in Figure 2.3. Solid squares denote NPV for tree built with full dataset (n=178) and boxes/outlying open circles denote NPV for 178 trees built with jack-knifed datasets for predictor sets removing predictor given on Y axis (except 'Null' & 'Full'). Grey diamonds denote NPV for tree built with single predictor given on Y axis only.

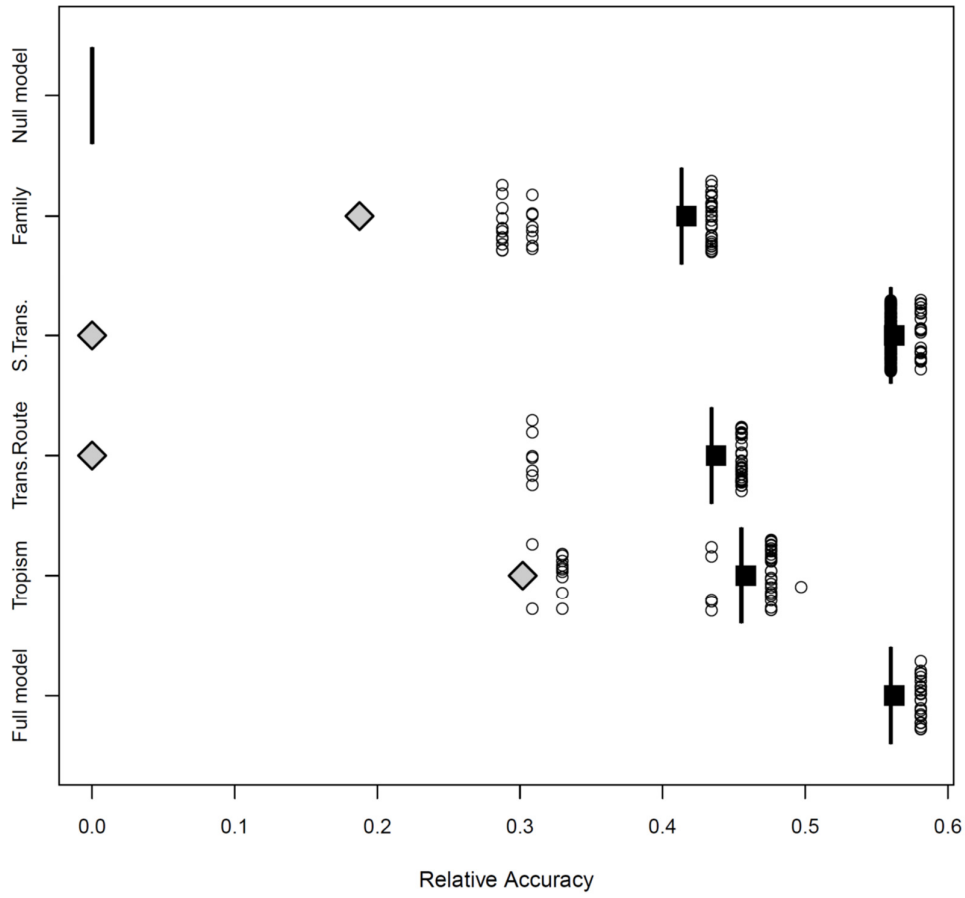


Figure A.7. Resulting tree accuracies relative to null model (extent of remaining variation explained beyond null model, calculated as (accuracy-null accuracy)/(1 – null accuracy)) for resulting trees when specifying different predictor sets and jack-knifed datasets as in Figure 2.3. Solid squares denote relative accuracy for tree built with full dataset (n=178) and boxes/outlying open circles denote relative accuracy for 178 trees built with jack-knifed datasets for predictor sets removing predictor given on Y axis (except 'Null' & 'Full'). Grey diamonds denote relative accuracy for tree built with single predictor given on Y axis only.

A.3. Supplementary tables

Table A.1. Virulence data for the 180 human RNA virus species ordered by genome type and taxonomy. Disease severity rating and supporting criteria for viruses rated 'severe' are given, following literature search protocol (see Chapter 2). Subsequent columns refer to whether virus is known to have caused fatalities in vulnerable individuals and/or otherwise healthy adults, and whether virus is known to have 'severe' strains if species is rated 'nonsevere'. CFR = case fatality ratio, HPS = hantavirus pulmonary syndrome, HFRS = hantavirus haemorrhagic fever with renal syndrome, HTLV = human T-lymphotropic virus, AIDS = acquired immunodeficiency syndrome.

Family	Genus	Species	Severity rating	Severity notes	Fatalities (vulnerable)	Fatalities (healthy adults)	Severe strains
-ssRNA viruses							
Arenaviridae	*Arenavirus*	*Chapare virus*	Severe	Haemorrhagic fever, CFR 100% (1 case)	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Guanarito virus*	Severe	Haemorrhagic fever, CFR 25%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Junin virus*	Severe	Haemorrhagic fever, CFR 15-30%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Lassa virus*	Severe	Haemorrhagic fever, CFR 15%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Lymphocytic choriomeningitis virus*	Nonsevere		Yes	Yes	No
Arenaviridae	*Arenavirus*	*Machupo virus*	Severe	Haemorrhagic fever, CFR 5-35%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Pichinde virus*	Nonsevere		No	No	No
Arenaviridae	*Arenavirus*	*Sabia virus*	Severe	Haemorrhagic fever, CFR 33% (3 cases)	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Whitewater Arroyo virus*	Severe	Haemorrhagic fever, CFR 100% (3 cases)	Yes	Yes	-
Bornaviridae	*Bornavirus*	*Borna disease virus*	Nonsevere		No	No	No
Bunyaviridae	*Hantavirus*	*Andes virus*	Severe	HPS	Yes	Yes	-
Bunyaviridae	*Hantavirus*	*Bayou virus*	Severe	HPS	Yes	No	-
Bunyaviridae	*Hantavirus*	*Black creek canal virus*	Severe	HPS	No	No	-
Bunyaviridae	*Hantavirus*	*Dobrava-Belgrade virus*	Severe	HFRS, CFR 5-35%	Yes	Yes	-

Bunyaviridae	*Hantavirus*	*Hantaan virus*	Severe	HFRS, CFR 5-15%	Yes	Yes	-
Bunyaviridae	*Hantavirus*	*Laguna Negra virus*	Severe	HPS	Yes	Yes	-
Bunyaviridae	*Hantavirus*	*New York virus*	Severe	HPS	Yes	Yes	-
Bunyaviridae	*Hantavirus*	*Puumala virus*	Nonsevere		Yes	Yes	No
Bunyaviridae	*Hantavirus*	*Saaremaa virus*	Nonsevere		No	No	No
Bunyaviridae	*Hantavirus*	*Seoul virus*	Nonsevere		Yes	Yes	No
Bunyaviridae	*Hantavirus*	*Sin Nombre virus*	Severe	HPS, CFR 32-75%	Yes	Yes	-
Bunyaviridae	*Hantavirus*	*Thailand virus*	Nonsevere		No	No	No
Bunyaviridae	*Hantavirus*	*Tula virus*	Nonsevere		No	No	No
Bunyaviridae	*Nairovirus*	*Crimean-Congo haemorrhagic fever virus*	Severe	Haemorrhagic fever, CFR 30%	Yes	Yes	-
Bunyaviridae	*Nairovirus*	*Dugbe virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Bunyamwera virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Bwamba virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*California encephalitis virus*	Severe	High frequency of severe symptoms (seizures, coma)	Yes	Yes	-
Bunyaviridae	*Orthobunyavirus*	*Caraparu virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Catu virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Guama virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Guaroa virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Kairi virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Madrid virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Marituba virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Nyando virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Oriboca virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Oropouche virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Shuni virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Tacaiuma virus*	Nonsevere		No	No	No
Bunyaviridae	*Orthobunyavirus*	*Wyeomyia virus*	Nonsevere		No	No	No
Bunyaviridae	*Phlebovirus*	*Candiru virus*	Nonsevere		No	No	No
Bunyaviridae	*Phlebovirus*	*Punta Toro virus*	Nonsevere		No	No	No

Bunyaviridae	*Phlebovirus*	*Rift Valley fever virus*	Nonsevere		Yes	Yes	No
Bunyaviridae	*Phlebovirus*	*Sandfly fever Naples virus*	Nonsevere		No	No	No
Bunyaviridae	*Phlebovirus*	*Uukuniemi virus*	Nonsevere		No	No	No
Filoviridae	*Ebolavirus*	*Reston ebolavirus*	Nonsevere		No	No	No
Filoviridae	*Ebolavirus*	*Sudan ebolavirus*	Severe	Haemorrhagic fever , CFR 41-65%	Yes	Yes	-
Filoviridae	*Ebolavirus*	*Tai forest ebolavirus*	Severe	Haemorrhagic fever	No	No	-
Filoviridae	*Ebolavirus*	*Zaire ebolavirus*	Severe	Haemorrhagic fever , CFR 61-90%	Yes	Yes	-
Filoviridae	*Marburgvirus*	*Lake Victoria Marburgvirus*	Severe	Haemorrhagic fever, CFR 23-90%	Yes	Yes	-
Orthomyxoviridae	*Influenzavirus A*	*Influenza A virus*	Nonsevere		Yes	Yes	Yes
Orthomyxoviridae	*Influenzavirus B*	*Influenza B virus*	Nonsevere		Yes	Yes	No
Orthomyxoviridae	*Influenzavirus C*	*Influenza C virus*	Nonsevere		No	No	No
Orthomyxoviridae	*Thogotovirus*	*Dhori virus*	Nonsevere		No	No	No
Orthomyxoviridae	*Thogotovirus*	*Thogoto virus*	Nonsevere		Yes	No	No
Paramyxoviridae	*Avulavirus*	*Newcastle disease virus*	Nonsevere		Yes	No	No
Paramyxoviridae	*Henipavirus*	*Hendra virus*	Severe	CFR 57% (7 cases)	Yes	Yes	-
Paramyxoviridae	*Henipavirus*	*Nipah virus*	Severe	CFR 10-75%	Yes	Yes	-
Paramyxoviridae	*Metapneumovirus*	*Avian metapneumovirus*	Nonsevere		No	No	No
Paramyxoviridae	*Metapneumovirus*	*Human metapneumovirus*	Nonsevere		Yes	No	No
Paramyxoviridae	*Morbillivirus*	*Measles virus*	Nonsevere		Yes	Yes	No
Paramyxoviridae	*Pneumovirus*	*Human respiratory syncytial virus*	Nonsevere		Yes	No	No
Paramyxoviridae	*Respirovirus*	*Human parainfluenza virus 1*	Nonsevere		No	No	No
Paramyxoviridae	*Respirovirus*	*Human parainfluenza virus 3*	Nonsevere		Yes	No	No
Paramyxoviridae	*Rubulavirus*	*Human parainfluenza virus 2*	Nonsevere		No	No	No
Paramyxoviridae	*Rubulavirus*	*Human parainfluenza virus 4*	Nonsevere		No	No	No

Paramyxoviridae	*Rubulavirus*	*Mumps virus*	Nonsevere		Yes	Yes	No
Paramyxoviridae	*Rubulavirus*	*Parainfluenza virus 5*	Nonsevere		No	No	No
Paramyxoviridae	*Rubulavirus*	*Simian virus 41*	Nonsevere		No	No	No
Rhabdoviridae	*Lyssavirus*	*Australian bat lyssavirus*	Severe	Rabies-like disease, CFR 100% (3 cases)	Yes	Yes	-
Rhabdoviridae	*Lyssavirus*	*Duvenhage virus*	Severe	Rabies-like disease, CFR 75% (4 cases)	Yes	Yes	-
Rhabdoviridae	*Lyssavirus*	*European bat lyssavirus 1*	Severe	Rabies-like disease, CFR 100% (3 cases)	Yes	No	-
Rhabdoviridae	*Lyssavirus*	*European bat lyssavirus 2*	Severe	Rabies-like disease, CFR 20-50%	Yes	Yes	-
Rhabdoviridae	*Lyssavirus*	*Irkut virus*	Severe	Rabies-like disease, CFR 100% (1 case)	Yes	Yes	-
Rhabdoviridae	*Lyssavirus*	*Mokola virus*	Severe	Neurologic disease different to classical rabies, CFR 33% (3 cases)	Yes	No	-
Rhabdoviridae	*Lyssavirus*	*Rabies virus*	Severe	Rabies disease, CFR 60%	Yes	Yes	-
Rhabdoviridae	*Vesiculovirus*	*Chandipura virus*	Nonsevere		Yes	No	No
Rhabdoviridae	*Vesiculovirus*	*Isfahan virus*	Nonsevere		No	No	No
Rhabdoviridae	*Vesiculovirus*	*Maraba virus*	Nonsevere		No	No	No
Rhabdoviridae	*Vesiculovirus*	*Piry virus*	Nonsevere		No	No	No
Rhabdoviridae	*Vesiculovirus*	*Vesicular stomatitis Alagoas virus*	Nonsevere		No	No	No
Rhabdoviridae	*Vesiculovirus*	*Vesicular stomatitis Indiana virus*	Nonsevere		No	No	No
Rhabdoviridae	*Vesiculovirus*	*Vesicular stomatitis New Jersey virus*	Nonsevere		No	No	No
Unassigned	*Deltavirus*	*Hepatitis delta virus*	NA	Reliant on Hepatitis B virus coinfection	Yes	Yes	-
+ssRNA viruses							
Astroviridae	*Mamastrovirus*	*Human astrovirus*	Nonsevere		Yes	No	No
Caliciviridae	*Norovirus*	*Norwalk virus*	Nonsevere		Yes	No	No
Caliciviridae	*Sapovirus*	*Sapporo virus*	Nonsevere		No	No	No
Coronaviridae	*Alphacoronavirus*	*Human coronavirus 229E*	Nonsevere		Yes	No	No
Coronaviridae	*Alphacoronavirus*	*Human coronavirus NL63*	Nonsevere		Yes	No	No
Coronaviridae	*Betacoronavirus*	*Betacoronavirus 1*	Nonsevere		Yes	No	No

Coronaviridae	*Betacoronavirus*	*Human coronavirus HKU1*	Nonsevere		No	No	No
Coronaviridae	*Betacoronavirus*	*Severe acute respiratory syndrome-related coronavirus*	Severe	CFR 9-12%	Yes	Yes	-
Coronaviridae	*Torovirus*	*Human torovirus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Aroa virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Bagaza virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Banzi virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Dengue virus*	Nonsevere		Yes	Yes	No
Flaviviridae	*Flavivirus*	*Edge Hill virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Gadgets Gully virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Ilheus virus*	Severe	Described as potentially severe, one subspecies (Rocio virus) has CFR 10%	Yes	Yes	-
Flaviviridae	*Flavivirus*	*Japanese encephalitis virus*	Severe	CFR 5-40%	Yes	Yes	-
Flaviviridae	*Flavivirus*	*Kokobera virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Kyasanur forest disease virus*	Severe	CFR 3-25%	Yes	Yes	-
Flaviviridae	*Flavivirus*	*Langat virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Louping ill virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Murray Valley encephalitis virus*	Severe	CFR 18-42%, frequent neurologic sequelae	Yes	Yes	-
Flaviviridae	*Flavivirus*	*Ntaya virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Omsk haemorrhagic fever virus*	Nonsevere		Yes	Yes	No
Flaviviridae	*Flavivirus*	*Powassan virus*	Severe	CFR 10-15%, frequent neurologic sequelae	Yes	No	-
Flaviviridae	*Flavivirus*	*Rio Bravo virus*	Severe	Two of the five known cases required hospital admission	No	No	No
Flaviviridae	*Flavivirus*	*St. Louis encephalitis virus*	Severe	CFR 7-30%	Yes	Yes	-

Flaviviridae	*Flavivirus*	*Tembusu virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Tick-borne encephalitis virus*	Severe	CFR 1-54%	Yes	Yes	-
Flaviviridae	*Flavivirus*	*Uganda S virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Usutu virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*Wesselsbron virus*	Nonsevere		No	No	No
Flaviviridae	*Flavivirus*	*West Nile virus*	Nonsevere		Yes	Yes	No
Flaviviridae	*Flavivirus*	*Yellow fever virus*	Severe	CFR 11-50%, haemorrhagic fever	Yes	Yes	-
Flaviviridae	*Flavivirus*	*Zika virus*	Nonsevere		No	No	No
Flaviviridae	*Hepacivirus*	*Hepatitis C virus*	Severe	Chronic disease including cirrhosis	Yes	Yes	-
Flaviviridae	*Pestivirus*	*Bovine viral diarrhea virus 1*	Nonsevere		No	No	No
Hepeviridae	*Hepevirus*	*Hepatitis E virus*	Nonsevere		Yes	Yes	No
Picornaviridae	*Aphthovirus*	*Equine rhinitis A virus*	Nonsevere		No	No	No
Picornaviridae	*Aphthovirus*	*Foot-and-mouth disease virus*	Nonsevere		No	No	No
Picornaviridae	*Cardiovirus*	*Encephalomyocarditis virus*	Nonsevere		No	No	No
Picornaviridae	*Cardiovirus*	*Theilovirus*	Nonsevere		Yes	Yes	Yes
Picornaviridae	*Enterovirus*	*Bovine enterovirus*	Nonsevere		No	No	No
Picornaviridae	*Enterovirus*	*Human enterovirus A*	Nonsevere		Yes	Yes	Yes
Picornaviridae	*Enterovirus*	*Human enterovirus B*	Nonsevere		Yes	Yes	Yes
Picornaviridae	*Enterovirus*	*Human enterovirus C*	Nonsevere		Yes	Yes	Yes
Picornaviridae	*Enterovirus*	*Human enterovirus D*	Nonsevere		Yes	No	Yes
Picornaviridae	*Enterovirus*	*Human rhinovirus A*	Nonsevere		Yes	No	No
Picornaviridae	*Enterovirus*	*Human rhinovirus B*	Nonsevere		No	No	No
Picornaviridae	*Enterovirus*	*Human rhinovirus C*	Nonsevere		Yes	No	No
Picornaviridae	*Erbovirus*	*Equine rhinitis B virus*	Nonsevere		No	No	No
Picornaviridae	*Hepatovirus*	*Hepatitis A virus*	Nonsevere		Yes	Yes	No
Picornaviridae	*Kobuvirus*	*Aichi virus*	Nonsevere		No	No	No
Picornaviridae	*Parechovirus*	*Human parechovirus*	Nonsevere		Yes	No	Yes

Picornaviridae	*Parechovirus*	*Ljungan virus*	Severe	Associated with birth defects, sudden infant death syndrome and adult myocarditis	No	No	-
Togaviridae	*Alphavirus*	*Barmah forest virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Chikungunya virus*	Nonsevere		Yes	Yes	No
Togaviridae	*Alphavirus*	*Eastern equine encephalitis virus*	Severe	CFR 30-70%	Yes	Yes	-
Togaviridae	*Alphavirus*	*Everglades virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Getah virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Highlands J virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Mayaro virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Mucambo virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*O'nyong-nyong virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Pixuna virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Rio Negro virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Ross River virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Semliki Forest virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Sindbis virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Tonate virus*	Nonsevere		Yes	No	No
Togaviridae	*Alphavirus*	*Una virus*	Nonsevere		No	No	No
Togaviridae	*Alphavirus*	*Venezuelan equine encephalitis virus*	Nonsevere		Yes	Yes	Yes
Togaviridae	*Alphavirus*	*Western equine encephalitis virus*	Severe	CFR 3-10%	Yes	Yes	-
Togaviridae	*Alphavirus*	*Whataroa virus*	Nonsevere		No	No	No
Togaviridae	*Rubivirus*	*Rubella virus*	Nonsevere		Yes	Yes	No
dsRNA viruses							
Picobirnaviridae	*Picobirnavirus*	*Human picobirnavirus*	Nonsevere		No	No	No
Reoviridae	*Coltivirus*	*Colorado tick fever virus*	Nonsevere		Yes	No	No
Reoviridae	*Orbivirus*	*Changuinola virus*	Nonsevere		No	No	No
Reoviridae	*Orbivirus*	*Great Island virus*	Nonsevere		No	No	No

Reoviridae	*Orbivirus*	*Lebombo virus*	Nonsevere		No	No	No
Reoviridae	*Orbivirus*	*Orungo virus*	Nonsevere		No	No	No
Reoviridae	*Orthoreovirus*	*Mammalian orthoreovirus*	Nonsevere		No	No	No
Reoviridae	*Orthoreovirus*	*Nelson Bay orthoreovirus*	Nonsevere		No	No	No
Reoviridae	*Rotavirus*	*Rotavirus A*	Nonsevere		Yes	No	No
Reoviridae	*Rotavirus*	*Rotavirus B*	Nonsevere		Yes	No	No
Reoviridae	*Rotavirus*	*Rotavirus C*	Nonsevere		No	No	No
Reoviridae	*Seadornavirus*	*Banna virus*	Nonsevere		No	No	No
ssRNA-RT viruses							
Retroviridae	*Deltaretrovirus*	*Primate T-lymphotropic virus 1*	Severe	Chronic neoplastic and neurologic disease including lymphoma, leukaemia, and HTLV-associated myelopathy	Yes	Yes	-
Retroviridae	*Deltaretrovirus*	*Primate T-lymphotropic virus 2*	Severe	Chronic neurologic disease including HTLV-associated myelopathy	Yes	Yes	-
Retroviridae	*Deltaretrovirus*	*Primate T-lymphotropic virus 3*	NA	Suspected chronic neurologic disease associations similar to other T-lymphotropic viruses though cohort studies not yet long enough to confirm	No	No	No
Retroviridae	*Lentivirus*	*Human immunodeficiency virus 1*	Severe	AIDS	Yes	Yes	-
Retroviridae	*Lentivirus*	*Human immunodeficiency virus 2*	Severe	AIDS	Yes	Yes	-
Retroviridae	*Spumavirus*	*African green monkey simian foamy virus*	Nonsevere		No	No	No
Retroviridae	*Spumavirus*	*Macaque simian foamy virus*	Nonsevere		No	No	No
Retroviridae	*Spumavirus*	*Simian foamy virus*	Nonsevere		No	No	No

Table A.2. Virulence data for 30 newly reported human RNA viruses not yet ratified as species, ordered by genome type and taxonomy. Disease severity rating and supporting criteria for viruses rated 'severe' are given, following literature search protocol (see Chapter 2). Predicted severity refers to prediction from the final classification tree (Figure 2.2). Bold type denotes where predictions do not match literature-based ratings. CFR = case fatality ratio, HPS = hantavirus pulmonary syndrome.

Family	Genus	Species	Severity rating	Severity notes	Fatalities (vulnerable)	Fatalities (healthy adults)	Severe strains
-ssRNA viruses							
Arenaviridae	*Arenavirus*	*Chapare virus*	Severe	Haemorrhagic fever, CFR 100% (1 case)	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Guanarito virus*	Severe	Haemorrhagic fever, CFR 25%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Junin virus*	Severe	Haemorrhagic fever, CFR 15-30%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Lassa virus*	Severe	Haemorrhagic fever, CFR 15%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Lymphocytic choriomeningitis virus*	Nonsevere		Yes	Yes	No
Arenaviridae	*Arenavirus*	*Machupo virus*	Severe	Haemorrhagic fever, CFR 5-35%	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Pichinde virus*	Nonsevere		No	No	No
Arenaviridae	*Arenavirus*	*Sabia virus*	Severe	Haemorrhagic fever, CFR 33% (3 cases)	Yes	Yes	-
Arenaviridae	*Arenavirus*	*Whitewater Arroyo virus*	Severe	Haemorrhagic fever, CFR 100% (3 cases)	Yes	Yes	-
Bornaviridae	*Bornavirus*	*Borna disease virus*	Nonsevere		No	No	No
Bunyaviridae	*Hantavirus*	*Andes virus*	Severe	HPS	Yes	Yes	-
Bunyaviridae	*Hantavirus*	*Bayou virus*	Severe	HPS	Yes	No	-
Bunyaviridae	*Hantavirus*	*Black creek canal virus*	Severe	HPS	No	No	-
Bunyaviridae	*Hantavirus*	*Dobrava-Belgrade virus*	Severe	HFRS, CFR 5-35%	Yes	Yes	-

Paramyxoviridae	*Rubulavirus*	*Tioman virus*	Nonsevere		Nonsevere
Rhabdoviridae	(novel unnamed)	*Bas-Congo virus*	Severe	Haemorrhagic fever, CFR 67% (3 cases)	Severe
+ssRNA viruses					
Astroviridae	*Mamastrovirus*	*Mamastrovirus 6*	Nonsevere		Nonsevere
Astroviridae	*Mamastrovirus*	*Mamastrovirus 8*	Nonsevere		Nonsevere
Astroviridae	*Mamastrovirus*	*Mamastrovirus 9*	Nonsevere		Nonsevere
Coronaviridae	*Betacoronavirus*	*Middle East respiratory syndrome-related coronavirus*	Severe	CFR 27-56%	**Nonsevere**
Picornaviridae	*Cosavirus*	*Human cosavirus A*	Nonsevere		Nonsevere
Picornaviridae	*Cosavirus*	*Human cosavirus B*	Nonsevere		Nonsevere
Picornaviridae	*Cosavirus*	*Human cosavirus C*	Nonsevere		Nonsevere
Picornaviridae	*Cosavirus*	*Human cosavirus D*	Nonsevere		Nonsevere
Picornaviridae	*Cosavirus*	*Human cosavirus E1*	Nonsevere		Nonsevere
Picornaviridae	*Cosavirus*	*Human cosavirus F*	Nonsevere		Nonsevere
Picornaviridae	*Klassevirus*	*Human klassevirus*	Nonsevere		Nonsevere
Picornaviridae	*Rosavirus*	*Rosavirus 2*	Nonsevere		Nonsevere

Table A.3. Six-rank system of classifying human RNA virus virulence with available data, along with example viruses and number of viruses fitting each exclusive rank's criteria

Rank	Definition	Example virus species	No. virus species
1	Fits any of 'severe' criteria outlined in main text (≥5% case fatality ratio, frequent reports of hospitalisation, significant morbidity from certain symptoms, otherwise explicitly described as "severe" or causing "severe disease")	*Rabies virus*	48
2	Those not fitting 'severe' criteria, but are reported to have caused fatalities in healthy adults	*Dengue virus*	14
3	Those not fitting 'severe' criteria, but have severe strains or subspecies reported to cause fatalities in healthy adults	*Influenza A virus*	6
4	Those not fitting 'severe' criteria, but are reported to have caused fatalities in vulnerable individuals (age 16 and below or 60 and above, immunosuppressed, having co-morbidities, or otherwise 'at-risk')	*Rotavirus A*	17
5	Those not fitting 'severe' criteria, but have severe strains or subspecies reported to cause fatalities in vulnerable individuals	*Human parechovirus*	2
6	Those not fitting 'severe' criteria that have never been reported to cause fatalities	*Human parainfluenza virus 1*	91

Table A.4. Accuracy and 95% confidence interval, sensitivity, and specificity for final pruned classification trees run with same methodology/risk factors as in Materials and Methods, except with alternative two-category definitions of "virulent" as the predicted variable. Vulnerable individuals are defined as those age 16 and below, age 60 and above, immunosuppressed, having co-morbidities, or otherwise cited as being 'at-risk'. Ranks follow those given in Table A.3

"Virulent" criteria	Accuracy (95% confidence interval)	Sensitivity	Specificity
Reported to have caused fatalities in healthy adults (n = 60)	75.3% (68.3%, 81.4%)	0.350	0.958
Reported to have caused fatalities in vulnerable individuals (n = 83)	78.7% (71.9%, 84.4%)	0.759	0.811
Fits 'severe' criteria outlined in main text, including viruses having severe strains or subspecies (n = 56)	84.3% (78.1%, 89.3%)	0.643	0.934
Virulence ranks 1 - 2 (n = 62)	83.7% (77.5%, 88.8%)	0.629	0.948
Virulence ranks 1 – 3 (n = 68)	82.0% (75.6%, 87.4%)	0.691	0.900
Virulence ranks 1 – 4 (n = 85)	77.0% (70.1%, 82.9%)	0.765	0.774

Table A.5. Posterior mean coefficients from Bayesian mixed logistic regression predicting disease severity with viral taxonomic family as a random effect (n = 161 after removing viruses with missing predictor data). Baselines for predictors were specified as follows: genome type: -ssRNA, human-to-human transmissibility: nonsustained, transmission route: nonvector-borne, tropism; single-organ system. 95% credible intervals that exclude zero are highlighted in bold.

Model component	Posterior mean coefficient (95% credible interval)
(intercept)	-0.71 (-2.06, 0.62)
Genome type: +ssRNA	0.10 (-1.71, 1.89)
Genome type: dsRNA	-2.34 (-5.65, 0.89)
Genome type: ssRNA-RT	1.53 (-1.27, 3.94)
Transmissibility: sustained	**-1.74 (-3.30, -0.24)**
Transmission route: vector-borne	**-1.93 (-3.27, -0.65)**
Tropism: multi-organ system	**3.38 (1.94, 4.90)**

Appendix B. Supplementary material for: Evolutionary routes of RNA virus emergence and human adaptation

B.1. Supplementary figures

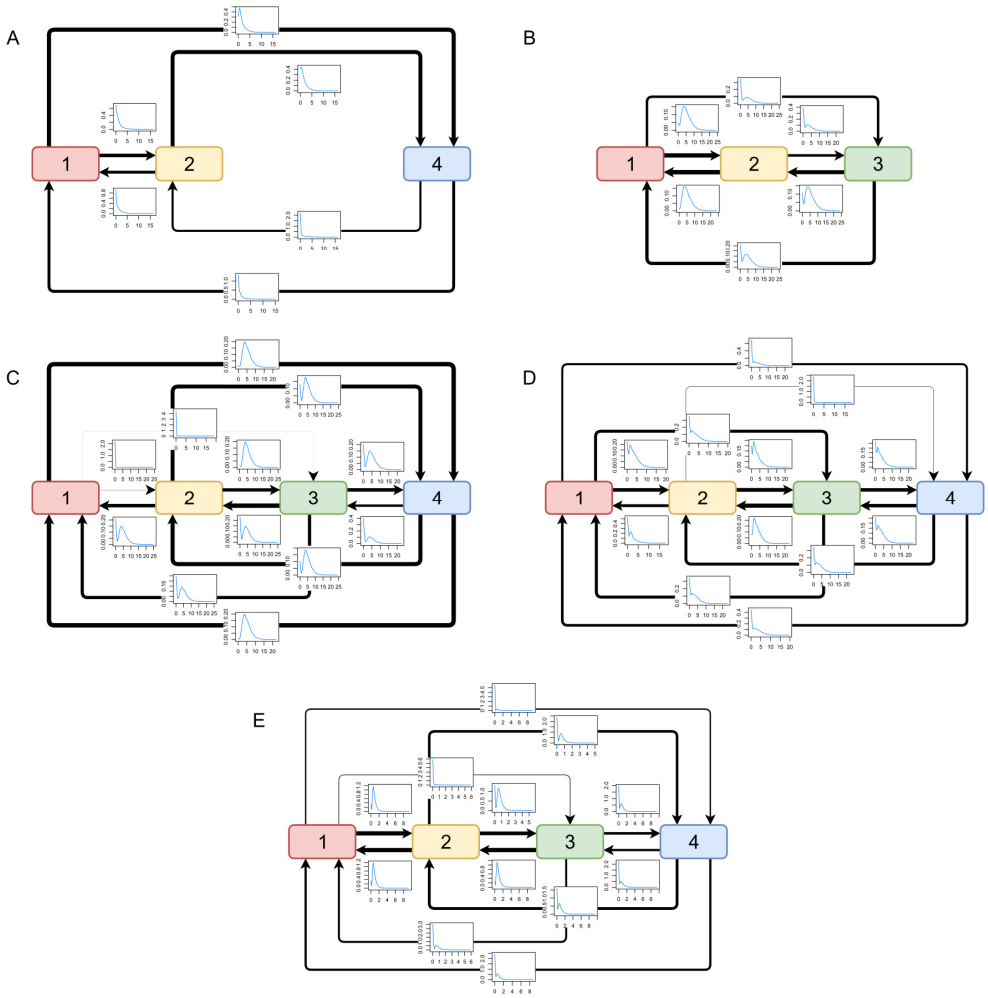


Figure B.1. Graphical summary of RJ-MCMC state-switching model configurations over existing family or genus-level phylogenies. A) family *Picornaviridae*, B) family *Rhabdoviridae*, C) family *Paramyxoviridae*, D) genus *Alphavirus*, E) genus *Flavivirus*. Line widths are proportional to frequency of transition parameter inclusion (i.e. not being fixed at zero) over RJ-MCMC chain, with transitions having zero frequency not presented. Graphs are density plots of transition parameter values over RJ-MCMC chain.

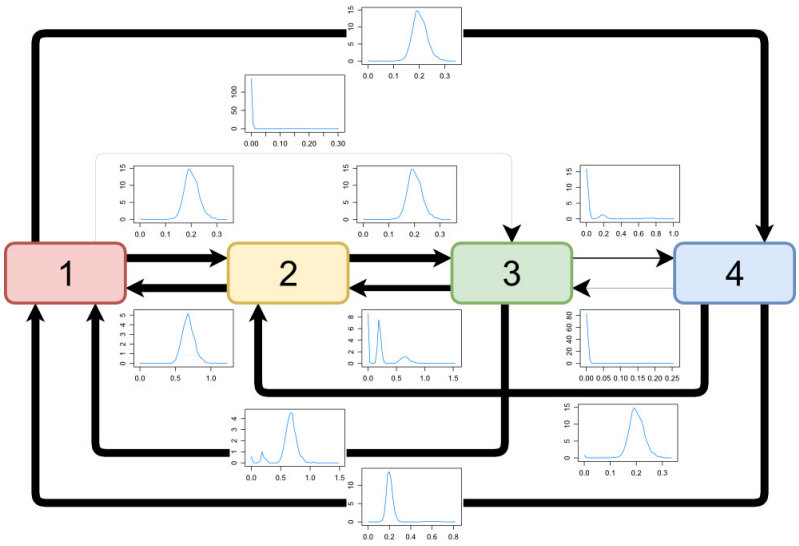


Figure B.2. Graphical summary of RJ-MCMC state-switching model configurations across taxonomically-structured cladogram under branch length assumption set (a). Line widths are proportional to frequency of transition parameter inclusion (i.e. not being fixed at zero) over RJ-MCMC chain, with transitions having zero frequency not presented. Graphs are density plots of transition parameter values over RJ-MCMC chain.

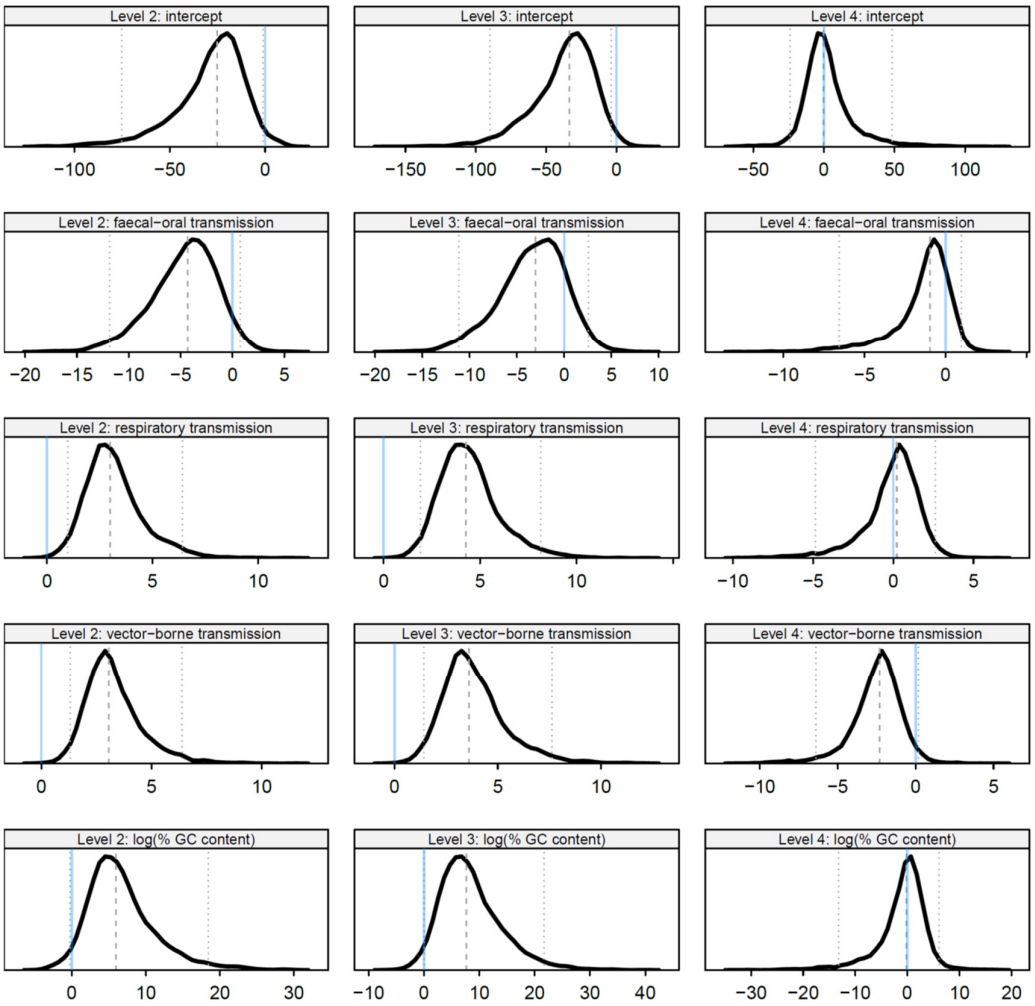


Figure B.3. Posterior density plots of log odds from multinomial phylogenetic mixed regression for viruses with complete data (n = 221). Dashed grey lines represent posterior medians, dotted lines represent posterior 95% credible intervals, and solid blue lines highlight zero for clarity. '(intercept)' denotes log odds of a virus having the specified Pathogen Pyramid level compared to level 1 (no human infection), Transmission variables denote additional log odds of a virus having the specified Pathogen Pyramid level given the specified transmission route compared to a baseline of direct-contact transmission.

B.2. Supplementary tables

Table B.1. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for viral phylogeny of family *Picornaviridae*. Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters					
Frequency	**No. parameters**	**1→2**	**1→4**	**2→1**	**2→4**	**4→1**	**4→2**
24.30%	1	A	A	-	A	A	-
18.14%	1	A	A	A	A	A	-
14.39%	1	-	A	A	A	-	A
13.70%	1	A	A	A	A	-	A

Table B.2. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for viral phylogeny of family *Rhabdoviridae*. Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters					
Frequency	**No. parameters**	**1→2**	**1→3**	**2→1**	**2→3**	**3→1**	**3→2**
29.27%	1	A	-	A	A	A	A
24.60%	1	A	A	A	-	A	A
10.30%	1	A	A	A	-	-	A
9.55%	1	A	-	A	A	-	A
6.95%	1	A	A	A	-	A	-
6.35%	1	A	A	A	A	A	A

Table B.3. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for viral phylogeny of family *Paramyxoviridae*. Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters											
Frequency	**No. parameters**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
13.87%	1	-	-	A	A	-	A	A	A	A	A	A	A
12.72%	1	-	-	A	A	A	A	A	A	A	A	A	-
11.45%	1	-	-	A	A	A	A	-	A	A	A	A	-
7.42%	1	-	-	A	A	A	A	A	A	A	A	A	A
6.46%	1	-	-	A	-	A	A	A	A	A	A	A	-
5.73%	1	-	-	A	A	A	A	-	A	-	A	A	-

Table B.4. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for taxonomically-structured cladogram under branch length assumption set (a), where deep taxonomic branches (root to genome type, genome type to family) are ten-fold the length of shallow taxonomic branches (family to genus, genus to species). Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters											
Frequency	**No. parameters**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
38.46%	2	A	-	A	B	A	-	B	A	-	A	A	-
23.40%	2	A	-	A	B	A	-	B	-	-	A	A	-
16.51%	2	A	-	A	B	A	-	B	B	-	A	A	-

Table B.5. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for taxonomically-structured cladogram under branch length assumption set (b), where deep taxonomic branches (root to genome type, genome type to family) are equal lengths to shallow taxonomic branches (family to genus, genus to species). Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters											
Frequency	**No. parameters**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
27.16%	2	A	-	A	B	A	-	B	A	-	A	A	-
12.64%	2	A	-	A	B	A	-	B	B	-	A	A	-
12.49%	2	A	-	A	B	A	-	B	-	-	A	A	-
5.19%	2	A	-	A	B	A	-	B	A	-	B	A	-

Table B.6. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for taxonomically-structured cladogram, under branch length assumption set (c), where deep taxonomic branches (root to genome type, genome type to family) are a hundred-fold the length of shallow taxonomic branches (family to genus, genus to species). Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters											
Frequency	**No. parameters**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
35.08%	2	A	-	A	B	A	-	B	A	-	A	A	-
21.72%	2	A	-	A	B	A	-	B	-	-	A	A	-
15.72%	2	A	-	A	B	A	-	B	B	-	A	A	-
7.27%	2	A	-	A	B	A	-	A	A	B	A	A	-

Table B.7. All model configurations with at least 5% frequency featured in RJ-MCMC analysis for taxonomically-structured cladogram, under branch length assumption set (d), where branch lengths were scaled to be proportional to the number of descendent nodes, following Grafen (1989). Run presented is that with the highest marginal likelihood. Letters in parameter columns show which transitions shared the same parameter values. "-" denotes a transition fixed at zero by the RJ-MCMC chain. Those parameters defining the stepwise or 'off-the-shelf' models are shaded.

		Parameters											
Frequency	**No. parameters**	**1→2**	**1→3**	**1→4**	**2→1**	**2→3**	**2→4**	**3→1**	**3→2**	**3→4**	**4→1**	**4→2**	**4→3**
27.78%	2	A	A	A	B	A	-	B	B	B	B	B	-
11.47%	2	A	A	A	B	A	-	A	B	B	B	B	A
8.96%	2	A	A	A	B	A	-	B	B	A	B	B	-
5.30%	2	A	A	A	B	A	-	B	B	B	B	B	A

Appendix C. Supplementary material for: Breadth and specificity of mammal host range predict dynamics of RNA virus emergence in humans

C.1. Supplementary methods

C.1.1. Human specialist sensitivity reanalysis

Please note that the following analysis was conducted post-thesis submission.

To test independence of model conclusions regarding human-to-human transmissibility to the inclusion of human-specialist viruses, final models predicting any transmissibility (Table 4.2) and sustained transmissibility (Table 4.3) were applied to the secondary dataset which excluded these specialist viruses (n = 239). Additionally, model building paths predicting these traits and featuring host range metrics a) number of nonhuman mammal host species, b) number of nonhuman mammal host orders, and c) phylogenetic breadth of nonhuman mammal host species were reconstructed using the secondary dataset. Unless stated otherwise, all model protocols followed the methods described in Chapter 4.

When human specialists were removed, reconstructions of the same final models as in Chapter 4 showed only very tentative or no evidence for an association between human-to-human transmissibility and breadth of mammal host orders (any transmissibility: LRT = 1.47, p = 0.226; sustained transmissibility: LRT = 3.55, p = 0.059). When model building paths were reconstructed, paths for all host range metrics (a) – (c) converged upon the same best model, which did not contain any host range predictors. Instead, bat-infective viruses were more likely to exhibit any

human-to-human transmissibility (Table C.6), and primate-infective viruses were more likely to exhibit any or sustained human-to-human transmissibility (Table C.6, C.7), consistent with the effects of these predictors in the main analysis of Chapter 4.

C.2. Supplementary figures

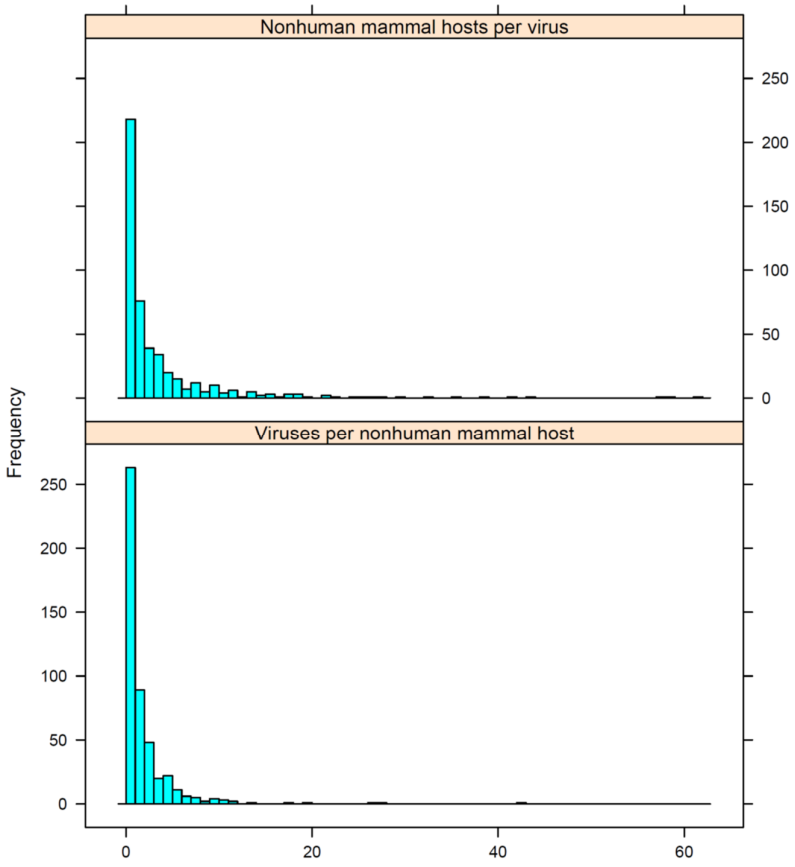


Figure C.1. Histograms of number of known hosts for each virus and number of known viruses in each host across 280 virus species and 683 nonhuman mammal host species. *Rabies virus* is omitted for clarity, having a much larger number of known host species than any other virus (147 host species).

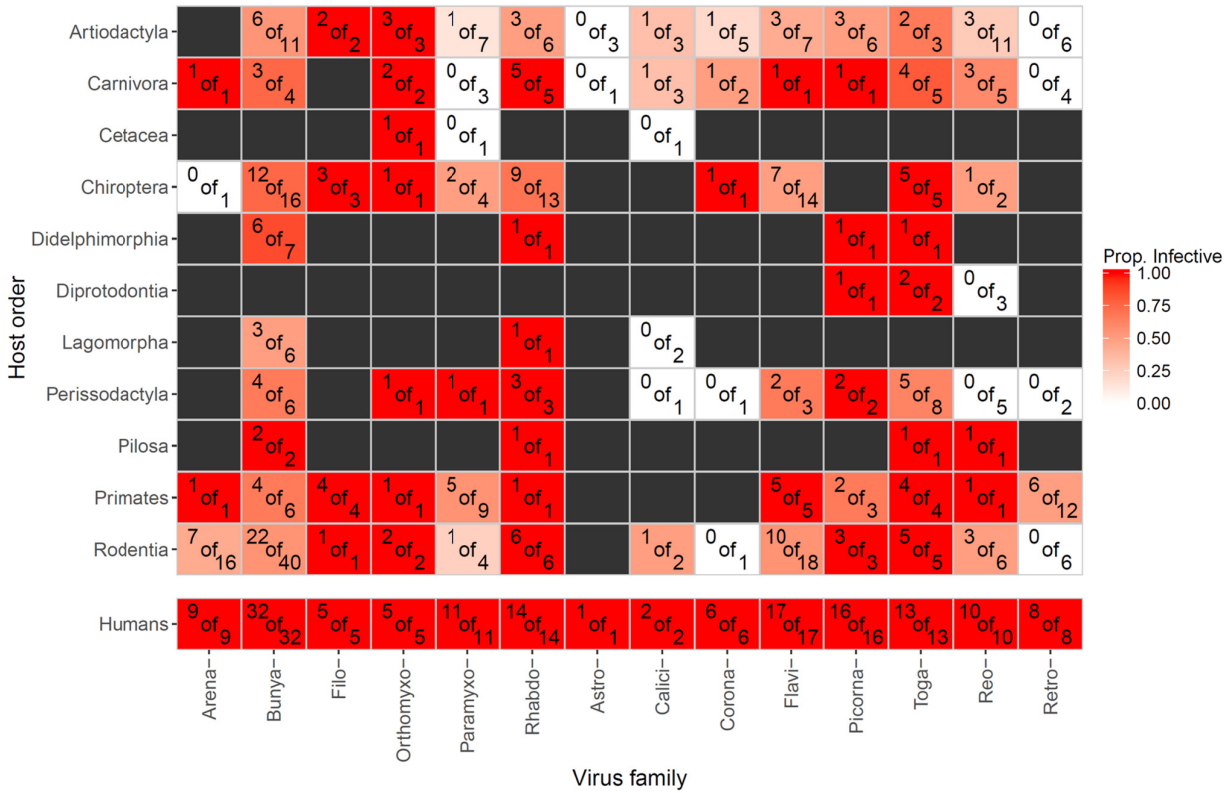


Figure C.2. Heatmap of nonhuman mammal host taxonomic order versus virus taxonomic family, with number in cells denoting number of human-infective viruses out of the total number of viruses in each viral family infecting each mammal order. A separate row indicates the number of human-infective viruses in each viral family. Colour scale denotes number of human-infective viruses as a proportion. Grey cells with no text indicate no viruses in that family infect that mammal order. Viral families are arranged by genome type (-ssRNA: Arena- to Rhabdo-; +ssRNA: Astro- to Toga; dsRNA: Reo-, ssRNA-RT: Retro-).

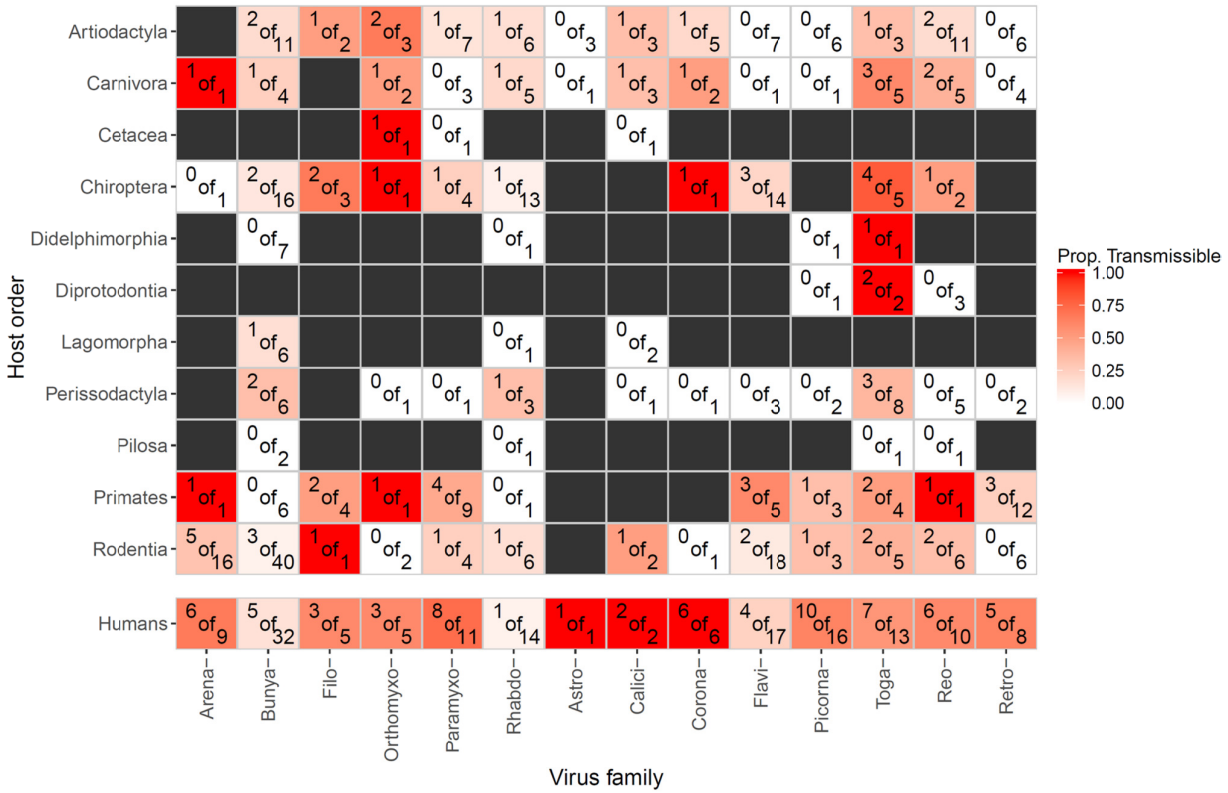


Figure C.3. Heatmap of nonhuman mammal host taxonomic order versus virus taxonomic family, with number in cells denoting number of human-transmissible viruses out of the total number of viruses in each viral family infecting each mammal order. A separate row indicates the number of human-transmissible viruses among human-infective viruses. Colour scale denotes number of human-transmissible viruses as a proportion. Grey cells with no text indicate no viruses in that family infect that mammal order. Viral families are arranged by genome type (-ssRNA: Arena- to Rhabdo-; +ssRNA: Astro- to Toga; dsRNA: Reo-, ssRNA-RT: Retro-).

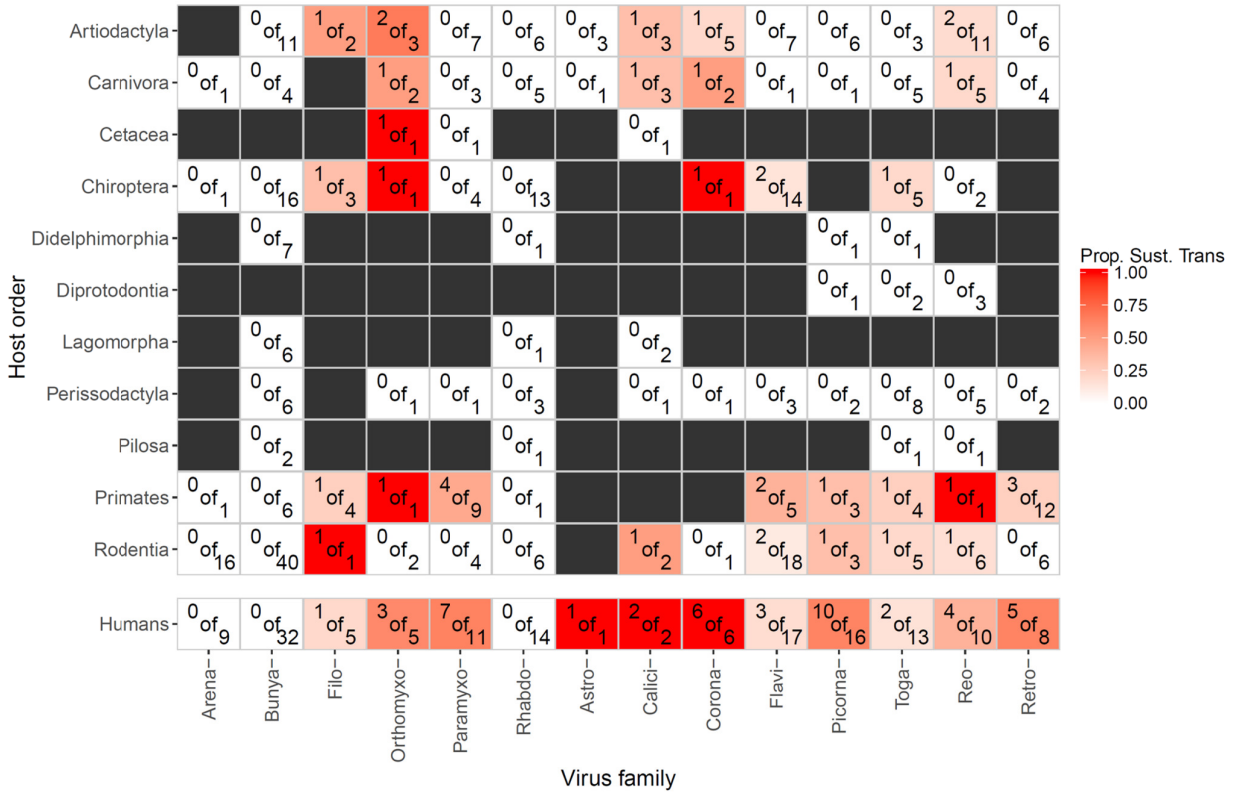


Figure C.4. Heatmap of nonhuman mammal host taxonomic order versus virus taxonomic family, with number in cells denoting number of sustained human-transmissible viruses out of the total number of viruses in each viral family infecting each mammal order. A separate row indicates the number of sustained human-transmissible viruses among human-infective viruses Colour scale denotes number of sustained human-transmissible viruses as a proportion. Grey cells with no text indicate no viruses in that family infect that mammal order. Viral families are arranged by genome type (-ssRNA: Arena- to Rhabdo-; +ssRNA: Astro- to Toga; dsRNA: Reo-, ssRNA-RT: Retro-).

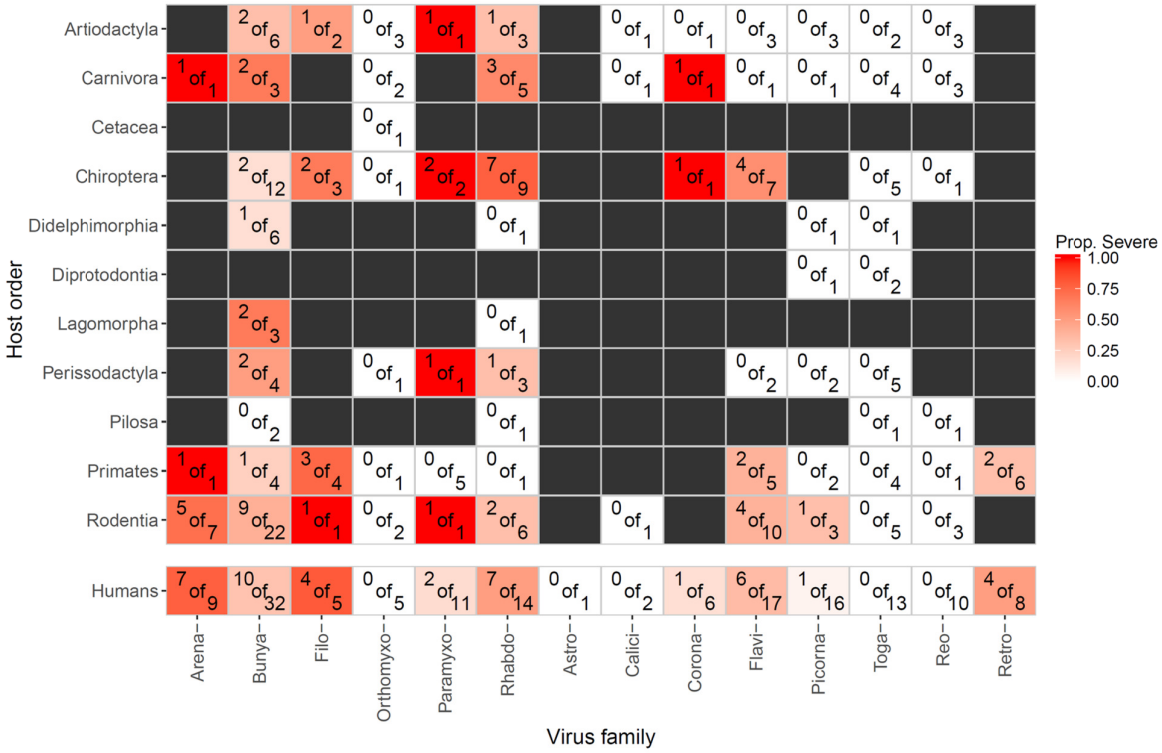


Figure C.5. Heatmap of nonhuman mammal host taxonomic order versus virus taxonomic family, with number in cells denoting number of viruses causing severe human disease out of the total number of human-infective viruses in each viral family infecting each mammal order. A separate row indicates the number of viruses causing severe human disease among human-infective viruses. Note that denominators differ from Figure C.2-4 as virulence was rated only for human-infective viruses (n = 148). Colour scale denotes number of sustained human-transmissible viruses as a proportion. Grey cells with no text indicate no human-infective viruses in that family infect that mammal order. Viral families are arranged by genome type (-ssRNA: Arena- to Rhabdo-; +ssRNA: Astro- to Toga; dsRNA: Reo-, ssRNA-RT: Retro-).

C.3. Supplementary tables

Table C.1. Outputs from best logistic mixed regression models of each model building path predicting human infectivity among mammalian RNA viruses excluding human-specialist viruses, using different host range metrics (a) – (d). The final model featured in main text and selected based on AIC is highlighted in bold. LRT = likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(a) No. nonhuman mammal host species model (n = 239, AIC = 275.31)			
(intercept)	0.086 (0.031, 0.234)	-	-
log(Virus citations)	1.376 (1.153, 1.642)	14.70	< 0.001
Chiroptera infection	4.169 (1.826, 9.515)	2.03	0.001
Primate infection	7.124 (2.479, 20.469)	6.66	< 0.001
Rodentia infection	2.801 (1.361, 5.765)	8.05	0.005
(b) No. nonhuman mammal host orders model (n = 239, AIC = 275.31)			
(intercept)	0.086 (0.031, 0.234)	-	-
log(Virus citations)	1.376 (1.153, 1.642)	14.70	< 0.001
Chiroptera infection	4.169 (1.826, 9.515)	12.03	0.001
Primate infection	7.124 (2.479, 20.469)	16.66	< 0.001
Rodentia infection	2.801 (1.361, 5.765)	8.05	0.005
(c) Phylogenetic breadth model (n = 239, AIC = 275.31)			
(intercept)	0.086 (0.031, 0.234)	-	-
log(Virus citations)	1.376 (1.153, 1.642)	14.70	< 0.001
Chiroptera infection	4.169 (1.826, 9.515)	12.03	0.001
Primate infection	7.124 (2.479, 20.469)	16.66	< 0.001
Rodentia infection	2.801 (1.361, 5.765)	8.05	0.005
(d) Minimum phylogenetic distance to humans model (n = 239, AIC = 271.22)			
(intercept)	1.874 (0.376, 9.348)	-	-
log(Virus citations)	1.388 (1.161, 1.659)	15.05	< 0.001
Min. phylogenetic distance to humans among hosts	0.984 (0.976, 0.992)	20.75	< 0.001
Chiroptera infection	4.257 (1.844, 9.829)	12.14	< 0.001
Rodentia infection	2.630 (1.274, 5.430)	7.00	0.008

Table C.2. Outputs from best logistic mixed regression models of each model building path predicting human-to-human transmissibility among mammalian RNA viruses, using different host range metrics (a) – (d). The final model featured in main text and selected based on AIC is highlighted in bold. LRT = likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(a) No. nonhuman mammal host species model (n = 280, AIC = 255.68)			
(intercept)	0.075 (0.029, 0.194)	-	-
log(Virus citations)	1.671 (1.399, 1.995)	43.29	< 0.001
Artiodactyla infection	0.265 (0.115, 0.610)	11.26	0.001
Rodentia infection	0.400 (0.184, 0.868)	5.66	0.017
(b) No. nonhuman mammal host orders model (n = 280, AIC = 234.49)			
(intercept)	0.156 (0.053, 0.457)	-	-
log(Virus citations)	1.685 (1.386, 2.048)	35.99	< 0.001
log(No. mammal host orders)	0.032 (0.009, 0.108)	38.08	< 0.001
Carnivora infection	4.942 (1.487, 16.429)	6.71	0.010
Chiroptera infection	9.514 (2.998, 30.194)	16.09	< 0.001
Primate infection	5.731 (1.965, 16.709)	10.66	0.001
(c) Phylogenetic breadth model (n = 280, AIC = 247.49)			
(intercept)	0.001 (0.000, 0.021)	-	-
log(Virus citations)	1.561 (1.298, 1.878)	27.69	< 0.001
Phylogenetic breadth binary variable (breadth = zero)	137.275 (5.714, 3297.918)	11.57	0.001
log(Phylogenetic breadth)	2.294 (1.276, 4.125)	9.08	0.003
Artiodactyla infection	0.195(0.077, 0.495)	14.04	< 0.001
Rodentia infection	0.392 (0.159, 0.969)	4.31	0.038
(d) Minimum phylogenetic distance to humans model (n = 239, AIC not compared)			
(intercept)	0.162 (0.032, 0.828)	-	-
log(Virus citations)	1.430 (1.180, 1.732)	15.06	< 0.001
Min. phylogenetic distance to humans among hosts	0.989 (0.982, 0.996)	9.24	0.002
Chiroptera infection	3.162 (1.156, 8.651)	5.29	0.021

Table C.3. Outputs from best logistic mixed regression models of each model building path predicting sustained human-to-human transmissibility among mammalian RNA viruses, using different host range metrics (a) – (d). The final model featured in main text and selected based on AIC is highlighted in bold. LRT = likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(a) No. nonhuman mammal host species model (n = 280, AIC = 180.47)			
(intercept)	0.053 (0.015, 0.184)	-	-
log(Virus citations)	1.613 (1.305, 1.993)	23.43	< 0.001
log(No. mammal host species)	0.315 (0.188, 0.528)	24.77	< 0.001
Primate infection	5.903 (1.883, 18.502)	9.79	0.002
(b) No. nonhuman mammal host orders model (n = 280, AIC = 157.32)			
(intercept)	0.106 (0.023, 0.490)	-	-
log(Virus citations)	1.631 (1.288, 2.065)	20.29	< 0.001
log(No. mammal host orders)	0.012 (0.002, 0.057)	49.74	< 0.001
Chiroptera infection	9.101 (1.670, 49.586)	6.68	0.010
Primate infection	21.424 (4.955, 92.634)	20.82	< 0.001
(c) Phylogenetic breadth model (n = 280, AIC = 189.98)			
(intercept)	0.005 (0.000, 0.169)	-	-
log(Virus citations)	1.549 (1.259, 1.906)	19.35	< 0.001
Phylogenetic breadth binary variable (breadth = zero)	11.740 (0.349, 394.882)	2.07	0.150
log(Phylogenetic breadth)	1.273 (0.669, 2.424)	0.56	0.454
Artiodactyla infection	0.206 (0.067, 0.639)	9.00	0.003
(d) Minimum phylogenetic distance to humans model (n = 239, AIC not compared)			
(intercept)	0.455 (0.065, 3.196)	-	-
log(Virus citations)	1.226 (0.969, 1.551)	2.91	0.088
Min. phylogenetic distance to humans among hosts	0.982 (0.972, 0.992)	18.81	< 0.001

Table C.4. Outputs from best logistic mixed regression models of each model building path predicting severe disease among human RNA viruses, using different host range metrics (a) – (d). The final model featured in main text and selected based on AIC is highlighted in bold. LRT = likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(a) No. nonhuman mammal host species model (n = 148, AIC = 153.64)			
(intercept)	0.120 (0.023, 0.620)	-	-
log(Virus citations)	1.004 (0.796, 1.267)	< 0.01	0.972
log(No. mammal host species)	2.574 (1.584, 4.183)	17.53	< 0.001
Artiodactyla infection	0.139 (0.030, 0.643)	7.65	0.006
Primate infection	0.250 (0.061, 1.013)	4.33	0.037
(b) No. nonhuman mammal host orders model (n = 148, AIC = 164.34)			
(intercept)	0.200 (0.050, 0.801)	-	-
log(Virus citations)	0.993 (0.814, 1.212)	< 0.01	0.946
Chiroptera infection	3.073 (1.171, 8.065)	5.26	0.022
(c) Phylogenetic breadth model (n = 148, AIC = 157.30)			
(intercept)	0.167 (0.005, 5.225)	-	-
log(Virus citations)	1.018 (0.813, 1.274)	0.02	0.877
Phylogenetic breadth binary variable (breadth = zero)	0.612 (0.019, 19.870)	0.08	0.783
log(Phylogenetic breadth)	1.285 (0.689, 2.398)	0.63	0.427
Artiodactyla infection	0.186 (0.042, 0.817)	5.67	0.017
(d) Minimum phylogenetic distance to humans model (n = 107, AIC not compared)			
(intercept)	0.612 (0.180, 2.084)	-	-
log(Virus citations)	0.918 (0.744, 1.133)	0.64	0.424

Table C.5. Outputs from likelihood ratio tests comparing model fits when adding random slopes upon host range metrics with respect to viral family. Cells denote LRT results for best models of each modelled virus trait in humans (columns) and each host range metric (rows) as presented in Table C.1 – C.4. Results where LRT ≈ 0 denote a negligible difference in likelihoods following addition of random slopes.

	Modelled virus trait			
Host range metric	**1. Infectivity**	**2. Any transmissibility**	**3. Sustained transmissibility**	**4. Severe disease**
(a) No. nonhuman mammal host species	*(host range not featured)*	*(host range not featured)*	LRT = 0.09, p = 0.762	LRT ≈ 0
(b) No. nonhuman mammal host orders	*(host range not featured)*	LRT ≈ 0	LRT = 0.25, p = 0.618	*(host range not featured)*
(c) Phylogenetic breadth	*(host range not featured)*	LRT = 0.22, p = 0.638	LRT = 0.50, p = 0.482	LRT ≈ 0
(d) Minimum phylogenetic distance to humans	LRT ≈ 0	LRT = 2.42, p = 0.120	LRT = 4.48, p = 0.034	*(host range not featured)*

Table C.6. Output from best logistic mixed regression model predicting human-to-human transmissibility among mammalian RNA viruses, excluding human-specialist viruses (n = 239). Model was convergently reached within each model building path using different host range metrics (a) – (c). LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	0.022 (0.007, 0.068)	-	-
log(Virus citations)	1.447 (1.196, 1.749)	16.52	< 0.001
Chiroptera infection	3.089 (1.157, 8.249)	5.34	0.014
Primate infection	3.182 (1.254, 8.072)	5.99	0.021

Table C.7. Output from best logistic mixed regression model predicting sustained human-to-human transmissibility among mammalian RNA viruses, excluding human-specialist viruses (n = 239). Model was convergently reached within each model building path using different host range metrics (a) – (c). LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
(intercept)	0.015 (0.004, 0.060)	-	-
log(Virus citations)	1.244 (0.988, 1.567)	3.46	0.063
Primate infection	9.736 (2.669, 35.511)	15.49	< 0.001

Appendix D. Supplementary material for: Allometry of mammal species predicts ability to host virulent human RNA viruses

D.1. Supplementary methods

D.1.1. Virus-centric mixed regression model analysis

To confirm conclusions were not biased by non-independencies arising from the data structure (i.e. the same virus may be hosted by and therefore represented within the virus richness of several different mammal species), analyses were repeated at the level of virus species. Virus-centric analyses were conducted a) using the same final dataset of 125 zoonotic viruses and 524 mammal host species as described in Chapter 5, and b) restricting data to host-virus relationships presenting stringent diagnostic evidence suggestive of within-host replication (defined as those identified via virus isolation or PCR-based methods), resulting in 95 zoonotic viruses and 298 mammal host species, within 494 host-virus pairs.

For each virus species, I calculated the median log-transformed adult body mass among all known host species. I also obtained the number of literature citations in PubMed search results for the virus species name. I then constructed mixed logistic regression models where outcomes were specified as categorical variables as to whether viruses caused severe human disease and whether viruses were capable of human-to-human transmission, respectively. Model predictors were specified as the median log-transformed host body mass, and log-transformed virus citation count. Models were fitted with a random intercept term upon viral taxonomic family, and additional model fits were tested by applying random slopes to the median body mass predictor. Unless otherwise stated, models were constructed, assessed, verified, and plotted following the methods described in Chapter 5.

All model conclusions were concordant with the primary analysis for both a) the complete dataset, and b) the restricted dataset. Specifically, viruses with a smaller median body mass among their mammal hosts were more likely to cause severe disease (Table D.1, Figure D.5), and no relationship was detected between median host body mass and human-to-human transmissibility (Table D.2). Additionally, virus sampling effort did not predict disease severity (Table D.1), but positively predicted human-to-human transmissibility (Table D.2), in agreement with results elsewhere within the thesis (see Chapter 4). In no cases did introducing random slopes improve model fit (highest observed LRT statistic = 1.12, $p = 0.570$), nor was any influence of collinearity upon fitted model variances detected (highest observed VIF value = 1.008).

D.2. Supplementary figures

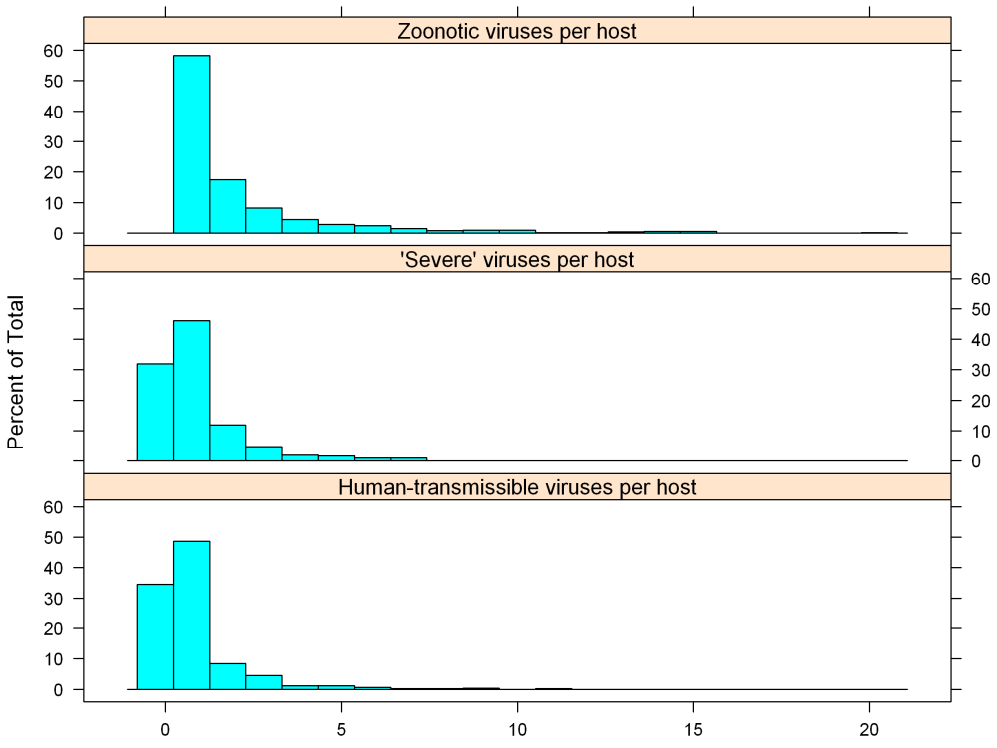


Figure D.1. Histograms of number of known RNA viruses with zoonotic status, causing 'severe' human disease, and capable of human-to-human transmission per mammal host species (n = 524).

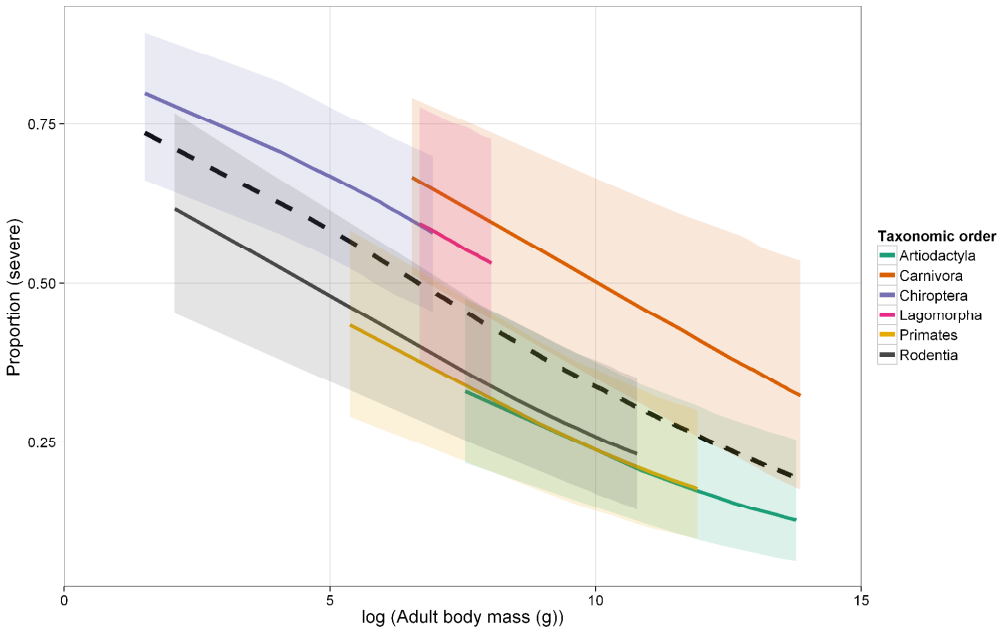


Figure D.2. Relationship between adult body mass in grams and proportion of zoonotic RNA viruses causing 'severe' human disease among 524 mammal species within six taxonomic orders as in Figure 5.1, within a single panel. Solid lines denote fitted effect from logistic mixed regression model with random intercepts for each taxonomic order, holding mammal species citations constant at the median value. Shaded areas denote 95% confidence interval from 1000 simulations accounting for variances of both the regression slope and random intercepts. Dashed black line denotes the overall regression slope with marginalised random intercepts (i.e. independent of mammal order).

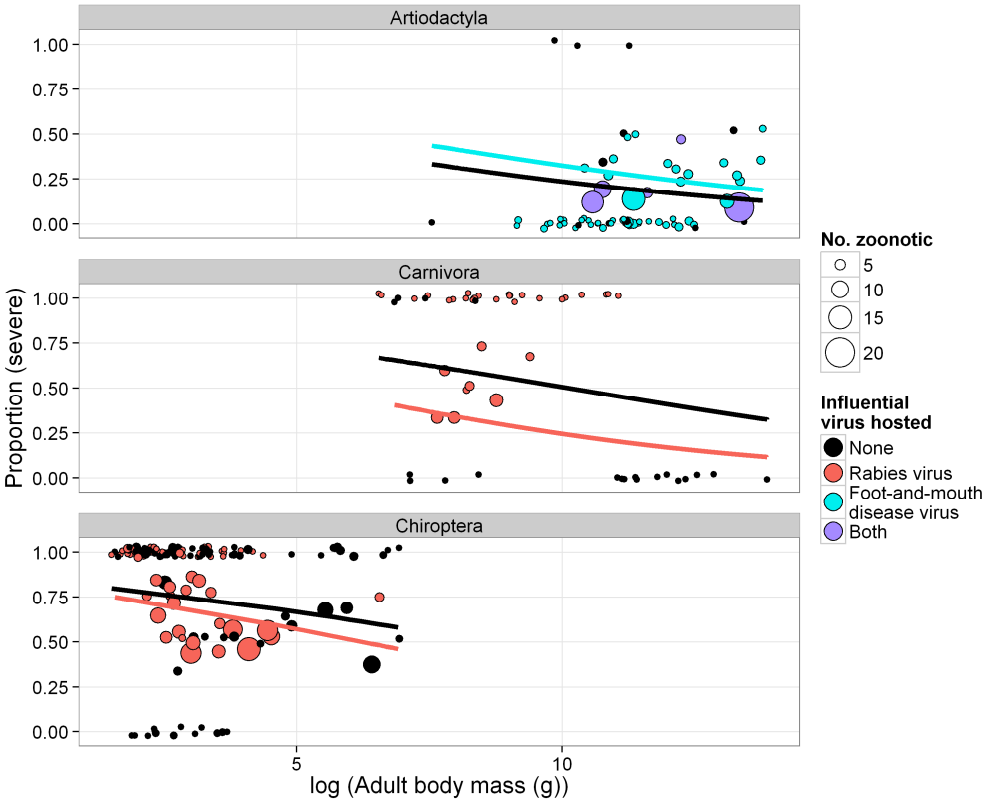


Figure D.3. Relationship between adult body mass in grams and proportion of zoonotic RNA viruses causing 'severe' human disease among mammal species as in Figure 5.1, for the three taxonomic orders primarily hosting viruses suspected as influential within logistic mixed regression models. Filled circles denote individual host species with sizes denoting number of zoonotic viruses and solid lines denote fitted effect from logistic mixed regression model as in Figure 5.1. Circle colours denote whether host species is a known host of *Rabies virus*, *Foot-and-mouth-disease virus*, or both. Hosts with the smallest circle size therefore are only known to host the single respective virus denoted by their colour. Lines denote fitted effect from logistic mixed regression models within sensitivity analysis, with line colours denoting virus removals (black: full dataset as in Figure 5.1, red: *Rabies virus* removed, blue: *Foot-and-mouth-disease virus* removed).

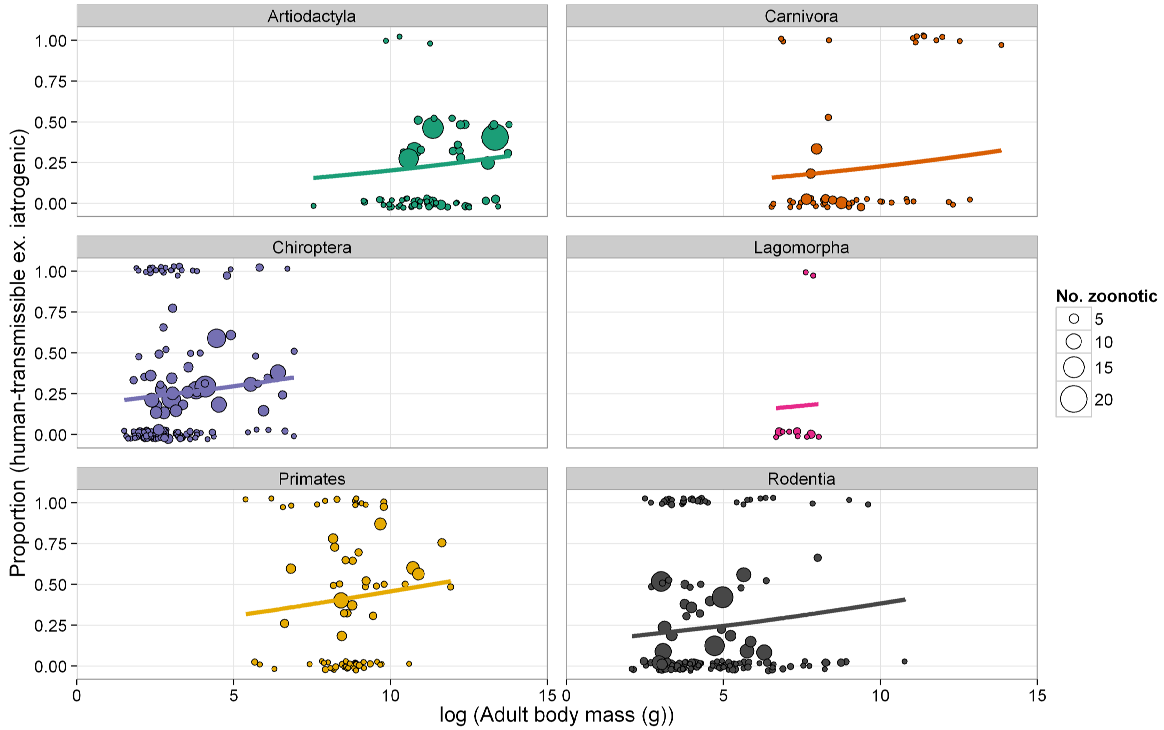


Figure D.4. Relationship between adult body mass in grams and proportion of zoonotic RNA viruses capable of human-to-human transmission (excluding iatrogenic transmission) among 524 mammal species within six taxonomic orders as in Figure 5.1, illustrated using separate panels and colour schemes. Filled circles denote individual host species where circle size denotes number of zoonotic viruses (i.e. denominator in calculated proportion on the y axis) and model weighting in logistic mixed regression. Filled circles are jittered for visibility. Lines denote fitted effect from logistic mixed regression model with random intercepts for each taxonomic order, holding mammal species citations constant at the median value.

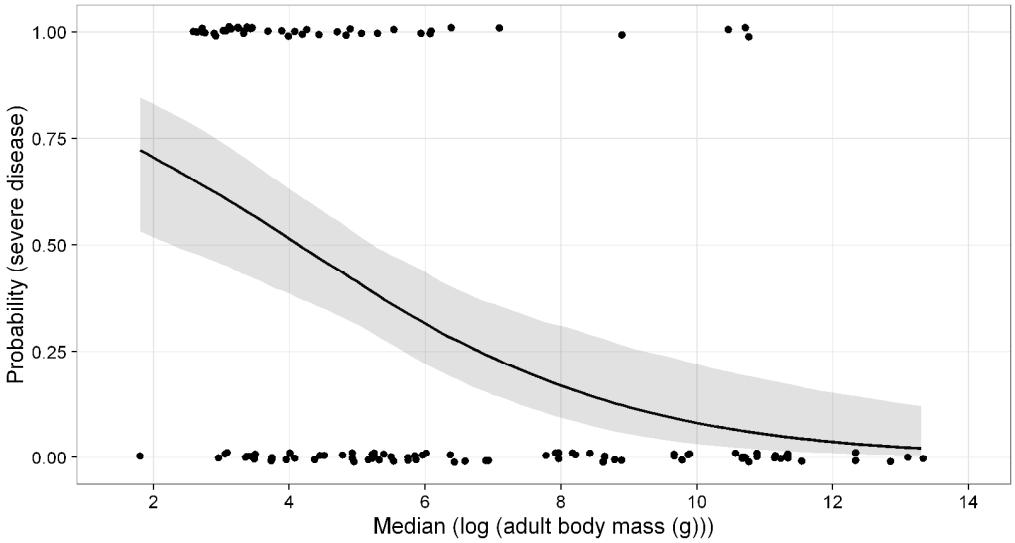


Figure D.5. Relationship between viral disease severity and median adult body mass of mammal host species for complete dataset containing 125 zoonotic RNA viruses. Solid line denotes fitted effect from logistic mixed regression model, holding virus citations constant at the median value and marginalising over random intercepts. Shaded areas denote 95% confidence interval based on 1000 simulations.

D.3. Supplementary tables

Table D.1. Outputs from logistic mixed regression models predicting severe disease among zoonotic RNA viruses based on A) complete dataset following Chapter 5 and B) restricted dataset containing only host-virus relationships with stringent evidence of within-host replication. Variance in random intercept term of virus taxonomic family for A) = 0.058, B) = 0.714. LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
A) Complete zoonotic RNA virus dataset (n = 125)			
(intercept)	5.887 (1.537, 27.419)	-	-
median log(Adult body mass of mammal hosts (g))	0.668 (0.537, 0.795)	23.52	< 0.001
log(Virus citations)	0.990 (0.805, 1.208)	0.01	0.923
B) Restricted zoonotic RNA virus dataset (n = 95)			
(intercept)	12.837 (1.616, 77.623)	-	-
median log(Adult body mass of mammal hosts (g))	0.708 (0.547, 0.860)	13.57	< 0.001
log(Virus citations)	0.898 (0.716, 1.119)	0.93	0.335

Table D.2. Outputs from logistic mixed regression models predicting human-to-human transmissibility among zoonotic RNA viruses based on A) complete dataset following Chapter 5 and B) restricted dataset containing only host-virus relationships with stringent evidence of within-host replication. Variance in random intercept term of virus taxonomic family for A) = 0.704, B) = 2.426. LRT = Likelihood ratio test.

Covariate	Odds ratio (95% CI)	LRT statistic	p(LRT)
A) Complete zoonotic RNA virus dataset (n = 125)			
(intercept)	0.094 (0.008, 0.229)	-	-
median log(Adult body mass of mammal hosts (g))	1.112 (0.953, 1.291)	1.87	0.172
log(Virus citations)	1.545 (1.246, 2.001)	17.57	< 0.001
B) Restricted zoonotic RNA virus dataset (n = 95)			
(intercept)	0.061 (0.001, 0.254)	-	-
median log(Adult body mass of mammal hosts (g))	1.057 (0.859, 1.288)	0.30	0.584
log(Virus citations)	1.943 (1.392, 3.096)	20.58	< 0.001

Appendix E. Publication: RNA viruses: a case study of the biology of emerging infectious diseases

The following publication is a review addressing emerging RNA viruses that describes a comprehensive list of viruses recognised to infect humans and their characteristics, which was updated through the systematic literature reviews conducted by myself as part of this thesis.

Citation: Woolhouse MEJ, Adair K, Brierley L. RNA viruses: a case study of the biology of emerging infectious diseases. Microbiol Spectrum. 2013;1(1):OH-0001-2012.

Microbiology Spectrum
American Society for Microbiology Press

RNA Viruses: A Case Study of the Biology of Emerging Infectious Diseases

MARK E. J. WOOLHOUSE, KYLE ADAIR, and LIAM BRIERLEY
Centre for Immunity, Infection & Evolution, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

ABSTRACT There are 180 currently recognized species of RNA virus that can infect humans, and on average, 2 new species are added every year. RNA viruses are routinely exchanged between humans and other hosts (particularly other mammals and sometimes birds) over both epidemiological and evolutionary time: 89% of human-infective species are considered zoonotic and many of the remainder have zoonotic origins. Some viruses that have crossed the species barrier into humans have persisted and become human-adapted viruses, as exemplified by the emergence of HIV-1. Most, however, have remained as zoonoses, and a substantial number have apparently disappeared again. We still know relatively little about what determines whether a virus is able to infect, transmit from, and cause disease in humans, but there is evidence that factors such as host range, cell receptor usage, tissue tropisms, and transmission route all play a role. Although systematic surveillance for potential new human viruses in nonhuman hosts would be enormously challenging, we can reasonably aspire to much better knowledge of the diversity of mammalian and avian RNA viruses than exists at present.

INTRODUCTION

Viruses account for only a small fraction of the 1400 or more different species of pathogen that plague humans —the great majority are bacteria, fungi, or helminths (1). However, as both the continuing toll of childhood infections such as measles and recent experience of AIDS and influenza pandemics illustrate, viruses are rightly high on the list of global public health concerns (2). Moreover, the great majority of newly recognized human pathogens over the past few decades have been viruses (3) and a large fraction of emerging infectious disease "events" have involved viruses (4).

There are two kinds of viruses: RNA viruses and DNA viruses. The latter largely consist, with the exception of a handful of pox- and herpesviruses, of viruses that have probably been present in and coevolved with humans for long periods of time. RNA viruses are very different. The majority of RNA viruses that infect humans are zoonotic, meaning that they can infect vertebrate hosts other than humans. Many of those that are not regarded as zoonotic are believed to have had recent (in evolutionary terms) zoonotic origins. So it is the RNA viruses that are of greatest interest in the context of One Health.

In this chapter, we review current knowledge of how RNA viruses in humans and other vertebrates are related, in terms of both of their evolution and their ecology, with the intention of trying to understand where human RNA viruses came from in the past and where new ones might emerge in the future. Until recently, research on these topics was essentially a series of case studies. Extraordinary work has been done detailing events such as the historical emergence of HIV-1 in Central Africa (5) and the more recent emergence of Nipah virus in Southeast Asia (6). But while every emergence event is a fascinating story in its own right, our aim here is to look beyond the specifics and to try to

Received: 14 June 2012, **Accepted:** 23 May 2013,
Published: 25 October 2013
Editors: Ronald M. Atlas, University of Louisville, Louisville, KY, and Stanley Maloy, San Diego State University, San Diego, CA
Citation: Woolhouse MEJ, Adair K, Brierley L. 2013. RNA viruses: a case study of the biology of emerging infectious diseases. *Microbiol Spectrum* 1(1):OH-0001-2012. doi:10.1128/microbiolspec.OH-0001-2012.
Correspondence: Mark E. J. Woolhouse, mark.woolhouse@ed.ac.uk

identify any underlying generalities that tell us something useful about the emergence of RNA viruses as a biological process.

We begin by comparing the RNA viruses reported to infect humans with RNA virus diversity as a whole and exploring the overlap between viruses in humans and viruses in other kinds of hosts. Next, we refine the analysis by distinguishing among viruses according to their ability not just to infect humans but also to transmit from one human to another, which is a prerequisite for a virus being able to cause major epidemics and/or become an established, endemic human pathogen. We then consider in more detail the subset of human RNA viruses that can persist in human populations without the need for a nonhuman reservoir. Next, we attempt to identify characteristics of RNA viruses that allow them to cross the species barrier and those that predispose them to cause severe disease, as such viruses are of particular public health concern. We go on to discuss how new human RNA viruses arise (sometimes to subsequently disappear again). From the information assembled we construct a conceptual model of the relationship between RNA viruses in humans and other hosts. We consider how this model might be of practical value, concentrating on risk assessments for newly discovered viruses and also the much discussed topic of the design of surveillance programs for emerging infectious diseases.

DIVERSITY OF HUMAN RNA VIRUSES

The diversity of human RNA viruses was recently surveyed using a formal methodology (3), and we update that information here. All RNA viruses known to infect humans were included, with the exception of those only known to do so as the result of deliberate laboratory exposures.

In this chapter, we use virus species as designated by the Ninth Report of the International Committee for the Taxonomy of Viruses (ICTV) (7) (noting that this differs from earlier ICTV reports used in previous work and that it will doubtless change again in the not-too-distant future). ICTV designations may not always accurately reflect the biological meaning of a "species," i.e., reproductive isolation. The operational criteria used for RNA viruses may include any or all of (i) phylogenetic relatedness based on sequence data, (ii) serological cross-reactivity, (iii) host range, and (iv) transmission route. It is also important to note that any analysis at the level of a virus species implicitly ignores a great deal of biomedically relevant diversity. This point is best illustrated by the influenza A viruses: the epidemiology and public health importance of seasonal influenza A and the H5N1 or H7N9 "bird flu" variants are very different, but all are included within a single species. Less variable virus species than influenza A may still contain multiple serotypes and other functionally distinct subtypes. Despite these limitations, the species remains the most useful unit for studying virus diversity currently available.

Updating the earlier survey (3) with new taxonomic information (7) reveals 180 recognized species of RNA viruses that have been reported to infect humans. These viruses represent 50 genera and 17 families (with one genus, *Deltavirus*, currently unassigned to a family). It is not immediately obvious what we should make of this. Is 180 a large number or a small one? Should we be surprised that it is not much higher or that it is not much lower? We consider such questions further below. We can, at least, be sure that 180 is an underestimate. New human RNA virus species are still being discovered or recognized at a rate of approximately 2 per year, although recent work (8) has suggested that the pool of undiscovered species could be much smaller than previously proposed (3). Even if we still have very little idea of the number of species "out there," it is, as we will consider in detail later on, possible to say something about where "out there" is.

The possibility of large numbers of as yet unrecognized viruses also raises the specter of ascertainment bias. Certain kinds of RNA viruses may be underrepresented, perhaps dramatically so, among those currently recognized. These might be viruses from particular taxonomic groups, those associated with less severe disease or certain kinds of symptoms, or simply those that are rare and/or occur in less studied regions of the world. While this is clearly an issue, it is worth pointing out that both the rates and kinds of RNA viruses being discovered or recognized have been remarkably consistent for the past half century, despite massive changes in the technologies for virus detection and identification and considerable variability in the effort put into virus discovery in different places and at different times (3).

RNA VIRUSES OF HUMANS AND NONHUMANS

One striking observation is that 160 species of human-infective RNA virus species (89% of the total) are regarded as zoonotic; i.e., they can also infect other kinds of vertebrate hosts. (The definition of "zoonotic" ignores arthropod vectors; these are regarded as specialized transmission routes rather than alternative host

species.) The nonhuman hosts usually (>90% of all zoonotic RNA virus species) include other mammals and less commonly (<40%) birds. Humans rarely, if ever, share their RNA viruses with anything else. Although the bias toward sharing viruses with other mammals is obvious, it is less clear whether we preferentially share viruses with particular kinds of mammals. Many human viruses (both RNA and DNA) are shared with ungulates, carnivores, rodents, primates, or bats (3), but our knowledge of the host range of most viruses is too incomplete for us to be confident about any underlying patterns. The remaining 20 RNA viruses are not known to naturally infect nonhuman hosts. However, most of these have close relatives that can infect other mammals. The only exceptions are hepatitis C, hepatitis delta, and rubella virus.

The overlap between the ability to infect humans and the ability to infect other mammals can be illustrated in other ways, too. Of the 62 recognized RNA virus genera containing species that can infect at least one kind of mammal, 50 (81%) contain species that can infect humans. And of the 19 recognized RNA virus families that contain species reported to infect mammals, all but 2 include species found in humans. The exceptions are the *Nodaviridae*, which are essentially insect viruses, and the *Arteriviridae*, which include species infecting a range of different mammals, notably including simian hemorrhagic fever virus.

The fact that human-infective species are distributed so widely among the RNA viruses of mammals strongly suggests that, in evolutionary terms, the ability to infect humans is very easily acquired by these viruses. It also implies that many, perhaps most, human RNA viruses need not have arisen by evolving from other human RNA viruses. This idea is supported by a recent analysis of the relationship between phylogeny and host range for three RNA virus families—*Paramyxoviridae*, *Caliciviridae*, and *Rhabdoviridae*—and two genera—*Alphavirus* and *Flavivirus*—which concluded that the majority of speciation events were associated with host species jumps (9). Note that this pattern contrasts markedly with the human DNA viruses, among which taxa such as the *Papillomaviridae* and the *Anelloviridae* appear to have undergone extensive diversification within humans.

THE PATHOGEN PYRAMID

The categorization of viruses based simply on their ability to infect humans fails to distinguish between a vast range of epidemiologies, from occasional very mild cases of Newcastle disease virus infection to pandemics of influenza A or HIV-1. A useful conceptual framework for thinking about this issue is the pathogen pyramid (10). The version of pyramid used here has four levels (Fig. 1).

Level 1 corresponds to human exposure, whether via ingestion, inhalation, the bite of an arthropod vector, or any other route. As discussed in the previous section, the most important sources of exposure are other mammals and, to a lesser degree, birds. There are no good estimates of the total diversity of mammal and bird viruses, but it seems likely that the human population is exposed to hundreds, perhaps thousands, of species on a regular basis. The major determinants of the rate of exposure to new viruses are the ecology and behavior of humans, the nonhuman virus reservoir(s), and (in some cases) arthropod vectors.

Level 2 corresponds to human infection, which we take to mean the ability to enter and replicate in human cells in vivo. For all (known) RNA viruses there are associated host responses, although not all infections necessarily lead to clinical symptoms of disease. Key determinants of the ability to infect humans include the route of entry (e.g., needle sharing has created a new entry route for blood-borne viruses) and the molecular biology of the human-virus interaction (discussed in more detail below). Of the 180 recognized species of RNA viruses that can infect humans, almost 60% (107 species) are restricted to level 2 (Fig. 2).

Level 3 corresponds to the ability both to infect humans *and* to transmit from one human to another. The ability to transmit refers to all kinds of transmission routes, including vectors. Less than half of human-infective RNA viruses (73 species in all) are able to transmit between humans. A minority of these (26 species) are restricted to level 3 (Fig. 2).

Level 4 corresponds to the ability to transmit sufficiently well that the virus can invade human populations, causing epidemics and/or establishing itself as an endemic human pathogen. In epidemiological parlance, this corresponds to the condition that R_0 is >1 within the human population, where R_0 is the basic reproduction number, defined as the number of secondary cases generated by a single primary case introduced into a large population of naïve hosts. In contrast, level 3 viruses have an R_0 of <1 in humans, which implies that although self-limiting outbreaks are possible, the infection cannot "take off" and cause a major epidemic. Although R_0 is partly determined by the transmissibility of the virus, it is also a function of the behavior and demography of the human host population; for example,

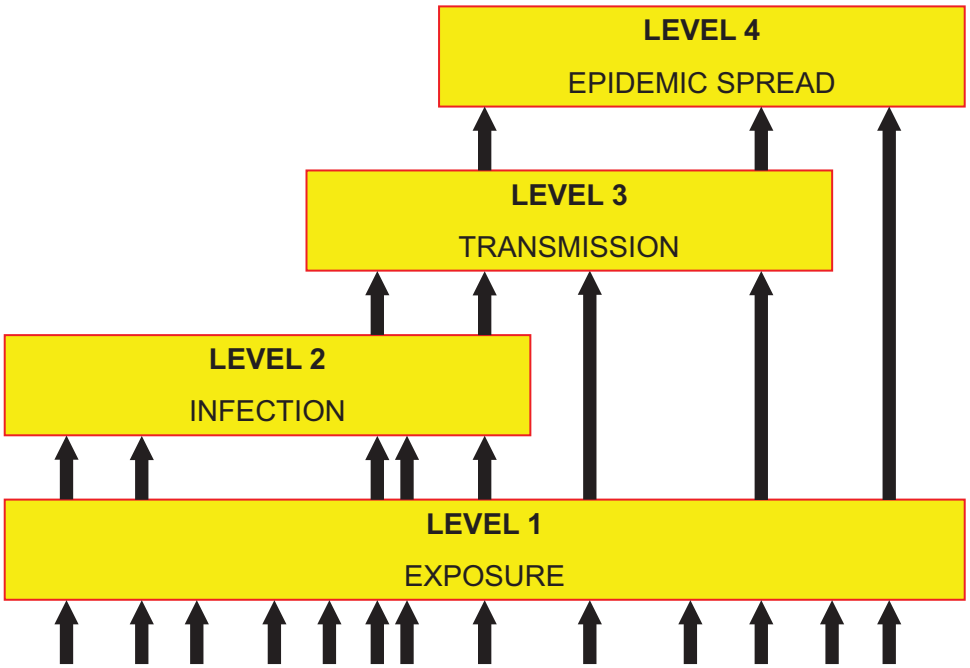


FIGURE 1 A representation of the pathogen pyramid. Each level of the pyramid represents a different degree of interaction between a virus and a human host. Level 1 corresponds to exposure of humans, level 2 to the ability to infect humans, level 3 to the ability to transmit from one human to another, and level 4 to the ability to cause epidemics or persist as an endemic infection. Arrows indicate pathways that viruses may take to reach each level. For example, a level 4 virus may arrive at that state directly, simply by exposure to the virus from a nonhuman reservoir. This is known as an "off-the-shelf" virus. Alternatively, it may initially enter the population as a level 2 or 3 virus—not capable of sustained transmission—but evolve the ability to transmit between humans at a sufficiently high rate to persist within a human population. This is known as a "tailor-made" virus. Adapted from reference 25. doi:10.1128/microbiolspec.OH-0001-2012.f1

changes in living conditions, travel patterns, and sexual behavior (for sexually transmitted viruses) can all greatly influence R_0. More generally, the term "crowd diseases" implies that certain human viruses (and other pathogens) can only become established once critical host population densities have been reached (10). Our best estimate is that there are 47 level 4 RNA virus species in humans (Fig. 2).

A useful exercise is to consider what kinds of viruses are found at levels 2, 3, and 4 in the pyramid. There appear to be three major determinants of this: (i) taxonomy (at the level of both family and genus), (ii) transmission route (especially the distinction between vector-borne transmission and other routes), and (iii) host range (expressed here as the ability to infect different mammalian orders). These three factors are not independent (1); in particular, there are very few vector-borne viruses with narrow host ranges (11).

Nonetheless, several patterns can be identified. First, only two vector-borne viruses are found at the top of the pyramid (level 4): yellow fever and dengue (Fig. 2). It is not immediately apparent why this should be so; we will consider this point further later on. Second, viruses with a host range that is, as far as we know, restricted to primates are rarely found lower down on the pyramid (levels 2 and 3), with a few exceptions such as the simian foamy viruses. The obvious implication is that if a virus is capable of infecting and transmitting from our closest relatives, then it is very likely to have the same capabilities in us. Patterns are also apparent in the taxonomy of human-infective viruses: for example, the *Bunyaviridae*, *Rhabdoviridae*, *Arenaviridae*, and *Togaviridae* (with the exception of rubella, which is atypical of that group) are not represented at level 4 at all. This reflects the fact that these four families are made up of viruses that are vector borne and/or are not primate specialists.

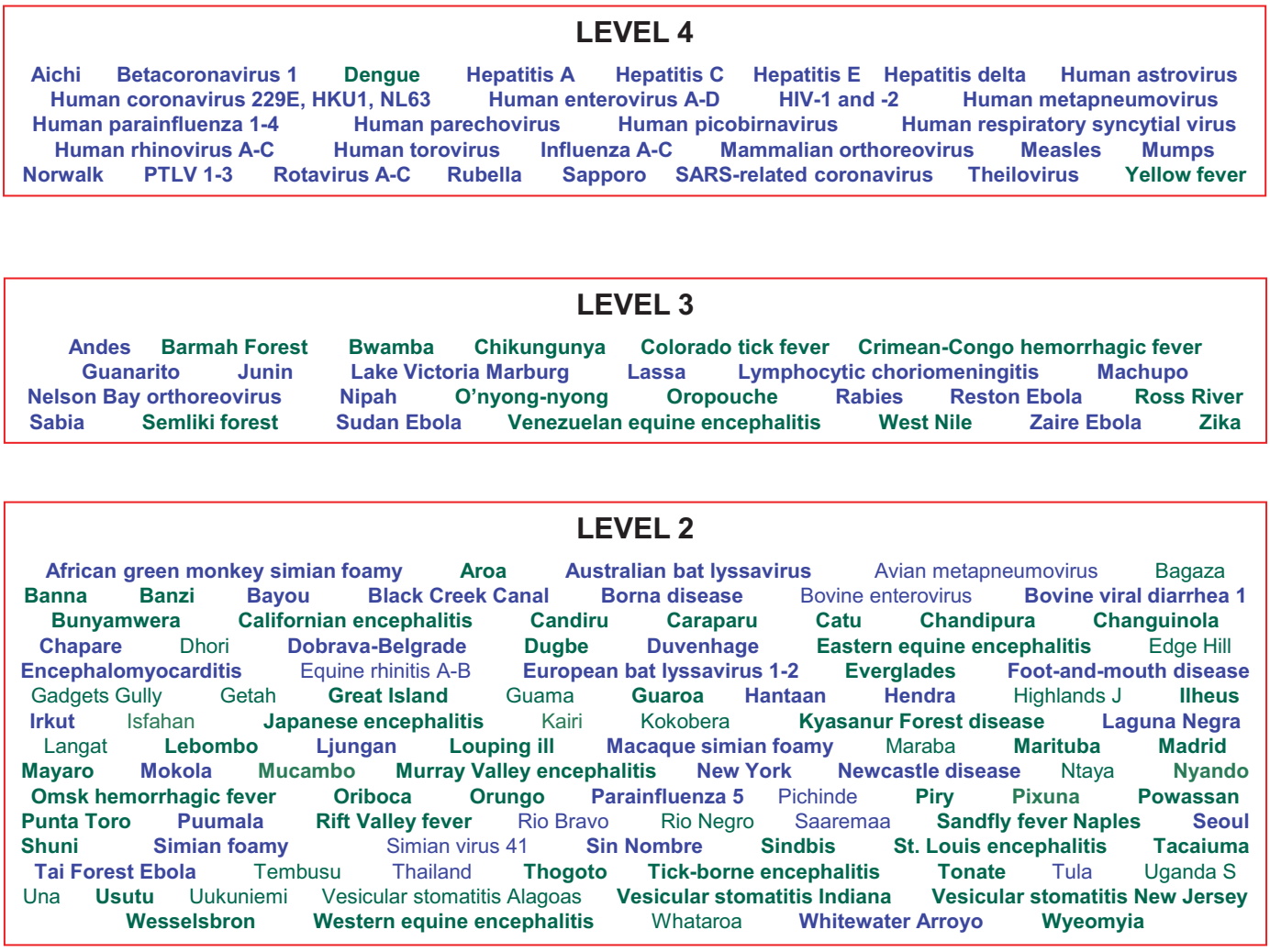


FIGURE 2 All currently recognized human-infective RNA viruses categorized with respect to their ability to infect and transmit from humans (levels 2, 3, and 4 of the virus pyramid—see Fig. 1) and distinguished in terms of transmission route (green for vector-borne transmission, blue for other routes) and nature of diagnostic evidence (the viruses not in boldface type have only been reported in humans using serology-based methods). doi:10.1128/microbiolspec.OH-0001-2012.f2

Finally, it is worth noting that the "shape" of the pathogen pyramid for RNA viruses differs from that for nonviral human pathogens. Most strikingly, much smaller fractions of recognized species of bacteria, fungi, protozoa, or helminths are capable of extensive spread in human populations (i.e., are found at level 4). On the other hand, human DNA viruses are even more concentrated at the top of the pyramid, with almost 90% of species at level 4. These patterns could simply be an artifact of our incomplete knowledge of virus diversity at lower levels of the pyramid, but they could also reflect real biological differences between viruses and other kinds of pathogens: viruses (especially DNA viruses) may be more likely to speciate within humans, or viruses (especially RNA viruses) that jump the species barrier into humans may be more capable of spreading in human populations or of rapidly evolving that capability (see below).

HUMAN-ADAPTED RNA VIRUSES

There is a semantic argument that only those viruses that are capable of persisting in human populations in the absence of a nonhuman reservoir should be described as

"human" viruses. In our terminology these are, by definition, the level 4 viruses, comprising 47 species, 20 of which are not known to have any natural hosts other than humans. These 47 viruses—referred to here as "human adapted"—represent 12 families and 29 genera. Their most striking common characteristic is that almost all of them are transmitted by ingestion, inhalation, or direct contact; just 2 are transmitted by vectors.

There are several possible routes for a virus to reach level 4 on the pathogen pyramid (indicated by the arrows in Fig. 1). One possibility is that humans are exposed to a virus that is already capable of effective transmission between humans; i.e., the virus is preadapted to humans (noting that this does not preclude further adaptation once the virus has entered the human population). These have been termed "off-the-shelf" viruses. Such viruses may be rare, perhaps extremely rare, variants of the population in the nonhuman reservoir, in which case the main determinants of the rate at which such viruses enter the human population is the amount of genetic variability within the reservoir and the rate at which humans are exposed to the preadapted variants.

Another possibility is that the virus first enters the human population with limited ability to transmit between humans (i.e., level 3) but that it is able to evolve that ability before the otherwise self-limiting chain of infections dies out (12). These have been termed "tailor-made" viruses. Key determinants of the rate at which such viruses invade the human population are the frequency of primary infections and the virus mutation rate. We note that for a level 2 virus to evolve human transmissibility, this would have to happen during the course of a primary infection. Such infections presumably give evolution relatively little material to work with, and it may be that level 2 viruses are "dead ends" in an evolutionary sense as well as an epidemiological sense. For example, rabies infections are relatively common in humans and are likely to have been so for thousands of years, but human-transmissible variants have failed to materialize (with the proviso that rabies is technically a level 3 pathogen because of rare instances of human-to-human transmission via organ transplants).

The origins of the human-adapted RNA viruses are of considerable interest, not least as a possible pointer to the likely sources of future viral threats to human health. It has previously been noted (10) that we have information on the origins of only a small minority of human pathogens, including RNA viruses. However, as stated above, it seems likely that many of them arose by species jumps from other mammals or (less often) birds, perhaps followed by some diversification within humans (e.g., human enteroviruses or parainfluenza viruses). The direct transmission routes used by most of these viruses are consistent with their being crowd diseases; that is, in contrast to vector-borne viruses, the basic reproduction number increases with human population density.

MECHANISMS

As explained above, whether a virus is found at level 2, 3, or 4 of the pyramid reflects its ability to transmit from one human to another. Human demography and behavior play a key role in this, but of course, intrinsic properties of the virus are also crucial.

The first consideration is the ability of the virus to infect humans at all. Given the importance of this topic, we know surprisingly little about it. In effect, the question comes down to factors that restrict host range. Empirically, it does seem that the species barriers between different mammals, including humans, are very leaky: the majority of known mammal RNA viruses are capable of infecting multiple species. Only two studies (3, 13), however, have looked systematically at mechanisms, showing that use of a phylogenetically conserved receptor to gain entry to host cells is a necessary but not sufficient condition for a virus to be able to infect both humans and nonprimates. This result appears robust, but the data are incomplete because the cell receptor has yet to be identified for the majority of human viruses.

Gaining entry to host cells is only the first step in initiating an infection. The virus must also be capable of replicating in host cells, being released from host cells, evading the innate immune response, and perhaps becoming systemic. All of these processes depend on the specifics of the molecular interplay between virus and host, and all can contribute to the species barrier and host range restriction (14). The species barrier may be quantitative rather than qualitative, perhaps expressed by the need for a higher infective dose. In one of very few experimental studies of the species barrier (15), it was found that the 50% lethal dose for rabies virus obtained from foxes was up to a million times lower for foxes than it was for cats and dogs. Similarly, there is evidence that human influenza A viruses can replicate in chimpanzees but do so at a much lower rate (14).

The ability to get into (i.e., infect) a host does not equate with the ability to get out of (i.e., transmit from) that host. A key determinant of the ability to transmit is the virus's capacity to invade and replicate in cells of

particular tissues, notably the lower gastrointestinal tract, the upper respiratory tract, the urogenital tract, or possibly the blood or skin. In a few cases, the determinants of tissue tropisms are well understood. For example, H5N1 influenza A transmits well from ducks and poultry but not from humans. This is because it utilizes a variant of the sialic acid receptor in the host cell membrane that occurs in the upper respiratory tract of ducks and poultry but is confined to the lower respiratory tract of humans (14).

Tissue tropisms inevitably play a key role in determining the route of virus transmission (e.g., respiratory, fecal-oral, or arthropod vector). It has been suggested that altering tissue tropism is harder for a virus to achieve than switching host species (9). This idea is borne out by the observation that transmission route tends to be a relatively deep-rooted trait in virus phylogenies, often to the level of family, in marked contrast to host range, which tends to be far more labile.

These few mechanistic and ecological insights fall well short of a proper understanding of why some kinds of viruses tend to occur at higher or lower levels of the pathogen pyramid. Host relatedness seems to play a role; hence, viruses from other primates do seem more likely to be transmissible in humans than those acquired from nonprimates, an idea supported by other studies of host relatedness and pathogen transmissibility (16). But not all highly transmissible human viruses have been acquired from other primates. Transmission route is also important; vector-borne viruses in particular seem to be relatively good at infecting humans but relatively poor at being transmitted by humans (17). It is possible that although humans are frequently exposed to vector-borne viruses, some of which are capable of setting up an infection, these viruses are not easily able to adapt to a new host (perhaps because any adaptation to a new vertebrate host must not compromise their interaction with the invertebrate vector [14]). Those that have adapted to humans—dengue and yellow fever—are ones that probably originated in other primates.

VIRULENCE

In public health terms the ability of a virus to spread through human populations is, of course, only part of the story; human RNA viruses also vary enormously in the degree of harm they cause, a characteristic referred to as virulence. In the context of human infections we generally regard a pathogen as virulent if it has a high case-fatality ratio or if infection routinely results in severe clinical disease. On this basis, HIV-1, severe acute respiratory syndrome coronavirus (SARS-CoV), and rabies would be regarded as virulent, whereas parainfluenza and rhinoviruses would not.

Pathogen virulence is a very complex phenomenon, reflecting properties of the pathogen, the host, and the interaction between them. It has been variously proposed that virulence is influenced by transmission route, host range, level of the pathogen pyramid, and the time that the pathogen and the host have had to coevolve (see reference 18 for an introduction to a large body of literature). These characteristics are not independent, so hypothesis testing is not straightforward, although some theories do look promising. For example, the only two recent instances of newly emerging level 4 pathogens—HIV-1 and SARS-CoV—are/were both spectacularly virulent, in line with ideas that the virulence of novel host-pathogen combinations need not be near any evolutionary optimum. The only two level 4 RNA viruses that are vector borne—dengue and yellow fever—are also relatively virulent, in line with ideas that vector-borne diseases can be more virulent because an ambulant host is not needed for transmission. There are also good examples of very virulent RNA viruses, such as rabies, for which humans are effectively dead-end hosts, in line with ideas that such infections are not subject to any evolutionary constraints because they do not contribute to the next generation of infections. On the other hand, many level 2 viruses, such as Newcastle disease virus, Sindbis virus, and others, result in only mild infections, so rabies may just lie at one end of a broad spectrum.

Another idea is that viruses acquired from particular kinds of reservoirs, primates versus nonprimates or mammals versus birds, might be especially virulent. The evidence, however, is inconsistent in this regard. It is true that some highly virulent human viruses, such as HIV-1 and dengue, were acquired from or are shared with other primates, our closest relatives. On the other hand, some highly virulent viruses are ultimately acquired from hosts much more distantly related to humans, such as H5N1 influenza A from birds or SARS and Nipah viruses from bats.

This important topic would clearly benefit from a systematic survey of the virulence of human RNA viruses (none has been published to date), which could be used to construct formal tests of the various hypotheses about pathogen virulence to be found in the evolutionary biology literature.

EMERGENCE AND THE CHANGING CAST OF RNA VIRUSES

New RNA virus species continue to be discovered, identified, or recognized in humans. Recent examples include Nelson Bay orthoreovirus, Irkut virus, primate T-lymphotropic virus 3, human coronavirus HKU1, and human rhinovirus C. Moreover, there is usually a backlog of reports of new human viruses that have yet to be formally recognized as species. Not all of these viruses will have recently invaded human populations; many will turn out to be long-standing human pathogens that have only recently been recognized or accepted as "species."

It is therefore important to understand that the continued accumulation of recognized human RNA virus species may reflect less the possibility that genuinely new viruses are continually emerging, most likely acquired from nonhuman reservoirs, than the fact that we are still getting to grips with the taxonomic diversity of viruses that have been with us for some time. This distinction between viruses that we have only just discovered and viruses that have only just discovered us is, of course, crucial in the context of emerging infectious diseases. If most of the so-called new viruses are not new at all, then this implies that events such as the advent of HIV/AIDS in the early 1980s or the curtailed SARS epidemic in 2003 may be just unusual, one-off occurrences with their own specific causes. If, on the other hand, genuinely new viruses are appearing all the time, then the HIVs and SARS-CoVs are more accurately regarded as just the highly visible tip of a much larger iceberg. Without a much more detailed and thorough understanding of the phylogenies and origins of all human viruses, not just those with high public health profiles, we cannot resolve this question.

Perhaps the most striking feature of recently discovered RNA viruses is that they tend to be much like the RNA viruses that we already knew about. They are members of the same virus families, have the same transmission routes, and share the same kinds of nonhuman hosts. If these newly recognized viruses are indeed emerging, then it seems as though there is nothing special about emergence, at least from a biologist's perspective. Even if this is correct, it is still often suggested that the *rate*, if not the biology, of pathogen emergence is higher in the early 21st century than it has been in the past. This reflects the notion that a variety of so-called drivers of emergence, ranging from human population growth to changes in farming methods, are combining to create a "perfect storm." This idea is difficult to evaluate critically. Arguably there have been only a handful of global emergence events in the past century, notably those involving HIV-1, variants of influenza A, and SARS-CoV. This is not a strikingly large number given that many of the other 40 or so human-adapted RNA viruses may have emerged only in the past few millennia. Of course, it could be argued that less dramatic events such as the geographical spread of West Nile virus or outbreaks of Ebola are more frequent now than they have been in the past, but that claim is even harder to test with any rigor.

Another side to this issue is rarely discussed. One recent study (8) reports that while the number of virus species accumulates, at the same time many of those recognized in past years or decades seem to have disappeared, these making up about one-third of the total. There is, of course, one well-known example of the eradication of a human RNA virus through human intervention, SARS-CoV, accompanying the even more impressive story of the eradication of smallpox, a DNA virus. However, there are many more examples of viruses that seem to have disappeared of their own accord, an unexpected observation worthy of careful consideration. There are several possibilities. First of all, rare infections, especially those with mild or common clinical presentations, may simply have been missed or no one has bothered to report them. Another possibility is that reports from earlier times are unreliable; for example, it is striking that no human cases of foot-and-mouth disease have been noticed since a handful of reports in the mid-1960s. But it seems likely that many of the missing viruses have indeed disappeared, at least temporarily, from humans, even if they are still present in nonhuman reservoirs. Some, of course, could reappear in humans at some point in the future: this has happened for the bat lyssaviruses, for example, and is a worrying possibility for SARS-CoV.

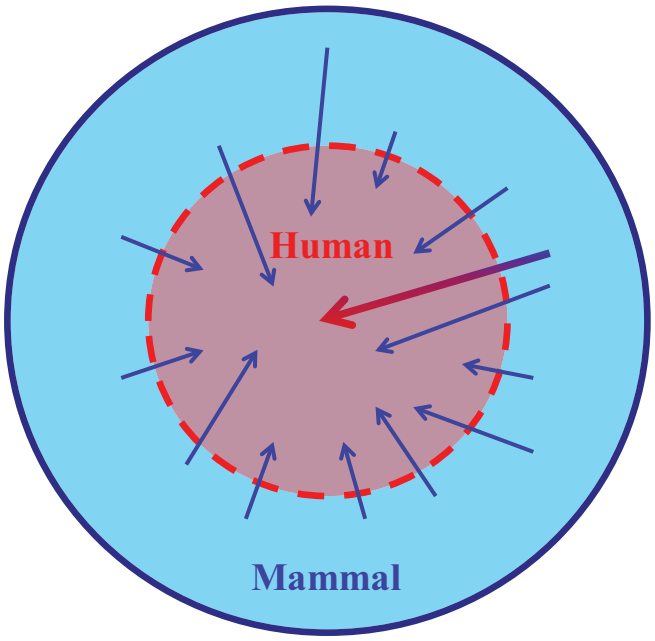
The implication of these missing viruses is that the extant diversity of human RNA viruses is perhaps closer to 100 species than the figure of 180 given earlier. The number of missing species corresponds, very roughly, to an average loss rate of 1 per year (8). Another way of expressing this is that there would have to be one new or rediscovered species of human RNA virus reported every year just to maintain the level of diversity that we are aware of at present.

A CONCEPTUAL MODEL

All of the above is consistent with the following conceptualization of the relationship between RNA viruses that can infect humans and those found in other kinds of

hosts, particularly other mammals. Rather than being distinct groups, viruses of humans and viruses of other mammals are readily interchanged over evolutionary time. Some of the viruses that cross the species barrier into humans persist and may become human-adapted viruses, though this seems to be a relatively rare occurrence. Many of the others remain as zoonoses, and yet others disappear again. The repertoire of human viruses is therefore not fixed but is dynamic, over time scales measured in decades (8). However, this process is far from random. Although humans share their RNA viruses with many different mammalian taxa, those from other primates appear most likely to be capable of spreading through human populations. Similarly, although almost every family of viruses found in mammals contains species found in humans, some virus families seem to be capable of, at best, limited spread in human populations. This conceptual model is illustrated diagrammatically in Fig. 3.

FIGURE 3 A schematic representation of the relationship between human viruses and viruses from other mammals. Human viruses are depicted as a subset of mammal viruses, only partially protected by a species barrier. There are frequent minor incursions of zoonotic viruses (small arrows), and many of these may not persist in human populations. Occasionally there may be a much more significant event (large arrow) whereby a mammal virus proves capable of establishing itself as a new human virus, perhaps involving adaptation to infect and transmit from humans. doi:10.1128/microbiolspec.OH-0001-2012.f3



SURVEILLANCE AND RISK ASSESSMENT

Our conceptual model has practical implications for both disease surveillance and risk assessment, especially in the context of newly emerging infectious diseases.

The importance of early detection of potential epidemics or pandemics cannot be overstressed, a point made by several major studies (2). The early detection through clinical surveillance of SARS, coupled with effective intervention based on case isolation and quarantine, prevented a potentially catastrophic pandemic (19). A matter of some debate is whether or not surveillance should be extended into the nonhuman reservoirs of infection from which novel human pathogens are most likely to emerge—a concept sometimes referred to as "getting ahead of the curve."

It helps, of course, if we know what we are looking for and where best to look for it (20). We currently have only the beginnings of answers to these questions. Viruses, especially respiratory viruses, are often picked out as the most obvious threat to global public health (2). New viruses are very likely to have a zoonotic origin, almost certainly acquired from mammals or birds. Emergence events are most likely to occur in regions—so-called hot spots—that combine high human population densities with high densities of domestic animals and/or a high diversity of wildlife (4). All of this information is useful but falls well short of a recipe for designing a feasible global surveillance system (20).

One strategy to increase the likelihood of early detection is to implement sentinel surveillance in people in close, high-risk contact with animal populations, such as bush meat hunters or slaughterhouse workers. In tandem with recent advances in the technologies available for virus detection—especially those based on high-throughput nucleic acid sequencing—such programs should at least improve our knowledge of the diversity of viruses "out there" that humans are exposed to, a process sometimes referred to as "chatter" (10). Pathogen discovery programs, particularly in understudied taxa such as wild rodents and bats (21), should also add greatly to our knowledge of potential threats to human health.

Once a novel or previously unknown virus is identified, it is obviously important to assess any potential risk to public health. Initial assessments are generally based on the kind of comparative biology approach discussed here. A recent example of this is Schmallenberg virus, a novel virus first detected in sheep and cattle in northern Europe in 2011. Schmallenberg is a member of *Orthobunyavirus*, a diverse genus of vector-borne bunyaviruses that are found in a variety of

hosts but especially in ungulates. Given these characteristics, and despite the fact that some distantly related orthobunyaviruses—notably Oropouche virus—do cause disease in and may even be transmitted by humans, Schmallenberg was provisionally designated low risk to humans and no human cases have yet been found (22). The even more recently reported Middle East respiratory syndrome (MERS) coronavirus (23) has rightly caused much more concern.

CONCLUDING REMARKS

Emerging diseases caused by RNA viruses are a One Health issue. There is a continuous interchange, over both epidemiological and evolutionary time scales, between viruses in humans and viruses in other animals that we cannot ignore. RNA viruses that pose serious threats to global public health have arisen repeatedly by jumping into humans from other animals. This has been going on for millennia and it continues today, as fast as ever and perhaps even faster. We have to anticipate that new viral threats will emerge in coming years or decades and we need to be prepared to rise to these new challenges as they appear.

It is worth pointing out that the first virus was discovered in nonhuman animals (foot-and-mouth disease virus at the very end of the 19th century) before they were identified in humans. The same is true (24) for several important kinds of viruses, such as retroviruses (and lentiviruses specifically), rotaviruses, papillomaviruses, and coronaviruses. A corollary of this is that veterinary rather than medical expertise may, at least initially, be our best source of knowledge about newly discovered viruses.

We have discussed the need for more effective surveillance for novel viruses but concluded that although attempts to characterize the kinds of viruses most likely to emerge are useful, precise prediction is not a realistic objective, for now at least. On the other hand, there could be considerable benefit from a better understanding of RNA virus diversity in the most important host species. At present we do not even have a complete inventory of the viruses in humans, and while we have some knowledge of the viruses in major livestock species, we know very little about the viruses present in wild mammals or birds. These gaps can and should be filled: we need to know what is out there now, and what might be waiting around the corner.

ACKNOWLEDGMENTS

We thank Conor O'Halloran for assistance with data collation. We are grateful to past and present members of Epigroup and numerous collaborators for many fruitful discussions. This work is part of the VIZIONS project funded by a Wellcome Trust Strategic Award.

REFERENCES

1. Taylor LH, Latham SM, Woolhouse MEJ. 2001. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci* **356:**983–989.

2. King DA, Peckham C, Waage JK, Brownlie J, Woolhouse MEJ. 2006. Epidemiology. Infectious diseases: preparing for the future. *Science* **313:**1392–1393.

3. Woolhouse MEJ, Scott FA, Hudson Z, Howey R, Chase-Topping M. 2012. Human viruses: discovery and emergence. *Philos Trans R Soc Lond B Biol Sci* **367:**2864–2871.

4. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008. Global trends in emerging infectious diseases. *Nature* **451:**990–993.

5. Sharp PM, Hahn BH. 2010. The evolution of HIV-1 and the origin of AIDS. *Philos Trans R Soc Lond B Biol Sci* **365:**2487–2494.

6. Epstein JH, Field HE, Luby S, Pulliam JR, Daszak P. 2006. Nipah virus: impact, origins, and causes of emergence. *Curr Infect Dis Rep* **8:**59–63.

7. King AM, Adams MJ, Carstens EB, Lefkowitz EJ (ed). 2012. *Virus Taxonomy: Ninth Report of the International Committee for the Taxonomy of Viruses*. Elsevier, Amsterdam, The Netherlands.

8. Woolhouse MEJ, Adair K. 2013. The diversity of human RNA viruses. *Future Virol* **8:**159–171.

9. Kitchen A, Shackelton LA, Holmes EC. 2011. Family level phylogenies reveal modes of macroevolution in RNA viruses. *Proc Natl Acad Sci USA* **108:**238–243.

10. Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature* **447:**279–283.

11. Woolhouse MEJ, Taylor LH, Haydon DT. 2001. Population biology of multihost pathogens. *Science* **292:**1109–1112.

12. Antia R, Regoes RR, Koella JC, Bergstrom CT. 2003. The role of evolution in the emergence of infectious diseases. *Nature* **426:**658–661.

13. Bae SE, Son HS. 2011. Classification of viral zoonosis through receptor pattern analysis. *BMC Bioinformatics* **12:**96. doi:10.1186/1471-2105-12-96.

14. Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. 2006. Host species barriers to influenza virus infections. *Science* **312:**394–397.

15. Blancou J, Aubert MF. 1997. [Transmission of rabies virus: importance of the species barrier]. *Bull Acad Natl Med* **181:**301–312. (In French.)

16. Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE. 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science* **329:**676–679.

17. Woolhouse MEJ, Adair K. 2013. Ecological and taxonomic variation among human RNA viruses. *J Clin Virol* [Epub ahead of print.] doi:10.1016/j.jcv.2013.02.019.

18. Ebert D, Bull J. 2008. The evolution and expression of virulence, p 153–167. *In* Stearns SC, Koella JC (ed), *Evolution in Health and Disease*, 2nd ed. Oxford University Press, Oxford, United Kingdom.

19. World Health Organization Multicentre Collaborative Network for Severe Acute Respiratory Syndrome Diagnosis. 2003. A multicentre collaboration to investigate the cause of severe acute respiratory syndrome. *Lancet* **361:**1730–1733.

20. Morse SS, Mazet JA, Woolhouse M, Parrish CR, Carroll D, Karesh WB, Zambrana-Torrelio C, Lipkin WI, Daszak P. 2012. Prediction and prevention of the next pandemic zoonosis. *Lancet* **380:**1956–1965.

21. Drexler JF, Corman VM, Müller MA, Maganga GD, Vallo P, Binger T, Gloza-Rausch F, Rasche A, Yordanov S, Seebens A, Oppong S, Adu Sarkodie Y, Pongombo C, Lukashev AN, Schmidt-Chanasit J,

Stöcker A, Carneiro AJ, Erbar S, Maisner A, Fronhoffs F, Buettner R, Kalko EK, Kruppa T, Franke CR, Kallies R, Yandoko ER, Herrler G, Reusken C, Hassanin A, Krüger DH, Matthee S, Ulrich RG, Leroy EM, Drosten C. 2012. Bats host major mammalian paramyxoviruses. *Nat Commun* **3:**796. doi:10.1038/ncomms1796.

22. Ducomble T, Wilking H, Stark K, Takla A, Askar M, Schaade L, Nitsche A, Kurth A. 2012. Lack of evidence for Schmallenberg virus infection in highly exposed persons, Germany, 2012. *Emerg Infect Dis* **18:**1333–1335.

23. Cotten M, Lam TT, Watson SJ, Palser AL, Petrova V, Grant P, Pybus OG, Rambaut A, Guan Y, Pillay D, Kellam P, Nastouli E. 2013. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. *Emerg Infect Dis* **19:**736–742.

24. Palmarini M. 2007. A veterinary twist on pathogen biology. *PLoS Pathog* **3:**e12. doi:10.1371/journal.ppat.0030012.

25. Woolhouse M, Antia R. 2008. Emergence of new infectious diseases, p 215–228. *In* Stearns SC, Koella JC (ed), *Evolution in Health and Disease*, 2nd ed. Oxford University Press, Oxford, United Kingdom.

Appendix F. Publication: Assessing the epidemic potential of RNA and DNA viruses

The following publication is a review that quantifies and critically assesses evidence of human-to-human transmissibility for each human RNA virus. The systematic literature review and assessment protocol described by this manuscript was conducted by myself as part of this thesis, and is further described by the chapters herein.

Citation: Woolhouse MEJ, Brierley L, McCaffery C, Lycett S. Assessing the epidemic potential of RNA and DNA viruses. Emerg Infect Dis. 2016;22(12):2037–44.

Assessing the Epidemic Potential of RNA and DNA Viruses

Mark E.J. Woolhouse, Liam Brierley, Chris McCaffery, Sam Lycett

Many new and emerging RNA and DNA viruses are zoonotic or have zoonotic origins in an animal reservoir that is usually mammalian and sometimes avian. Not all zoonotic viruses are transmissible (directly or by an arthropod vector) between human hosts. Virus genome sequence data provide the best evidence of transmission. Of human transmissible virus, 37 species have so far been restricted to self-limiting outbreaks. These viruses are priorities for surveillance because relatively minor changes in their epidemiologies can potentially lead to major changes in the threat they pose to public health. On the basis of comparisons across all recognized human viruses, we consider the characteristics of these priority viruses and assess the likelihood that they will further emerge in human populations. We also assess the likelihood that a virus that can infect humans but is not capable of transmission (directly or by a vector) between human hosts can acquire that capability.

A series of recent emerging infectious disease outbreaks, including the 2014 Ebola virus disease (EVD) epidemic in West Africa and the continuing Zika virus disease epidemic in the Americas, have underlined the need for better understanding of which kinds of pathogens are most likely to emerge and cause disease in human populations. Many, although not all, emerging infectious diseases are caused by viruses, and these frequently emerge from nonhuman host reservoirs (*1*–*3*). The enormous diversity (*4*) and high rates of evolution (*5*) of viral pathogens discourage attempts to predict with any precision which ones are most likely to emerge in humans. However, there is some consensus, at least in general terms, regarding the kinds of traits that are most essential in determining the capacity of a virus to infect, cause disease, and spread within human populations (Table 1). We focus on one of these traits, the capacity of a virus to spread from one human to another (by any transmission route other than deliberate laboratory exposure), a key determinant of the epidemic potential of a virus.

A theoretical framework for studying the dynamics of infectious disease outbreaks is well established (*6*). The capacity of an infectious disease to spread in a host population can be quantified in terms of its basic reproduction number, R_0. R_0 is defined as the average number of secondary cases generated by a single primary case in a large, previously unexposed host population, and its value tells us a great deal about the epidemiology of a pathogen. $R_0 = 0$ indicates no spread in that population; this value would apply to zoonotic infections that do not spread between humans. R_0 in the range $0<R_0\leq1$ indicates that chains of transmission are possible but that outbreaks will ultimately be self-limiting. $R_0>1$ indicates that major epidemics can occur or that the disease may become endemic in that host population. A higher value of R_0 also indicates that a greater reduction in transmission rates must be achieved to control an epidemic (*6*). R_0 values have been estimated for >60 common human pathogens (*7*), including human influenza A virus ($R_0\leq2$), measles virus ($R_0\leq18$), and dengue virus ($R_0\leq22$).

R_0 is determined by a combination of pathogen traits, such as its transmission biology, which is itself a complex interplay between the within-host dynamics of the pathogen and the host response to infection, and host traits, such as demography, behavior, genetics, and adaptive immunity. Consequently, for any given infectious disease, R_0 can vary between host species and between host populations. Infectious diseases with R_0 close to 1 are a particular concern because small changes in their epidemiologies can lead to major changes in the threat they pose to public health (*8*).

R_0 is closely related to another conceptual approach to disease emergence, the pathogen pyramid. There are different versions of this scheme (*3*,*9*). We consider a pyramid of 4 levels (Figure 1). Level 1 represents the background chatter of pathogens to which humans are continually or sporadically exposed but most of which are not capable of causing infection. Other levels can be considered in terms of the R_0 of the pathogen in humans: level 2 corresponds to $R_0 = 0$, level 3 to $0<R_0\leq1$, and level 4 to $R_0>1$.

Data and Analysis

Identifying and Characterizing Level 3 and 4 Viruses

We updated our previous systematic literature review (*10*) of the capacity of virus species to transmit between humans (i.e., level 3 and level 4 viruses; online Technical Appendix, http://wwwnc.cdc.gov/EID/article/22/12/16-0123- Techapp1.pdf). Such viruses are

Author affiliation: University of Edinburgh, Edinburgh, UK

DOI: http://dx.doi.org/10.3201/eid2212.160123

Table 1. Virus traits potentially relevant for capacity to emerge and cause disease in human populations*

Trait	Definition
Reservoir host relatedness	Viruses derived from specific host taxa (e.g., other primate species might be of increased concern)
Virus relatedness	Particular virus taxa might be predisposed to infect, cause disease, and transmit among humans
Virus host range	Viruses with a broad or narrow host range might be of greatest concern
Evolvability	Higher substitution rates might make it easier for some viruses to adapt to human hosts
Host restriction factors	Host factors, many still to be identified, are a barrier to viral infection and help determine which viruses can and cannot emerge
Transmission route	Certain transmission routes might predispose viruses to emerge in humans
Virulence	Certain virus or host factors might determine whether a virus causes mild or severe disease in humans
Host–virus coevolution	Lack of a shared evolutionary history might be associated with higher virulence

*Adapted from Morse et al. (*3*).

found in 25 of 29 families containing viruses that infect mammals or birds (discounting 2 reports of family *Nodaviridae* species in mammals/birds). The 4 exceptions comprise 2 families that have no known human-infective viruses (*Arteriviridae* and *Birnaviridae*) and 2 with species that have been reported in humans but only at level 2 (*Asfarviridae* and *Bornaviridae*).

A total of 22 of these families contain level 4 viruses with epidemic potential in humans (sometimes described as human-adapted viruses) (*11*). This finding indicates that this capability is widely distributed among virus taxa. The 3 families with level 3 viruses but no level 4 viruses are the *Arenaviridae*, *Bunyaviridae*, and *Rhabdoviridae*.

A list of 37 presumptive level 3 virus species is provided in Table 2. These species cover a wide taxonomic range and a variety of transmission routes, including vectorborne. Several level 3 viruses have historically been associated with sizeable outbreaks (>100 cases) in human populations: Bwamba, Oropouche, Lake Victoria Marburg, Sudan Ebola, and o'nyong-nyong viruses. For some other viruses, including Guanarito, Junin, lymphocytic choriomeningitis, and Sabia (all arenaviruses); simian virus 40; Titi monkey virus; and influenza A(H5N1) virus, human-to-human transmission is rare or merely suspected. In addition, several viruses are only known or believed to transmit between humans by iatrogenic routes or vertical transmission; this group (Table 2) might be regarded as unlikely epidemic threats.

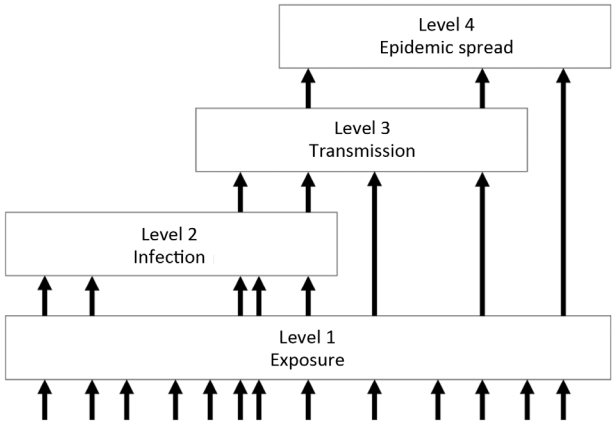


Figure 1. Pathogen pyramid for RNA and DNA viruses. Level 1 indicates viruses to which humans are exposed but which do not infect humans. Level 2 indicates viruses that can infect humans but are not transmitted from humans. Level 3 indicates viruses that can infect and be transmitted from humans but are restricted to self-limiting outbreaks. Level 4 indicates viruses that are capable of epidemic spread in human populations. Transitions between levels (indicated by arrows) correspond to different stages of virus emergence in human populations. Reprinted from Woolhouse et al. (*10*).

When a virus is transmitted by a vector, it can be particularly difficult to confirm or exclude the infectiousness of human cases, as with Semliki forest, Barmah forest, and Rift Valley fever viruses. Similarly, even when human–vector–human transmission is believed to occur, it is often difficult to quantify its contribution to a given outbreak, as with Venezuelan equine encephalitis virus.

Level 2 viruses are those that can infect humans (>100 species) but have never been reported to be transmitted by humans (*10*). In at least some instances, such as influenza A(H5N1) virus (*12*), this finding is attributable to tissue tropisms during human infection that are incompatible with onward transmission.

Shifts in pyramid level equate to shifts in the public health threat posed by a virus. We consider possible shifts in the following sections.

Level 1 to Levels 3 and 4

Virus species of mammalian and, more rarely, avian origin are sometimes observed to be transmissible between humans when first found in humans, which constitutes a jump from level 1 straight to level 3 or 4 (Figure 1), and events of this kind have been reported regularly. Recent examples that appear on the basis of available evidence to fit this model include severe acute respiratory syndrome coronavirus (first reported in humans in 2003), Bundibugyo Ebolavirus (2008), Lujo virus (2009), severe fever with thrombocytopenia syndrome virus (2011), and Middle East respiratory syndrome coronavirus (MERS-CoV) (2012).

We still have incomplete knowledge of the diversity of viruses that infect mammals and birds; the few hundred recognized species (*4*) surely represent only a small fraction of the total (*3*). Moreover, we have few predictors of potential human-to-human transmissibility. One possible indicator

Table 2. Viruses (n = 37) that are known or suspected of being transmissible (directly or indirectly) between humans but to date have been restricted to short transmission chains or self-limiting outbreaks

Genome, virus family	Virus name
Single-stranded RNA (ambisense)	
Arenaviruses	Guanarito, Junin, Lassa, Lujo, Machupo, Sabia, Dandenong,* lymphocytic choriomeningitis*
Bunyaviruses	Andes, Bwamba, Crimean-Congo hemorrhagic fever, Oropouche, Rift Valley, severe fever with thrombocytopenia syndrome
Single-stranded RNA (positive sense)	
Flaviviruses	Japanese encephalitis,* Usutu,* West Nile*
Coronaviruses	Middle East respiratory syndrome
Togaviruses	Barmah Forest, o'nyong-nyong, Ross River, Semliki Forest, Venezuelan equine encephalitis
Single-stranded RNA (negative sense)	
Filoviruses	Bundibugyo Ebola, Lake Victoria Marburg, Sudan Ebola
Paramyxoviruses	Nipah
Rhabdoviruses	Bas-Congo, rabies*
Double-stranded RNA	
Reoviruses	Nelson Bay, Colorado tick fever*
Double-stranded DNA	
Adenoviruses	Titi monkey
Herpesviruses	Macacine herpesvirus 1
Polyomaviruses	Simian virus 40
Poxviruses	Monkeypox, Orf, vaccinia

*Human transmission of these viruses is known only by iatrogenic or vertical routes.

is emergence from nonhuman primates, with suggestions that primate viruses are more likely to be able to, or to acquire the ability to, spread in human populations (*13,14*). However, emergence of human transmissible viruses from bat (e.g., severe acute respiratory syndrome coronavirus) or bird (e.g., influenza) reservoirs indicates that this trait is associated with a wide range of reservoirs.

Level 2 to Levels 3 and 4

The possibility that level 2 viruses might acquire the capacity to be transmitted between humans (i.e., move into level 3 or 4) is a major concern, especially in the context of influenza A(H5N1) virus and other avian influenza virus subtypes. However, there are few examples of this transition throughout the entire recorded history of human viruses going back to 1901. One possible example involves the simian immunodeficiency virus (SIV) and HIV. A SIV_{smm}-derived laboratory strain of SIV has been reported to infect humans, but without onward transmission (*15*). SIV_{smm} is related to HIV-2. SIV_{cpz}, which is related to HIV-1, has not been directly observed in humans. However, different HIV-1 lineages, independently derived from SIV_{cpz}, are variably transmissible in humans, and the pandemic HIV-1 M lineage was the only virus to overcome a key host restriction factor (human tetherin) (*16*). The only other examples of viruses newly transmitted between humans relate to rare instances of iatrogenic transmission (e.g., Colorado tick fever or rabies viruses).

Epidemiologic and phylogenetic considerations routinely inform our assessment of the likelihood of human-to-human transmission being observed in the future. For example, there is markedly less concern about rabies virus than about avian influenza virus, and we suggest 2 reasons for this observation. First, rabies virus has a much longer history of and a much higher incidence of human infection, but human-to-human transmission is extremely rare. Second, there is no evidence that other rhabdoviruses viruses (with the possible exception of Bas-Congo virus, which represents a novel genus) are transmissible in humans (or primates more generally).

Level 3 to Level 4

Level 3 viruses can also become level 4 viruses. We note that virus evolution is not (necessarily) required for R_0 to become >1 in human populations. Differences in host (or vector) behavior, ecology, or demography might be sufficient (*8*).

Instances of shifts from level 3 to level 4 in recent times have been infrequent. Three candidates are Ebolavirus, Zika virus, and chikungunya virus. However, although these viruses have caused epidemics of unprecedented size in humans populations in the past decade, the condition $R_0>1$ in human populations might had been previously met for all 3 viruses (*17–19*).

For Ebola virus, the epidemic in West Africa in 2014 constituted the first appearance of this virus in high-density, urban populations, which is expected to correspond to a higher value of R_0. The chikungunya virus epidemic in the Indian Ocean region in 2005 was associated with a vector species jump (from *Aedes aegypti* to *Ae. albopictus* mosquitoes) that has been linked to a mutation in the virus envelope 1 protein gene (*18*). The chikungunya virus epidemic in the Caribbean region in 2013 followed the first appearance of chikungunya in the Americas and infected

populations that had no history of exposure to the virus. The current Zika virus epidemic in South America appears to be another example of a transition from a level 3 to a level 4 arbovirus associated with geographic spread into areas with high densities of vectors (*19*). Occasional Zika virus transmission directly from infected humans to other humans are of considerable interest, but probably contribute little to R_0.

Chikungunya, Zika and the other level 4 arboviruses (yellow fever and dengue viruses) illustrate that, for arboviruses, a high potential for spread in human populations is linked to carriage by anthropophilic vector species, particularly mosquitoes of the genus *Aedes*. In contrast, no tick species are regarded as anthropophilic, and there are no level 4 and few level 3 tickborne arboviruses.

Epidemiologic Patterns

The preceding sections illustrate that identifying transitions of viruses between level 2 and level 3 or between level 3 and level 4 is not always straightforward. Standard epidemiologic theory can help clarify our expectations.

As we discussed, pyramid level is related to the basic reproduction number R_0 in human populations. In turn, the value of R_0 is indicative of expected outbreak dynamics. Some key results (Figure 2) are the probability that a single primary case will generate ≥1 secondary cases (for any value of R_0), the expected average size of an outbreak generated (over the range $0 \leq R_0 < 1$), and the probability that an epidemic will spread in the human population (for $R_0 > 1$). These results strictly apply to homogeneous infections in a homogeneous host population, although more general frameworks can accommodate host or pathogen heterogeneity (*20*–*22*). Nonetheless, the key predictions

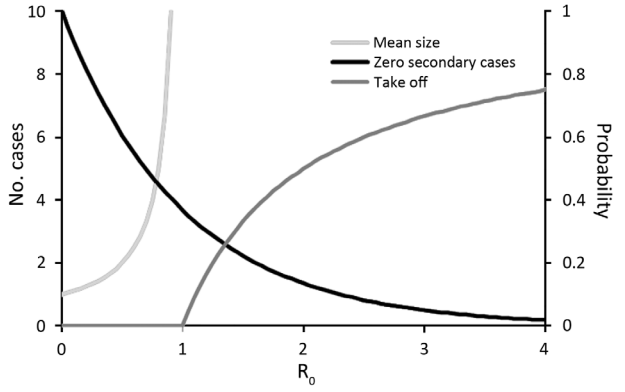


Figure 2. Expected outbreak dynamics for RNA and DNA viruses given a single primary case in a large, previously unexposed host population, as a function of the basic reproduction number R_0. Mean size of outbreak as total number of cases (N) is given by $N = 1/(1 - R_0)$ for $R_0 < 1$ (light gray line, left axis). Probability of 0 secondary cases (i.e., outbreak size N = 1) is given by $P_1 = \exp(-R_0)$ (black line, right axis). Probability of a major outbreak is given by $P_{takeoff} = 1 - 1/R_0$ for $R_0 > 1$ (dark gray line, right axis).

that secondary cases do not always occur even if $R_0 > 0$ and that major epidemics do not always occur even if $R_0 > 1$ are robust.

From an epidemiologic perspective, our confidence that a putatively level 2 virus is truly incapable of human-to-human transmission is thus a function of the number of index cases observed. The transition between level 3 and level 4 can be studied in terms of the expected distribution of outbreak sizes (*23*). In the range $0 \leq R_0 < 1$, an overdispersed distribution of outbreak sizes is expected: most outbreaks are small (often just single cases) with a long tail of larger outbreaks. This pattern has been reported for a range of emerging viral diseases (Figure 3). As the critical threshold $R_0 = 1$ is approached, this value is signaled in the outbreak size distribution (Figure 3). This framework has been used successfully to monitor the epidemiology of measles virus in the United Kingdom after a decrease in childhood vaccination rates in the late 1990s and indicated the approach to the critical threshold that corresponded to loss of herd immunity (*23*).

Outbreak size distribution analysis has been applied to human case data for Andes virus (*24*), monkeypox virus (*20*), and MERS-CoV (*25*) (Figure 3). For EVD up to 2013, data are clearly inconsistent with theoretical expectation for $R_0 < 1$ (Figure 3), which suggests that large numbers of small outbreaks have remained undetected or that R_0 was already >1 in at least some settings. Either way, $R_0 \approx 1$ for EVD in humans implies that small differences in the biology or epidemiology of the virus would lead to large changes in scale of outbreaks (*8*), which could make events such as the EVD epidemic in 2014, if not predictable, then much less unexpected.

Evolution

Changes in pyramid level might be mediated by virus evolution or changes in virus ecology (*28*). A major issue is whether the capacity of a virus to spread in human populations arises as a result of adaptation (evolution of transmissibility that occurs during human infection) or preadaptation (genetic variation within nonhuman reservoirs that predisposes a virus not only to infect humans but also transmit between humans, noting that RNA viruses often show high levels of genetic variation such that they are sometimes described as quasi-species [*29*]). These alternatives have been characterized as tailor-made and off-the-shelf, respectively (*28*). The first alternative implies a progression from no or low transmissibility between humans to moderate or high transmissibility. The second alternative implies moderate or high transmissibility at first infection of humans.

We consider that our survey of documented changes of pyramid level is most consistent with the off-the-shelf model of virus emergence. In particular, we can find no

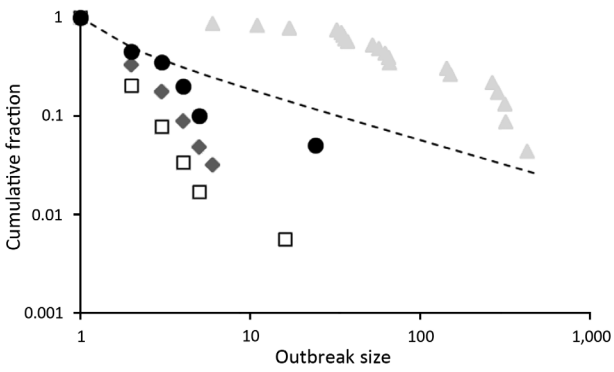


Figure 3. Distribution of outbreak sizes for RNA and DNA viruses as plots of outbreak size x (horizontal axis) versus fraction of outbreaks of size $\geq x$ (vertical axis), both on logarithmic scales. Data are shown for 4 infectious diseases. Squares indicates Andes virus disease in South America (*24*); diamonds indicate monkeypox in Africa (*26*); circles indicate Middle East respiratory syndrome in the Middle East (*25*); and triangles indicate filovirus (all species) diseases in Africa before 2013 (*27*). For comparison, expected values for the case $R_0 = 1$, obtained from the expression for the probability of an outbreak of size $\geq x$, $P(x) = \Gamma(x - ½)/\sqrt{\pi}\Gamma(x)$, are also shown (dashed line). Data for filoviruses are not consistent with expectation for $R_0<1$.

convincing examples of level 2 viruses becoming level 3 or 4 viruses, which suggests that, if this happens at all, it typically happens sufficiently rapidly (i.e., requires a sufficiently small number of introductions) that we fail to observe the level 2 phase. In contrast, we regularly observe viruses at levels 3 or 4 the first time they are detected in human populations.

Nonetheless, the possibility of virus evolution of transmissibility in a new host has been demonstrated experimentally for influenza A(H5N1) virus in ferrets (*30*). A theoretical study (*31*) suggested that the fact that this virus subtype has been circulating widely in poultry populations, with frequent human exposure and sporadic human infection for almost 20 years, provides little or no reassurance about its future evolutionary trajectory.

HIV lineages show clear evidence of adaptation to humans (*16*), but as discussed earlier, it is not clear whether the SIV lineages that gave rise to HIV-1 or HIV-2 were capable of transmission between human hosts. We speculate that extended infection times make tailor-made emergence more likely for retroviruses.

Transmission

Demonstrating that an infected human has the potential to transmit the infection to another human is not always straightforward. High virus titers in body secretions and excretions, blood, or skin are considered indicative. Case clusters are suggestive, but if persons occupy the same environment (e.g., household), then it might be difficult to rule out common exposure. Case clusters must be epidemiologically plausible (i.e., delimited in space and time in a manner consistent with the known or assumed epidemiology of the virus). Genotyping techniques are useful tools for confirming a cluster but do not resolve the source of infection.

For several of the viruses we studied (e.g., Bas-Congo, Lujo, Nelson Bay, and severe fever with thrombocytopenia syndrome viruses) (Table 2), the evidence for human-to-human transmission is best regarded as tentative, particularly where putative clusters were small. Such assessments can be even more difficult for vectorborne viruses. In many situations, the best evidence for the human-to-human transmission will come from analysis of virus genome sequences.

Phylogenetic Analysis

One approach to resolving the question of human-to-human transmission is analysis of nucleotide sequence data, sometimes referred to as forensic phylogenetics. Nucleotide substitution rates in fast-evolving RNA viruses, such as MERS-CoV and Ebola virus, are $\approx 1–5 \times 10^{-3}$/site/year (*32,33*), making it possible to use sequences isolated from different hosts at different times to estimate time-resolved phylogenetic trees. Estimates of the transmission chain from temporal sequence data can be improved by incorporating additional information on the date of onset of individual cases, duration of latent and infectious periods, and overall prevalence (*34*).

We provide some example phylogenetic trees generated from simulated epidemics (Figure 4). In an epidemic in an animal reservoir with occasional transmission to humans (Figure 4, panel A), for each human sequence, the most closely related next sequence is of animal origin. Clusters of closely related human sequences are shown, and the distribution of the expected cluster sizes is a function of R_0 (Figure 4, panels B, C) (*35*).

In an outbreak, it might be difficult to find and sample the putative source animal cases. However, estimating the time to most recent common ancestor (TMRCA) of the human cases will indicate how long the infection has been spreading. For sporadic zoonoses (Figure 4, panel A), most transmission has occurred unobserved in the animal reservoir, and the TMRCA of pairs of human cases will be long because these sequences are not closely related. For outbreaks involving human-to-human transmission (Figure 4, panels B, C), the TMRCA of the cluster of human cases will be closer to the date of the first human infection (whether sampled or not) and provides the estimated date of the zoonotic event.

Use of sequence data to distinguish between multiple instances of human infection from a common animal source and human-to-human transmission in the early stages of an

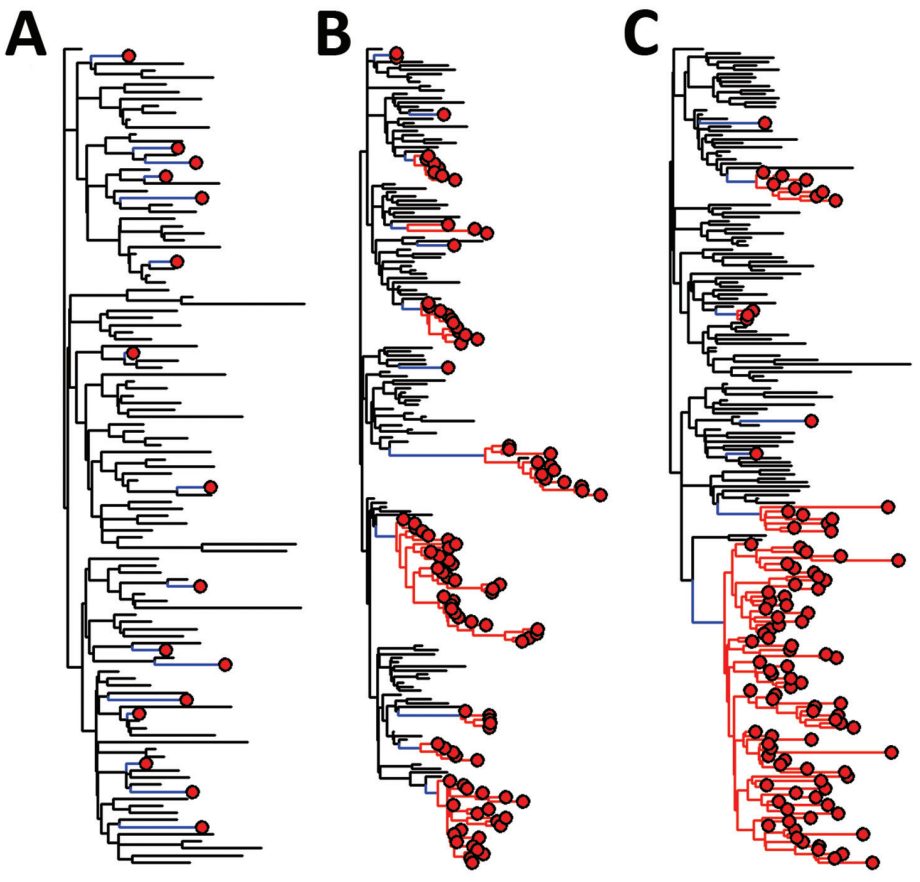


Figure 4. Phylogenetic trees for simulated emerging infectious disease outbreaks caused by RNA and DNA viruses in a mixed population of 1,000 human and 5,000 nonhuman hosts. Trees were constructed by using a standard susceptible–infected–removed model (*6*). For each of 3 infection scenarios in nonhuman hosts (black lines), rare zoonotic transmission events (blue lines), human-to-human transmission (red lines), and human cases (red circles) are indicated. For the nonhuman population $R_0 = 2$ throughout. Transmissibility within the human populations varies from A) spillover: no human–human transmission ($R_0 = 0$); B) limited human–human transmission with $R_0 = 1$; and C) epidemic spread within humans ($R_0 > 1$). A maximum of 100 infections are randomly sampled from each population in each simulated outbreak.

outbreak is extremely challenging because of short timescales, and involvement of few mutations. However, genetic differences and phylogenetic evidence show that at least 2 of the first 3 reported cases of influenza A (H7N9) virus infection in humans were believed to originate from distinct domestic avian sources (*36*).

Further sequencing of avian samples implied that a low-pathogenicity influenza A(H7N9) virus strain had been spreading in domestic birds for ≈1 year before sporadic cases were detected in humans (*37*). Similarly, detection of genetically distant lineages of MERS-CoV, which persisted for only a few months each, suggest multiple introductions from an animal reservoir and only limited human-to-human transmission to date (*32*). In contrast, the influenza A(H1N1) pandemic in 2009 and the EVD epidemic in West Africa in 2014 were believed to be the results of single zoonotic events, followed by sustained human-to-human transmission (*33*), as shown by a single rapidly expanding lineage.

Conclusions

Our survey of the capacity of RNA and DNA virus infections to be transmitted, directly or indirectly, between humans leads to several conclusions and practical suggestions for improving surveillance of emerging infectious diseases and targeting efforts to identify future public health threats. In support of these conclusions, the World Health Organization recently published list of priority emerging infectious diseases and corresponding viruses (*38*) included 6 of the viruses in Table 2.

A major observation is that the taxonomic diversity of viruses that are possible threats to public health is wide, but bounded. Most human infective viruses are closely related to viruses of other mammals and some to viruses of birds. There are no indications that humans acquire new viruses from any other source. However, diversification within human populations occurs and is a prominent feature of some DNA virus taxa (e.g., family *Papillomaviridae*) (*4*).

In general, however, our knowledge of origins of human viruses is still incomplete. Although the origins of HIV-1 have been extensively investigated (*16*), for most other viruses, even level 4 viruses, little or no research has occurred. An origins initiative (*9*) would help establish the routes into human populations that have been used by other viruses.

Transmissibility within human populations is a key determinant of epidemic potential. Many viruses that can infect humans are not capable of being transmitted by humans;

most human transmissible viruses that emerge already have that capability at first human infection or acquire it relatively rapidly. If transmission from humans would require a change in a phylogenetically conserved trait, such as tissue tropism or transmission route (*4*), then such viral paradigm shifts will probably be extremely rare (*39*).

Even when a virus is capable of transmission between humans, the critical threshold $R_0>1$ is not always achieved. However, because changes in virus traits or host population characteristics can influence R_0, level 3 viruses (Table 2) are of special interest from a public health perspective, and of special concern when, like MERS-CoV, they also cause severe illness. Demonstrating human transmissibility is often difficult, but essential. The best evidence is likely to come from virus genome sequencing studies. These studies should be a public health priority (*40*).

We currently have few clues to help us predict which mammalian or avian viruses might pose a threat to humans and, especially, which might be transmissible between humans. One argument in favor of experimental studies of these traits, including controversial gain of function experiments (*30*), is that they could help guide molecular surveillance for high-risk virus lineages in nonhuman reservoirs.

The first line of defense against emerging viruses is effective surveillance (*40*). A better understanding of which kinds of viruses in which circumstances pose the greatest risk to human health would enable evidence-based targeting of surveillance efforts, which would reduce costs and increase probable effectiveness of this endeavor.

Acknowledgments

We thank David McCulloch, Cristina Moreno, and Matthew Hall for assistance during this study.

This study was supported by the Wellcome Trust Vietnam Initiative on Zoonotic Infections Project grant 093724/B/10/Z to M.E.J.W. and European Union 2020 grant COMPARE #643476. L.B. was supported by a Natural Environment Research Council PhD studentship, and S.L. was supported by a University of Edinburgh Chancellor's Fellowship and the Biotechnology and Biological Sciences Research Council Strategic Programme grant BB/J004227/1.

Dr. Woolhouse is professor of Infectious Disease Epidemiology at the Centre for Immunity, Infection and Evolution at the University of Edinburgh, Edinburgh, UK. His primary research interests are pathogen emergence and antimicrobial drug resistance.

References

1. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci. 2001;356:983–9. http://dx.doi.org/10.1098/rstb.2001.0888
2. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451:990–3. http://dx.doi.org/10.1038/nature06536
3. Morse SS, Mazet JA, Woolhouse M, Parrish CR, Carroll D, Karesh WB, et al. Prediction and prevention of the next pandemic zoonosis. Lancet. 2012;380:1956–65. http://dx.doi.org/10.1016/S0140-6736(12)61684-5
4. King AM, Lefkowitz E, Adams MJ, Carstens EB. Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses. Amsterdam: Elsevier; 2012.
5. Woolhouse ME, Haydon DT, Antia R. Emerging pathogens: the epidemiology and evolution of species jumps. Trends Ecol Evol. 2005;20:238–44. http://dx.doi.org/10.1016/j.tree.2005.02.009
6. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. New York: Oxford University Press; 1991.
7. Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, et al. Global mapping of infectious disease. Philos Trans R Soc Lond B Biol Sci. 2013;368:20120250. http://dx.doi.org/10.1098/rstb.2012.0250
8. Woolhouse ME. Population biology of emerging and re-emerging pathogens. Trends Microbiol. 2002;10(Suppl):S3–7. http://dx.doi.org/10.1016/S0966-842X(02)02428-9
9. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447:279–83. http://dx.doi.org/10.1038/nature05775
10. Woolhouse ME, Adair K, Brierley L. RNA viruses: a case study of the biology of emerging infectious diseases. Microbiol Spectr. 2013;1:OH-0001–2012. http://dx.doi.org/10.1128/microbiolspec.OH-0001-2012
11. Woolhouse ME, Adair K. The diversity of human RNA viruses. Future Virology. 2013;8:159–71 http://dx.doi.org/10.2217/fvl.12.129.
12. Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. Host species barriers to influenza virus infections. Science. 2006;312:394–7. http://dx.doi.org/10.1126/science.1122818
13. Davies TJ, Pedersen AB. Phylogeny and geography predict pathogen community similarity in wild primates and humans. Proc Biol Sci. 2008;275:1695–701. http://dx.doi.org/10.1098/rspb.2008.0284
14. Cooper N, Nunn CL. 2013 Identifying future zoonotic disease threats. Evol Med Public Health. 2013;1:27–36. http://dx.doi.org/10.1093/emph/eot001
15. Khabbaz RF, Heneine W, George JR, Parekh B, Rowe T, Woods T, et al. Brief report: infection of a laboratory worker with simian immunodeficiency virus. N Engl J Med. 1994;330:172–7. http://dx.doi.org/10.1056/NEJM199401203300304
16. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med. 2011;1:a006841. http://dx.doi.org/10.1101/cshperspect.a006841
17. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. J Theor Biol. 2004;229:119–26. http://dx.doi.org/10.1016/j.jtbi.2004.03.006
18. Schwartz O, Albert ML. Biology and pathogenesis of chikungunya virus. Nat Rev Microbiol. 2010;8:491–500. http://dx.doi.org/10.1038/nrmicro2368
19. Fauci AS, Morens DM. Zika virus in the Americas: yet another arbovirus threat. N Engl J Med. 2016;374:601–4. http://dx.doi.org/10.1056/NEJMp1600297
20. Blumberg S, Lloyd-Smith JO. Inference of $R_{(0)}$ and transmission heterogeneity from the size distribution of stuttering chains. PLOS Comput Biol. 2013;9:e1002993. http://dx.doi.org/10.1371/journal.pcbi.1002993
21. Lord CC, Barnard B, Day K, Hargrove JW, McNamara JJ, Paul RE, et al. Aggregation and distribution of strains in

microparasites. Philos Trans R Soc Lond B Biol Sci. 1999; 354:799–807. http://dx.doi.org/10.1098/rstb.1999.0432
22. Yates A, Antia R, Regoes RR. How do pathogen evolution and host heterogeneity interact in disease emergence? Proc Biol Sci. 2006;273:3075–83. http://dx.doi.org/10.1098/rspb.2006.3681
23. Jansen VA, Stollenwerk N, Jensen HJ, Ramsay ME, Edmunds WJ, Rhodes CJ. Measles outbreaks in a population with declining vaccine uptake. Science. 2003;301:804. http://dx.doi.org/10.1126/science.1086726
24. Woolhouse ME, Gaunt E. Ecological origins of novel human pathogens. In: Relman DA, Hamburg MA, Choffnes ER, Mack A, editors. Microbial evolution and co-adaptation, Washington (DC): National Academies Press; 2009. p. 208–29.
25. Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. Lancet. 2013;382:694–9. http://dx.doi.org/10.1016/S0140-6736(13)61492-0
26. Fine PE, Jezek Z, Grab B, Dixon H. The transmission potential of monkeypox virus in human populations. Int J Epidemiol. 1988;17:643–50. http://dx.doi.org/10.1093/ije/17.3.643
27. Centers for Disease Control and Prevention. Outbreaks chronology: Ebola virus disease [cited 2015 Feb 1]. http://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html#modalIdString_outbreaks
28. Woolhouse M, Antia R. Emergence of new infectious diseases. In: Stearns SC, Koella JK, editors. Evolution in health and disease. 2nd ed. Oxford (UK): Oxford University Press; 2008. p. 215–28.
29. Domingo E, Martínez-Salas E, Sobrino F, de la Torre JC, Portela A, Ortín J, et al. The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance: a review. Gene. 1985;40:1–8. http://dx.doi.org/10.1016/0378-1119(85)90017-4
30. Herfst S, Schrauwen EJ, Linster M, Chutinimitkul S, de Wit E, Munster VJ, et al. Airborne transmission of influenza A/H5N1 virus between ferrets. Science. 2012;336:1534–41. http://dx.doi.org/10.1126/science.1213362
31. Arinaminpathy N, McLean AR. Evolution and emergence of novel human infections. Proc Biol Sci. 2009;276:3937–43. http://dx.doi.org/10.1098/rspb.2009.1059
32. Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. MBio. 2014;5:e01062–13. http://dx.doi.org/10.1128/mBio.01062-13
33. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369–72. http://dx.doi.org/10.1126/science.1259657
34. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLOS Comput Biol. 2014;10:e1003457. http://dx.doi.org/10.1371/journal.pcbi.1003457
35. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. Genetics. 2009;183:1421–30. http://dx.doi.org/10.1534/genetics.109.106021
36. Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W, et al. Human infection with a novel avian-origin influenza A (H7N9) virus. N Engl J Med. 2013;368:1888–97. http://dx.doi.org/10.1056/NEJMoa1304459
37. Lam TT, Wang J, Shen Y, Zhou B, Duan L, Cheung CL, et al. The genesis and source of the H7N9 influenza viruses causing human infections in China. Nature. 2013;502:241–4. http://dx.doi.org/10.1038/nature12515
38. World Health Organization. Blueprint for R&D preparedness and response to public health emergencies due to highly infectious pathogens, 2015 [cited 2016 Aug 21]. http://www.who.int/csr/research-and-development/meeting-report-prioritization.pdf?ua=1
39. Belshaw R, Gardner A, Rambaut A, Pybus OG. Pacing a small cage: mutation and RNA viruses. Trends Ecol Evol. 2008;23:188–93. http://dx.doi.org/10.1016/j.tree.2007.11.010
40. Woolhouse ME, Rambaut A, Kellam P. Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. Sci Transl Med. 2015;7:307rv5. http://dx.doi.org/10.1126/scitranslmed.aab0191

Address for correspondence: Mark E.J. Woolhouse, Centre for Immunity, Infection, and Evolution, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Rd, Edinburgh EH9 3FL, UK; email: mark.woolhouse@ed.ac.uk