Characterization of Emergent and Emerging Human-Infective RNA Viruses



Jeremy Ratcliff

Exeter College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Acknowledgements

I owe an immense debt of gratitude to many individuals without whom I would have been unable to produce this thesis.

Principal academic supervision was provided by Prof. Peter Simmonds, Prof. Adrian Smith, and Dr. Heli Harvala. Informal academic supervision at various stages of the project was provided by Dr. Tanya Golubchik, Dr. Craig Thompson, Dr. Valerie Odon, and Dr. Dung Nguyen. I also want to thank Prof. Nicole Zitzmann and Prof. Jan Rehwinkel for serving as transfer of status and confirmation of status assessors and providing strategic direction in the early stages of this thesis research. Thank you to members of the Simmonds Lab (Dr. Sarah Williams, Dr. Ahmad Ibrahim, Dr. Lisa Loew, Dr. Everlyn Kamau, Grace Hood, Dr. Visa Nurmi, Chanice Knight, Dr. Julien Andreani) for their general support and advice.

Thank you to all of the members of my immediate (and distant) family who supported me throughout this degree program - sorry for all of the times I tried explaining conceptual models of viral evolution over FaceTime. Members of the Medawar Building (Peny Zacharopoulou, Nicholas Lim, Oliver Sampson, Jai Bolton, Dr. Lucia Parolini, Danny Hambidge, Daniel Coneyworth, Dr. Matthew Pace, Dr. Timothy Tipoe), OUAFC teammates (particularly Kaiyang Song, Benjamin Putland, Dr. David Burton, Dr. Raphael Reinbold, Harry Lamaison, Dr. Laurence Wroe, Luke Smith, and Liam Elbourne), and "Marshall friends and partners" (Kevin Morris, Laura Hallas, Katie O'Nell, Andrew Brown, Maddy Dietrich, Christopher Crum, Hannah Stewart, Dr. Vaibhav Mohanty, Emily Steffke, Tommy Brettschneider, Dr. Julius Bright-Ross, Taylor Vandick, Josh Lappen, and Lyndon Henley Hanrahan) were essential in keeping me both sane and out of the laboratory (on occasion). I have been incredibly fortunate to have Cecilia Jay as a loving partner; I am grateful for her every day and especially for her support while I was writing this thesis.

My time at Oxford was only possible due to the foundations built as an undergraduate at Johns Hopkins University, and I am very grateful for the friends and mentors there who helped me along my academic journey. In no particular order, I want to provide particular thanks to Dr. Jeffrey Freeman, Dr. Jeannette Miller, Prof. Emily Fisher, Dr. David Graham, Prof. Eric Rice, Prof. Trina Schroer, Dr. Katherine Henry, Dr. Cassidy Rist, Kelsie Gory, Craig Appleby, Sebastian Salathe, Matthew Williams, Kayla Blair, Kristen Hori, Arden Chew, Maxwell Song,

Elena Andreou, Nigel Joseph, Achim Younker, Michael Swiercz, Cole Rosenberger, Nikki Knudsen, Hayden Helm, Ellie Clawson, Alex Doran, and Emily Kwan.

Large portions of the technical work of this thesis relied on the work of the clinical staff supporting the REMAP-CAP Immunoglobulin Domain and the COV003 vaccine clinical trial and the participants taking part in the research. Full lists of these individuals are included in the supplementary material of PMID: 35321556. I am particularly grateful to the team of Professor Manu Shankar-Hari at King's College London (Matthew Fish, Jennifer Rynne, and Aislinn Jennings) for cataloging the REMAP-CAP samples and handling sample transportation and for Dr. Sagida Bibi from the Oxford Vaccine Group for coordinating sample transportation for COV003.

Lastly, I thank the Marshall Aid Commemoration Commission and the Clarendon Fund for financial support I received during my DPhil. The research described in this thesis relied on the financial support of a large number of funding bodies listed here: the NIH, the European Commission, the European Union Platform for European Preparedness Against (Re-) Emerging Epidemics (PREPARE) consortium, the European Union Horizon 2020 Research and Innovation Programme Rapid European COVID-19 Emergency Research response (RECOVER) consortium, UKRI, NIHR, CEPI, the Wellcome Trust, the Lemann Foundation, Rede D'OR, the Brava and Telles Foundation, NIHR Oxford Biomedical Research Centre, NIHR Imperial Biomedical Research Centre, and AstraZeneca.

Contributions to this thesis

All work presented in this thesis was principally supervised by Prof. Peter Simmonds and co-supervised by Prof. Adrian Smith. Substantial supervision and direction were also often provided by Dr. Heli Harvala and Dr. Tanya Golubchik.

Chapter 2

Large portions of the content summarizing the evolution and function of the APOBEC family of proteins have been published previously as a review article:

- Ratcliff J, Simmonds P. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. Virology. 2021 Apr;556:62-72. doi: 10.1016/j.virol.2020.12.018. Epub 2021 Jan 7. PMID: 33545556; PMCID: PMC7831814.

The review article has, in essence, been updated to include all evidence in favor of the hypermutation phenotype that has been published since that review article was accepted.

Chapters 3, 4, 5, 6, and 7

VeSeq on the SARS-CoV-2 clinical samples was performed by George Macintyre-Cockett and the viromics team at the Wellcome Centre for Human Genetics. Dr. Tanya Golubchik processed the raw reads from the Illumina system and produced consensus sequences and read tables for each patient sample. The method used for sequencing (veSeq) was designed by Dr. David Bonsall and the AMPHEUS project.

The qualitative ELISA for REMAP-CAP samples was performed by the team of Prof. David Roberts at the Radcliffe Department of Medicine, University of Oxford.

For both REMAP-CAP and COV003, patient recruitment and sample acquisition was performed by the clinical teams for each trial. Full lists of these individuals are included in the supplementary material of PMID: 35321556.

Dr. Dung Nguyen and I extracted RNA from the REMAP-CAP baseline samples in tandem (over many long nights and weekends, for which I am extraordinarily grateful). Ms. Cecilia Jay assisted in extracting samples from the intensively sampled cohort and Dr. Sarah Williams assisted in running the PCR-RFLP.

Portions of material presented in these chapters have been published in several peer-reviewed articles and presented at scientific conferences including:

- Ratcliff, J et al. Virological Characterization of Critically Ill Patients With COVID-19 in the United Kingdom: Interactions of Viral Load, Antibody Status, and B.1.1.7 Infection. J Infect Dis. 2021 Aug 16;224(4):595-605. doi: 10.1093/infdis/jiab283. PMID: 34031695; PMCID: PMC8241475.
- Clemens, S.A.C., Folegatti, P.M., Emary, K.R.W. et al. Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 lineages circulating in Brazil. Nat Commun 12, 5861 (2021). https://doi.org/10.1038/s41467-021-25982-w. PMID: 34615860 PMCID: PMC8494913
- Ratcliff, J., Nguyen, D., Simmonds, P., and Harvala, H. Heterogeneous virological and serological characteristics of severely ill COVID-19 patients in the United Kingdom: considerations for therapeutic clinical trial design and subgroup analysis. European Society for Clinical Virology 2021 (Virtual). 2021 Sep 17.
- Ratcliff, J. et al. Highly Sensitive Lineage Discrimination of SARS-CoV-2 Variants through Allele-Specific Probe PCR. Journal of Clinical Microbiology, e02283-21 (2022). PMID: 35321556 PMCID: PMC9020347

Chapter 8

Several items presented in this chapter were formulated through discussion with Mr. Spyros Lytras, a PhD student of the University of Glasgow, and Prof. Jelke Fros of Wageningen University.

Material presented in this chapter has been presented at scientific conferences including:

- Ratcliff, J. and Simmonds, P. Dinucleotide Suppression Predicts Transmissibility of Arthropod-Borne Viruses in Humans. Wellcome Connecting Science: Virus Genomics and Evolution (Virtual). 2021 Sep 14.

Abstract

Improved understanding of the early adaptation of zoonotic, human-infective RNA viruses and the genetic factors leading to spillover will inform public health approaches attempting to limit the morbidity and mortality associated with these events. In this thesis, I use clinical microbiological methods to characterize the intra-host replicative and evolutionary dynamics of SARS-CoV-2 and bioinformatic approaches to investigate genetic factors associated with spillover of arboviruses.

Chapter one broadly reviews the field of virology.

Chapter two reviews the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide (APOBEC) family of proteins, focusing on their emergence, diversification, and function, and summarizes the bioinformatic and functional evidence of their role in SARS-CoV-2 evolution.

Chapter three details the methods underlying the subsequent results chapters.

Chapter four describes the development and validation of single nucleotide polymorphism-based genotyping assays for identifying variants of concern in COVID-19 clinical samples. Their performance is compared to next-generation sequencing, demonstrating the superiority of the genotyping assays at a wide range of viral loads.

Chapter five uses amplification data generated from two COVID-19 clinical trials to explore factors associated with viral load and within-patient viral load dynamics, uses a novel polymerase chain reaction approach to quantify subgenomic RNA transcripts, and investigates the emergence and spread of the Alpha variant in the United Kingdom and the Zeta, Gamma, and Delta variants in Brazil.

Chapter six investigates the interaction between SARS-CoV-2 virological features and COVID-19 clinical outcomes leveraging patients admitted to intensive care within the REMAP-CAP convalescent plasma clinical trial. The null results in this chapter support severe COVID-19 pathology being primarily immune-mediated.

Chapter seven quantitatively tests the hypothesis that convalescent plasma treatment was responsible for the emergence of SARS-CoV-2 variants of concern using an intensively sampled cohort within the REMAP-CAP convalescent plasma trial. Using a variety of approaches, no evidence in support of convalescent plasma therapy as a general driver of immune-escape substitutions was found.

Chapter eight probes the genetic factors associated with human-human transmissibility of arboviruses, finding strong evidence for the role of CpG suppression in the host optimization of flaviviruses. The strong genetic barriers to spillover enable more focused surveillance of this informal classification of viruses.

Contents

List of Figures **ix**

List of Tables **xi**

List of Abbreviations **xii**

1 Introduction **1**
- 1.1 Introduction . . . 1
- 1.2 Virus Classification and Relevant Viruses . . . 5
- 1.3 Virus Evolution . . . 8
- 1.4 Zoonotic Virus Emergence . . . 14
- 1.5 Aims of this Thesis . . . 15

2 A Review of Potential APOBEC-mediated Hypermutation of SARS-CoV-2 **16**
- 2.1 Introduction . . . 17
- 2.2 Deaminase-Dependent Restriction of Viruses and Mobile Genetic Elements . . . 22
- 2.3 APOBEC Origins, Diversification, and Evolution . . . 26
- 2.4 Bioinformatic Evidence for RNA Editing in SARS-CoV-2 . . . 31
- 2.5 Functional Evidence for a Role for APOBEC in Editing and Antiviral Restriction of Coronaviruses . . . 34
- 2.6 Evolutionary Consequences of C>U Mutational Bias in Coronaviruses 35
- 2.7 A Perspective on Future Functional Investigations of APOBEC-Mediated Editing of SARS-CoV-2 and Other RNA Viruses . . . 38

3 Methods **40**
- 3.1 Methods used in Chapters Four, Five, Six, and Seven . . . 41
- 3.2 Methods used in Chapters Four, Five, and Six . . . 45
- 3.3 Methods used in Chapter Four . . . 48
- 3.4 Methods used in Chapter Five . . . 49
- 3.5 Methods used in Chapter Six . . . 51
- 3.6 Methods used in Chapter Seven . . . 52
- 3.7 Methods used in Chapter Eight . . . 55

4 Rapid Methods for Detecting SARS-CoV-2 Mutations and Typing Variants 59
4.1 Introduction . . . 60
4.2 Results . . . 64
4.3 Discussion . . . 80

5 Viral Load Diversity and Dynamics in Patients Infected with SARS-CoV-2 89
5.1 Introduction . . . 90
5.2 Results . . . 93
5.3 Discussion . . . 111

6 Interactions of SARS-CoV-2 Virological Characteristics with COVID-19 Severity and Outcomes 118
6.1 Introduction . . . 119
6.2 Results . . . 120
6.3 Discussion . . . 127

7 SARS-CoV-2 Within-Host Evolution in Convalescent Plasma Treated Patients in the REMAP-CAP Immunological Domain 130
7.1 Introduction . . . 131
7.2 Results . . . 134
7.3 Discussion . . . 146

8 Dinucleotide Suppression and Host Species Optimization in Arboviruses 150
8.1 Introduction . . . 151
8.2 Results . . . 155
8.3 Discussion . . . 169

9 General Discussion 176

10 Appendix 180
10.1 Supplementary Tables . . . 180
10.5 References . . . 192

List of Figures

2.1 Emergence of APOBEC proteins 28
2.2 Expansion of A3 Paralogs in Placental Mammals 29

4.1 Theoretical Binding Dynamics of Matched and Mismatched Probes 66
4.2 Optimization of Settings for the ORF1ab:L3468V ASP-PCR 67
4.3 Performance of ASP-PCR, PCR-RFLP, and NGS in REMAP-CAP Trial 70
4.4 Performance of ASP-PCR and NGS in COV003 Trial 71
4.5 Distribution of Sample Viral Loads Tested by Screening Methods . 72
4.6 Impact of Degraded RNA in COV003 Samples on Method Performance 74
4.7 Detection of N:KR Mutation in REMAP-CAP Samples 75
4.8 Performance of Individual ASP-PCR Assays on Nondegraded Samples 77
4.9 Impact of Molecular Modification on S:D1118H/Alpha ASP-PCR Performance 78

5.1 Diversity in Viral Load between Trials 95
5.2 Impact of Seroconversion on Viral Load in REMAP-CAP 96
5.3 Viral Loads by Lineage of Infection in REMAP-CAP 97
5.4 Viral Loads by Presence of the N:KR Mutation in REMAP-CAP . . 98
5.5 Viral Loads by Lineage of Infection in COV003 99
5.6 Emergence of Variants in COVID-19 Clinical Trials 101
5.7 Impact of Patient Demographic and Clinical Factors on Patient Baseline Viral Load 103
5.8 Relationship between Ct values for N gene and sgRNA N gene RT-qPCR assays 105
5.9 N Gene sgRNA Abundance by Lineage of Infection 106
5.10 Viral Dynamics Classification System 108
5.11 Representative Plots of Viral Dynamics Classifications 109
5.12 Date of Final Sampling in ITU among Convalescent Plasma-treated patients 110

6.1 Alpha and N:KR in REMAP-CAP versus General Population in England 121

7.1 Description of Samples Included in Analysis 135
7.2 Whole Genome Mutation Rates in Control and Convalescent Plasma Treated Patients 140
7.3 Correlation Between Total SNPs Accumulated and Time in ITU . . 141
7.4 Gene-Specific Weighted Substitution Rates 142
7.5 Influence of Base Context on Transitions 145
7.6 Impact of Dinucleotide Context of Observed C>U Substitutions Compared to Expected Dinucleotide Frequencies 146

8.1 Dinucleotide Odds Ratios for Human-Infective Arbovirus Families . 158
8.2 Interaction of GC Content, CpG Suppression, and Arbovirus Family 160
8.3 CpG Suppression and Transmissibility in *Flaviviridae* and *Togaviridae* 162
8.4 Corrected Dinucleotide Odds Ratios for Flavivirus Phylogenetic Groupings 166
8.5 Corrected CpG O/E in Flavivirus Phylogenetic Groups 168
8.6 Influence of Infection of Birds with Corrected CpG Values in Mosquito-Borne Flaviviruses 169

List of Tables

1.1 Human-Infective Coronaviruses 8

2.1 Overview of APOBEC Family Proteins 31

3.1 Oligonucleotides Used in this Thesis 57
3.2 ASP-PCR settings 58

4.1 Country-Specific Performance of Lineage-Defining SNPs 65
4.2 Relative Method Performance on Matched Samples 73
4.3 Predicted Performance of S:D1118H/Alpha ASP-PCR , PCR-RFLP, NGS-Pangolin, and NGS-SNP 80

6.1 Association of disease severity at admission and virological factors . 125
6.2 Association of risk of death in ITU and patient virological factors . 125
6.3 Discharge rates and median ages for viral dynamics classifications . 127

7.1 Specific Spike Substitutions of Interest 136
7.2 Ratios of Total Nonsynonymous to Synonymous Substitutions Observed 142
7.3 Unremarkable SARS-CoV-2 Evolution in Immunocompromised Patients Receiving Convalescent Plasma Therapy 144

8.1 Human-Infective Arbovirus Families and their Members 156
8.2 Comparison between Simple and Corrected Methods for Estimating Dinucleotide Observed/Expected Ratios in Human-Infective Arboviruses 157
8.3 Flaviviruses Included and their Phylogenetic Groupings 163
8.4 Host Range of Mosquito-Borne Flaviviruses 167

10.1 Global Performance of Lineage-Defining SNPs 180
10.2 Accession Numbers for Viruses Included in Chapter 8 183
10.3 Accession Numbers and Characteristics for ZIKV Genomes Used in Section 8.2.4 188
10.4 Accession Numbers and Type for DENV Genomes Used in Section 8.2.4 191

List of Abbreviations

A1 . . . . . APOBEC1

A2 . . . . . APOBEC2

A3 . . . . . APOBEC3

A4 . . . . . APOBEC4

ACE2 . . . . . Angiotensin-Converting Enzyme 2

ACF . . . . . A1 Complementation Factor

ADAR . . . . . Adenosine Deaminases Acting on RNA

AID . . . . . Activation Induced Deaminase

APACHE . . . Acute Physiology and Chronic Health Evaluation II

APOBEC . . . Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like

AS-PCR . . . Allele-Specific Polymerase Chain Reaction

ASP-PCR . . Allele-Specific Probe Polymerase Chain Reaction

BALF . . . . . Bronchoalveolar Lavage Fluid

BMI . . . . . Body Mass Index

CBFβ . . . . . Core-Binding Factor Beta

CD . . . . . Cytidine Deaminase Fold

CD4 . . . . . Cluster of Differentiation 4

cDNA . . . . . Complementary Deoxyribonucleic Acid

CHIKV . . . . Chikungunya Virus

cISF . . . . . Classical Insect-Specific Flavivirus

COVID-19 . . Coronavirus Disease 2019

CP . . . . . Convalescent Plasma

Ct . . . . . Cycle Threshold

DENV . . . . . Dengue Virus

dISF . . . . . Divergent Insect-Specific Flavivirus

DNA . . . . . Deoxyribonucleic Acid

ds . . . . . Double-Stranded

E . . . . . Envelope

EBOV . . . . . Ebola Virus

Eif4g2 . . . . . Eukaryotic Translation Initiation Factor 4 Gamma 2

ELISA . . . . . Enzyme-Linked Immunosorbent Assay

EHV . . . . . Edge Hill Virus

gRNA . . . . . Genomic RNA

GVP . . . . . Global Virome Project

HBV . . . . . Hepatitis B Virus

HCoV . . . . . Human Coronavirus

HCV . . . . . Hepatitis C Virus

HEK . . . . . Human Embryonic Kidney

HIV . . . . . Human Immunodeficiency Virus

HPV . . . . . Human Papillomaviruses

HSV-1 . . . . . Herpes Simplex Virus 1

HTLV-1 . . . . Human T-Cell Leukemia Virus Type 1

IAV . . . . . Influenza A Virus

IBV . . . . . Influenza B virus

ICTV . . . . . International Committee for Taxonomy of Viruses

IFN . . . . . Interferon

IFR . . . . . Infection Fatality Rate

IgG . . . . . Immunoglobulin G

IRES . . . . . Internal Ribosome Entry Site

ITU . . . . . Intensive Treatment Unit

IU . . . . . International Units

kb . . . . . Kilobase

LECA . . . . . Last Eukaryotic Common Ancestor

LINE-1 . . . . Long Interspersed Element-1

LNA . . . . . Locked Nucleic Acid

M . . . . . Membrane

MBF . . . . . Mosquito-Borne Flavivirus

MERS . . . . . Middle Eastern Respiratory Syndrome

mL . . . . . Milliliter

MOI . . . . . Multiplicity of Infection

mRNA . . . . Messenger Ribonucleic Acid

N . . . . . Nucleocapsid

nm . . . . . Nanometer

NAAT . . . . . Nucleic Acid Amplification Test

NIBSC . . . . National Institute of Biological Standards and Control

NGS . . . . . Next-Generation Sequencing

NKVF . . . . . No-Known-Vector Flavivirus

NPA . . . . . Negative Percent Agreement

NPV . . . . . Negative Predictive Value

nt . . . . . Nucleotide

O/E . . . . . Observed/Expected

OAS . . . . . 2'-5'-Oligoadenylate Synthetase

ORF . . . . . Open Reading Frame

PCR . . . . . Polymerase Chain Reaction

PCR-RFLP . Polymerase Chain Reaction Restriction-Fragment Length Polymorphism

PPA . . . . . Positive Percent Agreement

PPV . . . . . Positive Predictive Value

PTEN . . . . . Phosphatase and Tensin Homolog

R_0 . . . . . Basic Reproductive Number

RBD . . . . . Receptor-Binding Domain

Rbx2 . . . . . RING-box subunit 2

RdRP . . . . . RNA-dependent RNA Polymerase

REMAP-CAP Randomised, Embedded, Multi-factorial, Adaptive Platform Trial for Community-Acquired Pneumonia

REPR . . . . . ApoB mRNA Editing Protein

RNA . . . . . Ribonucleic Acid

RSV . . . . . Respiratory Synctial Virus

RT-qPCR . . . Reverse Transcription Quantitative Polymerase Chain Reaction

S . . . . . Spike

ss . . . . . Single-Stranded

SARS-CoV-2 . Severe Acute Respiratory Syndrome Coronavirus 2

sgRNA . . . . Subgenomic Messenger RNA

SGTF . . . . . Spike Gene Target Failure

SINE . . . . . Short Interspersed Nuclear Element

SNP . . . . . Single Nucleotide Polymorphism

TBF . . . . . Tick-Borne Flavivirus

TE . . . . . Transposable Element

$\mathbf{T}_m$ . . . . . Melting Temperature

TRS . . . . . Transcription Regulatory Sequence

TSO . . . . . Template Switching Oligo

TTV . . . . . Transfusion Transmitted Virus

USAID . . . . United States Agency for International Development

UDI . . . . . Unique Dual Index

UTR . . . . . Untranslated Region

Vif . . . . . Viral Infectivity Factor

VIZIONS . . . The Vietnam Initiative on Zoonotic Infections

VoC . . . . . Variant of Concern

VoI . . . . . Variant of Interest

WNV . . . . . West Nile Virus

YFV . . . . . Yellow Fever Virus

ZAP . . . . . Zinc Antiviral Protein

ZDD . . . . . Zinc-Dependent Deaminase Domain

ZIKV . . . . . Zika Virus

1 Introduction

Contents

1.1 Introduction 1
1.1.1 Virus life cycle 2
1.1.2 Intra-host dynamics 4
1.2 Virus Classification and Relevant Viruses 5
1.2.1 *Coronaviridae* 6
1.2.2 Arboviruses 7
1.3 Virus Evolution 8
1.3.1 Conceptual Models of RNA Virus Evolution 9
1.3.2 Adaptive (Darwinian) Evolution 10
1.3.3 Non-Darwinian Evolution and the Neutral Theory . . . 12
1.4 Zoonotic Virus Emergence 14
1.5 Aims of this Thesis 15

1.1 Introduction

Viruses are a diverse set of obligate intracellular parasites whose members infect (probably) all known domains of life. They are the most abundant form of "life" on the planet, and estimates state the total length of all virions on Earth, laid end to end, would span 65 galaxies.[1] They are diverse both in terms of genome length, ranging from the puny 1.75 kilobase (kb) Hepatitis D virus to the 1.2 megabase Megavirus, and in physical diameter, from the 20 nanometer (nm) Picornaviruses

again to the >400 nm Megavirus.[2–4] Viruses themselves can be even be 'infected' by other viruses; virophage propagation requires co-infection with a giant virus during which time they siphon resources from the giant virus.[5] Substantial motivations for researching viruses is their negative impacts on human health and agriculture and their potential as platforms for biotechnological development. This thesis will begin with an overview of the field, exploring general concepts directly and indirectly related to the content presented in the results chapters. The concepts may not be explicitly mentioned later, but they were essential to the development of experimental plans and broader interpretation of the results.

While diverse, (practically all) viruses have some common general properties.[3] Their genomes are composed of either ribonucleic acid (RNA) or deoxyribonucleic acid (DNA), which can either be single-stranded (ss) or double-stranded (ds) (but not both). ssRNA genomes can either be positive-sense (+) or negative-sense (-). (+)ssRNA genomes can be directly translated as mRNA while (-)ssRNA genomes are used as templates to synthesize messenger RNA (mRNA).[6, 7] (-)ssRNA genomes can also be segmented where the genomic content is stored on multiple gene segments rather than a continuous RNA molecule (some (+)ssRNA viruses, especially among those that infect plants, can also be segmented).[8] Virus genomes are encapsulated by a protein shell made of repetitive, identical protein subunits called a capsid that gives an individual virion their distinctive, symmetrical shape. Some viruses also have a lipid bilayer membrane (envelope) encircling the capsid protein, which influences viral entry and exit and environmental stability. The virus life cycle also moves through general stages, as discussed below:

1.1.1 Virus life cycle

From parent virion to progeny virion, the intracellular virus life cycle involves three stages – entry, replication, and exit.[9]

Viral Entry

This first step begins with recognition of cell-associated target receptors and attachment factors by surface proteins on the virion. Examples of these include angiotensin-converting enzyme 2 (ACE2) for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), sialic acid for influenza A and influenza B viruses (IAV and IBV), and cluster of differentiation 4 (CD4) for human immunodeficiency virus (HIV).[10–12] Using these receptors, the virion attaches to the target cell membrane and penetrates the cell through direct fusion or receptor-mediated endocytosis. Direct fusion for enveloped viruses involves fusion of the viral envelope with the cellular plasma membrane, often via a fusion peptide. Receptor-mediated endocytosis facilitates entry for both enveloped and nonenveloped viruses. The virus is brought into the cytosol via an endosome. Endosomal degradation proceeds through either pH-dependent effects, mediated by a fusion peptide for enveloped viruses, or endosomal lysis by the capsid proteins of nonenveloped viruses, both of which deliver the intact viral particle inside the cytoplasm. Viruses then hijack cell trafficking machinery to move to an intracellular location where the viral genome is uncoated. This location can be either the cytosol or the nucleus.

Replication and Expression

Genome replication and viral gene expression proceed through distinct mechanisms for different viruses (see the Baltimore classifications below). For example, ds-DNA viruses uncoat within the cell nucleus, transcribe mRNA using host RNA polymerases, and replicate using either host DNA polymerases or, in the case of larger viruses such as those in the *Herpesviridae* family, self-encoded DNA polymerases.[13] RNA viruses encode an RNA-dependent RNA polymerase (RdRP) for genome replication. For gene expression, all (+)ssRNA viral genomes can be read as mRNA and transcribed directly without further modification.[6] (-)ssRNA genomes must be transcribed into (+)RNA transcripts by the RdRP to be translated into viral proteins. The RdRPs of some (-)ssRNA viruses, such as IAV, cannot cap their viral transcripts and instead engage in “cap-snatching” where the 5’-m7G

caps of cellular mRNAs are stolen to prime translation of viral RNA.[14, 15] All viruses leverage host ribosomes for translation of their gene products.

Exit

Following replication and translation of nonstructural and structural viral proteins, viral capsids aggregate intracellularly. Viral genomes are either packaged during capsid assembly, as with Tobacco Mosaic virus, or after the capsid has been fully-formed, as with members of the *Adenoviridae* family.[16] For nonenveloped viruses, the fully-formed virions are released into the extracellular space upon host cell lysis, completing the life cycle. Enveloped viruses, on the other hand, must first commandeer their envelope from host cell membranes (either the plasma membrane or endomembrane system) and then bud from the host cell via exocytosis. For both pathways, release into the extracellular space is generally the end of the viral life cycle for lytic viruses, although some viruses will 'mature' extracellularly to prime infectivity.[17]

1.1.2 Intra-host dynamics

For a virus to persist in a population, individual virions must evade host antiviral mechanisms for a sufficient amount of time to transmit progeny to a new host before being cleared. Many viruses achieve this through rapid generation of virions and transmission in a short window before the host immune system can eliminate the pathogen. These viral infections are characterized by an acute infection of relatively short duration (days to weeks) with viral clearance achieved by the activity of the innate and/or adaptive immune system. These include pathogens with varying cell and tissue tropisms and other properties such as respiratory viruses (rhinoviruses, IAV, IBV, and respiratory syncytial virus (RSV)), gastrointestinal viruses (rotavirus and norovirus), and hemorrhagic fever viruses (Ebola virus (EBOV) and Marburg virus).[18–24] Other viruses establish persistent, long-lasting infections through a variety of mechanisms.[25] Along with epithelial cells, Herpes Simplex Virus 1 (HSV-1) can infect terminally differentiated, long-lived neuronal cells. HSV-1 then

shuts down protein synthesis and survives uncleared in the host cell as a non-lytic (latent) infection unless reactivated.[26–28] Retrovirus proviral integration into host genomes is an obligatory step in the retroviral life cycle and virtually ensures presence of the virus in the integrated cell and any of its progeny.[29] Other viruses, such as Hepatitis C virus (HCV) and Cytomegalovirus, successfully modulate host immune responses to evade clearance and continuously replicate.[30, 31] Lastly, some typically acute viral infections, such as EBOV, can achieve persistence by infecting cells in immune privileged sites such as the eyes or testes.[32]

1.2 Virus Classification and Relevant Viruses

For macroscopic life, designations at the species level are commonly based on the ability to produce fertile offspring. For microorganisms that reproduce asexually, alternative means of classification have to be deployed. Both viruses and bacteria are organized taxonomically based on a combination of phenotypic traits and, as sequencing has become more prevalent, genetic relatedness.[33] Like the taxonomic rankings for macroscopic life, the concept of a "virus species" is an abstraction.[34] Nevertheless, the classification of viruses into distinct taxonomic ranks assists the interpretability of the scientific literature.[35]

Organizing these individual virus species into higher order classifications has been ongoing for over 50 years. David Baltimore's *Expression of Animal Virus Genomes*, first published in 1971, was one of the earliest taxonomic systems for viruses.[36] His taxonomy classified viruses based on the means by which they transcribed mRNA, which eventually defined seven groupings mainly based on the organization of genomic content. The Baltimore classification system roughly defines an individual viruses' genome structure, size, and presence of segmentation.[37] The International Committee on Taxonomy of Viruses (ICTV), the body tasked with formalizing the taxonomic relationships between viruses, has defined a species as:

"... a monophyletic group of viruses, whose properties can be distinguished from those of other species by multiple criteria.... These criteria may include, but are not limited to, natural and experimental host range, cell and tissue tropism, pathogenicity, vector specificity, antigenicity, and the degree of relatedness of their genomes or genes..."[38]

The framework raises some potential complications: what are the relative weights given to each of the criteria? What is the branch length considered when defining monophyletic groupings? Members of the Louping Ill virus species phylogenetically lie entirely within the clade of the species Tick-Borne Encephalitis virus and are thus monophyletic.[39] However, members of the two species are markedly different in terms of host range, geographical location, and pathogenicity, motivating their classification as separate species.

Further complicating taxonomy is how to assign species for viruses only identified via reconstruction of metagenomic sequencing. This is challenging for those viruses that have not been isolated and thus have no phenotypic information available. The full discussions are beyond the scope of this thesis, but in summary it is recognized that the current taxonomy is inconsistent in its treatment of genetic relatedness, delineation of species based on metagenomic-derived sequences will need to be based on an established working definition of genetic relatedness (of which currently there is none), and the number of viruses being identified from metagenomic sequencing, particularly from ocean samples, far outpaces the ability of the ICTV to ratify each member.[39–43]

For the purpose of the content covered in this thesis, there are two classifications/taxonomic groups of interest: the family *Coronaviridae* and the polyphyletic members of the informal classification "arboviruses".

1.2.1 *Coronaviridae*

Coronaviridae are a family within the order *Nidovirales* composed of enveloped (+)ssRNA viruses. Their genomes are around 30 kb in length and encode nonstruc-

tural polyproteins in open reading frame (ORF) 1a and ORF1b in the 5' two-thirds of the genome and the structural proteins spike (S), envelope (E), membrane (M), and nucleocapsid (N) in the 3' one-third of the genome.[44] Translation of the ORF1b polyprotein, which encodes the viral RdRP, is dependent on a programmed -1 ribosomal frameshift.[45] Translation of the 3' structural proteins and a host of small accessory proteins is dependent on the generation of subgenomic mRNAs (sgRNA). These sgRNAs are discontinuously synthesized during negative-strand RNA synthesis when the RdRP encounters a transcription regulatory sequence (TRS) upstream of the ORF.[44, 46–48] Synthesis pauses and is reinitiated at the TRS in the 5'-untranslated region (UTR) of the genome. Thus, all sgRNAs contain a common leader sequence at the 5' end of the transcript.

Coronavirus viral entry is dependent on S mediated receptor binding on the surface of the target cell and subsequent cleavage of S by host proteases.[44, 49] The presence of receptors on the surface of target cells determine the virus' cell tropism. There are seven coronaviruses known to infect and transmit in human populations (Table 1.1). Human coronavirus (HCoV)-NL63 and HCoV-229E are seasonal alphacoronaviruses and HCoV-HKU1 and HCoV-OC43 are seasonal betacoronaviruses that all cause cold-like symptoms and circulate during the winter months.[50] Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV), SARS-CoV, and SARS-CoV-2 are recently emergent betacoronaviruses that are responsible for outbreaks of viral disease associated with high levels of morbidity and mortality.[51, 52]

1.2.2 Arboviruses

Arboviruses (arthropod-borne viruses) are an informal classification for a diverse set of polyphyletic viruses that transmit to non-insect hosts via arthropod vectors. Commonly, this transmission is due to the feeding behavior of hematophagous arthropods such as mosquitoes, ticks, and sand mites.[61] Practically all arboviruses have genomes that are composed of ss or dsRNA (a single DNA arbovirus, African

	Virus	Classification	Year of Discovery/Emergence
Alphacoronavirus	HCoV-229E	Seasonal	1965[53]
	HCoV-NL63	Seasonal	2004[54]
Betacoronavirus	HCoV-HKU1	Seasonal	2005[55]
	HCoV-OC43	Seasonal	1967[56]
	MERS-CoV	Emergent	2012[57]
	SARS-CoV-1	Emergent	2003[58]
	SARS-CoV-2	Emergent	2019[51, 59, 60]

Table 1.1: Human-Infective Coronaviruses Viruses classified as described in Zhu et al. (2020)[60]

swine fever virus, has been isolated)[62]. Arboviruses can have either segmented or nonsegmented genomes.

1.3 Virus Evolution

In biology, evolution is the progressive alteration of the genotypic and phenotypic characteristics of a population over time.[63] RNA viruses similarly evolve and adapt and typically display rates of molecular evolution several orders of magnitude greater than eukaryotes.[64] While molecular evolution is generally due to the misincorporation of nucleotides during genome replication, other routes for generating diversity are also available to RNA viruses. Both segmented and nonsegmented RNA virus genomes can engage in recombination when an individual cell is coinfected with two or more viruses from the same species.[65] Recombination can occur through a variety of methods but all ultimately generate an RNA transcript of mixed ancestry. Segmented viruses can also undergo reassortment during coinfection, where segments from different parental genomes are packaged into a single virion. A substantial source of the observed diversity of IAV and IBV is due to reassortment.[66] Conceptual models of RNA virus evolution provide a framework through which to better comprehend their remarkable speed of diversification. Here, two conceptual models are briefly summarized:

1.3.1 Conceptual Models of RNA Virus Evolution

Virus Seascape

The concept of the "fitness landscape" has been present in evolutionary biology for over 90 years.[67] In these three-dimensional topological maps, phenotypic fitness is represented by height and genotypes are represented by x- and y-coordinates on a cartesian plane. Individual mutations represent single steps in the x or y direction of the plane. The primary theoretical use of fitness landscapes is to demonstrate that there exist local minima and maxima for fitness and that, at times, substantial genetic "distance" would need to be covered (i.e., many mutations introduced) to reach an overall increased fitness – and that, crucially, organisms may need to cross a "fitness valley" to reach a higher local maxima. A "fitness seascape" is an extension of the fitness landscape model that allows for the fitness "height" to be dynamic over time (like the heights of waves in an ocean) as evolutionary pressures ebb and flow.[68] Relevant to RNA viruses, strong evidence in favor of the seascape comes from the impacts of vaccine- and infection-derived immunity in a population of potential hosts. For a review on the potential role of population-level neutralizing antibody responses on the conformational preferences of the SARS-CoV-2 S receptor binding domain (RBD), please see Wolf et al. (2022).[69]

Prisoners of War

The "Prisoners of War" model was developed by Simmonds, Aiewsakun, and Katzourakis in 2019 to explain the conflict between high RNA virus evolution rates when viewed over the short-term versus much lower rates when viewed over the long-term.[70] In this host-driven evolutionary model, the neutral sequence space available for a virus to explore over the long-term is constrained by selective pressures that are governed by the host's mutation rate. Over thousands of generations, viruses will continually sample and cycle throughout this neutral space, periodically gaining and losing the same adaptations, and thus appearing to mutate rapidly over the short-term but at a slower rate in the long-term. This concept of cycling through restricted neutral evolutionary space is supported by a study of periodic gain and

loss of cross-reactivity in sera derived from children infected with IAV.[71] In that study, sera from children infected with IAV in 2006/2007 efficiently neutralized historical IAV isolates from 1933, 1977, 1999, and 2006 but failed to neutralize isolates from 1918, 1934, and (prospectively) 2009.

1.3.2 Adaptive (Darwinian) Evolution

Adaptive evolution is analogous to Darwin's theory of natural selection. In this model, mutations are gained and lost by chance. Those random alleles that increase host fitness are selected for and fixed due to the increased likelihood of being passed on to progeny and persist in the population. Those that are deleterious are selected against and purged. For RNA viruses, one of the main research topics on viral evolution is the adaptation of viral proteins.

Detection of selective pressures for viruses is similar to that used for other species and genes. One of the most common methods for detecting divergence is through the calculation of nonsynonymous/synonymous mutation rate ratios (dN/dS).[72] Synonymous substitutions are assumed to be selectively neutral (although this assumption is not perfect, see below) and thus their rate of accumulation can function as a molecular clock and a stable approximation of the substitutions/site/year.[73, 74] Nonsynonymous substitutions, on the other hand, are subject to selective pressures on the protein product that they encode. In a system absent of selective pressures and over sufficient time, the ratio should approximate to one as mutations are fixed randomly (no bias to either type of mutation). A ratio exceeding one indicates positive (directional) selection where change is advantageous; a ratio less than one indicates negative (purifying) selection where change is suppressed. The dN/dS ratio is normalized to the total number of synonymous and nonsynonymous sites in the basal sequence. There have been multiple computational methods published for their calculation, mostly slightly differing in their approach to counting the number of synonymous and nonsynonymous changes when codon pairs differ by more than one base.[72, 75–77] The contexts in which certain methods for calculating dN/dS ratios are more appropriate than others is beyond the scope

of this thesis. Another useful test for detecting selection for specific genes is the McDonald-Kreitman Test which compares the ratio of nonsynonymous to synonymous polymorphisms within a species to dN/dS values between species.[78] Both dN/dS ratios and the McDonald-Krietman test statistically treat neutrality as the null hypothesis.

Selective pressures on RNA viruses are not solely limited to the genome's protein-coding content. Instances of selection on the RNA content itself are highlighted below:

Maintenance of non-coding elements

Internal base pairing of ssRNA can form complex physical structures (RNA secondary structure). This structure is determined by the sequence of bases and therefore can be impacted by substitutions. RNA secondary structure can form discrete elements that serve functional roles for the virus. The primary sequence of these non-coding regions are subject to similar selective pressures as protein-coding elements; as the selective pressures are generally to maintain existing structures, they serve as a constraint of viral evolution and limit the sequence space that can be explored by the virus.[79]

For several ssRNA virus species, RNA secondary structure elements serve discrete roles in gene translation and genome replication. Internal ribosome entry sites (IRES) recruit host ribosomes to facilitate gene translation in a 5'-m7G independent manner.[80] IRES elements have been described in members of the families *Flaviviridae* (Genera: *Hepacivirus* and *Pestivirus*) and *Picornarividae*, among others, and are located primarily within the 5'-UTR but can also be located within protein-coding elements.[81, 82] Other secondary structure elements are essential for packaging of the virion, as demonstrated in members of the *Togaviridae*, *Coronaviridae*, *Hepadnaviridae*, and *Retroviridae* families.[83]

As described earlier, SARS-CoV-2, like other members of the *Coronaviridae* family, encodes a -1 ribosomal frameshift in ORF1. This frameshift is governed by a pseudoknot structure.[84] Multiple reading frames can increase protein diversity

and can tightly regulate the relative quantities of individual proteins produced. Mutations compromising the pseudoknot would prohibit transcription of the ORF1ab genes, ultimately preventing genome replication as this portion of the ORF encodes the RdRP and helicase genes.[47]

Dinucleotide Suppression

In both vertebrate and viral genomes, the dinucleotides UpA (TpA in DNA) and CpG occur less frequently than would be expected by chance.[85] Artificially increasing the frequency of these dinucleotides in viral genomes through synonymous recoding leads to substantial loss of fitness *in vitro* and *in vivo*.[86–88] In vertebrates, the loss of CpG is generally attributed to spontaneous deamination of methylated cytosines to thymines leading to the progressive loss of the CpG dinucleotide.[89] This explanation is clearly insufficient for RNA viruses as their genomes are not targets of DNA methyltransferases. Over a decade of work leveraging synonymously recoded infectious clones, researchers in multiple labs have pinpointed anti-CpG selection being due to the action of the interferon (IFN)-induced zinc antiviral protein (ZAP) binding CpG motifs in viral RNA transcripts.[90–92] Neither invertebrate-infecting or dsRNA viruses demonstrate suppression of CpG.[85] UpA suppression is more ubiquitous and is present in vertebrates, invertebrates, and viruses infecting both hosts. UpA is a target of the ribonuclease RNAse L, likely driving its suppression.[91, 93]

1.3.3 Non-Darwinian Evolution and the Neutral Theory

Adaptive evolution fails to perfectly reflect observed evolutionary changes. Thus, a model of "Non-Darwinian evolution" began to be developed in the late 1960s.[94] One of the prevailing models for molecular evolution is the neutral theory first posited by Motoo Kimura in 1968.[95, 96] The neutral theory states that differences within species are governed by mutation-drift equilibrium rather than the 'balancing selection' that dominated prior to the theory's development. However, accumulated evidence since the initial publication of the theory has contradicted many of the

original assumptions and conclusions on which it was based.[97] Still, the neutral theory remains a useful statistical framework through which to detect selection as it supports "neutral selection" as a null hypothesis.

Genetic Drift and Population Bottlenecks

Genetic drift, predicted by the neutral theory to be the major driver of observed mutations, is the random fluctuation of alleles in a population.[98] The impact of genetic drift on a population is a function of its effective population size; smaller populations are more prone to its effect.[99] Relevant to RNA viruses, which intracellularly exist as a 'mutant swarm' or 'quasispecies' following genome replication, both intra-host dissemination and inter-host transmission are associated with tight population bottlenecks.[100] These bottlenecks limit the effective population size at the new site or in the new host, which can lead to the random fixation of neutral or slightly deleterious mutations even if at low frequency in the index cell or host.

Direct RNA editing

A less appreciated but still impactful method for generating diversity in RNA viruses is direct editing of the genome by host enzymes. The adenosine deaminases acting on RNA (ADAR) protein family catalyzes adenosine to inosine mutations in dsRNA, which lead to A>G substitutions during translation or replication.[101] The proteins are IFN inducible and have antiviral activity on a large set of RNA and DNA viruses. The apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) family of proteins are well regarded for their mutational burden on HIV transcripts and are reviewed in extensive detail in Chapter Two. If the RNA editing events are nonrandomly distributed in the genome and take place in so-called 'mutational hotspots', then those sites will mistakenly appear to be under selection in standard divergence analyses.

1.4 Zoonotic Virus Emergence

A 'zoonosis' describes the incursion of an animal pathogen into the human population. The majority of infectious diseases currently circulating in humans that cause illness are zoonotic in origin, including those responsible for explosive outbreaks such as pandemic influenza, HIV, MERS-CoV, SARS-CoV, SARS-CoV-2, EBOV, and Nipah Virus.[102, 103] Beyond their propensity to enter human population, zoonotic pathogens are also harder to control due to the possibility of the original host population serving as a reservoir. Variola virus, the causative agent of smallpox, was likely zoonotic in origin but crucially lost the ability to infect animals during its thousands of years of circulating in humans.[104] This restriction in host range was crucial toward successful public health efforts to eliminate the pathogen globally. The poliovirus serotypes (of the species Enterovirus C), which may be the second human pathogen successfully eliminated, are also restricted to humans.[105]

The above factors motivate the intensive study of this class of virus. The earliest work to quantitatively assess the risk of zoonotic disease emergence was not completed until the early 2000s.[102] Since then, substantial conceptual progress in the field has come from the lab of Professor Mark Woolhouse. First articulated in Wolfe et al. (2007),[106] the Woolhouse lab refined the concept of the pathogen pyramid over several publications.[107–109] The pyramid categorizes pathogens based on their infection and transmission characteristics in humans; level 1 defines pathogens to which humans are exposed but not infected, level 2 defines pathogens for which humans can be infected but do not transmit (dead-end hosts), level 3 defines pathogens that transmit in human population but with an estimated basic reproductive number (R_0) < 1, and level 4 defines pathogens that transmit in human populations with an $R_0 > 1$ (epidemic pathogens). The categorization enables investigation of the pathogen factors associated with transmissibility. Many labs have also contributed bioinformatic and molecular work on the viral and host determinants of infection and transmissibility, which will be discussed in more detail in Chapter Eight. In general, the risk factors for zoonotic spillover and pathogen

success following spillover are poorly defined and not generalizable. Further research in the field is warranted given the potential impact of these events on global health.

1.5 Aims of this Thesis

Utilizing a large collection of nasopharyngeal swab samples collected from patients with coronavirus disease 2019 (COVID-19) as part of two clinical trials and publicly available sequence data for arboviruses, this thesis aims to:

1. Design and validate high-throughput, low-cost techniques to monitor RNA virus adaptation in human populations (Chapter Four)
2. Characterize the impacts of virus adaptation on viral dynamics and human health (Chapters Five and Six)
3. Identify potential drivers of SARS-CoV-2 evolution and adaptation (Chapters Two and Seven)
4. Identify genetic predictors of increased pandemic potential among arboviruses (Chapter Eight)

2

A Review of Potential APOBEC-mediated Hypermutation of SARS-CoV-2

Contents

2.1 Introduction 17
2.1.1 Discovery of the APOBEC Family. 18
2.1.2 Conservation of structure 19
2.1.3 Diversity of function in humans 19
2.2 Deaminase-Dependent Restriction of Viruses and Mobile Genetic Elements 22
2.2.1 Human Immunodeficiency Virus (HIV-1) 22
2.2.2 Retrotransposons and Mobile Genetic Elements 23
2.2.3 DNA viruses 24
2.2.4 RNA Viruses 25
2.2.5 Deaminase-Independent Restriction and Insensitivity . . 26
2.3 APOBEC Origins, Diversification, and Evolution . . . 26
2.3.1 Origin 27
2.3.2 A3 Paralog Expansion 27
2.3.3 APOBEC Protein Evolution 30
2.4 Bioinformatic Evidence for RNA Editing in SARS-CoV-2 31
2.4.1 C>U Mutational Asymmetry in SARS-CoV-2 31
2.5 Functional Evidence for a Role for APOBEC in Editing and Antiviral Restriction of Coronaviruses 34
2.6 Evolutionary Consequences of C>U Mutational Bias in Coronaviruses 35
2.7 A Perspective on Future Functional Investigations of APOBEC-Mediated Editing of SARS-CoV-2 and Other RNA Viruses 38

Abstract

Members of the APOBEC family of cytidine deaminases demonstrate antiviral activities in mammalian cells. A large body of work has established that APOBECs induce lethal editing in the DNA genomes of small DNA viruses, herpesviruses, and retroviruses. Their editing activity on RNA templates may have similar effects on coronaviruses and other RNA viruses but has been largely under-characterized. APOBEC-like directional C>U transitions of genomic (+)strand RNA are greatly overrepresented in mutations emerging in SARS-CoV-2 genome sequences during the COVID-19 pandemic, consistent with potential antiviral activity of APOBEC against coronaviruses. A C>U mutational process may leave evolutionary imprints on coronavirus genomes, including extensive homoplasy at sites of editing and reversion at targeted sites, the occurrence of driven amino acid sequence changes in viral proteins with little or no adaptive value, and, if sustained over longer periods, the previously reported marked global depletion of cytosine and excess of uracil bases in the genome of HCoV-NL63. Bioinformatic analysis supports the activity of APOBEC on a wide range of RNA viruses and functional evidence on transcripts derived from SARS-CoV-2 support APOBEC-coronavirus interactions leading to C>U transition. This review synthesizes the current knowledge on APOBEC evolution and function and the evidence of their role in APOBEC-mediated genome editing of SARS-CoV-2 and other coronaviruses.

2.1 Introduction

As pathogen host ranges expand into new species, they must adapt to the unique intracellular selective pressures of their new hosts to be able to successfully transmit within a population. The sustained spread of SARS-CoV-2, coupled with readily deployable and accessible sequencing technologies, has given researchers an unprecedented ability to track virus evolution from spillover through early

adaptation to human hosts. Mining this data, multiple authors have identified an overabundance of homoplastic cytidine to uridine transitions in the SARS-CoV-2 genome, particularly in the context of 5'-Up<u>C</u>-3', driving a large proportion of the observed nonsynonymous mutation burden throughout the genome early in the pandemic. This mutational pattern is similar to that induced by the cytidine deaminases of the APOBEC family of proteins. The ability of this enzyme family to modify retroviruses, retroelements, and DNA viruses is well established, while their potential to modify RNA viruses is exploratory. This thesis chapter will review the current knowledge of APOBEC function and evolution in depth, the evidence for this mutational phenomenon in coronaviruses, and its potential evolutionary consequences for RNA viruses.

2.1.1 Discovery of the APOBEC Family.

In 1987, two groups identified that tissue-specific expression of the proteins ApoB100 and ApoB48 from the ApoB gene depended on post-transcriptional modifications of mRNA rather than differences in the genes themselves.[110, 111] In both human and rat enterocytes, a cytosine to uracil modification (C6666U) generated an ochre stop codon and translation of a truncated version of ApoB100 (ApoB48) with modified biological function. In 1993, the protein catalyzing this reaction was isolated from rat enterocytes and designated apoB mRNA editing protein (REPR).[112] Shortly thereafter, the existence of other mRNA editing enzymes functioning through a similar catalytic subunit was hypothesized, and the enzyme responsible for ApoB48 was renamed Apolipoprotein B mRNA Editing Catalytic Polypeptide 1 (APOBEC1).[113] Since that time, many APOBEC1-related genes and splice variants have been identified in the genomes and transcriptomes of fish, mammals, non-human primates, and humans, which collectively are referred to as the APOBEC family of proteins. In humans, the family of cytidine deaminases has 11 members: APOBEC1 (A1), APOBEC2 (A2), seven APOBEC3s (A3A, A3B, A3C, A3D, A3F, A3G, A3H), APOBEC4 (A4), and activation induced deaminase (AID).[114, 115] The proteins possess a shared, highly conserved zinc-dependent deaminase domain

(ZDD) but are otherwise diverse in function, having been implicated in a plethora of diverse biological pathways including lipid metabolism, antibody diversification, restriction of viruses and transposable elements, and development of cancer.[116–121]

2.1.2 Conservation of structure

Members of the APOBEC family of proteins are readily identified through their possession of a highly conserved ZDD motif His-X-Glu-X23-28-Pro-Cys-X2-4-Cys that is present in the cytidine deaminase fold (CD).[122] In those maintaining enzymatic activity, the active site of the ZDD coordinates a zinc ion that activates a water molecule for nucleophilic attack of C4 of the pyrimidine ring of the cytidine nucleoside. This attack, in conjunction with the upstream glutamic acid residue, leads to the deamination of the base to uracil. The CD fold is nested in a conserved super-secondary structure of a five-stranded β-sheet surrounded by six α-helices, with minor variations in the X residues potentially leading to the diverse functions and targets of this family of proteins.[122] A3B, A3D, A3F, and A3G contain double deaminase domains where the N-terminal domain is no longer catalytically active for deamination of ssDNA.[122] In humans, all members of the family have multiple exons and some have been shown to produce splice variants with altered function.[123–125]

2.1.3 Diversity of function in humans

A1 primarily functions to increase diversity in host-expressed RNAs in a tissue-specific manner.[120] The RNA binding and catalytic activity of A1 is dependent on the binding of one of two cofactors, A1 complementation factor (ACF) or RBM47. In addition to ApoB, A1 edits NF1, eukaryotic translation initiation factor 4 gamma 2 (eIF4G2), and the 3' UTRs of other host mRNAs. [126–130] ACF binds an 11-nucleotide (nt) 'mooring sequence' 3' of the target cytosine residue that partially, but not completely, determines the specificity of the nucleotide target.[128, 131–133] A1 deaminates cytidines on RNA targets in the context of 5'-Ap<u>C</u>-3' with the 5' end of the mooring sequence 4-6 nt downstream of the target. The protein is

primarily found in the cytoplasm but can traffic to the nucleus. A1 is unusual in its strict targeting of RNA, as most human APOBEC proteins are regarded for their ability to deaminate cytidine to uridine in ssDNA.

AID and AID-like enzymes are an essential component of the adaptative immune system in all jawed vertebrate species.[134, 135] In humans, AID drives somatic hypermutation and class-switch recombination in B cells through recruitment of DNA repair enzymes following introduction of dU into immunoglobulin V and C genes.[136–139] The introduced dU can lead to stochastic introduction of insertions or deletions during mismatch repair or transitions during DNA replication, both of which lead to the diversity expanding the host antibody repertoire. AID preferentially deaminates in the 5'-WpRpC-3' context (W = A or T, R = A or G) and is basally present in the cytoplasm but traffics to the nucleus to perform immunoglobulin gene editing.

A2 has roles in mammalian skeletal and cardiac muscle differentiation.[140–143] Transcripts of eIF4G2 and phosphatase and tensin homolog (PTEN) have been suggested as mRNA targets in A2-overexpressing hepatocytes.[144] However, transgene expression of A1 in hepatocytes also leads to Eif4g2 editing, so it is unclear if this represents the *in vivo* function or is an artefact of the overexpression model.[130, 145] A4 was first discovered in humans bioinformatically and contains orthologs in other mammals, chickens, and frogs.[115] The function of the protein remains unclear but expression is upregulated in the testes of mice, leading the authors to speculate on a role in spermatogenesis. An A4 homolog was recently cloned from chicken cells and showed *in vitro* antiviral activity against Newcastle disease virus, although editing of the viral genome was not described.[146] Neither A2 nor A4 have described target sequences. A2 shows a dispersed intracellular distribution while the intracellular localization of A4 remains poorly characterized.

The principal activity of the seven members of the A3 subfamily in humans is the restriction of viruses and mobile genomic elements, which is covered extensively below. A3A, A3B, A3C, A3D, A3F, and A3H target sequences with a 5'-TpC-3' motif, while A3G targets 5'-CpC-3'. McDaniel et al. investigated the relevance of

sequence context adjacent to a TC motif and secondary structure elements to A3 activity on DNA oligonucleotides.[147] Flanking sequence preference varied between subfamily members and within difference structural contexts. In general, A3A, A3F, and A3H preferred A or T flanking the target motif and all members of the subfamily preferred the edited site to be presented outside of defined secondary structure elements (e.g., on the loop rather than the stem of a stem-loop structure). A3A, A3C, and A3H show a broadly distributed intracellular localization, A3B is predominantly nuclear, and A3D, A3F, and A3G are primarily localized in the cytoplasm.

Intriguingly, both A3A and A3G have recently been experimentally shown to edit host mRNAs *in vitro*, with A3A-driven editing in activated innate immune cells and A3G-driven editing in both hypoxic natural killer and $CD8^+$ T cells and overexpressing human embryonic kidney (HEK) 293T cells.[148–150] As all A3s have demonstrated RNA binding activity (but not deaminase activity),[122] it does open up the possibility of RNA deamination only in limited cellular contexts by members of the A3 subfamily. Perhaps reflecting the structure dependence of A3-meditated editing of DNA oligonucleotides, it was shown that editing of host mRNA in activated macrophages and monocytes by A3A occurred preferentially at sites flanked by short palindromic sequences.[147, 148] The authors demonstrated that 72% of edited sites in monocytes and 67% of edited sites in macrophages were present in the loop of a predicted stem-loop structure. In both cell lines, the most common flanking sequence was an upstream A and a downstream A/G. C was also present in the -3 position more often than expected by chance. While the context and structural preferences for ssDNA targets and editing by A3A appear consistent, it is not clear if the changing backbone of the nucleic acid target may impact the preferences for individual APOBECs. Investigation similar to McDaniel et al. but for RNA targets would be useful, however these may be difficult given the paucity of *in vitro* models in which to study A3 RNA editing.

Many of the A3 proteins are IFN-inducible. Treatment with IFN-α has been shown to upregulate A3B, A3C, A3F, and A3G in hepatocytes and A3A, A3F, and A3G in monocyte-derived dendritic cells and macrophages.[151–153] Treatment

with IFN-β, but not IFN-γ, induced A3A expression in uninfected decidual tissue organ cultures.[154] An IFN-driven phenotype may partially explain the challenge and cell type specificity in measuring A3 RNA deaminase activity *in vitro* if it is dependent on the expression of cofactors also upregulated by IFN.

2.2 Deaminase-Dependent Restriction of Viruses and Mobile Genetic Elements

2.2.1 Human Immunodeficiency Virus (HIV-1)

The inhibition of HIV-1 replication by A3D, A3F, A3G, and A3H is the most extensively studied of the viral restriction phenotype of the APOBEC family and has been well reviewed elsewhere.[155–160] During production of new virions in an infected host cell, A3 proteins form ribonucleoprotein complexes with HIV-1 RNA and the nucleocapsid domain of the Gag protein. The A3 proteins are subsequently packaged and encapsidated in the mature HIV-1 virions. The major hypermutation events take place during subsequent cellular infection. After viral entry and uncoating, A3 proteins co-localize with HIV-1 RNA during reverse transcription of nascent negative-strand viral complementary DNA (cDNA). This single-stranded cDNA acts as the template for plus-strand genomic DNA synthesis and is highly susceptible to the cytidine deamination activities of the packaged A3 proteins. The C>U transitions in the template strand (negative-sense strand) of the cDNA result in G>A transitions in the genomic strand (positive-sense strand).[161] Many hypermutated proviruses are degraded during or after generation of the cDNA template strand, potentially through a pathway mediated by uracil DNA glycosylase-2, and fail to integrate into the host cell genome.[162–165] A large proportion of those hypermutated proviruses that do successfully integrate are rendered defective due to an abundance of mutated start codons and nonsense mutations in several open reading frames.[166–168]

HIV-1 and other retroviruses encode a potent counteractor of APOBEC-mediated restriction – the accessory protein viral infectivity factor (Vif). The early 2000s were marked by a cascade of information regarding the interaction between Vif

and A3G.[169–177] Per the current model, Vif recruits an E3 ubiquitin ligase complex composed of Elongin B/C, cullin 5, RING-box subunit 2 (Rbx2), and core-binding factor beta (CBFβ) to polyubiquitinate A3G, leading to degradation via the host proteasome.[178–181] Vif-like proteins have been described for other lentiviruses including simian immunodeficiency virus, feline immunodeficiency virus, bovine immunodeficiency virus, jembrana disease virus, visna/maedi virus, and caprine arthritis encephalitis virus.[182–187] *Spumaretrovirinae*, another subfamily of retroviruses, infect a diverse number of mammals and encode the accessory protein Bet, which impairs the activity of A3G through direct binding and sequestration rather than targeted degradation.[188–190] APOBEC antagonism by lentivirus-encoded proteins strongly suggests active antiviral APOBEC activity in these representative species. As an alternative mechanism, Human T-cell leukemia virus type 1 (HTLV-1) avoids APOBEC activity simply through exclusion of A3G from the packaged virus.[191] The diverse approaches of retroviruses to avoid A3 activity highlight the importance of this subfamily for restricting this class of viruses.

2.2.2 Retrotransposons and Mobile Genetic Elements

Transposable elements (TEs) are selfish genetic elements that are able to move within the genome via one of two methods.[192–194] Class I TEs (retrotransposons) replicate through a "copy and paste" mechanism wherein reverse transcription of an RNA intermediate produces a cDNA copy of the genetic element that integrates elsewhere in the genome. Class II TEs (DNA transposons) utilize a "cut and paste" mechanism via the activity of an encoded transposase enzyme and do not require RNA intermediaries. Therefore, the class I TEs represent a suitable model through which to understand APOBEC restriction of RNA targets. The evolution of the human genome has been shaped by mobile genetic elements, and estimates suggest that 46% of it may be comprised of inactivated TEs.[195, 196] In humans, only a subset of class I TE elements (long interspersed element-1s (LINE-1s) and short interspersed nuclear elements (SINEs)) remain retrotransposition-competent.[197–199] The majority of APOBEC activity on retroelements has been

demonstrated on LINE-1 elements *in vitro*. A3A inhibits LINE-1 retrotransposition through a deaminase-dependent mechanism while AID, A1, A3B, A3C, and A3F act through a deaminase-independent pathway.[200–203] It has been hypothesized that the massive reduction in active LINE-1 elements in the primate taxon compared to Rodentia could be the result of extensive duplication and expansion of the A3 gene.[199, 204] However, analysis of the depletion of A3 targeted sequence motifs in human retroelements only demonstrated the "footprint" of A3 activity on a subset of endogenous retroviruses and not in LINE-1 or SINE elements.[205]

2.2.3 DNA viruses

Beyond retroviruses and TEs, APOBEC-driven C>U/G>A mutations have been demonstrated in a diverse set of ssDNA and dsDNA viruses.[206] A3B, A3C, A3F, and A3G have been demonstrated to deaminate both the ssDNA negative-sense and positive-sense strand of the pararetrovirus Hepatitis B virus (HBV) *in vitro* and *in vivo* resulting in a low proportion of G>A hypermutated genomes.[207] *In vitro* work has also demonstrated that AID and A1 can edit HBV DNA and HBV DNA and RNA, respectively.[208, 209] Murine A1 is canonically expressed in hepatic cells and, in a mouse model of HBV infection, was shown to edit transgenic HBV genomes.[210] The low endogenous expression of AID and A1 in the human cells canonically infected with HBV limits the interpretability of their contribution to hypermutation, although AID been shown to be upregulated in human hepatocytes upon infection with HCV *in vitro*.[165, 211, 212] Human papillomaviruses (HPV), members of the *Papillomaviridae* family of dsDNA viruses, have also been shown *in vitro* to be edited by A3A, A3C, and A3H in transfected HEK293T cells, while overexpressed A3A, but not A3B or A3C, was able to restrict HPV replication in immortalized keratinocytes.[208, 213] Bioinformatic analysis of the HPV genome suggests that the preferred target of A3 enzymes (5'-Tp<u>C</u>-3') is significantly depleted, consistent with long-term adaptation to infection of A3 expressing tissues.[214] Other DNA viruses shown to be hypermutated by A3 proteins included the single-stranded

circular Transfusion Transmitted Virus (TTV) and the dsDNA HSV-1 and Epstein-Barr Viruses.[215, 216] Direct A3 interaction with TTV was not demonstrated, but G>A hypermutation was found preferentially in the 5'-TpC-3' or 5'-CpC-3' contexts on the template strand, consistent with A3 activity.

2.2.4 RNA Viruses

Although the first description of APOBEC editing activity was on the cellular ApoB mRNA template, direct evidence of APOBEC-driven deamination in RNA viruses has been problematic to demonstrate. The earliest evidence came in 2004 when a rodent A1 homolog was shown to edit HIV-1 RNA *in vitro*.[217] Recently, it was shown that the restriction of HCoV-NL63 by A3C, A3F, and A3H was deaminase-dependent, but, in this case, the authors were unable to demonstrate that this resulted from genome hypermutation even after serial passaging of HCoV-NL63 in A3 overexpressing cells.[218] Furthermore, while bioinformatic analysis of a large dataset of human-infective viruses demonstrated significant depletion of A3 target motifs (5'-TpC-3') among members of the the DNA virus families *Parvoviridae*, *Herpesviridae*, and *Papillomaviridae*, there was little evidence for this in RNA virus genomes with the exception of the seasonal coronaviruses HCoV-NL63, HCoV-OC43, HCoV-HKU1, and HCoV-229E (this signature was not detected in MERS-CoV, SARS-CoV, SARS-CoV-2, or their zoonotic source viruses).[219] An alternative analysis investigating the overabundance of C>U changes compared to U>C found strong evidence of the mutational asymmetry in HCV (all genotypes), Foot-and-Mouth Disease virus type A, canine kobuvirus, rubella virus, and all the human-infective coronaviruses.[220] The authors found a 5' U was associated with the greater C>U/U>C asymmetry, consistent with the context preference of most A3 proteins. Analysis of immunodeficiency-related vaccine-derived rubella virus in the granulomas of infected children did demonstrate a preponderance of C>U transitions in the context of 5'-UpC-3', accounting for 30% of all mutations when compared to the virus strain used in vaccine manufacture.[221] The enrichment for transitions was strongest in the 5'-UpCpA-3' context.

2.2.5 Deaminase-Independent Restriction and Insensitivity

Deaminase-independent restriction has been demonstrated primarily using overexpression models of catalytically inactive A3 mutants. This phenomenon has been reported for many viruses, including HIV-1's RNA stage, measles, mumps, RSV, HCoV-NL63, adeno-associated virus, enterovirus A71, and minute virus of mice.[222–224] LINE-1 is also inhibited by A3B and A3F in a deaminase-independent manner.[225] In HIV-1, the prevailing model for this restriction is binding of genomic RNA inhibiting the initiation of reverse transcription, which may also explain the phenotype in LINE-1. The mechanism in other RNA and DNA viruses is not well described. Influenza virus and vaccinia virus are not restricted by A3 overexpression and avoid targeting through unknown mechanisms.[226, 227]

Both APOBEC deaminase-dependent and -independent mechanisms impact the effective host range of a pathogen and will apply a selective pressure driving evolution. However, only deaminase-dependent restriction will leave identifiable footprints on the virus genomes; the loss of cytosines to deamination over the long-term will decrease the GC content of the virus genome and specific dinucleotides (such as 5'-Up<u>C</u>-'3) would be depleted over time, as discussed later.

2.3 APOBEC Origins, Diversification, and Evolution

The APOBEC family of proteins contain deaminase homologs throughout the tree of life. Understanding the role of APOBEC-driven natural selection on pathogens requires knowledge of in which species the selective pressure may or may not be present. Those species that share intracellular selective pressures may represent hosts more readily amenable for spillover events. Accurate characterization of the specific components of the APOBEC proteins responsible for antiviral activity against RNA viruses would enable prediction of the extant species containing APOBEC homologs retaining this activity. In turn, this will enable better understanding of the natural

host species and effective host range of SARS-CoV-2. The APOBEC family is highly diverse and not well characterized in the majority of species, limiting this effort.

2.3.1 Origin

The APOBEC family is related to bacterial, yeast, and plant deaminases all possessing highly conserved amino acid motifs responsible for coordination of zinc in the active site (Figure 2.1).[228–230] Bioinformatic analysis suggests that A4 may be the ancient member of the APOBEC family, emerging as early at the last eukaryotic common ancestor (LECA), as homologs have been identified in the phylum Cnidaria and occasionally in algal lineages.[231] Another relatively recent study identified, cloned, and demonstrated deaminase activity in AID homologs from sea urchins and Brachiopods.[232] The AID homologs showed increased expression in barrier tissues and were upregulated in response to bacterial challenge, suggesting a role in antimicrobial defense. Ancestral members of the family at the point of vertebrate speciation include both AID and A2, and homologs of the two proteins have been identified bioinformatically in birds, amphibians, and ray-finned fish (zebrafish and pufferfish), with an AID homolog responsible for diversity in variable lymphocyte receptors of jawless vertebrates.[228, 233] A1 and A3 both arose from independent duplication events of AID.[228] A1 has been identified in birds, reptiles, amphibians, and lungfish, suggesting it arose prior to the tetrapod-lungfish divergence.[231] A3 emerged from AID duplication in placental mammals and has undergone frequent duplication, fusion, and loss in different branches of the clade.[231, 234]

2.3.2 A3 Paralog Expansion

A3 homologs are not present in marsupials or monotremes and appear to be exclusive to placental mammals.[235–237] The A3 proteins contains three paralogs (A3Z1, A3Z2, A3Z3) that are designated according to conservation of amino acid residues within the ZDD (Figure 2.2A).[238] From the most recent common ancestor of the three A3 paralogs, A3Z3 may be the most ancient having diverged approximately 250 million years ago with A3Z1 and A3Z2 diverging from each other 50 million years

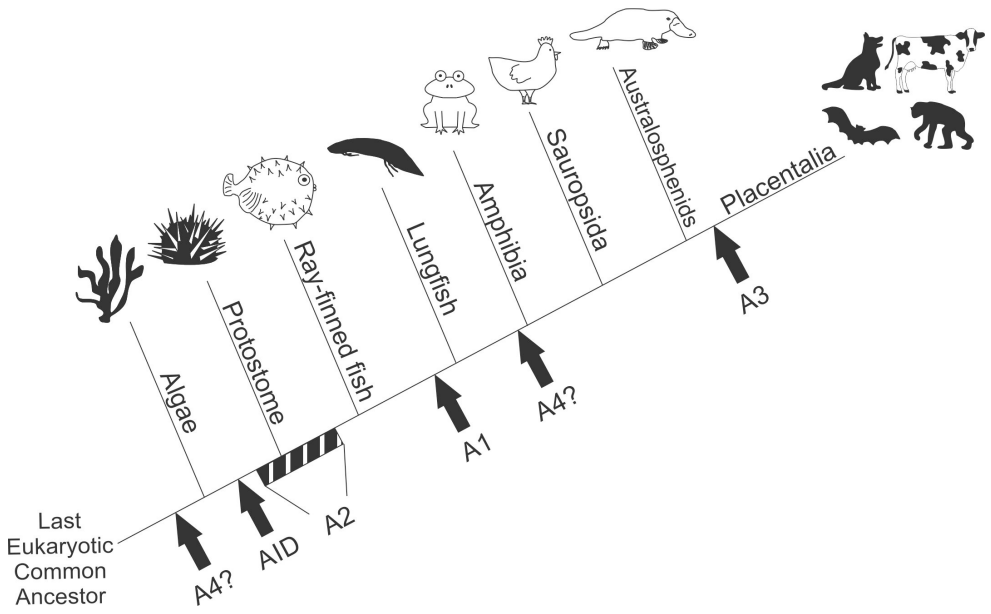


Figure 2.1: Emergence of APOBEC proteins. Branch length is not drawn to timescale. Question marks for A4 protein emergence represent a lack of concrete evidence at the divergence of amphibians and putative evidence in algae species. Black for A2 represents uncertainty on emergence prior to protostome – deuterostome split. A1 and A3 both arose from duplication of AID.

later.[234] It is hypothesized that the most recent common ancestor of placental mammals contained a gene set of A3Z1-A3Z2-A3Z3 that, through multiple processes of differential evolution, have led to the diversity seen in extant species (Figure 2.2B). Mapping and identification of A3 genes takes advantage the presence of genes CBX6 and CBX7 upstream and downstream of the gene locus, respectively. Particular attention will be played to species known to be hosts for coronaviruses: bats and humans.

The complexity of APOBEC expression in bats has recently been revealed.[239] Within the pteropid family, eighteen A3 proteins were identified, of which twelve are single deaminase A3Z1, one is a single deaminase A3Z2, four are composed of a single deaminase of a novel paralog termed 'A3Z2B', and one is a double deaminase of A3Z2B-Z3. Active expression of 13 of these proteins was identified in spleen tissue extracted from a black flying fox. The large expansion of Z1 is particularly unusual as this paralog has been independently lost in the mammalian order Rodentia and the families *Felidae* and *Suidae* (domestic and wild pigs) (Figure 2.2B).[240]

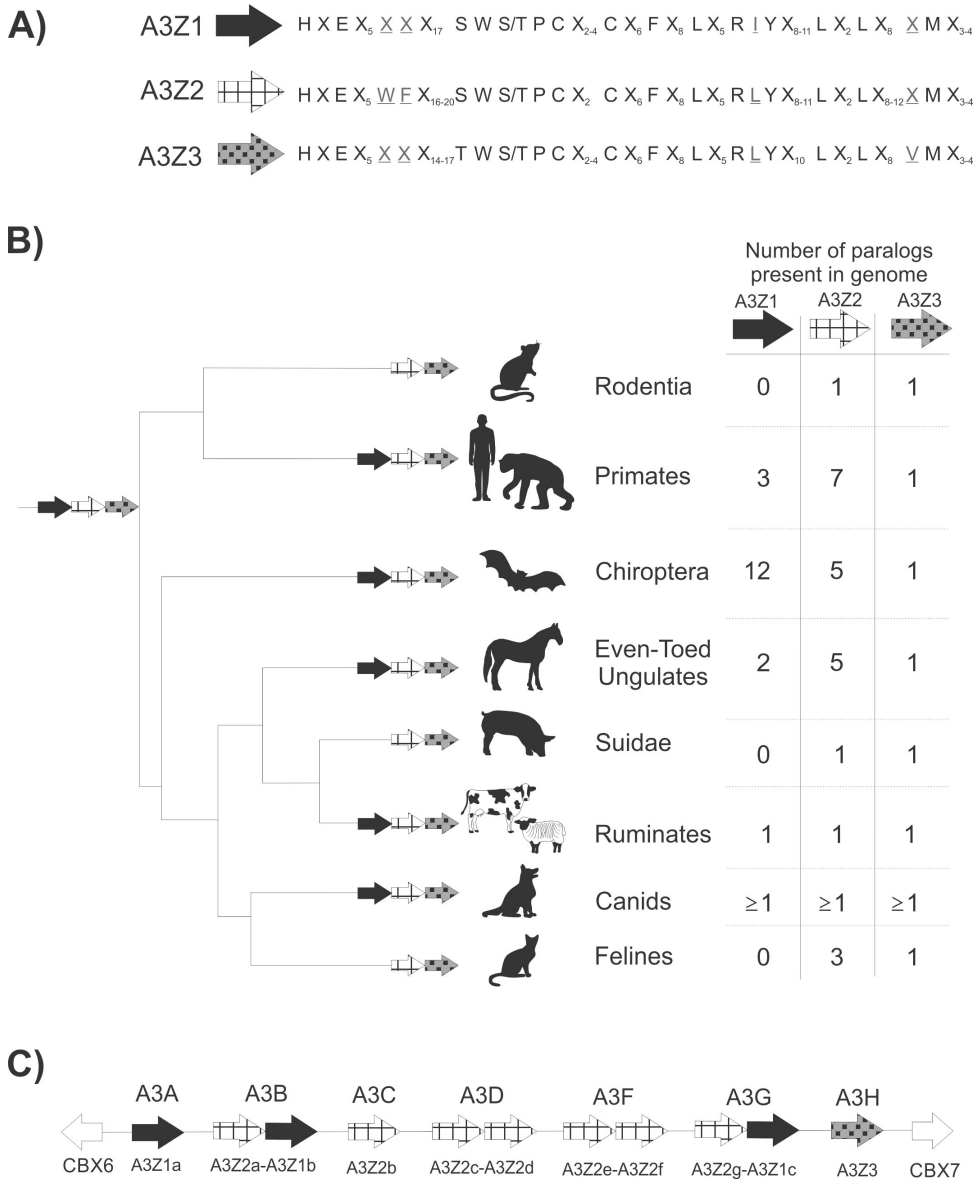


Figure 2.2: Expansion of A3 Paralogs in Placental Mammals. (A) Amino acid motifs within the ZDD of the three A3 paralogs. Differences highlighted in grey and underlined. (B) Expansion of paralogs in different branches of placental mammals. A3Z1 has been independently lost in several taxonomic classes. Branch length is not drawn according to timescale. Table lists number of paralogs and not number of gene products. Taxonomic relationships designed in consultation with the Tree of Life web project. (C) Diagram of the A3 gene locus in humans and primates. Adjacent paralogs (e.g., A3Z2a-A3Z1b) represent fused paralogs in a single gene product.

The A3 locus and its evolution in humans compared to other mammals is thoroughly described in Münk et al. 2012 (Figure 2.2C).[234] In primates, the A3Z2 paralog has seen the greatest expansion, being present in all dual-deaminase A3 proteins. The current organization likely arose through as many as 6 gene duplication events. In all known species, there is no evidence for duplication of the A3Z3 gene.

2.3.3 APOBEC Protein Evolution

The long-term evolution of the individual APOBEC proteins is of particular interest as it may indicate which perform essential activities for host survival (and demonstrate purifying/negative selection) and which may be involved in host-pathogen interactions (and demonstrate positive/directional selection). Current evidence for the overall evolutionary trajectory is summarized in Table 2.1. When evaluating the evolution of AID using representatives from seven mammalian clades, the protein was shown to be under strict purifying evolution.[234] Repeating the analysis for the A3Z1, A3Z2, and A3Z2 paralogs demonstrated multimodal evolution and sliding-window analysis indicated that certain regions of the gene are under purifying selection and others are under positive selection. This multimodal relationship for A1 was also present but to a lesser degree than with the A3 paralogs. It is likely that the regions under purifying selection are necessary for the protein's function but the regions under positive selection allow for diversity in targets or are the targets of antagonism by viral proteins. Similar analysis with a more limited subset of clades indicated that A3Z1, A3Z2, and A3Z3 all showed dN/dS ratios greater than 1, consistent with overall positive selection.[240] When comparing whole genes between humans and the closely-related chimpanzees, AID and A3A were under purifying selection, A3B, A3D, and A3G were under positive selection, and A3F was under positive selection in a sliding window analysis only.[241] Comparing humans to a broader range of primates, A2 was under purifying selection and A3C was under positive selection in a sliding window analysis. A1 was under purifying selection in whole gene comparisons and positive selection under sliding window, consistent with the multimodal results seen above. Curiously, the sites under

Name	Taxonomic Emergence	Evolutionary Pressure	Deaminase Domains	Cellular Localization[a]	Preferred target sequence (principally defined in ss-DNA)
A1	Chordata	Multimodal (all mammals), purifying (whole gene), positive (sliding window)	1	n/C	5'-ApC-3'
AID	Deuterostomes	Purifying	1	n/C	5'-WpRpC-3'
A2	Bony vertebrates	Positive	?	N/C	?
A4	LECA?	?	?	?	?
A3Z1	Placental Mammals	Multimodal	1	?	?
A3Z2	Placental Mammals	Multimodal	1	?	?
A3Z3	Placental Mammals	Multimodal	1	?	?
A3A	Primates	Purifying	1	N/C	5'-TpC-3'
A3B	Primates	Positive	2	N	5'-TpC-3'
A3C	Primates	Positive (sliding window)	1	N/C	5'-TpC-3'
A3D	Primates	Positive	2	C	5'-TpC-3'
A3F	Primates	Positive (sliding window)	2	C	5'-TC-3'
A3G	Primates	Positive	2	C	5'-CpC-3'
A3H	Primates	Unclear	1	N/C	5'-TpC-3'

Table 2.1: Overview of APOBEC Family Proteins. [a]adapted from Salter et al. 2016. C, predominantly cytoplasmic; n/C, can traffic to nucleus but predominantly cytoplasmic; N/C, even localization; N, predominantly nuclear.

positive selection for A3G were not isolated to the regions targeted by the HIV-1 Vif protein as one may expect if the positive selection limited retroviral antagonism escape only. Altogether, the evidence implicates the A3 proteins and potentially some functionality of A1 as being involved in the pathogens arms race and being suitable targets for further investigation in an RNA virus context.

2.4 Bioinformatic Evidence for RNA Editing in SARS-CoV-2

2.4.1 C>U Mutational Asymmetry in SARS-CoV-2

Coronaviridae, in common with other members of the *Nidovirales* order, encode a 3'-5' exonuclease that dramatically reduces the effective error rate of the viral RdRP.[242–244] Consequently, coronaviruses demonstrate a low substitution rate and minor differences in sequence diversity despite a genome size of nearly 30kb. Study of the short- and medium-term evolution of SARS-CoV-2 has been facilitated by the large number of complete genome sequences that have accumulated since the start of the COVID-19 pandemic on the GISAID platform.[245]

The accuracy of the sequences obtained by a range of whole genome sequencing technologies, combined with minimal evolutionary diversification in the few months after the start of its global spread, provide a unique opportunity to investigate mutational events at very high resolution early in the pandemic. Most observed published analyses of SARS-CoV-2 genomic sequence data report a preponderance of C>U transitions in the viral genome. An early analysis of 1,001 SARS-CoV-2 sequences indicated that 38 – 42% of all mutations in the queried datasets were due to C>U changes and that the ratio of C>U to U>C transitions was nearly six times higher than expected.[246] Demonstrating that this was not an artefact of the sequencing or bioinformatics methods used, a parallel analysis of datasets of EBOV from historical strains and the 2014-16 West African outbreak demonstrated no such mutational asymmetry. In Kilmczak et al. (2020), an analysis of 32,115 isolates deposited on GenBank by late June 2020 tabled 251,273 mutations when compared to a reference MN908947 genome.[247] Once duplicate mutations were removed and 12,156 independent mutational events identified, C>U and A>G mutations were shown to be the most prevalent single base substitutions. However, the removal of duplicate mutations may have biased the analysis against detection of independently occurring but identical (homoplastic) C>U transitions at context-favored sites, as documented extensively in other SARS-CoV-2 mutational analyses.[246] The phenomenon is not restricted to transmission in humans, as a 3.5-fold asymmetry in C>U/U>C mutations was similarly noted in samples from infected minks that had arisen during its transmission within mink farms in the Netherlands.[248]

Most early studies of mutation frequencies were based upon the characterization of substitutions away from the SARS-CoV-2 global consensus sequence in individual genome sequences. Other researchers have, instead, investigated intra-host diversity using read level data available from next-generation sequencing (NGS). These methods likely provide a better snapshot of the mutational spectrum of SARS-CoV-2 genomic sequences; for example, lethal mutations driven by C>U changes that were likely to be lethal can be readily identified. These methods are also absent of the tight bottlenecks of person-to-person transmission that may limit the

fixation of low frequency C>U changes, regardless of impact on fitness.[249, 250] Graundenzi et al. (2021) analyzed the accumulation of minor population variants from deep sequencing on a large dataset of 1,133 swab samples.[251] Far and away the most abundant mutation within their cohort was C>U; 28% of all cytosines demonstrated detectable C>U substitutions. In fact, C>U or G>A mutations were responsible for 81.2% of the observed mutational signatures following a non-negative matrix factorization approach. The authors hypothesized the observed mutational signatures were the combined impact of APOBEC, reaction oxygen species, and/or ADAR. Similarly, within their dataset of 1,181 samples Tonkin-Hill et al. (2021) found C>U mutations accounting for 47.9% of all within-host point mutations.[252] The authors also found a 9.6-fold increase in normalized C>U mutations rates on the positive strand compared to the negative strand, consistent with RNA editing or RNA damage mechanisms rather than replication errors.

In some of the analyses, the locations of C>U transition were loosely associated with the base and structure context of RNA deamination favored by A3 proteins. Base context in Simmonds (2020), which analyzed consensus sequences, and Di Giorgio et al. (2020), which used read level data from whole-genome sequencing of bronchoalveolar lavage fluid (BALF) samples, indicated high preference for C>U deamination in the presence of upstream and downstream A or U.[246, 253] Klimczak et al. (2020) analyzed preference in trinucleotide motifs and found strong preference for 5'-Up<u>C</u>-3' with total ablation of the phenotype in the context of an upstream C.[247] This extended context compared to the standard preference of several APOBECs for 5'-Up<u>C</u>-3' was similar to the A3A-driven editing of cellular mRNA sequences and in A3A-mediated mutations in ssDNA templates.[148, 254] Graudenzi at al. (2021), however, found no evidence for any base context effect in their analysis.[251] Lastly, in Tonkin-Hill et al. (2021), the trinucleotides 5'-Up<u>C</u>pG-3' and 5'-Up<u>C</u>pA had the highest accumulation rate of C>U changes when normalized by sequence composition.[252] 5'-Gp<u>C</u>pC and 5'-Cp<u>C</u>pU had the lowest rates.

The other factor appearing to influence site preference in RNA sequences is the context of the target cytosine in surrounding RNA secondary structure elements.

Over half of the sites in human mRNA that were edited by A3A were flanked by short palindromic sequences and typically located in the unpaired termini of stem-loops.[148] Supporting this, an analysis of RNA structural contexts of edited sites in SARS-CoV-2 and rubella identified preferential C>U mutations in terminal loop compared to stem sequences.[247, 255]

2.5 Functional Evidence for a Role for APOBEC in Editing and Antiviral Restriction of Coronaviruses

The observed excess of C>U transitions, the nature of the 5' and 3' contexts where such transitions occur, and the previously documented ability of several APOBEC paralogs to edit RNA templates make a compelling case for a role of APOBEC in editing of coronavirus genomes. However, functional studies have largely been deficient in demonstrating a role for APOBEC proteins in generated SARS-CoV-2 viral diversity.

In 2018, human A3 and HCoV-NL63 were convincingly demonstrated to interact but without clear evidence of systematically biased RNA editing.[218] Transfection of LLC-Mk2 cells (rhesus macaque kidney) with individual A3 proteins revealed deaminase-dependent virus restriction by A3C, A3F, and A3H but not A3A, A3E, or A3G. C>U and G>A point mutations were identified in virus serially passaged in A3A, A3C, A3E, A3F, A3G, and A3H overexpressing cells but did not result in hypermutation of progeny virus. Potentially recapitulating interactions of APOBEC during assembly and incorporation into retroviral particles, A3C, A3F, and A3H co-localized with the coronavirus nucleocapsid protein through a catalytically-independent mechanism, while A3A, A3D, and A3G did not co-localize, perfectly matching their differences in virus restriction.

Protein-protein interactions between A3 proteins and coronaviruses have furthermore been previously demonstrated in chimeric HIV-1 expressing SARS-CoV or HCoV-229E nucleocapsid protein.[256] In these experiments, A3G packaged in both chimeric virions; packaging was dependent on the carboxy-terminal self-association

domain of the coronavirus nucleocapsid protein and not through binding to genomic RNA. Viral-like particles composed of SARS-CoV nucleocapsid and membrane proteins also co-packaged A3G without the presence of viral RNA.

Direct C>U editing of 200 nt-long SARS-CoV-2 RNA fragments in vitro has recently been reported for A1 (with co-factor ACF), A3A, and A3G.[257] C>U mutation rates were elevated compared to all other mutations in cells transfected with A1 or A3A. C>U editing in cells transfected with A3G was not elevated above the rate of G>A. Both A1+ACF and A3A editing had strong context dependence consistent with their previously described preferences. A1 preferred a 5'-ACu/a-3' context and A3A preferred 5'-UCa/u-3'. Through the use of splice-specific sequencing primers, the authors demonstrated the editing was taking place on the RNA rather than the DNA template being transcribed. The impact on viral fitness was less clear. Doxycycline-induced expression of A1+ACF, A3A, and A3G did not limit viral replication in Caco-2 cells as measured by plaque assay or reverse transcription quantitative polymerase chain reaction (RT-qPCR). Perhaps counterintuitively, A3A expression led to a statistically significant increase in plaque-forming units after 72 hours. The viruses demonstrating increased titer in the presence of A3A were not sequenced, limiting the interpretability of the role of C>U editing on SARS-CoV-2 fitness *in vitro* or *in vivo*. Further studies identifying C>U editing sites across a broader range of A3 proteins in live virus assays would clarify the functional role of APOBEC on SARS-CoV-2 fitness.

2.6 Evolutionary Consequences of C>U Mutational Bias in Coronaviruses

Irrespective of the underlying mutational mechanism, there exists substantial evidence for a potent editing process that drives C>U mutations into the genomes of coronaviruses and other RNA viruses in certain favored sequence and structural contexts. This driver of virus evolution is quite distinct from the occurrence of random mutations through misincorporation errors during transcription. Apart

from the bias of mutational patterns to C>U changes (or G>A if the negative-sense strand is edited), these sequence changes may also primarily be distributed in mutational 'hotspots' that may lead to the appearance of convergent changes between otherwise genetically unlinked strains. For example, multiple C>U mutations occurred independently among variants of SARS-CoV-2 that were distributed into different parts of a phylogenetic tree, and therefore conflicted with relationships reconstructed from phylogenetically informative sites.[246] A high frequency of homoplastic mutations were similarly described in a larger analysis of 7,666 SARS-CoV-2 genome sequences.[258]

The occurrence of C>U hypermutation in SARS-CoV-2 also has a substantial skewing effect on relative frequencies of nonsynonymous to synonymous substitutions and dN/dS ratios. As introduced in Chapter One, dN/dS ratios can roughly be used to detect selection and, in a neutrally evolving system, should approximate to 1. Comparative analyses of virus sequence datasets typically reveal low dN/dS ratios indicative of strong purifying selection, such as the 0.1 ratio found in EBOV during its 2016 epidemic spread in Central and West Africa or the values between 0.1 and 0.2 for the seasonal HCoVs (see Table 1.1).[246] Contrastingly, the high dN/dS ratio between strains (∼0.6-0.7)[246, 258] or between SARS-CoV-2 and the most closely related bat sarbecovirus, RaTG13,[259] are indicative of both the high rates of positive selection as SARS-CoV-2 adapts to a new host and the contribution of elevated mutation rates possibly due to the effects of host innate antiviral effector mechanisms, such as APOBEC.[246, 258, 259]

Furthermore, a substantial proportion of the amino acid changes that emerge within SARS-CoV-2 populations originate from nonsynonymous changes induced by C>U transitions, rather than necessarily representing evolutionarily selected phenotypic changes that might be expected to occur as the virus adapts to replication and transmission in a new host.[260] In the Simmonds (2020) analysis, >50% of observed substitutions were due to C>U changes.[246] The occurrence of C>U driven amino acid changes with mildly deleterious phenotypes and associated pressures for reversion considerably complicate analyses of the critical protein changes

associated with human host adaptation. Importantly, the observed preponderance of nonsynonymous mutations induced by APOBEC-like transitions does not challenge the assumption that APOBEC activity would be randomly distributed throughout the genome. For all codon triplets, substitutions at the second position result in nonsynonymous mutations. In the first and third positions of a genome with equal nucleotide frequencies, only 5% of substitutions in the first position and 76% of substitutions in the third position would result in a synonymous mutation.[72] These numbers are elevated for C>U transitions primarily due to the wobble effect in the third position, being synonymous substitutions for 12.5% and 100%, respectively. Of the 48 possible C>U transitions for the 37 codons that contain a cytosine, only 18 of the 48 (37.5%) would result in a synonymous substitution. While these calculations must be applied to a given sequence's observed codon distribution, in almost all instances equivalent rates of nonsynonymous mutations/nonsynonymous site and synonymous mutations/synonymous site will result in a greater total count of nonsynonymous substitutions than synonymous substitutions.

The occurrence of large numbers of homoplastic sites similarly complicates analyses of sites under selection. Finally, the sheer number of excess C>U changes observed in SARS-CoV-2 genomes (>40% of observed substitutions in some datasets) has a distorting effect on estimates of SARS-CoV-2 substitution rates, particularly as many sites may rapidly revert over time and the number of target sites for C>U mutation may deplete over time.[258]

The long-term effects of the C>U/U>C transition asymmetry on virus diversification are difficult to predict. However, the compositional abnormalities of other coronaviruses, particularly seasonal HCoVs, might plausibly be attributed to the long term outcome of prolonged C>U editing.[261, 262] For example, the enrichment of uracil (40%), depletion of cytosine (13%) and intermediate frequencies of adenine and guanine bases (28% and 19%) in HCoV-HKU1 has been proposed to have been driven by APOBEC associated mutations.[262] Of potential relevance to exploring differences in APOBEC activity in varied hosts, there is no compositional asymmetry and a relatively higher GC content within bat sarbecoviruses (from which

SARS-CoV-2 likely derives) compared to HCoVs. The intensity of C>U mutational pressure observed in human SARS-CoV-2 sequences since its zoonotic emergence may therefore reflect a less permissive, hostile internal cellular environment in human cells than might be found in a better co-adapted, virus-tolerized immune system of the Chiropteran order.[263] The pronounced activity of APOBEC in human cells may derive from an inability of a bat-adapted virus to counteract the human APOBEC editing pathway. The varying degrees of base asymmetry and GC contents of seasonal HCoVs may then be indicative of the varying times they have circulated in humans following their acquisition from bats or other hosts without APOBEC-driven editing.[264] SARS-CoV-2 may follow their evolutionary trajectory as infections become permanently established in human populations.

2.7 A Perspective on Future Functional Investigations of APOBEC-Mediated Editing of SARS-CoV-2 and Other RNA Viruses

APOBECs have been shown to be important for host and retroviral function, and the possibility that they play a role in RNA viral evolution would substantially expand the scope of their antiviral activities. However, part of the difficulty in studying the process is the plethora of different APOBECs in humans with different RNA and DNA targeting specificities, rates of induction by chemokines and cytokines, and tissue and cellular expression. Furthermore, the number and identities of APOBECs is highly variable in mammals; for example, the mouse, which might otherwise represent a suitable *in vivo* model for APOBEC function in whole host defense, possesses only one A3 gene compared to seven or more in primates and bats.

Functional studies using human cell lines are similarly hampered by the current lack of information of which APOBEC protein might mediate the observed RNA editing in coronaviruses and in other RNA viruses, or whether their might exist currently uncharacterized defense pathways mediating the observed RNA editing. Among the members of the diverse family, A1 and A3, with their evidence for extensive positive selection during mammalian evolution, represent the frontrunners

for deaminase-dependent mutation of RNA viruses. A1 is an unlikely candidate for SARS-CoV-2 editing *in vivo*, as the restricted hepatic expression of the protein limits its ability to mutate what is principally a respiratory pathogen. Among the A3 proteins, it may be possible to exclude A3G from contributing to SARS-CoV-2 deamination as the base context favored by this protein does not match the observed phenotype. However, this exclusion may not be warranted for other viruses. Beyond that, one may consider intracellular distributions of this proteins as limiting factors as SARS-CoV-2 and other coronaviruses spend their entire intracellular life cycles in the host cytoplasm, but first the intracellular distributions of the A3 proteins in different activation contexts should be better defined. Further, it is conceivable that, similar to HIV-1, any editing of SARS-CoV-2 RNA may be enhanced by packaging of A3 proteins in the viral capsid with the bulk of deleterious editing taking place in the subsequent cellular infection. In this way, experiments aimed at identifying protein-protein interactions between the viral capsid and host proteins may assist in narrowing the breadth of proteins that should be investigated. In all cases, *trans* complementation of the editing phenotype by gene transfection may be difficult to achieve *in vitro* if there are functionally required interacting cofactors or activated intracellular states are necessary.

The knowledge gained from this line of inquiry would better define the discordant innate antiviral pathways between organisms. To spillover to new hosts, zoonotic viruses must be able to counteract any potent antiviral mechanisms in the new hosts' cells for sufficient time to produce viable progeny virions. APOBEC serves both as an innate immune pathway for RNA virus restriction and potentially a driver of their evolution as they circulate in human populations. Robust, functional evidence of this phenotype would represent a significant step forward in our understanding both of the barriers to zoonotic spillover and the short- and long-term evolution of recently emergent human-infective RNA viruses.

3 Methods

Contents

3.1 Methods used in Chapters Four, Five, Six, and Seven 41
3.1.1 Clinical Samples . . . 41
3.1.2 Viral RNA Extraction . . . 42
3.1.3 Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR) . . . 43
3.1.4 Next-Generation Sequencing (NGS) . . . 44
3.1.5 Bioinformatic Processing . . . 44
3.2 Methods used in Chapters Four, Five, and Six . . . 45
3.2.1 Nested Polymerase Chain Reaction Restriction-Fragment Length Polymorphism (PCR-RFLP) . . . 45
3.2.2 Allele-Specific Probe Polymerase Chain Reaction (ASP-PCR) . . . 46
3.2.3 SARS-CoV-2 Consensus Sequence Lineage Assignment . 47
3.3 Methods used in Chapter Four . . . 48
3.3.1 Assessment of NGS Success . . . 48
3.3.2 Probit Analysis . . . 48
3.3.3 Predictive Performance . . . 48
3.3.4 Statistical Analysis . . . 48
3.4 Methods used in Chapter Five . . . 49
3.4.1 Qualitative Enzyme-Linked Immunosorbent Assay (ELISA) 49
3.4.2 Subgenomic RNA Reverse Transcription Polymerase Chain Reaction . . . 49
3.4.3 Baseline Viral Load Analysis . . . 49
3.4.4 Intensive Viral Dynamics Analysis . . . 51
3.5 Methods used in Chapter Six . . . 51
3.5.1 Disease Severity and Outcome Analysis . . . 51
3.5.2 Variant and N:KR Emergence Analysis . . . 52
3.6 Methods used in Chapter Seven . . . 52

3.6.1 Consensus Sequence Analysis 52
3.6.2 Mutation Rate Statistical Analysis 52
3.6.3 Mutation and Context Statistical Analysis 54
3.6.4 Homoplastic SNP Occurrence Estimation 54
3.7 Methods used in Chapter Eight 55
3.7.1 Virus Sequences and Characteristics 55
3.7.2 Dinucleotide Analysis 56

3.1 Methods used in Chapters Four, Five, Six, and Seven

3.1.1 Clinical Samples

Samples used for method development and assessment in Chapters Four, Five, Six, and Seven were collected as part of two COVID-19 clinical trials:

REMAP-CAP

The Randomised, Embedded, Multi-factorial, Adaptive Platform Trial for Community-Acquired Pneumonia (REMAP-CAP) Immunological Domain was an international, open-label, randomized convalescent plasma (CP) trial enrolling patients aged 18 years or older receiving intensive care-level organ support. Patients were eligible if enrolled into the trial within 48 hours of admission to an intensive treatment unit (ITU) with a prior positive SARS-CoV-2 microbiological test. Exclusion criteria included receiving pretrial antibody therapy (monoclonal antibody, hyperimmune immunoglobulin, or CP) or if 14 days had passed from their initial hospitalization. At time of enrollment, demographic and clinical features were collected by research staff. A full trial protocol and efficacy results for the trial have been published.[265, 266] Nasopharyngeal swabs and blood samples were taken prior to randomization (baseline/day 1 samples), shipped to a central testing laboratory, and stored at -80°C until processing. A subset of patients were sampled repeatedly on days 1, 2, 3, 4, 6, 9, 12, 15, and 28, or until ITU discharge or death.

COV003

COV003 is a randomized, controlled phase 3 multi-site trial that began in Brazil on June 23, 2020 assessing the efficacy of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2 infection/COVID-19. Efficacy, safety data, the full study protocol, and an exploratory analysis of lineage-specific efficacy are available elsewhere.[267–269] Trial participants were 18 years or older and recruited to do being at high risk of exposure (e.g., healthcare workers). Participants who developed primary COVID-19 symptoms were asked to contact their study site for sampling. Nasopharyngeal samples that were separately confirmed SARS-CoV-2 positive using commercial nucleic acid amplification tests (NAATs) were shipped to a central testing laboratory where RNA was extracted as described below. At the behest of the ethics committee, all patients were unblinded and offerred vaccination once efficacy was established. Samples continued to be collected after this date, including >90% of the samples later established to be infections with the Gamma variant.

Ethics

REMAP-CAP and COV003 were both conducted according to the principles of the latest version of the Declaration of Helsinki (version Fortaleza 2013). REMAP-CAP was performed in accordance with regulatory and legal requirements (EudraCT number: 2015-002340-14) and was approved by London-Surrey Borders Research Ethics Committee London Centre (18/LO/0660). Written or verbal informed consent, in accordance with regional legislation, was obtained from all patients or their surrogates. COV003 was approved by the Brazilian National Research Ethics Committee (ref: 32604920.5.0000.5505) and the Oxford Tropical Research Ethics Committee (ref: 20-36).

3.1.2 Viral RNA Extraction

SARS-CoV-2 RNA from REMAP-CAP nasopharyngeal samples was extracted using either the QIAamp Viral RNA Mini Kit (QIAGEN) or the Quick-DNA/RNA Viral Kit (Zymo Research). SARS-CoV-2 RNA from CoV003 nasopharyngeal samples

were extracted using the Quick-DNA/RNA Viral Kit. For the QIAamp Viral RNA Mini Kit, 140 µL of sample was processed according to the manufacturer's spin protocol, substituting carrier RNA with linear acrylamide (ThermoFisher Scientific) at a final concentration of 20 µg/mL, and eluted into 60 µL of buffer AVE. For the Quick-DNA/RNA Viral Kit (Zymo Research), 200 µL of sample was processed according to the manufacturer's protocol and eluted into 50 µL of nuclease-free water. Research reagent 19/304 (National Institute of Biological Standards and Control (NIBSC)) containing an equimolar mix of four HIV-1 particles each encapsulating RNA representing a partial fragment of the SARS-CoV-2 genome (GenBank reference MN908947.3). Sequences of the four packaged constructs are available on GenBank as accession numbers MT299802, MT299803, MT299804, and MT299805. SARS-CoV-2 RNA was extracted using the QIAamp Viral RNA Mini Kit, serially diluted in nuclease-free water, and aliquoted into 10 µL single-use aliquots. Clinical RNA extracts and RNA products were stored at -80°C.

3.1.3 Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR)

SARS-CoV-2 RNA was quantified via RT-qPCR using the QuantiTect Probe RT-PCR kit (QIAGEN), CDC N1 oligonucleotides (ADTBio; Table 3.1), and research reagent 19/304. A 25 µL reaction mix was prepared containing 12.5 µL of RT-PCR mix, 0.25 µL of reverse transcriptase, 5 µL of 5X oligo mix, 2.25 µL of water, and 5 µL of RNA eluate and added to a MicroAmp Fast Optical 96-well reaction plate (ThermoFisher Scientific). RT-qPCR was carried out on an Applied Biosystems StepOnePlus Real-Time PCR system (ThermoFisher Scientific) with the following settings: 50°C for 30 minutes, 95°C for 15 minutes, and 40 cycles of 94°C for 15 seconds and 60°C for 1 minute. Intra-assay variation was standardized through use of a standard curve of NIBSC RNA control 19/304 single-use aliquots serially diluted from 10,000 copies/reaction to 100 copies/reaction. Cycle threshold (Ct) values were converted to international units/milliliter (IU/mL) by a conversion rate provided

by NIBSC.[270] Samples with no detectable RNA were given a representative Ct value of 43 and an IU/mL value of 10 for data analysis.

3.1.4 Next-Generation Sequencing (NGS)

These steps were performed by George MacIntyre Cockett.

SARS-CoV-2 NGS was performed using the veSeq protocol as described elsewhere.[250] Briefly, a total of 30 µl of RNA eluate was concentrated to 2 µl using RNAClean XP beads (Beckman). The SMARTer Stranded Total RNA-Seq Kit v2 (Pico mammalian input; Takara BioSciences) was used to construct total nucleic libraries, following the manufacturer's instructions without the RNA fragmentation or Ribo-depletion steps. The SMARTer Stranded kit uses random primers, tailed with the Illumina read 1 sequence, to start reverse transcription, and a Template Switching Oligo (TSO) to add the Illumina read 2 sequence at the 3' end of the synthesized cDNA. The 1st stranded cDNA was PCR amplified with 12 cycles, using indexed primers, to generate tagged Illumina libraries. Indexed libraries were pooled at equal volume and captured using 120-mer oligonucleotide xGen lockdown probes (IDT) designed in-house to match 60 nt sliding windows of the NC_045512 reference sequence. Hybridization of 500 ng of up to 96 pooled DNA libraries was performed overnight with the probes, 5 µg of Cot-1 DNA and 2 µl xGen Universal blocker-TS mix (IDT) using the SeqCap EZ Hybridization and Wash Kit (Roche). Biotinylated probes were bound to streptavidin M-270 Dynabeads, and virus enriched libraries were amplified with further 12 cycles of PCR using Post-LM Primers 1 and 2 (Roche). Captured pools were cleaned with 0.68x volume Ampure XP beads (Beckman), and up to 4 pools were combined with equal molarity and sequenced on a single lane of an Illumina SP500 flow cell, using v1.0 chemistry, on an Illumina NovaSeq 6000 system.

3.1.5 Bioinformatic Processing

These steps were performed by Dr. Tanya Golubchik.

Samples were demultiplexed using unique dual indexes (UDI), and read output was validated against Ct values to confirm sample integrity. Genomes were assembled from sequencing reads using the ShiverCovid pipeline v1.8 (`https://github.com/BDI-pathogens/ShiverCovid`) with variant frequencies calculated using shiver with default settings of no base alignment quality and maximum pileup depth of 1000000.[271] For analysis of consensus genomes, consensus calls required a minimum of two uniquely mapped (deduplicated) reads per position, equivalent to roughly more than 15 raw reads per position. Trimmed reads were mapped to the SARS-CoV-2 RefSeq genome of isolate Wuhan-Hu-1 (NC_045512) using MAFFT version 7.402.[272]

3.2 Methods used in Chapters Four, Five, and Six

3.2.1 Nested Polymerase Chain Reaction Restriction-Fragment Length Polymorphism (PCR-RFLP)

A nested Polymerase Chain Reaction Restriction-Fragment Length Polymorphism (PCR-RFLP) method was designed for identifying B.1.1.7/Alpha variant infection based on the S:D1118H single nucleotide polymorphism (SNP). The G>C transversion at site 24915 disrupts the restriction site of the restriction enzyme *SfcI* (5'-CTRYAG-3'). To assess *SfcI* activity, 5 µL of RNA eluate was reverse transcribed using SuperScript III Reverse Transcriptase (Thermofisher Scientific) and the S:1118 outer primers (Table 3.1) according to the manufacturer's protocol for gene-specific primers. cDNA was then amplified using a nested PCR protocol. Briefly, for the 1st round amplification, a 25 µL reaction mix containing 2 µL of cDNA, 5 µL of 5X GoTaq Green Master Mix (Promega), 0.125 µL of 5u/µL GoTaq G2 polymerase (Promega), 2 µL of 2.5 mM dNTP mix (Stratech Scientific), 5 µL of 2.5 µM S:1118 outer primer mix, and 11.875 µL of PCR-grade water was amplified on thermal cycler with the following settings: 95°C for 2m, 30 cycles of 95°C for 20s, 47°C for 20s, and 72°C for 45s, and a final extension of 72°C for 5m. For the 2nd round amplification, a 25 µL reaction mix containing 1 µL of 1st round product, 5 µL of 5X GoTaq Master Mix (Promega), 0.125 µL of 5u/µL GoTaq G2 polymerase

(Promega), 2 µL of 2.5 mM dNTP mix (Stratech Scientific), 5 µL of 5X S:1118 inner primer mix, and 12.875 µL of PCR-grade water was amplified on a thermal cycler with the following settings: 95°C for 2m, 30 cycles of 95°C for 20s, 56°C for 20s, and 72°C for 45s, and a final extension of 72°C for 5m. 6 µL of the 2nd round product was digested using *SfcI* (New England Biolabs) according to the manufacturer's instructions. PCR product was visualized by gel electrophoresis in a 2.5% agarose gel containing Sybr-Safe DNA Gel Stain (ThermoFisher Scientific) and a 100bp DNA ladder (M.Biotech). Uncut amplicons (lineage B.1.1.7/Alpha) ran near 200 bp and cut amplicons (WT) ran near 120 bp. Samples that failed to produce a second round PCR product were designated "undetermined".

3.2.2 Allele-Specific Probe Polymerase Chain Reaction (ASP-PCR)

SNP sites targeted for lineage discrimination were chosen based on their lineage-specific predictive value estimated using publicly available sequence data published on GISAID.[245]

Allele-Specific Probe Polymerase Chain Reaction (ASP-PCR) was performed using the Quantitect Probe RT-PCR kit (QIAGEN) with 5 µL of extracted RNA (or library preparation cDNA) in a 25 µL reaction volume on an Applied Biosystem StepOnePlus Real-Time PCR system using the genotyping program. SNPs were designated based on their clustering with discrimination controls. Serially diluted cDNA aliquots of sequence-confirmed samples were used as discrimination controls; ultrapure water served as negative controls. Samples that failed to achieve a change in signal in either probe greater than those of the no template controls or lacked evidence of amplification were designated "undetermined". Reaction conditions (annealing and extension temperature, length and oligonucleotide concentrations) were optimized using serially diluted cDNA generated from samples of known lineage. The cDNA was generated using the forward and reverse primers of the target of interest and SuperScript III Reverse Transcriptase (ThermoFisher Scientific)

according to the manufacturer's protocol for gene-specific primers. The RT-PCR settings for each ASP-PCR target are described in table 3.2.

To test the effects of modified oligonucleotides, the D1118H/Alpha ASP oligonucleotide set was redesigned using locked nucleic acids (LNA) over the SNP site (Table 3.1). To measure the impact of low-molecular weight amides on differentiation, 2-pyrrolidinone (Merck) was added to the qPCR master mix at a reaction concentration of 0.4M as recommended in Chakrabarti and Schutt (2001).[273]

3.2.3 SARS-CoV-2 Consensus Sequence Lineage Assignment

Phylogenetic reconstruction was performed by Dr. Tanya Golubchik.

In the first instance, SARS-CoV-2 lineages were assigned based on the Pangolin framework.[274] Both REMAP-CAP and COV003 samples had lineages assigned by the pangolin tool (v3.1.11 for REMAP-CAP and v.2.4.2 for COV003). In REMAP-CAP, where a large proportion of samples failed to return sufficient coverage for pangolin assignment, SNP analysis was used to differentiate Alpha vs non-Alpha ("WT"). Base calling at a minimum of two of the following Alpha-associated SNPs were used to type samples: S:A570D, S:P681H, S:T716I, S:S982A, S:D1118H, N:D3L, and N:S235F. If more than two bases were called and there were conflicting assignments, the lineage with the greater number of called bases was assigned. There were no ties. For COV003, where pangolin assigned multiple samples B.1.1 or B.1.1.28 (the parent lineages of P.1 and P.2), phylogenetic reconstruction was used to resolve discrepancies between pangolin assignment and ASP-PCR lineage. Reconstruction was performed using IQ-TREE version 1.6.12 with the generalized time reversible + FreeRate model and 1000 bootstrap replicates.[275]

COV003 samples were assigned as either WT, Gamma, Zeta, Alpha, or Delta. Sublineages P.1.* and AY.* (from Pangolin) were assigned to their parent lineage. For COV003, samples were assigned with the following priority: Phylogenetic > Pangolin > ASP-PCR. For REMAP-CAP, samples were assigned with the following priority: Pangolin = SNP Analysis > ASP-PCR > PCR-RFLP.

3.3 Methods used in Chapter Four

3.3.1 Assessment of NGS Success

Two approaches were used to assess and compare lineage assignment through NGS methods. NGS – Pangolin verified whether the sample met the default quality control parameters to be assigned to a lineage by the Pangolin tool. Samples unable to be processed by the Pangolin tool were assigned "undetermined" and considered unsuccessful. NGS - SNP evaluated variant calls at the same sites of interest targeted by the ASP-PCR method. Consensus sequences with no coverage over the site of interest were assigned "undetermined" and considered unsuccessful. For COV003, samples were defined as degraded if they had both high viral load ($\geq 10^6$ IU/mL) and poor whole genome coverage ($\leq 2*10^4$ deduplicated bases).

3.3.2 Probit Analysis

Probit regression models were generated using base function 'glm' with family = "binomial(link = "probit")" in R version 4.0.4.[276] Confidence intervals were generated in consultation with open-source code and methods published by Dr. Gavin Simpson on his personal website.[277]

3.3.3 Predictive Performance

Method performance on representative datasets was estimated by summing the probit function output for each individual sample.[278, 279] IU/mL, N gene Ct value, or average Ct value were used as the model input variable depending on metadata availability.

3.3.4 Statistical Analysis

Differences in viral load density between methods were compared via an Anderson-Darling k-Sample test using function ad.test from package kTests (https://cran.r-project.org/web/packages/kSamples/index.html).

3.4 Methods used in Chapter Five

3.4.1 Qualitative Enzyme-Linked Immunosorbent Assay (ELISA)

This method was performed by members of the laboratory of Prof. David Roberts.

Patient baseline blood samples were tested to qualitatively detect serum immunoglobulin G (IgG) antibodies against SARS-CoV-2 S at the time of enrollment using a high-throughput, 384-well enzyme-linked immunosorbent assay (ELISA) presented elsewhere.[280] In the validation study, the method demonstrated >98% sensitivity and specificity.

3.4.2 Subgenomic RNA Reverse Transcription Polymerase Chain Reaction

SARS-CoV-2 N gene sgRNA was targeted via RT-qPCR using the QuantiTect Probe RT-PCR kit (QIAGEN), an in-house designed sgRNA forward primer, and the CDC N1 probe and reverse primer (ADTBio; Table 3.1). As in RT-qPCR for the N gene, a 25 µL reaction mix was prepared containing 12.5 µL of RT-PCR mix, 0.25 µL of reverse transcriptase, 5 µL of 5X oligo mix, 2.25 µL of water, and 5 µL of RNA eluate and added to a MicroAmp Fast Optical 96-well reaction plate (ThermoFisher Scientific). RT-qPCR was carried out on an Applied Biosystems StepOnePlus Real-Time PCR system (ThermoFisher Scientific) with the following settings: 50°C for 30 minutes, 95°C for 15 minutes, and 40 cycles of 94°C for 15 seconds and 60°C for 1 minute. Attempts to clone a sgRNA transcript were unsuccessful, so no quantified standard was used; assay readout was the Ct value.

sgRNA relative abundance was calculated using $2^{-\Delta Ct}$ analysis. ΔCt was calculated as Ct_{sgRNA}-Ct_{gRNA} using the Ct values returned from the subgenomic RT-qPCR and the CDC N1 RT-qPCR.

3.4.3 Baseline Viral Load Analysis

Association studies of demographic and clinical features with log-transformed baseline viral loads were performed using base functions in R version 4.0.4. Four

specific hypotheses were tested using a collection of categorical and continuous variables. Correlation between each variable tested within each hypothesis and patient serostatus was calculated using Kendall's Rank Correlation. Variables whose correlations displayed a p value < 0.10 were deemed correlated; the high threshold was deployed to be more conservative regarding confounding. Multivariable analysis including all correlated variables, numerically recoded categorical variables, and patient serostatus was conducted using linear regression. To reduce the risk of committing a type I error, a Bonferroni correction was applied taking into account the seven total association studies performed. Therefore, a p value < 0.007 was deemed significant. Lastly, one multivariable linear regression test was performed incorporating all statistically significant variables.

Hypothesis I: Presence of co-morbidities will increase patient viral load.

Categorical variables tested: Any co-morbidity (asthma, other respiratory disease, liver failure, diabetes, chronic cardiovascular disease, immunosuppression).
Continuous variables tested: None.
Correlated variables: Asthma and diabetes; other respiratory disease, liver failure, and immunosuppression.

Hypothesis II: Prior treatment with immune-modulating drugs will increase patient viral load.

Categorical variables tested: Medication received prior to trial baseline sample (corticosteroid, remdesivir, treatment with an IL-6 inhibitor (sarilumab and/or tocilizumab)).
Continuous variables tested: None.
Correlated variables: Corticosteroid and remdesivir.

Hypothesis III: Patients with increased severity of illness will have higher viral loads.

Categorical variables tested: Specific interventions (high flow oxygen, invasive ventilation).
Continuous variables tested: C-reactive protein (µg/mL), Acute Physiology and Chronic Health Evaluation II (APACHE II) score, and Glasgow Coma Score.
Correlated variables: All five variables correlated.

Hypothesis IV: Patients with demographic characteristics associated with poorer immune responses will have higher viral loads.

Categorical variables tested: Sex at birth.
Continuous variables tested: Age (years), body mass index (BMI).
Correlated variables: All three variables correlated.

3.4.4 Intensive Viral Dynamics Analysis

Association studies between viral dynamics and patient and virological factors were performed using base functions in R version 4.0.4. As direction of effect was not prespecified, hypotheses were not developed a priori. Continuous variables (age, BMI, APACHE II, Glasgow coma score) were analyzed using Kruskal-Wallis Rank Sum Test. Categorical variables (sex, co-morbidities (listed above), corticosteroid treatment, remdesivir treatment, treatment with an IL-6 inhibitor, virus type) were analyzed using a Fisher's Exact Test. Due to the low number of samples classified as "Hard relapse", both relapse categories were combined for the association studies. Employing a Bonferroni correction across all tests, a p value < 0.0033 was considered significant.

3.5 Methods used in Chapter Six

3.5.1 Disease Severity and Outcome Analysis

All association and correlation tests were performed using R version 4.0.4. Binomial regressions were performed using function 'glm' with family = "binomial". A Bonferroni correction for multiple hypotheses was applied within each section (i.e., severity $p < 0.004$ and outcome $p < 0.0125$).

3.5.2 Variant and N:KR Emergence Analysis

Measuring the temporal emergence of the Alpha variant and the N:KR mutation in the general population used publicly available sequences from GISAID.[245] All sequences collected between April 20, 2020 and January 7, 2021 within England were downloaded. Lineages were assigned by Pangolin v4.0.4 integrated into the GISAID platform and Alpha was inclusive of all samples designated B.1.1.7 or Q.*. For the N:KR mutation, sequences were aligned to EPI_ISL_402124 using MAFFT version 7 leveraging the computing cluster of Osaka University.[281] N:203/204 was designated using the Biopython package and user-defined functions.[282] Dates of collection for GISAID samples were assigned to REMAP-CAP randomization weeks using function week() in Microsoft Excel.

Global differences in proportions were assessed in R using function ad.test from package kTests (https://cran.rproject.org/web/packages/kSamples/index.html).

3.6 Methods used in Chapter Seven

3.6.1 Consensus Sequence Analysis

Consensus fasta files aligned to NC_045512 from patients intensively sampled in the REMAP-CAP trial were imported into Python and analyzed using the Biopython package and user-defined functions.[282] Only samples demonstrating sufficient viral loads for sequencing (Ct < 30.5) were processed via NGS. Emergent consensus level SNPs were identified by comparing sequences at time t to a sequence at time $t+1$.

3.6.2 Mutation Rate Statistical Analysis

To compare weighted average mutation rates, Welch's T test was modified to account for weighted average population means and effective population sizes as described below and recommended in Madansky (n.d.).[283]

Welch's T test is given by:

$$T = \frac{M_A + M_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

$$d.f. = \frac{(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B})^2}{\frac{(\frac{S_A^2}{n_A})^2}{n_A - 1} + \frac{(\frac{S_B^2}{n_B})^2}{n_B - 1}}$$

where M is the population mean, s^2 is the population variance, n is the population size, and *d.f.* is the degrees of freedom.

To modify for weighted averages, Welch's T test needs to be able to use population weighted averages and effective population sizes.

Therefore, Welch's modified T test for weighted averages can be derived as:

$$T = \frac{\overline{X_A} + \overline{X_B}}{\sqrt{\frac{S_A^2}{e_A} + \frac{S_B^2}{e_B}}}$$

$$d.f. = \frac{(\frac{S_A^2}{e_A} + \frac{S_B^2}{e_B})^2}{\frac{(\frac{S_A^2}{e_A})^2}{e_A - 1} + \frac{(\frac{S_B^2}{e_B})^2}{e_B - 1}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}$$

$$e = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} = \frac{1}{\sum_{i=1}^{n} w_i^2}$$

Where $\bar{X}$ is the population weighted average, s^2 is the population variance, e is the effective population size, n is the total number of observations, w is the relative weight, and *d.f.* is the degrees of freedom. $(\sum_{i=1}^{n} w_i)^2$ reduces to 1 in all instances where the sum of the weights are equal to one. Per patient weights were calculated as an individual patient's proportionate contribution to the total number of days observed for all patients.

T tests were conducted in R version 4.0.4. using the base function pt with the above derived T statistics and degrees of freedom as inputs.

3.6.3 Mutation and Context Statistical Analysis

Mononucleotide and dinucleotide frequencies for SARS-CoV-2 were calculated based on the NC_045512 RefSeq sequence using the 'composition scan' function for noncoding regions in SSE v1.4.[284] The "simple" method was used to calculate expected dinucleotides in lieu of the corrected method as C>U transitions driven by the action of RNA editing would not be restricted to coding regions.

Normalized mutational assymtery ratios were calculated as described in Simmonds and Ansari (2021).[220] The normalized ratio (n(X>Y/Y>X)) is calculated by:

$$n(\frac{X>Y}{Y>X}) = (\frac{f(X>Y)}{f(Y>X)}) * (\frac{f(Y)}{f(X)})$$

where X and Y are any two nucleotides, *f(X>Y)* and *f(Y>X)* are the observed transition frequencies, and *f(X)* and *f(Y)* are mononucleotides frequencies in the reference sequence.

For the visualization in figure 7.5, ratios normalized to the base mononucleotide frequency are displayed. These ratios were calculated as:

$$n(5'X) = (\frac{f(5'X_{observed})}{f(X)})$$

where X is any nucleotide, *f(5'X$_{observed}$)* is the observed frequency of X as an upstream base, and *f(X)* is the global mononucleotide frequency. The dinucleotide base context analysis that compared counts of observed vs expected dinucleotides was performed using a Fisher's Exact Test.

3.6.4 Homoplastic SNP Occurrence Estimation

The likelihood of the occurrence of a homoplastic SNP can be approximated using a similar approach to solving the famous "Birthday Problem" first described in 1939.[285] Assuming every SNP is a mutually exclusive event, as would occur in a neutrally-evolving genome, the likelihood of any SNP occurring multiple times in a population can be approximated by:

$$p(n) \approx 1 - \frac{x-1}{x}^{\frac{n(n-1)}{y}}$$

Where n is the total number of SNPs observed, x is the genome length, and y is the number of occurrences (e.g., y = 2 when solving for occurrences in two people). Estimates for specific substitutions (e.g., C>U) are the product of the proportion of total substitutions of that type within the study dataset (e.g., 20.7%) and the likelihood of a homoplastic mutation occurring ($p(n)$).

3.7 Methods used in Chapter Eight

3.7.1 Virus Sequences and Characteristics

Human infectious RNA viruses with vector-borne transmission routes were identified from a database curated by Woolhouse and Brierley (2018).[103] Flaviviruses and phylogenetic groupings were identified from two systematic reviews produced by Blitvich and Firth and a research paper from Pandit et al. (2018).[286–288] Mosquito-borne flavivirus host ranges were determined using the Centers for Disease Control and Prevention "Arbocat" database (https://wwwn.cdc.gov/arbocat/). Host range for viruses not listed in Arbocat, the Blitvich and Firth reviews, or in Pandit et al. (2018) were gathered from original research using the following search strategy on PubMed: (virus name) AND (virus) AND ((Host) OR (Bird*) OR (Avian) OR (Mammal*)). References for host ranges are listed in table 8.4. A random samples of Zika virus genomes were downloaded based on the most recent modification date and are provided in table 10.3. 16 sylvatic DENV genomes were identified from Vasilakis et al. (2007).[289] A random sample of 337 DENV genomes were downloaded from GenBank. Accession numbers for sylvatic and urban DENV genomes are provided in table 10.4.

Genome sequences for the viruses used in this study were downloaded from GenBank using non-redundant sequences curated as part of the RefSeq project. Accession numbers are provided in tables 10.2 and 10.3. If no RefSeq sequence was available, a whole genome, polyprotein, or partial sequence was downloaded,

with preference for the genome with the most coverage. Non-Refseq sequences were used for the following viruses:

Alfuy virus, Alkhumra hemorrhagic fever virus, Bainyik virus, Barkedji virus, Batu Cave virus, Calbertado virus, California encephalitis virus, Crimean-Congo hemorrhagic fever virus, Iguape virus, Kunjin virus, Sand fever Naples-like virus, Stratford virus, and Torres virus.

This study used the human transmissibility levels as assigned by Woolhouse et al (2016).[108] Briefly, T1 viruses are unable to infect humans. T2 viruses are those known to infect humans either through direct (isolation, nucleic acid detection) or indirect (serology) evidence but without documented human-human transmission. T3 viruses can transmit in human populations but with an estimated R_0 value < 1. T4 viruses transmit in human populations with an estimated R_0 value > 1. Human-vector-human transmission is considered human-human transmission.

3.7.2 Dinucleotide Analysis

GC content, observed mononucleotide and dinucleotide frequencies, and expected dinucleotide frequencies were calculated for each virus genome using "Composition Scan" set to non-coding regions in SSE v1.4. Simple dinucleotide ratios are calculated using the following formula:

$$\frac{Observed}{Expected} = \frac{f(XpY)}{f(X) * f(Y)}$$

where *f(XpY)* is the observed frequency of any dinucleotide and *f(X)* and *f(Y)* are the observed mononucleotide frequencies. Corrected dinucleotide ratios were calculated by aligning virus genomes to the largest ORF and performing "Composition Scan" set to coding regions. All association studies were performed using two-tailed Mann-Whitney U tests for continuous data or Fisher's Exact Test for categorical data.

Target	Purpose	Amplicon length (bp)	Name	5'>3' Sequence[a]	Reaction concentration (nm)
N gene	Quantification	72	CDC_N1_FW	GACCCCAAAATCAGCGAAAT	500
			CDC_N1_RV	TCTGGTTACTGCCAGTTGAATCTG	500
			CDC_N1_P	FAM-ACCCCGCATTACGTTTGGTGGACC-BHQ1	125
sgRNA N gene	Relative quantification	<200 bp	sgRNA_FW	CGATCTCTTGTAGATCTGTTCTC	500
S:417	Genotyping (Gamma or other)	114	K417T_FW	GAGGTGATGAAGTCAGACAAATCGC	600
			K417T_RV	GATTGTTAGAATTCCAAGCTATAACGCAGC	600
			K417T_P_WT	HEX-TCAGCAATMTTTCCAGTTTGCCCTGG-BHQ1	150
			K417T_P_P1	FAM-TCAGCAATMGTTCCAGTTTGCCCTGG-BHQ1	150
ORF1ab: 3468	Genotyping (Zeta or other)	192	L3468V_FW	CACAAGCAGCTGGTACGGACACA	500
			L3468V_RV	GCAGAAAGAGGTCCTAGTATGTCAACATGG	500
			L3468V_P_WT	HEX-TTAATGTTTTAGCTTGGTTGTACGCTGCTG-BHQ1	125
			L3468V_P_P2	FAM-TTAATGTTGTAGCTTGGTTGTACGCTGCTG-BHQ1	125
S:1118	Genotyping (Alpha or other)	199	D1118H_FW	GCCATTTGTCATGATGGAAAAG	600
			D1118H_RV	CTAATTCAGGTTGCAAAGGATC	600
			D1118H_P_WT	FAM-CCACAAATCATTACTACAGACAACACATTTGTGT-BHQ1	150
			D1118H_P_B.1.1.7	TAMRA-CCACAAATCATTACTACACACAACACATTTGTGT-BHQ2	150
			D1118H_P_WT_LNA	FAM-TCATTACTAC+A+G+ACAACACATTTG-BHQ1	150
			D1118H_P_B.1.1.7_LNA	HEX-TCATTACTAC+A+C+ACAACACATTTG-BHQ1	150
N:KR	Genotyping (N:RG or N:KR)	160	NKR_FW	TCTACGCAGAAGGGAGCAGAGG	500
			NKR_RV	AGCAGCAGCAAAGCAAGAGCAG	500
			NKR_P_RG	FAM-CCAGGCAGCAGTAGGGGAACTTCTCCTGCT-BHQ1	125
			NKR_P_KR	FAM-CCAGGCAGCAGTAAACGAACTTCTCCTGCT-BHQ1	125
PCR-RFLP	1st round amplification	367	D1118H_FW_Outer	CTTGCTGCTACTAAAATGTCAG	500
			D1118H_RV_Outer	GCTCATACTTTCCAAGTTCTTG	500
	2nd round amplification	72	D1118H_FW	GCCATTTGTCATGATGGAAAAG	500
			D1118H_RV	CTAATTCAGGTTGCAAAGGATC	500

Table 3.1: Oligonucleotides Used in this Thesis. [a]Underlined characters indicate mismatched bases for allele-specific probes, (+) is 5' of an LNA-modified base.

<table>
<tr><th>ASP-PCR set</th><th>Pre-and post-amplification read</th><th>Reverse transcription</th><th>Heat activation</th><th>Denaturation</th><th>Annealing/ Extension</th><th>Number of cycles</th></tr>
<tr><td>S:D1118H/ Alpha</td><td rowspan="3">60°C for 30s</td><td rowspan="4">50°C for 30m</td><td>95°C for 15m</td><td>94°C for 15s</td><td>55°C for 20s, 60°C for 1m</td><td rowspan="3">45</td></tr>
<tr><td>S:K417T/ Gamma</td><td rowspan="3">95°C for 10m</td><td rowspan="3">95°C for 15s</td><td>58°C for 20s, 60°C for 45s</td></tr>
<tr><td>N:KR</td><td>60°C for 1m</td></tr>
<tr><td>ORF1ab: L3468V/ Zeta</td><td>66.5°C for 30s</td><td>66.5°C for 1m</td><td>50</td></tr>
</table>

Table 3.2: ASP-PCR settings. ORF1ab:L3468V ASP-PCR pre- and post-amplification reads were performed at the lowest temperature during cycling, others were read at 60°C. Variance in heat activation and denaturation temperatures for S:D1118H vs N:KR, S:K417T, and ORF1ab:L3468V was failure to correct default settings during initial optimization experiments for the former.

4

Rapid Methods for Detecting SARS-CoV-2 Mutations and Typing Variants

Contents

4.1 Introduction 60
4.2 Results 64
4.2.1 SNP Positive and Negative Percent Agreement 64
4.2.2 Optimization of ASP-PCR Reaction Conditions 64
4.2.3 PCR-RFLP and ASP-PCR Validation 67
4.2.4 PCR-RFLP, ASP-PCR, and NGS Performance for Lineage Assignment 68
4.2.5 ASP-PCR and NGS Performance for Detecting N:KR Mutation 74
4.2.6 Performance of ASP-PCR Across Diverse Targets 76
4.2.7 Effect of Molecular Modifications 76
4.2.8 ASP-PCR on NGS Library Preparations 78
4.2.9 Predicted Performance of Typing Methods on Larger Cohorts 79
4.3 Discussion 80

Abstract

Tools to differentiate SARS-CoV-2 variants of concern and track the ongoing evolution of the virus are necessary to support public health efforts and the design and evaluation of COVID-19 therapeutics and vaccines. Although NGS has been

adopted as the gold standard method for discriminating SARS-CoV-2 lineages, alternative methods may be required when processing samples with low viral loads or low RNA quality or when NGS is not available. To this aim, two rapid methods targeting lineage-specific single nucleotide polymorphisms (SNPs) were developed and used to screen samples from two clinical trials conducted in the United Kingdom and Brazil. A nested polymerase chain reaction restriction fragment length polymorphism assay (PCR-RFLP) (applied to 264 samples) and allele-specific probe polymerase chain reaction (ASP-PCR) (1641 total tests on 1082 total samples) were developed and validated to target three SNPs and a double substitution in the SARS-CoV-2 genome. Probit regression models were developed to compare ASP-PCR and PCR-RFLP performance against 1,771 NGS results for the same cohorts. PCR-RFLP and ASP-PCR were shown to have a higher likelihood of discriminating SARS-CoV-2 lineages than NGS over a wide range of viral loads. Comparative advantage for the rapid methods over NGS was most pronounced in samples with Ct values between 26-30 and in samples that showed evidence of degradation. Results for samples screened by ASP-PCR and NGS showed 99% concordant results while concordance was lower for PCR-RFLP and NGS (95.2%). Further, ASP-PCR was shown to have equivalent performance using intermediate products of NGS (library preparation) as with the paired primary RNA sample. ASP-PCR is more practical to implement than PCR-RFLP for laboratories with the necessary equipment and is well-suited to augment but not replace NGS. The method can differentiate SARS-COV-2 lineages with high accuracy and would be best deployed to screen samples with lower viral loads or that may suffer from degradation or in situations where high-throughput typing is necessary and NGS is unavailable or impractical. Future work should investigate further destabilization from target-oligo base mismatch through altered oligonucleotide chemistry or chemical additives.

4.1 Introduction

The ongoing evolution of SARS-CoV-2 leading to emergence of variants of concern (VoC) and variants of interest (VoI) has highlighted the need for broadly accessible

methods for detecting and tracking SARS-CoV-2 mutations. NGS with a collection of different methods and protocols has been adopted as the gold standard approach for detecting SARS-CoV-2 variants. Several targeted and untargeted whole-genome NGS methods have been deployed for sequencing SARS-CoV-2, including the tiling amplicon-based ARTIC protocol and targeted enrichment RNA-seq-based veSeq protocol.[250, 290]

Tools to identify the emergence of novel SARS-CoV-2 variants and track their spatial spread are necessary to support public health interventions. Single-nucleotide polymorphisms (SNPs) such as S:D614G may be associated with increased transmissibility.[291] VoCs Alpha (Pangolin: B.1.1.7 and Q.*), Beta (Pangolin: B.1.351), Gamma (Pangolin: P.1 and P.1.*), Delta (Pangolin: B.1.617.2 and AY.*), and Omicron (Pangolin: BA.*) and VoI Zeta (Pangolin: P.2) have constellations of mutations that, to varying degrees, appear to impact transmissibility, severity, and evasion of vaccine and prior infection-derived adaptive immune responses.[292–295] Rapid identification of novel lineages through NGS is necessary for accurate characterization and risk assessment. However, recovering whole-genome sequences for samples with low viral load is challenging and RNA quality must be high, conditions that may not be met in samples taken from patients late in infection or in settings where optimal sample handling and storage is not available.[296, 297] In particular, high levels of RNA degradation during storage or transport of primary sample or RNA can severely impact the effectiveness of amplicon-based protocols such as ARTIC, the most popular SARS-CoV-2 sequencing technique globally,[290] because of the generation of 400bp amplicons in its pre-amplification step.

Herein, we describe the development and application of two rapid methods for characterizing SARS-CoV-2 VoCs across two diverse patient cohorts and directly compare their performance to NGS deployed on the same samples. Both novel methods identify lineages by the presence of individual SNPs at lineage informative sites in the virus genome. Polymerase Chain Reaction - Restriction Fragment Length Polymorphism (PCR-RFLP) has been used since the late 1980s to detect point mutations in nucleic acid targets.[298, 299] The method takes advantage of

restriction enzyme cleavage of PCR products containing specific restriction sites. The presence of a point mutation can disrupt (or create) a specific restriction site preventing (or enabling) endonuclease digestion. The presence of the point mutation can then be readily evaluated using gel electrophoresis. In this study, a nested PCR-RFLP was applied to samples from one clinical trial.

Allele-specific probe PCR (ASP-PCR) leverages differential binding affinities of two fluorescently-labelled probes differing in a base overlapping a SNP site. A target-oligo mismatch between one of the two probes and the sample of interest result in an energetic cost destabilizing the mismatched target-oligo duplex. Subsequently, the mismatched probe will melt from the target cDNA at a lower temperature than the properly matched probe. By performing PCR cycling at a temperature between the melting temperatures (T_m) of the two probes, amplification of one probe at the exclusion of the other can indicate the correctly bound probe and hence the underlying sequence of the target. ASP-PCR uses similar amplicon length (120-200 bp) and technology as RT-qPCR and is therefore both more robust to degraded RNA than ARTIC and accessible to the majority of microbiology laboratories. In this study, ASP-PCR was applied to samples from two clinical trials to detect three SNPs and one double substitution in the SARS-CoV-2 genome.

The Randomized, Embedded, Multifactorial, Adaptive Platform Trial for Community-Acquired Pneumonia (REMAP-CAP) Immunological Domain trial recruited 2,097 critically ill COVID-19 patients to a convalescent plasma (CP) trial from March 9, 2020 to January 18, 2021 with principal patient representation from the United Kingdom.[265, 266] Crucially, reports indicate the emergence of the antigenically distinct Alpha variant in the United Kingdom from November 2020, overlapping trial recruitment.[300] To detect and differentiate those infections, the S substitution S:D1118H was targeted both by PCR-RFLP and ASP-PCR. This mutation was chosen as it was the only Alpha-associated S substitution identified that enabled PCR-RFLP. COV003 was a phase III trial of the ChAdOx1 nCoV-19 vaccine that recruited 10,416 patients from six trial sites throughout Southern and Eastern Brazil from June 23, 2020 to December 1, 2020. The trial overlapped the emergence of

both the Gamma and Zeta variants in the country, motivating efforts to determine the impact of variant of infection on vaccine efficacy.[301, 302] For these VoCs, ASP-PCR assays were designed targeting the substitutions S:K417T for Gamma and ORF1ab:L3468V for Zeta.

In both trials, the ability of ASP-PCR, PCR-RFLP, and the veSeq NGS method to provide any lineage designation was compared. While not as widely used as the ARTIC protocol, veSeq involves bait capture rather than amplicons and, as such, is more resistant (but not impervious) to RNA fragmentation and degradation, which was necessary given RNA quality in the COV003 trial. This method also allows for robust, quantitative assessment of viral minor populations. As not all samples were tested by both methods, probit regression models for each technique were derived to allow direct method comparison. SARS-CoV-2 lineages are an abstraction that are most commonly dynamically defined by the software package Pangolin.[274] As such, the "ground truth" lineage assignment is that provided by the Pangolin software; comparisons between method accuracy are best demonstrated through investigation of concordant and discordant lineage assignments. At lower sequence coverage, the confidence estimates outputted by Pangolin diminish, motivating the use of alternative methods for lineage designation from NGS outputs as demonstrated later.

Lastly, the polyphyletic N:R203K/G204R (N:KR) mutation has independently arisen multiple times in distantly related SARS-CoV-2 lineages including the VoCs Alpha, Gamma, and Omicron (but not Beta or Delta). Prior evidence demonstrated a role for the double substitution in increased viral fitness *in vitro*, higher viral loads in patients in Saudi Arabia and in Syrian hamster challenge studies, and a weak (but significant) correlation with increased severity.[303–305] The double substitution is due to three bases being substituted (GGG>AAC) and as such presented as a strong candidate for ASP-PCR due to the increased number of target-oligo mismatches that would be present. To that end, an ASP-PCR assay was designed and validated to detect the double substitution and used to screen a portion of non-Alpha samples in the REMAP-CAP trial. In total, the increased

sensitivity of PCR-RFLP and ASP-PCR over NGS for specific allele genotyping and lineage discrimination is clearly demonstrated.

4.2 Results

4.2.1 SNP Positive and Negative Percent Agreement

To assess the suitability of different SNP targets for lineage specification in the ASP-PCR platform, global sequencing data published as part of the GISAID repository was utilized, leveraging the collation performed by www.Outbreak.info.[245, 306] The country-specific positive percent agreement (PPA), negative percent agreement (NPA), positive predictive value (PPV), and negative predictive value (NPV) for various lineage-defining SNPs were calculated (Table 4.1). Despite being designed when little sequencing data was available, the SNPs used in this study were shown to be highly accurate for lineage discrimination when analyzed using sequences deposited up until September 26th, 2021 (prior to emergence of the Omicron VoC). Within the United Kingdom, the S:D1118H SNP was present in 99.90% of Alpha infections and detection of the SNP gave a 99.93% chance that the sample is Alpha. The ORF1ab:L3468V SNP proved to be the best Zeta-associated SNP with the highest PPV of 99.69%. For Gamma, the S:K417T SNP demonstrated the lowest PPA among the ten SNPs analyzed (93.72%), but still maintained very high PPV (99.42%). SNP selection for ASP-PCR cannot solely be on the measure of maximizing PPV and NPV and must also consider target suitability for ASP-PCR design (see Discussion). Analyses of potential target SNPS at a global scale for Alpha, Gamma, Zeta, Omicron, and other VoCs/VoIs are included in a separate table as reference (Table 10.1).

4.2.2 Optimization of ASP-PCR Reaction Conditions

For ASP-PCR, the combined annealing/extension temperature is the key determinant of assay performance. That step in the assay must be conducted at a temperature where the matched probe is largely bound and the mismatched probe is largely unbound for a difference in amplification of the two probes to be detected

Variant	Location	SNP	PPA[a]	NPA[b]	PPV[c]	NPV[d]
Alpha (n = 260,546)	United Kingdom (n = 895,220)	ORF1a:T1001I	99.93%	99.92%	99.82%	99.97%
		ORF1a:A1708D	99.54%	99.93%	99.82%	99.81%
		ORF1a:del3675/3677	99.62%	99.65%	99.15%	99.84%
		ORF1b:P314L	98.88%	2.23%	29.34%	82.93%
		S:del69/70	99.43%	99.41%	98.57%	99.77%
		S:A570D	99.97%	99.98%	99.95%	99.99%
		S:P681H	99.91%	99.84%	99.61%	99.96%
		S:S982A	99.97%	99.98%	99.94%	99.99%
		S:D1118H	99.90%	99.97%	99.93%	99.96%
		N:S235F	99.23%	99.97%	99.93%	99.69%
Gamma (n = 26,979)	Brazil (n = 42,948)	ORF1a:K1795Q	99.50%	99.87%	99.92%	99.17%
		ORF1b:P314L	98.45%	5.49%	63.77%	67.67%
		ORF1b:E1264D	98.28%	99.77%	99.86%	97.17%
		S:L18F	99.11%	99.47%	99.68%	98.51%
		S:P26S	98.72%	99.46%	99.67%	97.88%
		S:K417T	93.72%	99.07%	99.42%	90.33%
		S:H655Y	99.24%	99.30%	99.58%	98.73%
		S:V1176F	99.09%	66.02%	83.13%	97.73%
		ORF8:E92K	97.54%	99.61%	99.76%	95.99%
		N:P80R	99.03%	99.77%	99.87%	98.38%
Zeta (n = 2,694)	Brazil (n = 42,948)	ORF1a:L3468V	96.36%	99.98%	99.69%	99.76%
		ORF1a:L3930F	99.26%	96.08%	62.87%	99.95%
		ORF1b:P314L	97.74%	3.07%	6.32%	95.29%
		S:E484K	93.95%	35.06%	8.83%	98.86%
		S:D614G	98.78%	1.72%	6.30%	95.44%
		S:V1176F	98.85%	26.72%	8.28%	99.71%
		N:A119S	98.78%	99.65%	95.00%	99.92%
		N:R203K	98.70%	18.85%	7.53%	99.54%
		N:G204R	98.40%	18.99%	7.52%	99.44%
		N:M234I	96.99%	99.71%	95.71%	99.80%

Table 4.1: Country-Specific Performance of Lineage-Defining SNPs. VoC and VoI sublineages included as their parent lineages. Analysis includes 3,776,750 sequences deposited on GISAID up until 26/09/2021.
[a]Calculated as: Number of lineage with SNP / Total number of lineage
[b]Calculated as: Number of non-lineage without SNP / Total number of non-lineage
[c]Calculated as: Number of lineage with SNP / Total number with SNP
[d]Calculated as: Number of non-lineage without SNP / Total number without SNP .

(see Figure 4.1 for a theoretical example). However, the slope of the individual melt curves is unknown and the bioinformatically predicted T_M can be inaccurate. To optimize each assay, trials at multiple temperatures were conducted.

To ease interpretation, increase reproducibility, and preserve the primary RNA samples, serial dilutions of cDNA from sequence-confirmed samples were generated using the relevant forward and reverse primers and used as inputs during optimization. For each optimization, trialing of annealing/extension temperature

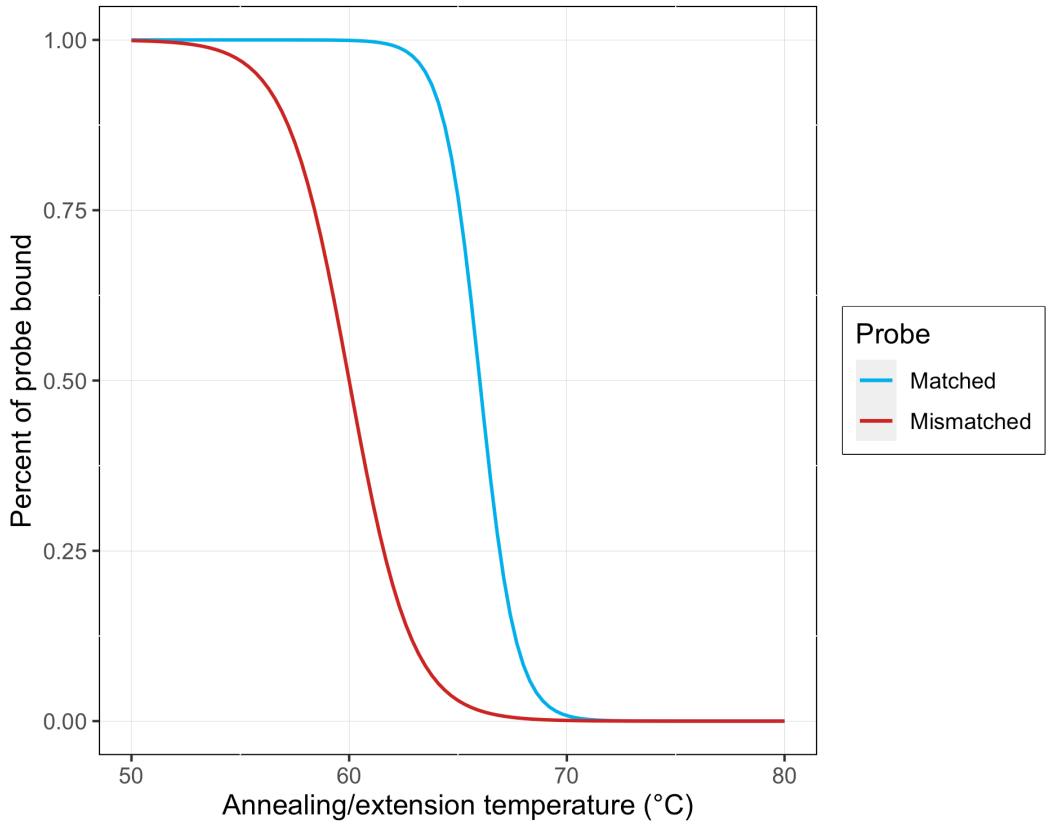


Figure 4.1: Theoretical Binding Dynamics of Matched and Mismatched Probes. The matched probe melts from the target DNA at a higher temperature than the mismatched probe. Annealing/extension in the temperature range where the matched probe is bound and unmatched probe is melted are suitable for ASP-PCR. The RT-qPCR chemistry of molecular probes (rather than integrated fluorescent dyes) precludes the experimental measurement of these values for given PCR products.

began at the recommended temperature for activity of the DNA polymerase in the assay kit of 60°C. The annealing/extension temperature was then raised or lowered on subsequent trials dependent on the deficiencies in the assay's performance. Lower temperatures increase the total amount of probe bound but may decrease the difference between the two probes and vice versa for higher temperatures. Three steps in the optimization strategy for the ORF1ab:L3468V ASP-PCR are visualized for reference (Figure 4.2); in total, optimization for that assay required eight total trials testing six annealing/extension temperatures, two oligonucleotide concentrations, and two total number of cycles. Importantly, the bioinformatically predicted T_M of the probes (based on perfect complementary binding) is not the most appropriate guide for optimization. The difference between the predicted

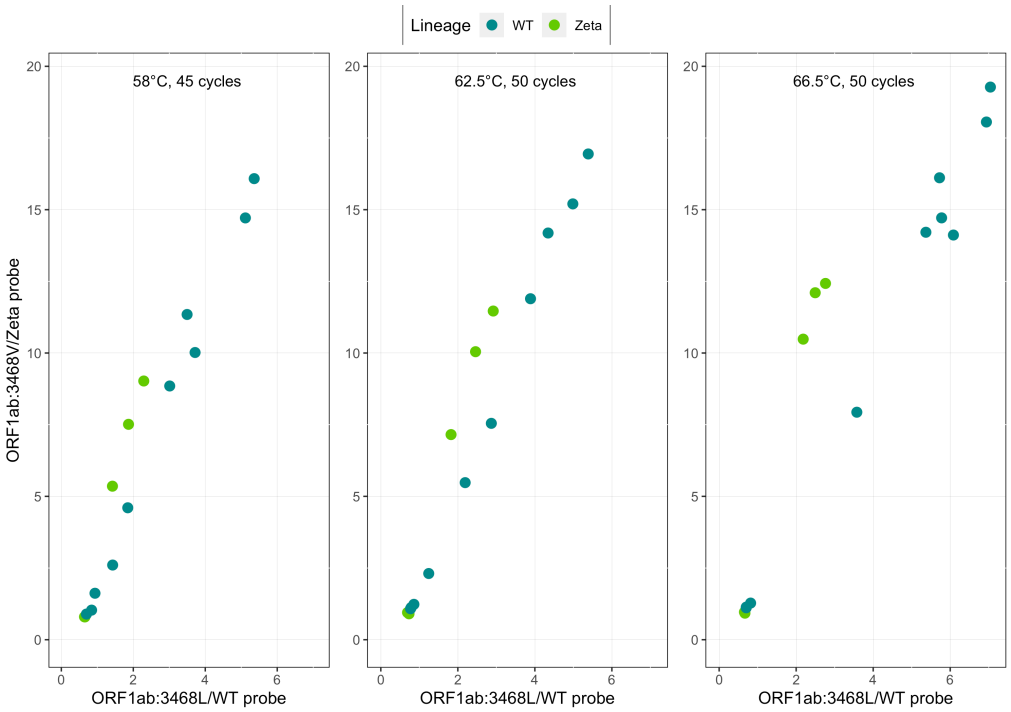


Figure 4.2: Optimization of Settings for the ORF1ab:L3468V ASP-PCR. Optimization performed using aliquots of cDNA generated from one sequence confirmed P2 sample and two sequence-confirmed WT samples, each diluted four times in a 10-fold serial dilution (fifteen samples total). Temperature is annealing/extension temperature for 1m per cycle. X- and Y-axis depict the overall change in signal for each probe from pre-amplification read to post-amplification read. Note the increasing clustering of the Zeta samples to the upper-left quadrant of the cartesian plane as the temperature is increased due to decreased signaling of the mismatched ORF1ab:3468L/WT probe.

T_M of the less thermodynamically stable probe and the optimized assay annealing-extension temperature ranged from 14.0°C for the S:D1118H ASP-PCR to only 2.6°C for the ORF1ab:L3468V ASP-PCR. Oligonucleotides used in this study, including reaction concentration, are available in Table 3.1; ASP-PCR reaction settings are available in Table 3.2.

4.2.3 PCR-RFLP and ASP-PCR Validation

Both of the rapid typing methods were validated using RNA extracted from sequence-confirmed samples of varying viral loads. The S:D1118H PCR-RFLP and ASP-PCR assays were evaluated against a panel of 30 WT samples and 10 Alpha samples from

the REMAP-CAP trial. The methods produced highly concordant results with 97.5% specificity and 93.5% sensitivity for PCR-RFLP and 100% specificity and 97.5% sensitivity for ASP-PCR. There was a single false positive for the PCR-RFLP; sanger sequencing demonstrated that this was the result of a U24921C mutation that similarly disrupted the activity of the *SfcI* restriction enzyme. The S:K417T ASP-PCR assay was evaluated with 10 Gamma and 39 non-Gamma samples from the COV003 trial and demonstrated 100% specificity and 93.9% sensitivity. The ORF1ab:L3468V ASP-PCR assay displayed 100% specificity and 95.7% sensitivity when tested with 25 Zeta and 22 non-Zeta samples from COV003. The N:KR ASP-PCR, tested on 10 N:KR containing and 30 N:RG containing samples, only maintained 80% sensitivity but had a high specificity of 100%.

4.2.4 PCR-RFLP, ASP-PCR, and NGS Performance for Lineage Assignment

RNA was extracted from primary nasopharyngeal samples from SARS-CoV-2 infected participants in the REMAP-CAP and COV003 trials and processed for lineage assignment via either PCR-RFLP, ASP-PCR, NGS, or multiple methods. For REMAP-CAP, this included 264 samples for PCR-RFLP, 717 samples for ASP-PCR, and 1,129 samples for NGS (51 by all three methods, 185 by PCR-RFLP and NGS only, and 610 by ASP-PCR and NGS only). For COV003, this included 365 samples for ASP-PCR and 641 samples for NGS (353 by the two methods). NGS results were assessed either for their ability to produce a lineage using the Pangolin software (NGS-Pangolin) or base calling over the SNPs targeted by ASP-PCR (NGS-SNP).

In terms of raw performance, PCR-RFLP, ASP-PCR, NGS-Pangolin, and NGS-SNP successfully typed 71.1%, 61.8%, 33.8%, and 39.5% of sampled tested in REMAP-CAP (Figure 4.3A-D). ASP-PCR, NGS-Pangolin, and NGS-SNP successfully typed 81.9%, 65.8%, and 59.8% in COV003, respectively (Figure 4.4A-C). The results for samples tested by NGS-Pangolin and PCR-RFLP and by NGS-Pangolin and ASP-PCR were compared. NGS-Pangolin was used for the

comparison rather than NGS-SNP as it is a more standard method. For PCR-RFLP vs NGS-Pangolin, the most common combination of results was successful designations by both methods (36.0%, Table 4.2A). Only 0.01% of the 236 samples failed on PCR-RFLP but were successful on NGS-Pangolin versus 31.4% with the converse relationship. For NGS-Pangolin vs ASP-PCR, the most common combination of results was being successfully typed by ASP-PCR and not typed by NGS-Pangolin (37.4% in REMAP-CAP, 48.4% in COV003, Table 4.2B and C). Samples were not randomized to screening method so direct comparison of raw performance may not be appropriate. High viral load samples successfully sequenced early in screening were diverted from the rapid methods; samples with insufficient viral load for sequencing, particularly in COV003, were only tested by PCR-RFLP or ASP-PCR. PCR-RFLP was developed earlier than ASP-PCR and only 51 samples that had failed typing with PCR-RFLP were screened by ASP-PCR. This selection bias led to significantly different viral load distributions between screening methods ($p = 1.6*10^{-6}$ for REMAP-CAP and $p = 6.3*10^{-6}$ for COV003, Anderson-Darling k-Sample Test, Figure 4.5A,B). REMAP-CAP samples tested by ASP-PCR had significantly lower median viral load than NGS ($4.40*10^4$ IU/mL vs $1.08*10^5$ IU/mL, $p = 0.002$, two-tailed Mann-Whitney U Test) as did those in COV003 ($4.72*10^6$ IU/mL vs $2.01*10^7$ IU/mL, $p = 8.85*10^{-6}$, two-tailed Mann-Whitney U Test). REMAP-CAP samples tested by PCR-RFLP had significantly higher median viral load than NGS ($2.38*10^5$ IU/mL vs $1.08*10^5$ IU/mL, $p = 0.007$, two-tailed Mann-Whitney U Test).

To facilitate direct comparisons of the methods, probit regression models were derived for both methods and trials to predict the likelihood of producing a lineage designation at a given viral load input (Figures 4.3D and 4.4D). Across the entire range of viral loads seen in either trial, the probit regression models demonstrated the clear superiority for making a lineage designation of ASP-PCR over both NGS-SNP and NGS-Pangolin. This advantage was most pronounced in samples with Ct values in the range of 26-30. ASP-PCR and PCR-RFLP largely showed near equivalent performance with overlapping confidence intervals for the majority of

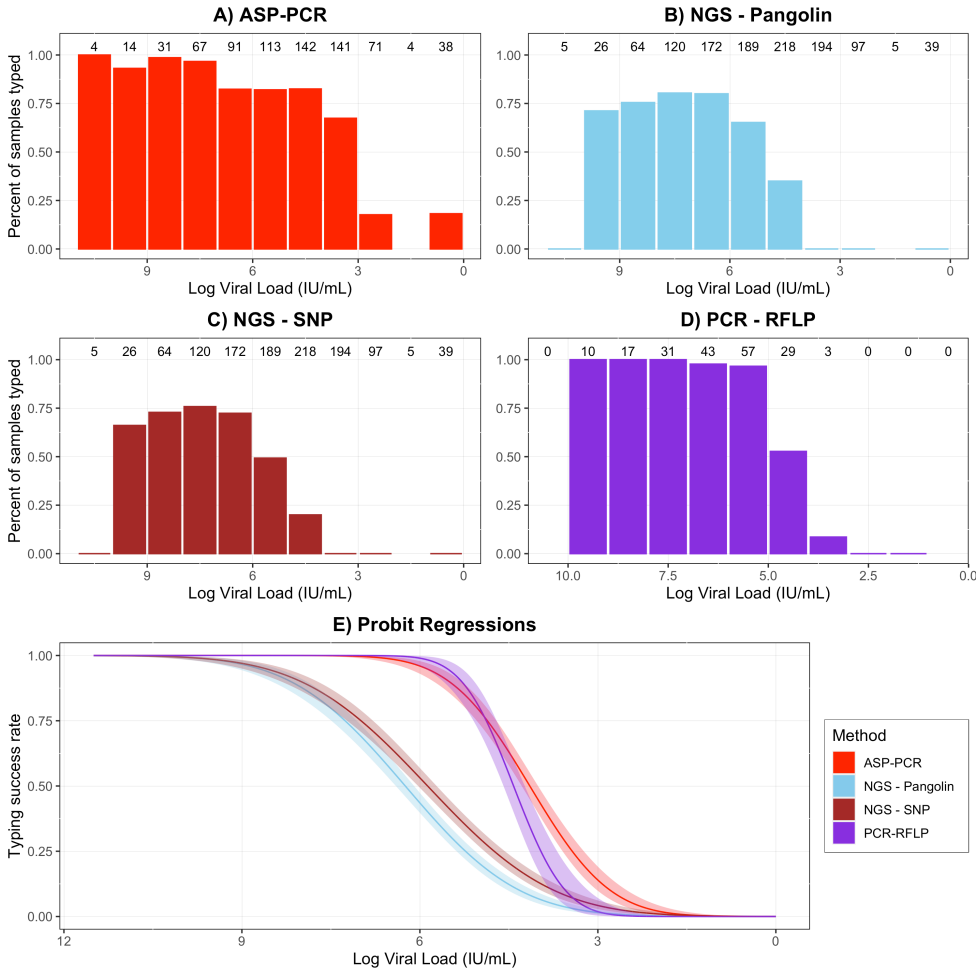


Figure 4.3: Performance of ASP-PCR, PCR-RFLP, and NGS in REMAP-CAP Trial. (A-D) Individual method performance on REMAP-CAP samples. Integers indicate total samples tested in one log data bins. (E) Probit regression of likelihood of lineage designation success for ASP-PCR, PCR-RFLP, NGS-SNP, and NGS-Pangolin with 95% confidence intervals derived from REMAP-CAP samples.

the viral load distribution. ASP-PCR did demonstrate moderate advantage over PCR-RFLP in the low viral load range ($<1.23*10^4$ IU/mL, Ct $\sim$ 32.3).

Concerns regarding the quality of RNA from several recruitment sites in COV003, motivated by shorter than expected average read lengths, prompted investigation of the utility of ASP-PCR for degraded samples. Degraded samples were defined as those that returned a viral load estimate $\geq 10^6$ IU/mL (Ct value $\sim$26 or less in CDC N1 RT-qPCR assay) but with fewer than $\leq 2*10^4$ bases with a read depth of at least two. Subsetting analysis to these samples (n = 113 for ASP-PCR, n = 118 for NGS-Pangolin and NGS-SNP) or those that were nondegraded (all samples

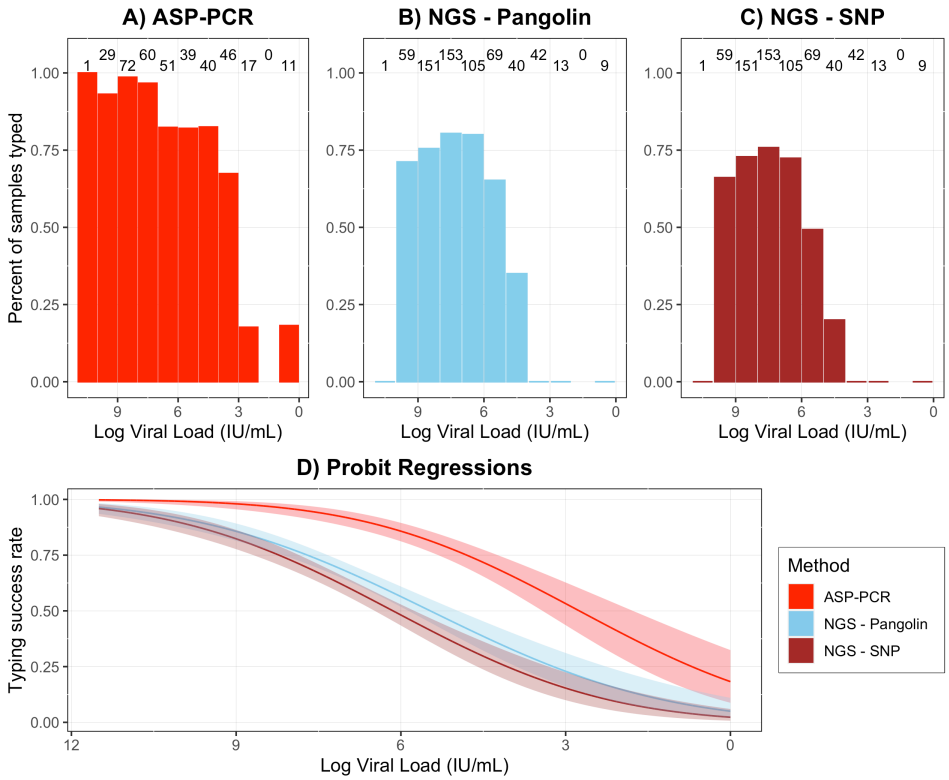


Figure 4.4: Performance of ASP-PCR and NGS in COV003 Trial. (A-C) Individual method performance on COV003 samples. Numbers above indicate total samples in one log data bins. (D) Probit regression of likelihood of lineage designation success for ASP-PCR, NGS-SNP, and NGS-Pangolin with 95% confidence intervals derived from COV003 samples.

that did not meet criteria for degradation; n = 252 for ASP-PCR, n = 523 for NGS-Pangolin and NGS-SNP), the advantages of ASP-PCR are clear (Figure 4.6). ASP-PCR was minimally impacted by the sample degradation status, successfully typing 95.6% of the 113 total degraded samples tested by ASP-PCR, while the utility of both NGS-Pangolin and NGS-SNP were massively compromised, only typing 11.0% and 0.0% of the 118 degraded samples processed by NGS, respectively. That degraded samples were defined by a parameter that itself determines the ability of NGS-Pangolin and NGS-SNP to make a lineage designation may overemphasize the impact of degradation status on the NGS based methods.

For the 85 REMAP-CAP samples with paired PCR-RFLP and NGS-Pangolin results, designations were concordant for 95.2% (81/85) of the samples. Two samples were called Alpha by PCR-RFLP and WT by NGS-Pangolin; two other samples

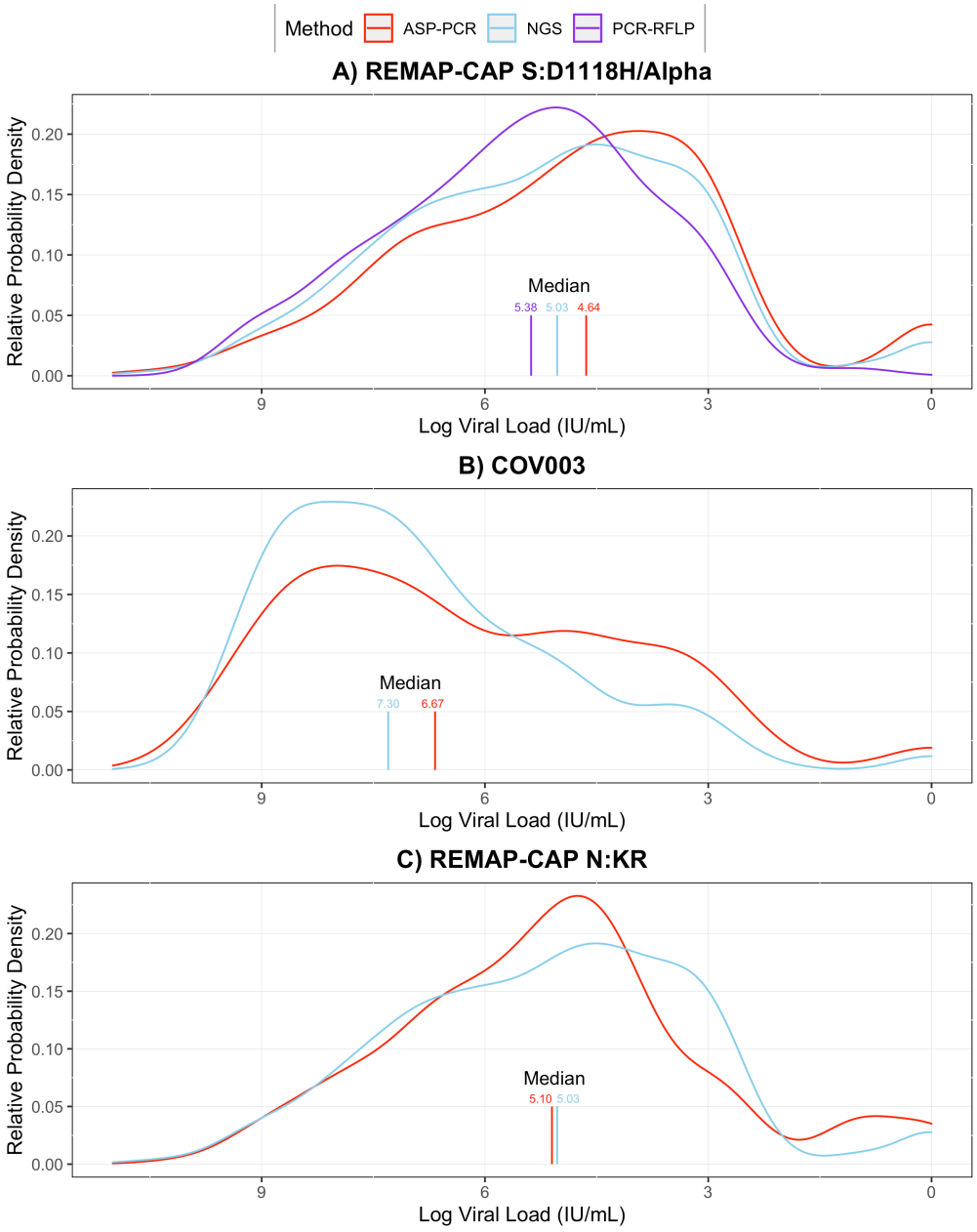


Figure 4.5: Distribution of Sample Viral Loads Tested by Screening Methods. Relative probability density and median viral loads of samples tested by each method in each trial. (A) REMAP-CAP S:D1118H/Alpha. (B) COV003. (C) REMAP-CAP N:KR. Distributions in (A) and (B) were significantly different (Anderson-Darling k-Sample Test).

A) REMAP-CAP: Performance of PCR-RFLP and NGS		PCR-RFLP	
		Successfully Typed	Not Typed
NGS-Pangolin	Successfully Typed	36.0% (85/236)	0.01% (3/236)
	Not Typed	31.4% (74/236)	31.4% (74/236)
B) REMAP-CAP: Performance of ASP-PCR and NGS		ASP-PCR	
		Successfully Typed	Not Typed
NGS-Pangolin	Successfully Typed	27.1% (179/661)	0.01% (5/661)
	Not Typed	37.4% (247/661)	34.8% (230/661)
C) COV003: Performance of ASP-PCR and NGS		ASP-PCR	
		Successfully Typed	Not Typed
NGS-Pangolin	Successfully Typed	34.6% (122/353)	0.04% (14/353)
	Not Typed	48.4% (171/353)	13.0% (46/353)

Table 4.2: Relative Method Performance on Matched Samples. Analysis only includes samples tested by both methods. Success defined as described in methods. Percentages may not sum to 100% due to rounding.

were called WT by PCR-RFLP and Alpha by NGS-Pangolin. In each case, SNP calling over the S:1118 amino acid using the NGS-produced sequence confirmed the NGS-Pangolin result. One of the samples incorrectly called Alpha by PCR-RFLP had the U24910C SNP that disrupts *SfcI* activity. For the 179 REMAP-CAP samples with paired ASP-PCR and NGS-Pangolin results, lineage designation was concordant for 99% (178/179) samples. A single sample was designated Alpha by ASP-PCR and the parent lineage B.1.1 by NGS-Pangolin – base-calling the sample indicated S:1118D and a false positive for ASP-PCR. For the 122 COV003 samples with paired ASP-PCR and NGS-Pangolin results, designations were concordant for 85% (104/122) of samples. All 18 discordant samples were called Gamma or Zeta by the ASP-PCR. However, this discrepancy was likely due to Pangolin miscalls due to low sequence coverage, as 17/18 of these discordant samples were assigned as a parent lineage of Gamma and Zeta by Pangolin (B.1.1 or B.1.1.28). These samples were subsequently typed as Gamma or Zeta by phylogenetic reconstruction, leading to a more accurate assessment of concordance for COV003 of 99% (121/122).

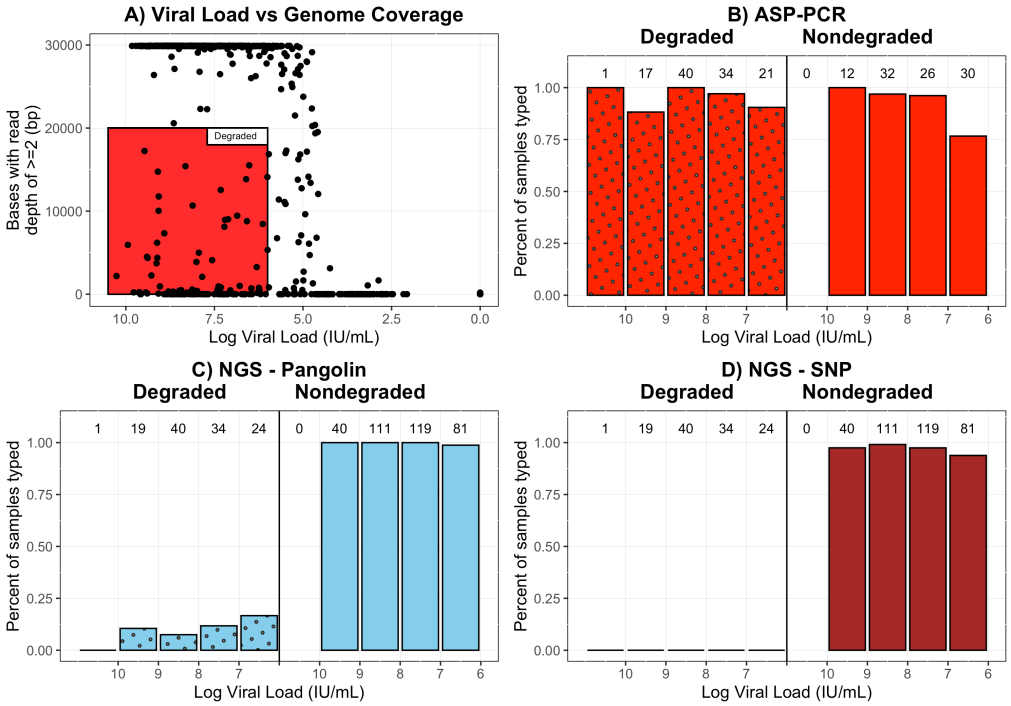


Figure 4.6: Impact of Degraded RNA in COV003 Samples on Method Performance. (A) Genome coverage of COV003 samples plotted vs sample viral load. Samples with $\leq 2*10^4$ bases with ≥ 2 reads and viral load $> 10^6$ IU/mL were defined as degraded (red box). (B-D) Individual method performance on COV003 degraded and nondegraded samples. Numbers above plots indicate total samples included in one log data bins.

4.2.5 ASP-PCR and NGS Performance for Detecting N:KR Mutation

To detect the double substitution N:KR, ASP-PCR and NGS-SNP approaches were used. All sequenced REMAP-CAP samples (n = 1,129) were evaluated using NGS-SNP and 285 samples were screened by ASP-PCR. NGS-SNP returned coverage over 46.8% of samples and ASP-PCR was successful for 67.7% of samples (Figure 4.7A,B). REMAP-CAP samples were screened instead of COV003 as clinical and demographic metadata was available for these patients, enabling other analyses (see Chapters Five and Six). In contrast to screening for the S:D1118H mutation, the distribution of viral loads tested by each method was not significantly different (p = 0.11, Anderson-Darling k-Sample Test, Figure 4.5C) and samples screened

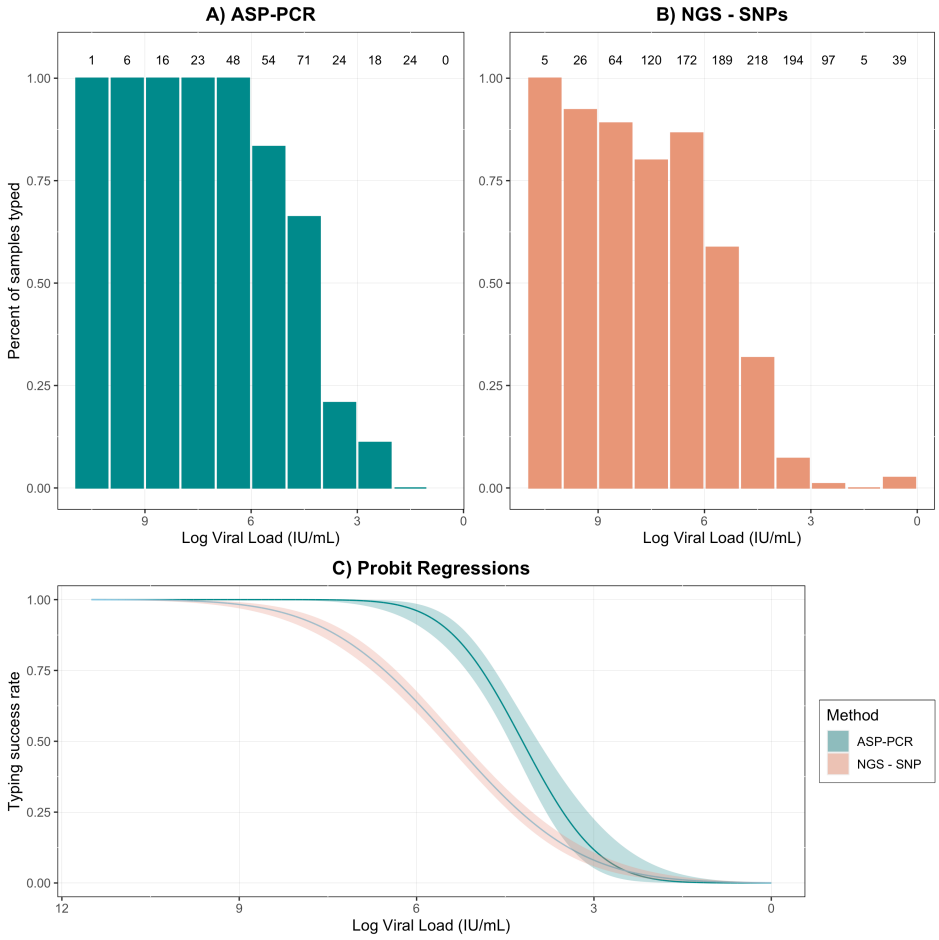


Figure 4.7: Detection of N:KR Mutation in REMAP-CAP Samples. (A-B) Individual method performance on REMAP-CAP samples. Numbers above indicate total samples in one log data bins. (C) Probit regression of likelihood of N:KR designation for ASP-PCR and NGS-SNPs with 95% confidence intervals derived from REMAP-CAP samples.

by ASP-PCR did not have significantly different median viral loads than NGS ($1.26*10^5$ IU/mL vs $1.08*10^5$ IU/mL, p = 0.71, two-tailed Mann-Whitney U Test). Similar to comparisons between NGS-Pangolin and the S:D1118H ASP-PCR, probit regressions displayed a substantial advantage for N:KR ASP-PCR over NGS-SNP in the Ct range of 26-30 (Figure 4.7C). Results were concordant for 95.3% (101/106) of samples with results for both approaches. All discordantly typed samples were called WT by NGS-SNP and N:KR by the N:KR ASP-PCR, suggestive of an issue with false positives in this assay.

4.2.6 Performance of ASP-PCR Across Diverse Targets

The performance of each individual ASP-PCR assay was compared using their respective probit regressions. In an attempt to avoid biases from degradation of samples in COV003 but not in REMAP-CAP, only samples not identified as degraded from COV003 were included in this analysis. In general, the performance for ASP-PCR assays was strongly dependent on the cohort in which the assay was tested, with the S:D1118H and N:KR probit regressions nearly overlapping for all viral loads (Figure 4.8). The S:D1118H regression, which had the narrowest confidence intervals, demonstrated slight superiority to both ORF1ab:L3468V and S:K417T at viral loads above $2.2*10^6$ IU/mL (Ct $\sim$ 25.7) and a clear deficit in performance to both assays at viral loads below $2.3*10^4$ IU/mL (Ct $\sim$ 31.5). However, this diminished performance at lower viral loads is due to both S:K417T and ORF1ab:L3468V ASP-PCR assays returning designations for some samples negative by N gene PCR. These samples were likely partially degraded (and hence artefactually low by N gene Ct), but the method used to identify degradation (NGS read depth) was unable to identify this characteristic in samples with viral loads $<10^6$ IU/mL.

4.2.7 Effect of Molecular Modifications

Altered reaction chemistry that increases the destabilization of a single base mismatch between target-oligo may improve the performance of ASP-PCR for low viral load samples and overcome the stable performance seen using kit reagents only. To measure the effects of the addition of low molecular amides (2-pyrrolidinone), oligonucleotides modified with LNAs, or both treatments, a panel of 83 samples of varying viral loads from REMAP-CAP were re-extracted from the primary samples and used to screen effects on ASP-PCR performance. In samples untreated with 2-pyrrolidinone, modification of oligonucleotides to LNAs was associated with a statistically significant increase in typing percentage ($p = 0.04$, Fisher's Exact Test; Figure 4.9). The improvement came from increased likelihood of typing low viral load samples; LNA-modified oligonucleotides typed 63.4% of samples with Ct values > 30 vs only 36.6% for original oligonucleotides. The general trend

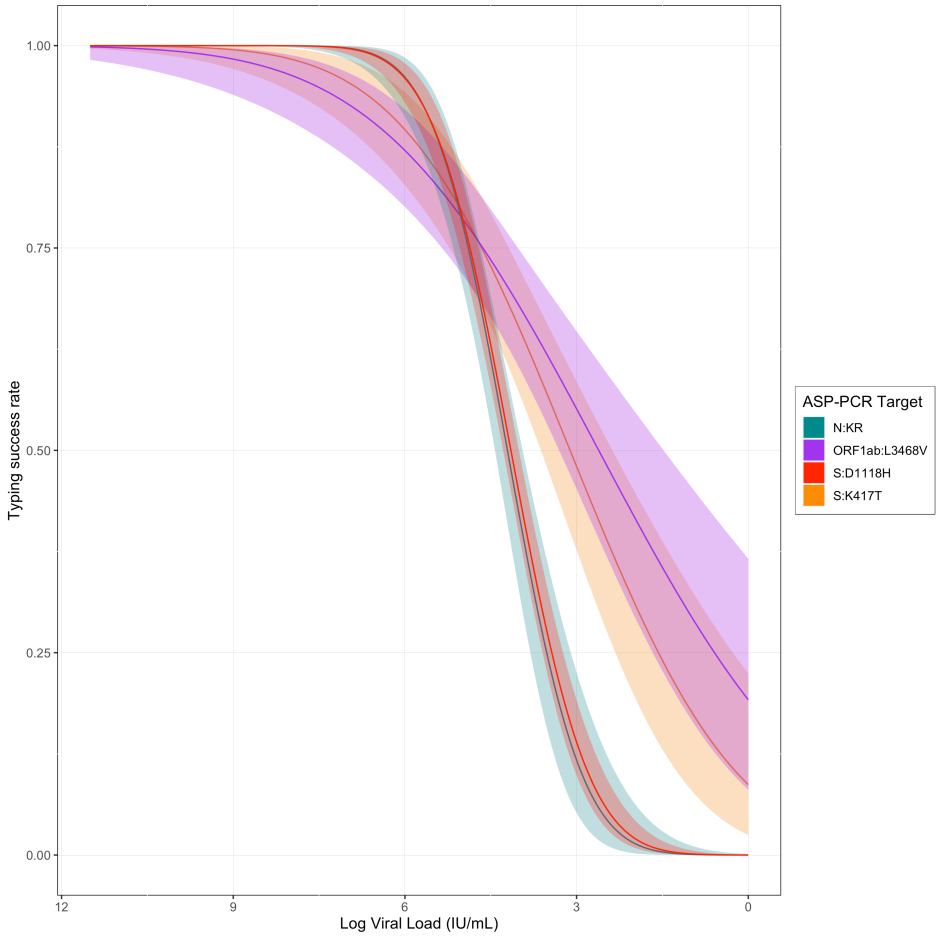


Figure 4.8: Performance of Individual ASP-PCR Assays on Nondegraded Samples. Probit regression of likelihood of N:KR or lineage designation for ASP-PCR with 95% confidence intervals. S:D1118H and N:KR regressions derived from REMAP-CAP samples. ORF1ab:L3468V and S:K417T derived from CoV003 samples not identified as degraded.

was to increase typing percentage with LNAs and decrease with 2-pyrrolidinone, but the other comparisons did not reach significance ($p > 0.05$, Fisher's Exact Test). While the addition of 0.4M 2-pyrrolidinone did decrease the raw signal observed for the discriminating probe, this improvement in discrimination came at the expense of lineage designation for viral load samples of all concentrations – the median raw change in signal in samples successfully typed by both treatments was 58.4% lower over the matched probe in the 2-pyrrolidinone reaction compared to the same sample in the untreated reaction. Additionally, LNA probes returned signals for several samples negative by N gene qPCR, which may represent false positives. There were no concerns related to degradation in the REMAP-CAP

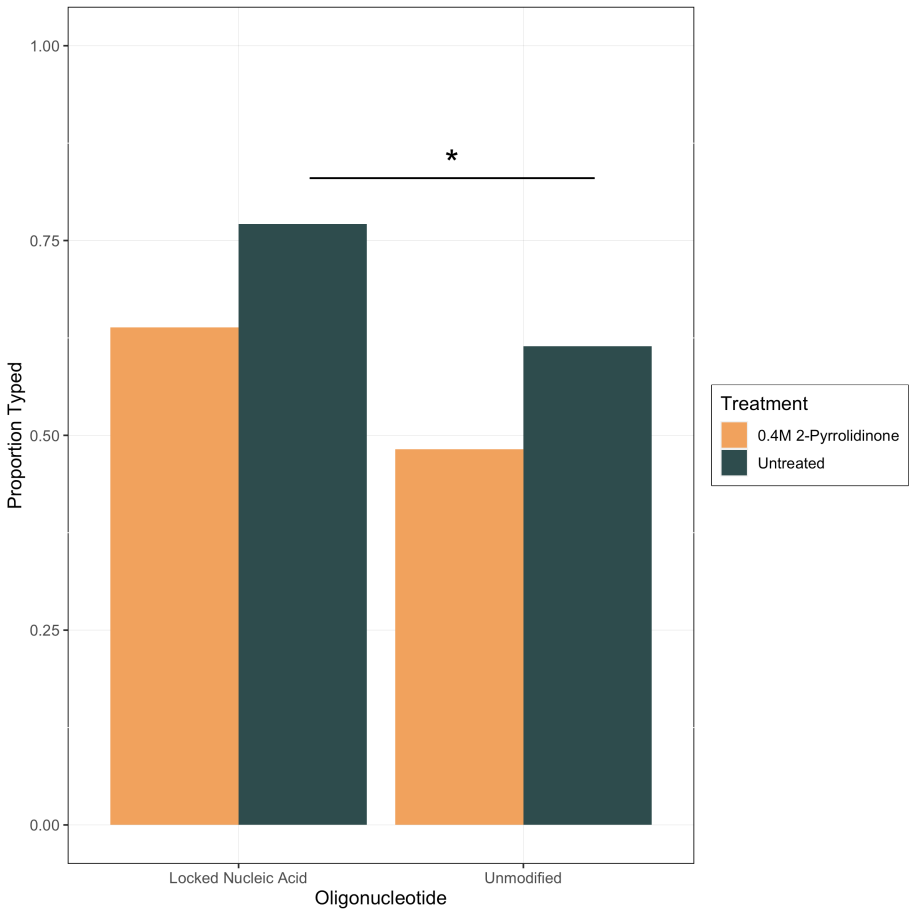


Figure 4.9: Impact of Molecular Modification on S:D1118H/Alpha ASP-PCR Performance. Percentage of total samples successfully typed using approach described in methods. * = $p < 0.05$ and from Fisher's Exact Test. ASP-PCR derived lineage designation (i.e., WT or Alpha) concordant for all samples by all methods.

samples on which these modifications were tested.

4.2.8 ASP-PCR on NGS Library Preparations

As in other NGS protocols, the pre-sequencing sample preparation of veSeq generates a library. This library has been enriched for SARS-CoV-2 RNA, reverse transcribed, amplified through 12 rounds of PCR, and purified to remove unbound primers using bead-based technology. To investigate whether this library preparation could be used as a direct input for ASP-PCR, 90 library preparation samples from the COV003 trial (hereafter "cDNA" samples) and paired RNA samples were processed using the S:K417T and L:3468V ASP-PCR methods. Cycling conditions were identical for both sample types, although the cDNA ASP-PCR did not include a reverse

transcription step in the RT-PCR protocol. The performance between sample types was remarkably consistent. On RNA samples, the S:K417T and ORF1ab:L3468V ASP-PCRs typed 88.9% (80/90) and 92.2% (83/90) of the samples, respectively. Using the cDNA samples, the S:K417T and ORF1ab:L3468V ASP-PCRs both typed 91.1% (82/90) of the samples.

The cDNA ASP-PCRs (S:K417T and ORF1ab:L3468V) returned a WT designation for a single sample that was negative by N gene PCR; the RNA ASP-PCR assays for this sample were both unsuccessful. Investigating the read data from the NGS processing of that sample revealed several (but very few) reads that mapped to the SARS-CoV-2 genome, lending evidence that this sample may have been highly degraded. These low reads may also be the result of contamination. The evidence points to cDNA ASP-PCR being more resilient to degradation that RNA ASP-PCR, possibly due to the PCR amplification during the library preparation in the veSeq protocol.

4.2.9 Predicted Performance of Typing Methods on Larger Cohorts

The output of a probit regression is a link function describing the likelihood of an outcome given an input variables. To estimate the utility of ASP-PCR over larger, potentially more representative populations, the REMAP-CAP S:D1118H/Alpha ASP-PCR, PCR-RFLP, NGS-Pangolin, and NGS-SNP probit regression models were applied over published datasets containing Ct data (Table 4.3). The Choudhuri et al. (2020) dataset included 1,044 patients admitted to hospital and sampled within 48 hours of admission in New York State during the first wave.[278] Median Ct values in Choudhuri et al. (2020) were lower than in REMAP-CAP, indicating a higher viral load (27.3 vs 29.8, $p = 5.18*10^{-12}$, two-tailed Mann-Whitney U Test). The Tso et al. (2021) dataset described viral loads in 57,517 patients sampled as part of community testing in the Texas.[279] Median Ct values in Tso et al. (2021) were higher than in CoV003 (23.8 vs 22.8, $p = 1.21*10^{-5}$, two-tailed Mann-Whitney U Test). ASP-PCR and PCR-RFLP had advantages over the NGS methods in both

Study description	Samples included	n	Viral load variable	Mean (SD)	Predicted ASP-PCR (95% CI)	Predicted PCR-RFLP (95% CI)	Predicted NGS-SNP (95% CI)	Predicted NGS-Pangolin (95% CI)
Patients sampled within 48 hours of hospital admission	All	1,044	Ct, unclear gene	27.63 (6.18)	74.1% (72.1, 76.0)	68.0% (64.3, 71.7)	47.2% (44.2, 50.1)	41.8% (39.0, 44.5)
Community testing in Texas	All	57,517	Ct, N gene	22.90 (5.97)	96.3% (94.9, 97.3)	91.7% (88.3, 94.1)	67.9% (65.0, 70.6)	62.9% (60.0, 65.6)
	Ct <26	34,718		18.83 (4.38)	99.9% (99.9, 100.0)	99.9% (99.5, 100.0)	88.6% (86.0, 90.7)	86.9% (84.3, 89.1)
	26 ≤ Ct ≤ 30	17,567		28.20 (1.10)	95.7% (92.9, 97.5)	87.1% (79.6, 92.2)	40.9% (37.4, 44.5)	30.5% (27.1, 34.1)
	Ct ≥ 30	5,232		31.05 (0.82)	73.4% (67.9, 78.4)	52.3% (42.8, 61.8)	21.7% (18.7, 24.9)	12.3% (9.87, 15.2)

Table 4.3: Predicted Performance of S:D1118H/Alpha ASP-PCR , PCR-RFLP, NGS-Pangolin, and NGS-SNP. Estimates generated by applying Ct-derived probit regression models of REMAP-CAP S:D1118H ASP-PCR over Ct values from each representative dataset. Results for community testing in Texas[279] shown by viral load range to demonstrate the majority of positive signal is derived from samples with Ct values <30.

the hospitalized and community testing cohorts, largely due to large proportions of samples in the Ct 26-30 range (30.5% of the Tso et al. (2021) samples were in this range, for instance). In the Tso et al. (2021) dataset, ASP-PCR was predicted to successfully type 96.3% of samples and PCR-RFLP 91.7% versus only 67.9% for NGS-Pangolin and 62.9% for NGS-SNP.

4.3 Discussion

The benefits of NGS extend far beyond lineage designation and can be applied to answer numerous research questions of public health importance including identifying genomic sites under positive selection, in-depth molecular epidemiology, and detection of novel variants.[290] Perhaps most importantly to this project, the design of ASP-PCR oligonucleotides and the very knowledge of what lineages to target is entirely dependent on the generation and dissemination of SARS-CoV-2 sequences produced via NGS. However, as evidenced here, there exist multiple use cases where rapid, SNP-typing based methods could be applied to improve variant

identification. In general, these methods would be best deployed in studies wherein the lineages of interest are known, requiring publicly available sequencing data in the study location; the expected frequency of the target lineage in the study population can be estimated; and there are not co-circulating SARS-CoV-2 variants that also contain the target SNP during the study period.

Appropriate SNP selection is the key to the interpretability of SNP-based typing assays. The global PPV and NPV estimates presented in table 10.1 are biased due to the analysis assuming random selection of any sequence that has been sequenced during the pandemic (ignoring sweeps of various lineages at different points in time) and equal sequencing coverage globally. Geographic and temporal restrictions for sequence analysis increase the accuracy of PPV and NPV estimates and could allow use of targets that appear less suitable in the global analysis. This impact is demonstrated in directly comparing results for Alpha, Gamma, and Zeta between tables 4.1 and table 10.1. In contrast, the global PPA and NPA values are less susceptible to these assumptions and should be the starting point for designing novel ASP-PCR oligonucleotide sets before refinement using temporal and geographic restrictions. A drawback of SNP-based typing assay deployment is that designs may need to be updated as new variants arise or global sequencing coverage improves. Selecting rare mutations (e.g. ORF1ab:L3468V for Zeta) rather than those potentially undergoing positive selection (e.g. epitopes in the RBD) may mitigate the risk posed by convergent evolution.

The necessary design of the S:D1118H PCR-RFLP assay specified that any disruption of the 5'-CTRYAG-3' restriction site of *SfcI* indicated an Alpha infection. While the G24915C mutation underlying S:D1118H was by far the most common mutation in the restriction site, the assay is unable to differentiate that SNP from any other and is therefore susceptible to false Alpha designations. Three of the 469 total sequenced REMAP-CAP samples that had coverage over the entire restriction site had SNPs other than G24915C that would abrogate restriction enzyme activity. These samples, had they been tested by PCR-RFLP, would have been false positives. PCR-RFLP assays where gain of activity is indicative of a

specific SNP would be more specific. However, this is entirely dependent on the largely stochastic nature of the surrounding base context of the genomic coordinates of lineage-defining SNPs. For PCR-RFLP, this requirement may be particularly challenging in the A/U rich genome of SARS-CoV-2 as restriction sites tend to associate with genomic regions with high GC content.[307] While infrequent these are not nonexistent, and the G28986T SNP in the N:M234I substitution defining Zeta leads to functionality for five restriction enzymes. In general, the specificity of ASP-PCR is higher than PCR-RFLP. In rare instances, other SNPs may also be present in the oligonucleotide binding sites for ASP-PCR. Importantly, these mismatches would impact both assay probes equally and would likely result in a decrease in sensitivity rather than specificity. This energetic impact can also be mitigated through the use of degenerate bases in the oligonucleotide sequence, such as the M base in the S:K417T probes for resilience to the S:K417N mutation indicative of the VoC Beta. As such, it is important to examine the diversity in the oligonucleotide binding sites when designing ASP-PCR assays.

Additionally, there are practical considerations for using ASP-PCR in place of PCR-RFLP for laboratories containing necessary equipment and personnel. ASP-PCR only requires addition of primary samples to prepared master mix and the reaction takes place on an automated system. PCR-RFLP, in contrast, has multiple steps that require manual handling of the samples, increasing risks of contamination particularly when transferring the 1st round PCR product to the 2nd round PCR master mix. The PCR-RFLP takes an absolute minimum of five hours to complete, not including time spent manipulating reagents, in contrast to around two hours for the ASP-PCR. Lastly, to screen samples at the magnitude relevant to the REMAP-CAP project ($\sim$2750 samples total), the reagent cost per sample for ASP-PCR was calculated to be $\sim$ £1.30 versus $\sim$ £4.65 for PCR-RFLP.

ASP-PCR may prove a useful alternative for analyzing stored samples where estimated viral loads are too low for NGS or where sample integrity suffers from known or suspected degradation. ASP-PCR, with lower implementation costs and quicker turnaround than NGS, may be most appropriately used in settings where

patients may present with lower viral loads and SARS-CoV-2 lineage data would be informative, such as in guiding treatment with monoclonal antibodies in hospitals and outpatient treatment centers. This use case is particularly important in the context of the emergence of Omicron where many monoclonal antibody therapies have seen large decreases in efficacy.[308] Lastly, the method may be an attractive alternative to NGS in resource-limited areas due to its reduced reagent costs, use of general microbiology equipment, and lower necessary operational technical expertise.

A further weakness of ASP-PCR compared to NGS is the inability of ASP-PCR to resolve mixed-lineage infections, recombinants, or samples with intra-host diversity at the site of interest. The assay generally performs with the differentiation between two types occurring due to dropout of one probe (i.e., type A has signal from one probe and type B has signal from both probes). Samples where there are mixed lineages will, in most instances, be detected as type B when extending the above example. For this reason, locations in the genome undergoing positive selection where diversity may be expected (such as the spike RBD) or that may be the targets of host editing enzymes (see Chapter Two) should be avoided.

Other attempts have been made to leverage RT-PCR to identify SARS-CoV-2 variants. The Thermo Fisher Scientific COVID-19 PCR TaqPath assay indirectly detects the S:H69/V70del through failure of one of three multiplex SARS-CoV-2 PCR targets (Spike Gene Target Failure (SGTF)) and as such can serve as both a primary diagnostic and variant screening tool.[309] SGTF is present in both Alpha and the Omicron sublineage BA.1. Lee et al. (2021) designed an Alpha allele-specific PCR (AS-PCR) based on the S:H69/V70del, S:Y144del, and S:A570D mutations with mismatches in the 3' end of the forward primer.[310] A potential weakness is that method does require at least two independent AS-PCR reactions to perform lineage identification for individual samples, although an approach using pooled samples, as described in the study, could be used for local lineage prevalence estimation. Babiker et al. (2021) constructed a multiplex, ASP-PCR based on SNP identification from dropout of probes targeting S:K417T/S:K417N and the S:E484K and S:N501Y mutations, although these are largely uninformative for

variant identification except on fine spatial scales (Tables 4.1 and 10.1).[311] The Vogels et al. (2021) multiplex RT-qPCR method discriminates Alpha, Beta, Gamma, and other by targeting the deletions ORF1ab:3675-3677del and S:69/70del.[312] It can also discriminate Delta, BA.1, and BA.2. Lastly, Harper et al. (2021) designed a genotyping panel for identifying 19 SNPS using PACE chemistry and allele-specific forward primers.[313] Multiplexed reactions to designate multiple SNPs in a single reaction would be highly advantageous but variable target performance (as seen in Harper et al. (2021)) would have to be carefully avoided or the assay sensitivity would be limited to the lowest performing oligonucleotide target. Lee et al. (2021) and Harper et al. (2021) both based their designs on 3' mismatches being lethal for PCR reactions, but experimental evidence challenges that assumption.[314] In addition, several commercial companies are now offering SARS-CoV-2 PCR assays based on the principles of ASP-PCR, but researchers will typically require knowledge of which SNPs have been targeted and their specificity for strain identification in their research areas. They may find greater cost-effectiveness and adaptability with in-house designed oligonucleotides.

The drop in performance for some high viral load samples for both NGS methods in COV003 versus REMAP-CAP was thought to be due to fragmentation of RNA during sample storage or shipment, as evidenced by shorter than expected library insert size and variable performance on samples from different recruitment sites.[267] The RT-qPCR and ASP-PCR amplicon length likely impacted their respective performance for COV003 samples, as highly fragmented RNA may not have had intact RNA spanning primer binding sites. For COV003, the shorter amplicon (S:K417T) performed slightly better for high viral load samples, but the confidence intervals from probit regression overlapped for all viral loads. On rare occasions in the COV003 trial, ASP-PCR designated lineages for samples that had tested negative on N gene qPCR. This phenomenon did not occur on negative REMAP-CAP samples, leading to the conclusion that these were likely samples where fragmentation disrupted the N gene target but not the other regions targeted by ASP-PCR. Retesting these samples with a multiplexed quantification reaction

with primers across the genome would provide additional clarity but was beyond the resource limits of this study.

The advantage of veSeq over the ARTIC protocol for degraded samples is due to their respective design and implementation. After a size selection removes nucleic acids shorter than 400 nt, veSeq uses a target capture approach based on 120 nt long molecular probes with sequence homology to the target pathogen to bind target nucleic acids.[250] Due to their size, the 120 nt probes are resistant to some diversity in the target. Provided fragmented nucleic acids pass the size exclusion, one or more molecular probes should bind with sufficient affinity to pull down the nucleic acid, which is then amplified non-specifically and sequenced. The ARTIC protocol, in contrast, uses two pools of primers to generate overlapping 400 nt amplicons.[315] These primers are sensitive to substantial drops in performance due to minor changes in primer binding sites (the same principle used by ASP-PCR), as has been demonstrated repeatedly when novel variants have arisen.[312, 316] Due to this, fragmentation at a primer binding site will result in loss of that amplicon. The current formulation of ARTIC primers (`https://github.com/artic-network/artic-ncov2019/blob/master/primer_schemes/nCoV-2019/V4/SARS-CoV-2.primer.bed`) cover 4929 bases in total, so there is roughly a one in six chance that an RNA break will remove a primer binding site. Further, breaks in the overlapping region between adjacent amplicons will result in neither amplicon being properly generated, as the forward and reverse primers are not able to rendezvous on a single amplicon.

In this study, four different mutations in the SARS-CoV-2 genome were targeted by ASP-PCR, including one transition (S:D1118H; G>C), two transversions (S:K417T and ORF1ab:L3468V; both T>G), and a three nucleotide double substitution (N:KR; GGG>AAC). Each mutation may lead to a different energetic cost between the target and oligo due to the mutation type (e.g., it was hypothesized that a transversion, which would lead to a purine-purine clash, would be more destabilizing than a transition) or the adjacent base context (e.g., adenines or thymines, having only two hydrogen bonds with their complementary base, may

have a lower ability to stabilize the steric clash from adjacent unmatched bases than cytosines or guanines). While all targets were intentionally chosen with an adenine or thymine up- and downstream of the target, the S:D1118H SNP is the only ASP-PCR with cytosines or guanines two bases 3' and 5' of the target. It was though that the diversity in targets may lead to lessons learned in the selection of target SNPs. However, the targets behaved remarkably similarly despite this diversity, lending support to the claim that these assays represent the theoretical maximum performance using standard RT-qPCR chemistry and de-emphasize the importance of adjacent base context when selecting between multiple candidate SNPs.

In an attempt to increase the performance of the ASP-PCR assay, two chemical additives/modifications were trialed. Previous research on improving the specificity of PCR has shown that addition of low-molecular weight amides can reduce the amplification of off-target PCR products.[273] It is theorized that the mechanism of action for organic additives is destabilization of template double-helices;[317] this effect would partially explain the drops in raw signal seen in samples treated with 2-pyrrolidinone and decreased amplification for most samples. Modifications that increase the relative destabilization caused by a single base mismatch (as with LNAs) rather than disrupting binding dynamics of all oligonucleotides (as with 2-pyrrolidinone) should be prioritized. These modifications would further support the development of multiplexed ASP-PCR assays as the effective reaction temperature range would be expanded and could accommodate a wider range of target primers.

The success of ASP-PCR on material from veSeq library preparations may have important consequences on expanding diagnostic access. In instances where remote laboratories have the personnel and equipment to generate a library but do not have a sequencer on site, completed libraries can be shipped to processing laboratories at 4°C. In contrast, extracted RNA must be shipped on dry ice, increasing costs and limiting accessibility, or the primary sample itself must be shipped, increasing complexity, applicable regulations for infectious substances, and time. Shipped library preparations can then be used for sequencing, ASP-PCR, and/or other PCR-based chemistry.

One major benefit of ASP-PCR is the flexibility and adaptability to target new SARS-CoV-2 SNPs as they arise. Based on this experience, SNP targets should fulfill most or all of the following criteria to ease design and optimization:

- Follow best practices for primer/probe design as specified by the RT-PCR mix being used, including for GC content, presence of GC clamp, and avoidance of dinucleotide repeats and runs of bases. Place the target mismatch in or near the middle of the probe.
- If possible, choose a target SNP with an adenine or uracil up- and downstream of the mismatched base to increase the energetic cost of the mismatch.
- All oligonucleotides for ASP-PCR should be designed to melt at similar temperatures (contrasting best practice for RT-qPCR).
- Optimization should begin with the annealing/extending temperature ∼5°C below the lowest probe T_m. Decrease the temperature to increase raw signal; increase the temperature to increase discrimination. Reaction conditions cannot easily be predicted computationally and optimization should leverage cDNA aliquots of sequence confirmed samples.
- Label both probes with bright fluorophores, such as 6-FAM and VIC/HEX.

Selection of SNPs should also consider the potential selective pressures on the mutated sites. Concerns of the use of sites under high positive selection have already been mentioned. On the other hand, sites that are neutral or undergoing purifying selection may be more likely to revert to their WT base. In general, the objections to using sites under positive selection due to convergent evolution in multiple VoCs/VoIs (and thus an inability to distinguish between the two, e.g., S:N501Y in Alpha, Beta, Gamma, and Omicron) can be mitigated by deploying ASP-PCR on temporally- and geographically-restricted patient populations where sequencing data specifies expected VoCs/VoIS to be encountered. In this study, sites were generally chosen for their suitability for ASP-PCR more than for their potential long-term use in light of ongoing evolution.

Understanding the emergent mutations and the properties of those mutations in novel human-infective RNA viruses requires tools to process samples suboptimal for traditional methods such as NGS. The principles underlying ASP-PCR and PCR-RFLP can be readily applied to novel pathogens. ASP-PCR, in particular, is a highly accurate and sensitive method for discriminating SARS-CoV-2 lineages and its application to support specific research aims is well supported by the data presented in this study. The flexibility of the assay will allow for novel designs for emerging variants (such as Omicron) that can be readily validated and implemented using the lessons learned presented here. Extending the assay to other viruses where established SNP-based typing is clinically or epidemiologically informative (such as HIV-1) should be prioritized.[318]

5

Viral Load Diversity and Dynamics in Patients Infected with SARS-CoV-2

Contents

5.1 Introduction 90
5.2 Results 93
5.2.1 Diversity of Viral Load 93
5.2.2 Emergence of SARS-CoV-2 variants in Patient Cohorts 100
5.2.3 Viral Load Associations with Patient Clinical and Demographic Factors 102
5.2.4 Subgenomic RNA 103
5.2.5 SARS-CoV-2 Viral load dynamics in intensively sampled patients hospitalized with COVID-19 107
5.2.6 Association of viral dynamics classification with patient and virological factors 109
5.3 Discussion 111

Abstract

SARS-CoV-2 viral loads influence patient- and population-level outcomes. Varying inherent replicative capacity may explain differences in viral load between SARS-CoV-2 lineages, which can, in turn, determine transmissibility. To better understand the virological and host factors that influence viral load, SARS-CoV-2 RNA from patient samples collected as part of two COVID-19 clinical trials was quantified

by RT-qPCR. Viral loads in the COV003 trial, which recruited patients at time of symptom onset, were significantly higher than in REMAP-CAP, which recruited patients at time of ITU admission. The Alpha variant emerged during recruitment in REMAP-CAP and the Zeta, Gamma, and Delta variants sequentially emerged in COV003. Viral loads were found to be influenced by infecting variant, with Delta and Gamma infections having higher viral loads than Zeta and WT in COV003 and Alpha having higher than WT in REMAP-CAP. The signal for Alpha in REMAP-CAP was driven almost entirely by seropositive patients. The polyphyletic N:R203K/G204R mutation was associated with reduced viral loads in WT REMAP-CAP infections. N gene sgRNA was quantified by RT-qPCR via targeting the conserved leader sequence common to all SARS-CoV-2 transcripts. Relative sgRNA abundance was stable across all viral loads and is unlikely to be useful as a marker for active cellular infection. Alpha samples demonstrated remarkably higher relative sgRNA abundance than WT, Zeta, or Gamma and this increase was likely not due to the S:N501Y substitution or N:R203K/G204R double substitution in isolation. Lastly, viral load dynamics in intensively sampled REMAP-CAP patients revealed substantial diversity in viral trajectories, with 19.0% of patients having no detectable SARS-CoV-2 RNA throughout treatment while conversely 21.5% of patients failed to clear the virus at any point during their stay in ITU. Patient viral load should be used for subgroup analysis in ongoing and new COVID-19 clinical trials. Future trials including a wider range of collected metadata, especially time since symptom onset, would allow for a more complete understanding of the host factors influencing viral load.

5.1 Introduction

Viral replication can be detected using semiquantitative and quantitative methods for measuring the generation of pathogen-specific material. These assays include NAAT methods for amplifying viral genomic material, immunochromatographic assays that bind viral antigens (e.g., lateral flow tests and ELISAs), and *in vitro* assays to detect virus-driven cytopathic effects.[319] Each of these methods allow for

direct or indirect assessment of the concentration of virions in a clinical sample and thus a patient's viral load. For a novel virus such as SARS-CoV-2, understanding the intra-host dynamics of the virus is necessary for clinical decision making (e.g., when a patient can safely be transferred from an isolation ward), implementation of non-pharmaceutical interventions (e.g., the length of time to wear face masks following infection), strategies for asymptomatic screening, and more.[320, 321]

SARS-CoV-2 was initially assumed to behave with an acute trajectory similar to other respiratory viruses. Indeed, the majority of studies of SARS-CoV-2 dynamics have shown rapid growth of viral load within several days of infection, followed by a gradual decrease temporally related to the onset of symptoms.[322–326] However, many studies have also demonstrated a prolonged shedding period where patients test positive for SARS-CoV-2 on an RT-qPCR test for close to three weeks post the initial positive test.[327, 328] Some case studies have described immunocompetent individuals testing positive for greater than 50 days.[329–331] This length of time is in contrast to the median durations of shedding for RSV and IAV of 9.5 and 5.7 days, respectively, although persistent shedding has been described, infrequently, after infection with bocavirus or rhinovirus.[328, 332] The prolonged shedding in a minority of patients is evidence of substantial inter-patient diversity and motivates directed study.

SARS-CoV-2 viral load has a role in both individual- and population-level outcomes. Transmissibility and therefore epidemic spread within a population are linked to viral load with clear mechanistic explanations.[333, 334] The role of viral load on COVID-19 clinical manifestations is less clear. In other viral infections, such as Chikungunya virus (CHIKV), H1N1 IAV, and other lower respiratory tract viruses, higher patient viral load at diagnosis has been associated with increased severity of symptoms.[335–337] However, many community surveillance-based studies of COVID-19 have found no difference in SARS-CoV-2 viral load between asymptomatic and symptomatic individuals.[338–341] There does appear to be an association between increased clinical severity and duration of PCR positivity, although the evidence is mixed.[323] Several studies have shown that viral load at

time of admission to hospital with COVID-19 is associated with increased mortality, but this may be an artefact of time since symptom onset.[342, 343]

Like other respiratory viruses, SARS-CoV-2 viral load has primarily been defined as the quantity of genomic material present within a nasopharyngeal or oropharyngeal swab sample. RT-qPCR allows for relative quantification based on the indirect detection of polymerase activity by the accumulation of fluorescent signal as PCR products are generated. The relative viral load is inversely related to the number of amplification cycles necessary to generate signal above a certain threshold (the cycle threshold (Ct)). If RT-PCR is performed on eluate from a known volume of primary sample and the assay includes calibrated standards, absolute quantification can be achieved, allowing for direct comparison with samples processed in other laboratories. Viral load is generally meant to reflect the total copies of genomic RNA (gRNA) and thus the number of virions (assuming one copy of gRNA per virion). However, as introduced in Chapter One, SARS-CoV-2 gene transcription necessitates the production of sgRNAs to express one-third of the viral proteome. If these shorter transcripts contain the target sites for the oligonucleotides used for qPCR, the assay is unable to discriminate between the two targets. As such, viral loads with qPCR targets in the structural proteins (e.g., E, M, and N) can be considered the sum of both gRNA and sgRNAs. Any shifts in sgRNA transcription may anomalously appear to be a change in gRNA and virion production, leading to incorrect interpretation of the changes in viral load within a patient.

The intra-host dynamics of SARS-CoV-2 are further complicated by the ongoing adaptive evolution and emergence of VoCs and VoIs that may have altered replicative capacity. Several mutations on the spike protein (e.g., N501Y, D614G, S477N, and P681H/R) have been implicated in increased infectivity of human cells *in vitro*, which may lead to altered tissue tropisms and more rapid spread within the host.[344, 345] As yet unidentified mutations in the Alpha variant increase its resistance to IFN, dampening the early innate immune response.[346, 347] Interactions between individual mutations and viral load alter the characteristics of the virus and may have roles in the increased severity and transmissibility of successive VoCs. As

introduced in Chapter Four, the REMAP-CAP convalescent plasma (CP) trial and COV003 COVID-19 vaccine trial are two well-characterized cohorts that collectively include >2000 patients with quantified SARS-CoV-2 viral loads. To investigate potential interactions between SARS-CoV-2 viral load, lineage of infection, and patient clinical and demographic factors, the amplification data generated from these two large patient cohorts was leveraged. Lastly, to better distinguish the role of sgRNA on total viral load detected, an sgRNA-specific qPCR assay targeting the N gene used for quantification was generated and used to screen a subset of samples from both trials.

5.2 Results

5.2.1 Diversity of Viral Load

To investigate the diversity of viral loads in patients infected with SARS-CoV-2, amplification data generated as part of broader efforts in two large clinical trials was leveraged. For REMAP-CAP, viral loads and anti-S gene IgG antibody status were determined using respiratory and serum samples collected from patients prior to treatment in the CP trial. All randomized patients were enrolled within 48 hours of admission to ITU and had a prior positive SARS-CoV-2 NAAT test. Demographic and other clinical information was collected, but time since symptom onset was not. In COV003, patients experiencing COVID-19-like respiratory symptoms presented to clinical recruitment sites for sampling. All patients had a positive NAAT test performed at the trial site before primary samples were shipped to a central laboratory for re-extraction and quantification. Clinical and demographic data, including vaccination status, was not available for the COV003 trial for the purpose of this thesis. Lineage of infection in both trials were determined using a combination of NGS, ASP-PCR, and PCR-RFLP (introduced in Chapter Four). Throughout this chapter (and others), apparent inconsistencies in population denominators between analyses are due to incomplete patient metadata.

As the COV003 trial sampled individuals close to symptom onset while REMAP-CAP enrolled individuals further along in the disease course, it was hypothesized

that the viral loads among swabbed COV003 patients (n = 905) would be higher than in REMAP-CAP (n = 1427). As expected, median viral loads in the COV003 trial were significantly higher ($3.20*10^7$ IU/mL vs $5.95*10^4$ IU/mL, p = $6.63*10^{-119}$, two-tailed Mann-Whitney U Test; Figure 5.1). Viral loads in REMAP-CAP patients that had seroconverted (n = 954) were significantly lower than in those who had not (n = 421) ($2.38*10^4$ IU/mL vs $1.23*10^5$ IU/mL, p = $1.08*10^{-31}$, two-tailed Mann-Whitney U Test; Figure 5.2). The difference between COV003 and REMAP-CAP was maintained when only including seronegative REMAP-CAP patients (n = 421) ($3.20*10^7$ IU/mL vs $1.23*10^5$ IU/mL, p = $2.29*10^{-129}$, two-tailed Mann-Whitney U Test; Figure 5.1). 10.8% (158/1458) of patients in REMAP-CAP had no detectable SARS-CoV-2 RNA at baseline; 88.2% (127/144) of these SARS-CoV-2 negative patients had seroconverted. 1.6% (15/920) of patients in COV003 had no detectable RNA.

Lineages detected in REMAP-CAP were delineated into WT (n = 573) and Alpha (n = 320) using ASP-PCR, PCR-RFLP, or NGS methods. Each lineage included a mix of seropositive and seronegative patients, with 60.4% (335/554) of WT infections of known serostatus having seroconverted at baseline along with 64.8% (195/301) of Alpha infections of known serostatus. Viral loads in patients infected with WT were marginally but significantly lower than those infected with Alpha, with median viral loads of $5.88*10^5$ IU/mL and $1.75*10^6$ IU/mL, respectively (p = 0.00012, two-tailed Mann-Whitney U Test; Figure 5.3A). Curiously, there was no difference in viral loads between seronegative WT-infected and seronegative Alpha-infected individuals (median viral loads of $3.52*10^6$ IU/mL and $4.65*10^6$ IU/mL, respectively, p = 0.39, two-tailed Mann-Whitney U Test; Figure 5.3B) and the difference in viral load between WT and Alpha lineages was driven almost entirely by the seropositive population (median viral loads of $1.58*10^5$ IU/mL and $1.23*10^6$ IU/mL, respectively, p = $2.25*10^{-8}$, two-tailed Mann-Whitney U Test; Figure 5.3C). While seroconversion in WT-infected individuals was associated with a 22-fold reduction in viral load, seroconversion in Alpha-infected individuals only reduced the viral load 4-fold. While minor in magnitude, a difference between serodiscordant

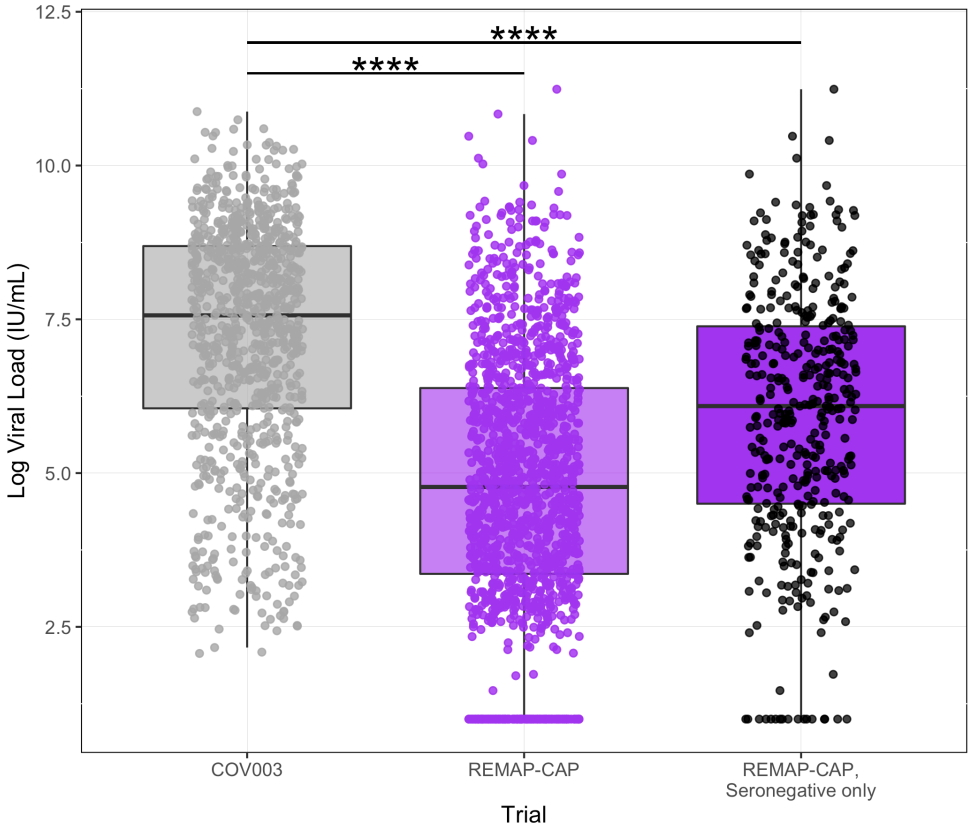


Figure 5.1: Diversity in Viral Load between Trials. Viral loads of patients enrolled in the COV003 trial (n = 905), REMAP-CAP trial (n = 1427), and those in the REMAP-CAP trial seronegative at time of enrollment (n = 421) as measured by RT-qPCR. **** = $p < 0.00005$ and from two-tailed Mann-Whitney U Test.

Alpha patients was still significant ($p = 0.005$, two-tailed Mann-Whitney U Test). As expected, a significantly higher proportion of patients where typing methods failed to identify the infecting lineage had seroconverted (364/453, 80.3%) compared to those successfully typed (530/855, 62.0%) ($p = 1.67*10^{-11}$, Chi-Squared Test).

As introduced in Chapter Four, the N:R203K/G204R (N:KR) mutation has previously been associated with an increase in viral load amongst hospitalized patients in Saudi Arabia.[304] To investigate the role of the N:KR mutation within our hospitalized REMAP-CAP cohort, patients were assigned as N:RG (WT) or N:KR based on sequencing coverage over genome coordinates 28881 though 28883 or the results of an N:KR-specific ASP-PCR assay. In total, 591 samples had sequencing coverage and 193 samples had ASP-PCR designations for a deduplicated total of 677 samples with N:203/204 classifications. Median viral loads between

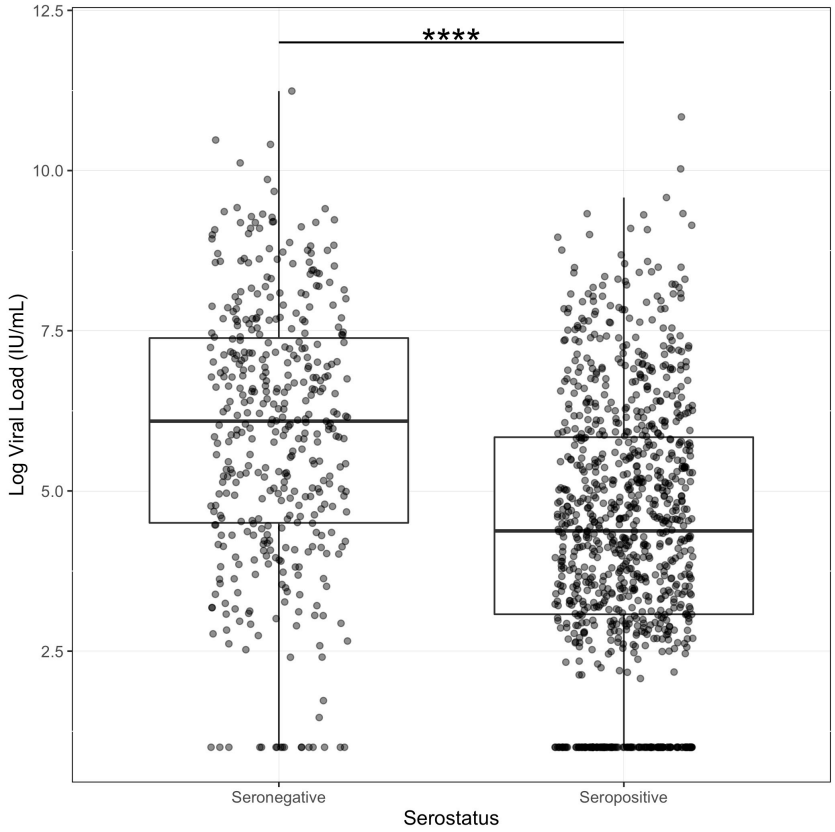


Figure 5.2: Impact of Seroconversion on Viral Load in REMAP-CAP. Difference in viral load at baseline for seronegative (n = 421) and seropositive (n = 954) REMAP-CAP patients. **** = $p < 0.00005$ and from two-tailed Mann-Whitney U Test.

samples containing N:RG (n = 360) or N:KR (n = 317) were not significantly different ($1.32*10^6$ IU/mL vs $1.90*10^6$ IU/mL, p = 0.60, two-tailed Mann-Whitney U Test; Figure 5.4A).

34.3% (227/661) of samples with both an N:203/204 classification and a designated lineage were Alpha. 100% (227/227) of N:203/204-classified Alpha samples had the N:KR mutation, leading to Alpha being 73.2% (227/310) of total N:KR samples. To limit the potential confounding of other adaptive mutations within the Alpha variant, the analysis of the N:KR mutation was limited to only those samples identified as WT (n = 434). Of these samples, 351 were N:RG and 83 were N:KR. In contrast to expectations, the N:KR mutation was associated with a significant decrease in median viral load compared to N:RG in WT lineage samples ($3.90*10^5$ IU/mL vs $1.43*10^6$ IU/mL, respectively, p = 0.007, two-tailed Mann-Whitney

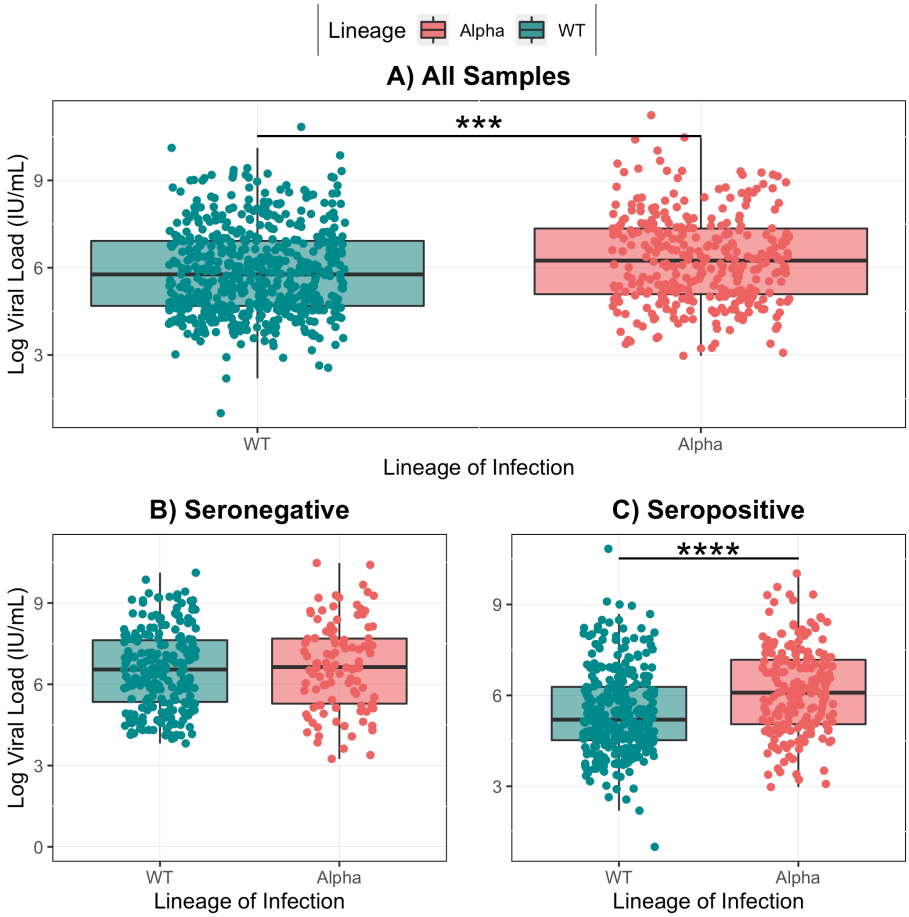


Figure 5.3: Viral Loads by Lineage of Infection in REMAP-CAP. Viral loads in REMAP-CAP trial for those infected by (A) WT (n = 573) and Alpha (n = 320). Analysis also limited to (B) seronegative WT (n = 219) and Alpha (n = 106) and (C) seropositive WT (n = 335) and Alpha (n = 301). *** = $p < 0.0005$, and **** = $p < 0.00005$ and from two-tailed Mann-Whitney U Test.

U Test; Figure 5.4B). This decreased viral load for N:KR vs N:RG was present comparing both seropositive (medians of $3.51*10^5$ IU/mL vs $6.83*10^5$ IU/mL) and seronegative (medians of $7.97*10^5$ IU/mL vs $4.85*10^6$ IU/mL) populations, although differences were not significant ($p = 0.073$ and 0.051, respectively, two-tailed Mann-Whitney U Test). These results suggest the N:KR mutation is associated with a reduction in viral load in intensely sick COVID-19 patients and that this is not an artefact of delayed admission to ITU.

Samples in COV003 were delineated into WT (n = 105), Zeta (n = 238), Gamma (n = 386), and Delta (n = 14). Viral loads across the four lineages were significantly different ($p = 1.98*10^{-18}$, Kruskal-Wallis Rank Sum Test; Figure

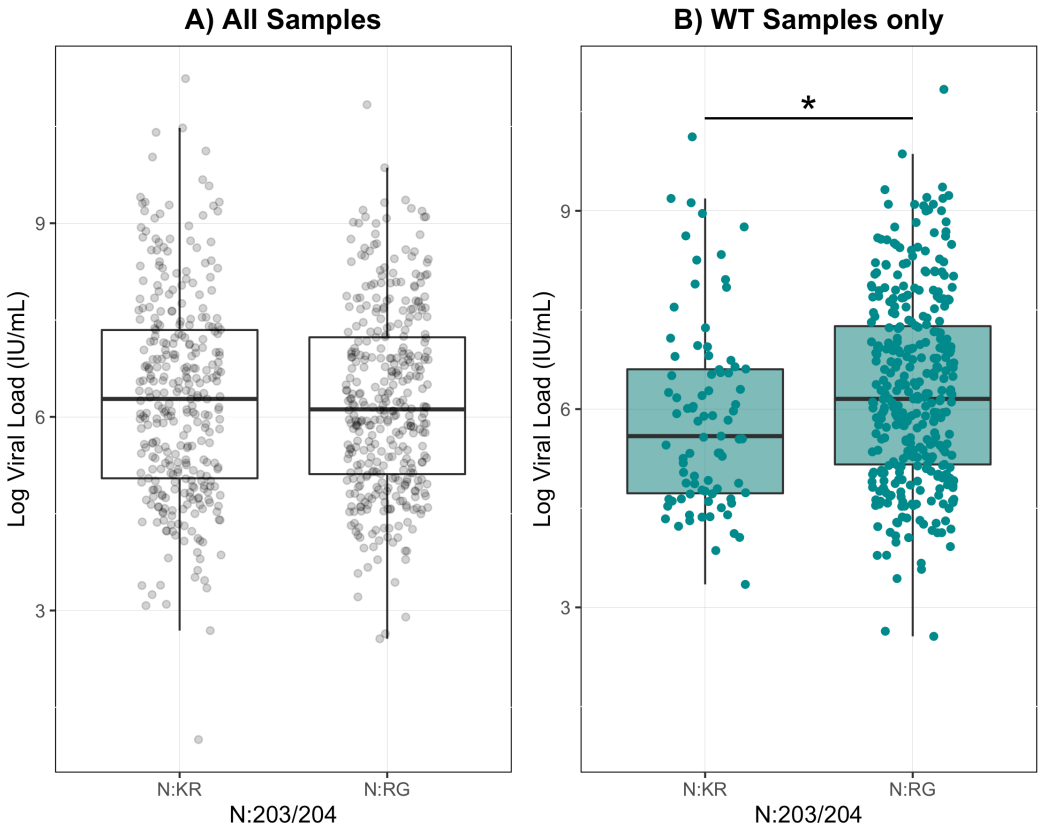


Figure 5.4: Viral Loads by Presence of the N:KR Mutation in REMAP-CAP. Presence or absence of N:KR mutation was confirmed by sequencing (n = 591) or ASP-PCR (n = 87). (A) Analysis includes all samples with designation (n = 677), including Alpha samples. (B) Analysis limited to only samples separately designated as WT (n = 434). * = $p < 0.05$ and from two-tailed Mann-Whitney U Test.

5.5). Median viral loads increased in the order of WT<Zeta<Gamma<Delta, although in direct comparisons with adjacent variants only the difference between Zeta ($1.75*10^7$ IU/mL) and Gamma ($2.22*10^8$ IU/mL) reached significance ($p = 8.93*10^{-15}$, two-tailed Mann-Whitney U Test). A limitation of the ASP-PCR assay and classification based on lineage-defining SNPs is that the broad "WT" designation invariably includes lineages that, while not labelled as VoC or VoI, have intrinsic differences in replicative capacity. For example, while not statistically significant, sequenced COV003 samples that clustered with the Pangolin lineage B.1.1.28 through phylogenetic reconstruction showed higher median viral loads than those that clustered with B.1.1.33 ($1.04*10^7$ IU/mL vs $1.34*10^6$ IU/mL, p = 0.082, two-tailed Mann-Whitney U Test). Further, COV003 WT samples had higher viral

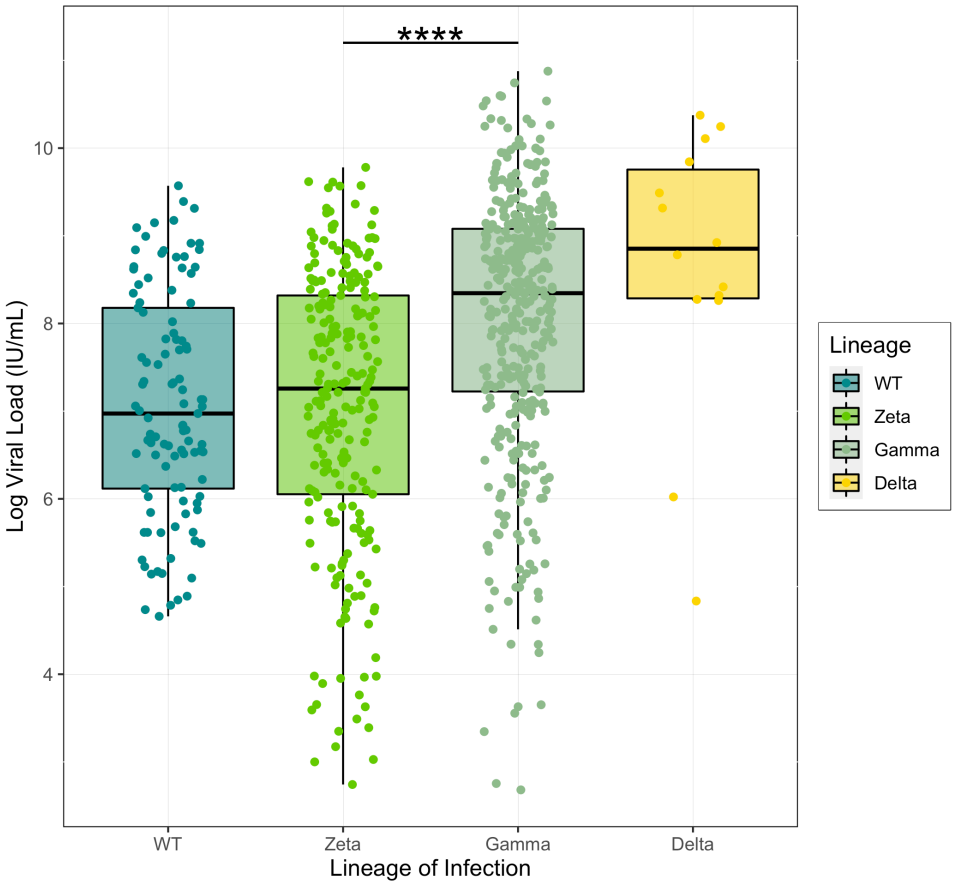


Figure 5.5: Viral Loads by Lineage of Infection in COV003. Viral loads in COV003 trial for those infected by WT (n = 105), Zeta (n = 238), Gamma (n = 386), and Delta (n = 14). **** = $p < 0.00005$ and from two-tailed Mann-Whitney U Test. Overall distribution significant by Kruskal-Wallis Rank Sum Test ($p = 1.98*10^{-18}$).

loads than seronegative WT samples in REMAP-CAP ($9.40*10^6$ IU/mL vs $5.88*10^5$ IU/mL, $p = 0.002$, two-tailed Mann-Whitney U Test).

The dramatic influence of lineage of infection on viral loads in COV003 suggests that these may distort comparisons between REMAP-CAP and COV003. Supporting this, median viral loads of seronegative, WT-infected REMAP-CAP patients were not significantly different than those of WT-infected COV003 patients ($4.64*10^6$ IU/mL vs $9.40*10^6$ IU/mL, $p = 0.087$, two-tailed Mann-Whitney U Test).

5.2.2 Emergence of SARS-CoV-2 variants in Patient Cohorts

REMAP-CAP samples were collected from March 2020 to the second week of January 2021, overlapping the emergence of the Alpha variant in the United Kingdom. Alpha began to stably emerge within the patient cohort in the last week of October 2020, and through gradual replacement of WT infections became responsible for 83.3% (120/144) of typable infections over the last two weeks of trial recruitment (Figure 5.6A). At this point, the trial was halted as it failed to meet its interim analysis efficacy benchmark. The earliest sampled patient assigned Alpha in our dataset (due to the presence of the S:A570D, S:D1118H, N:D3L, and N:S235F mutations) presented to ITU in North-East England and was randomized to the trial on September 3, 2020. As of April 2022, the earliest Alpha sample from the United Kingdom available on GISAID was collected in South-East England on September 20, 2020 (hence the "Kent" variant).[245, 348] A second sample collected in the Midlands on September 9, 2020 also had the S:A570D, N:D3L, and N:S235F mutations. Both of these early sequences were processed on the same sequencing plate and both contained the N:KR mutation (characteristic of Alpha). ASP-PCR was inconclusive on the sample collected September 3, 2020 and designated "WT" for the sample collected September 9, 2020. By the rank priority for typing methods, the NGS-based typing was considered more accurate than ASP-PCR. While these patients may provide evidence for the spread of Alpha throughout the United Kingdom prior to its identification in public data, the low sample viral load and the presence of other samples with Alpha genomes on the sequencing plate limit the ability to exclude contamination.

COV003 lineage emergence was characterized by selective sweeps of successive variants (Figure 5.6B). Zeta emerged early in the trial study period (October 15, 2020) and became stably dominant (>50% of successive weekly cases) by the middle of November 2020. The earliest high confidence Gamma sample was sampled on January 1st, 2021 and the variant was dominant eight weeks later by the end of February 2021. One sample collected in August 2020 was typed as Gamma in the

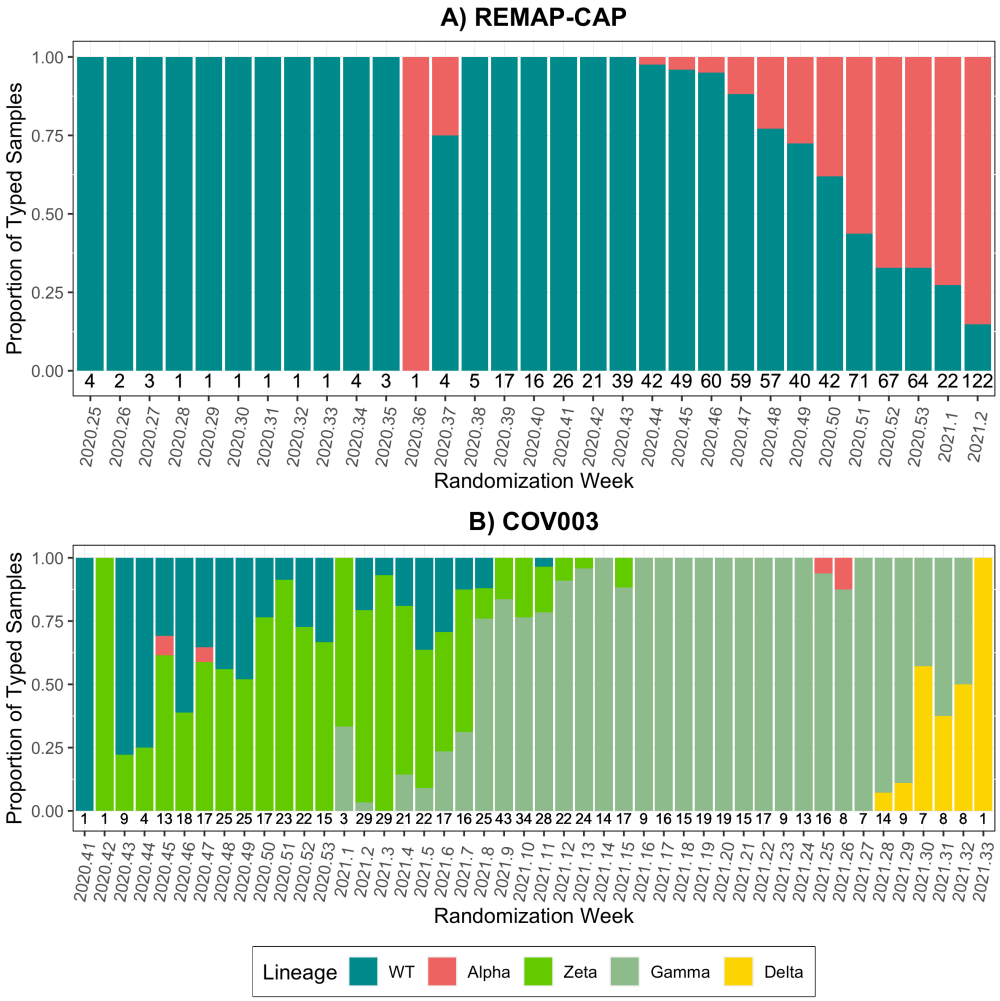


Figure 5.6: Emergence of Variants in COVID-19 Clinical Trials. For both plots, number labels on x-axis indicate total number of typed samples in that randomization week. (A) REMAP-CAP trial. (B) COV003 trial, limited to period after Zeta first emerged. Note the few patients randomized per week in REMAP-CAP until the final third of 2020.

S:K417T ASP-PCR and had a viral load precluding sequencing; re-extracting and retyping the sample gave an identical ASP-PCR designation. This sample predates the emergence of P.1 in Manaus, Brazil by several months and thus is more likely to represent a spontaneous S:K417T mutation than a true Gamma infection.[301] In the COV003 trial, Gamma stayed dominant being near 100% of samples for 21 weeks until the first appearance of Delta on July 8, 2021. Delta quickly became dominant only five weeks after it was first identified and maintained that status until the final trial sample was collected. Gamma was first identified in the Manaus

region of the North West of Brazil in November 2020; the delayed incursion of Gamma into the trial study population was likely due to the recruitment sites in Brazil (São Paulo, Rio de Janiero, Salvador, Santa Maria, Porto Alegre, and Natal)) being located in the Southern and Eastern regions of the country.

5.2.3 Viral Load Associations with Patient Clinical and Demographic Factors

To investigate the interactions between clinical and demographic factors and viral load, association studies in the REMAP-CAP cohort were performed. To mitigate type I errors, four hypotheses were defined: higher viral loads may be due to 1) the presence of co-morbidities; 2) prior treatment with immune-modulating drugs; 3) increased severity of illness; and 4) demographic characteristics associated with poorer immune responses, such as male sex at birth and older age. To mitigate any impact on confounding and decrease the total number of hypothesis tests run, correlation between patient factors in each hypothesis was determined prior to association studies. Correlated variables were analyzed with multivariate analysis using linear regression including patient serostatus, which was shown to strongly influence viral load in Figure 5.2, as a covariate. After applying a Bonferroni correction across all association studies, only patient age (controlled for sex, BMI, and serostatus) was shown to be associated with increased viral loads while invasive ventilation and high-flow oxygen treatment (controlled for Glasgow Coma Score, APACHE II score, C-reaction protein concentration, and serostatus) were shown to be associated with decreased viral loads (Figure 5.7A-C). Treatment with IL-6 inhibitors was not significant after the application of the Bonferroni correction ($p = 0.032$), however this may be a consequence of the small population treated ($n = 38$) (Figure 5.7D). Potentially of note, only 5.2% (2/38) patients treated with IL-6 inhibitors had no detectable RNA versus 10.6% of those not treated ($n = 140/1316$). This difference was again not significant ($p = 0.421$, Fisher's Exact Test). All significant variables (invasive ventilation, high-flow oxygen, patient

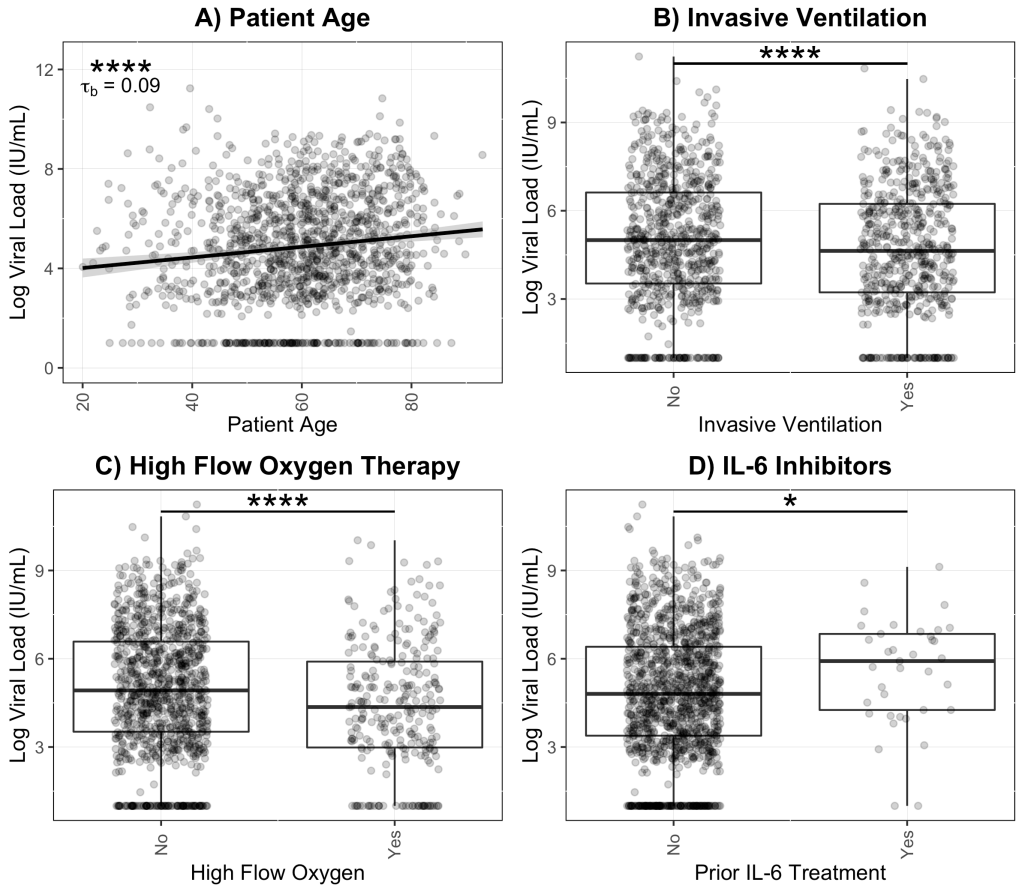


Figure 5.7: Impact of Patient Demographic and Clinical Factors on Patient Baseline Viral Load. τ_b value for patient age from Kendall's Rank Correlation. P values from multivariable linear regression of log-transformed viral loads accounting for patient age, baseline serostatus, invasive ventilation, high-flow oxygen therapy, and prior treatment with IL-6 inhibitors. * = $p < 0.05$ and **** = $p < 0.00005$.

age, and serostatus) remained strongly significant in a multivariable regression model incorporating these factors only.

5.2.4 Subgenomic RNA

The RT-qPCR assay used in these studies targets the SARS-CoV-2 N gene and thus the viral load measurements presented here are a sum of gRNA and sgRNA. Due to the unique inclusion of a common leader sequence in SARS-CoV-2 sgRNAs, these transcripts can be specifically targeted via RT-PCR using a leader sequence-specific forward primer first described elsewhere.[349] The probe and reverse primer of the CDC N1 oligonucleotide set are sufficiently close to the 5' end of the SARS-CoV-2

N gene to amplify an RT-PCR product using an sgRNA-specific forward primer. An sgRNA PCR oligonucleotide set was designed and utilized on a subset of samples from both the REMAP-CAP and COV003 trials to define the relationship between N gene sgRNA and total gRNA. Due to a lack of a quantified standard for the sgRNA target, Ct values returned for each assay were used for comparison.

Unsurprisingly, N gene Ct and sgRNA Ct values correlated strongly for both REMAP-CAP and COV003 (τ_b = 0.79 and 0.77, respectively, Kendall's Rank Correlation; Figure 5.8A). The targets of the N gene assay are inclusive of a broader range of SARS-CoV-2 genomic content (including the N gene sgRNA transcript) and thus the N gene Ct should be lower than the sgRNA Ct. This relationship was demonstrated in both trials using $2^{-\Delta Ct}$ analysis. Across all samples, median sgRNA abundance compared to total N gene RNA was 11.4% in REMAP-CAP and 6.7% in COV003. $2^{-\Delta Ct}$ weakly negatively correlated with N gene Ct in REMAP-CAP (τ_b = -0.155, p = 0.00035, Kendall's Rank Correlation) but did not correlate with N gene Ct in COV003 (τ_b = 0.029, p = 0.66, Kendall's Rank Correlation). Therefore, sgRNA was less abundant at decreased SARS-CoV-2 viral loads in patients enrolled in REMAP-CAP but not in COV003.

There was a stark difference between lineage of infection and $2^{-\Delta Ct}$ values in REMAP-CAP. sgRNA was significantly more abundant in Alpha infections making up 25.1% of total signal versus only 5.8% in WT infections (p = 4.1*10^{-12}, two-tailed Mann-Whitney U Test; Figure 5.9A). This difference can be visualized by the gap in the linear regressions for Alpha and WT in Figure 5.BB. No difference between lineage of infection and $2^{-\Delta Ct}$ values was detected in COV003 when comparing WT, Zeta, and Gamma infections (10.9%, 6.1%, and 6.7%, respectively, p = 0.20, Kruskal-Wallis Rank Sum Test; Figure 5.9B). WT infections in COV003 had significantly higher abundance than in REMAP-CAP (p = 0.015, two-tailed Mann-Whitney U Test). The reason for this was unclear, but may be due to varied biology in the WT group, as described earlier, or the different relative times of sampling in the course of infection between the two trials. sgRNA abundance in seropositive vs seronegative WT infections in REMAP-CAP were not significantly different (medians of 7.36%

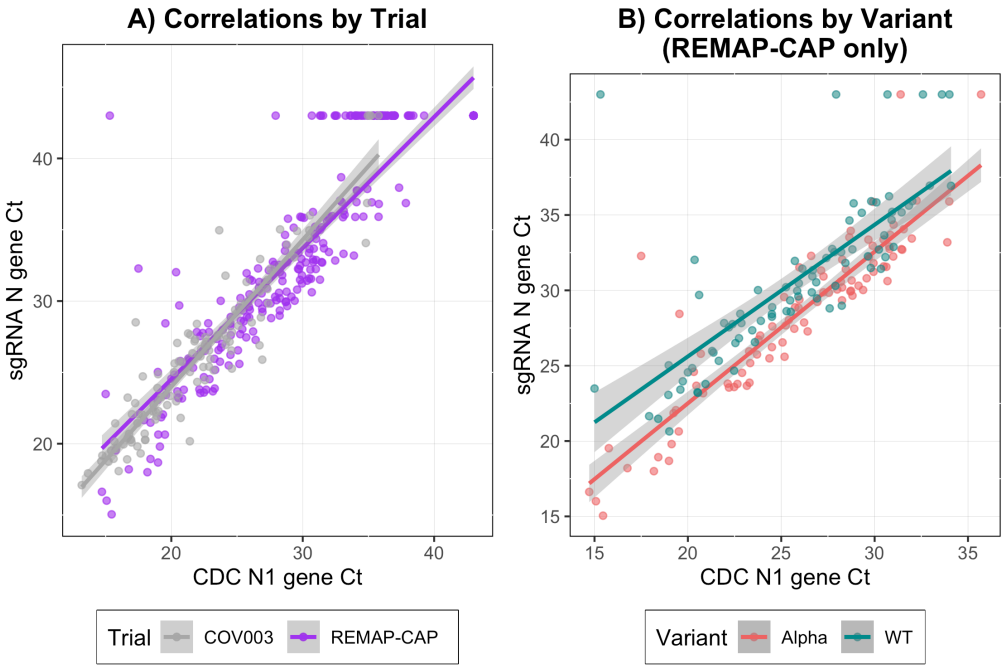


Figure 5.8: Relationship between Ct values for N gene and sgRNA N gene RT-qPCR assays. Correlations of CDC N1 gene Ct and sgRNA N gene Ct for REMAP-CAP and COV003 trials and for variant of infection in REMAP-CAP trial. REMAP-CAP (all) ($\tau_b = 0.79$), REMAP-CAP (WT) ($\tau_b = 0.74$), REMAP-CAP (Alpha) ($\tau_b = 0.78$), and COV003 ($\tau_b = 0.77$) all demonstrated strong linear dependence via Kendall's Rank Correlation. Samples negative by CDC N1 RT-qPCR excluded from visualization and correlation test. Slope for variant of infection (B) is less steep than for trials (A) as few low viral load samples with sgRNA Ct values of 43 could successfully be typed.

and 7.91%, respectively, p = 0.53, two-tailed Mann-Whitney U Test), casting doubt on that explanation (not all patients had determined serostatus, hence the incongruence between abundance values in all WT infections and WT infections with defined serostatus). Further, $2^{-\Delta Ct}$ values for WT infections in seronegative REMAP-CAP patients and all WT infections in COV003 were also significantly different (5.96% vs 10.88%, p = 0.049, two-tailed Mann-Whitney U Test),

Several lines of evidence suggest that the N:KR mutation generates a TRS and a novel N gene sgRNA transcript and that its presence may increase the relative expression of sgRNAs from other ORFs.[350, 351] To test the impact of the double substitution in isolation, analysis of the small subset of WT infections with

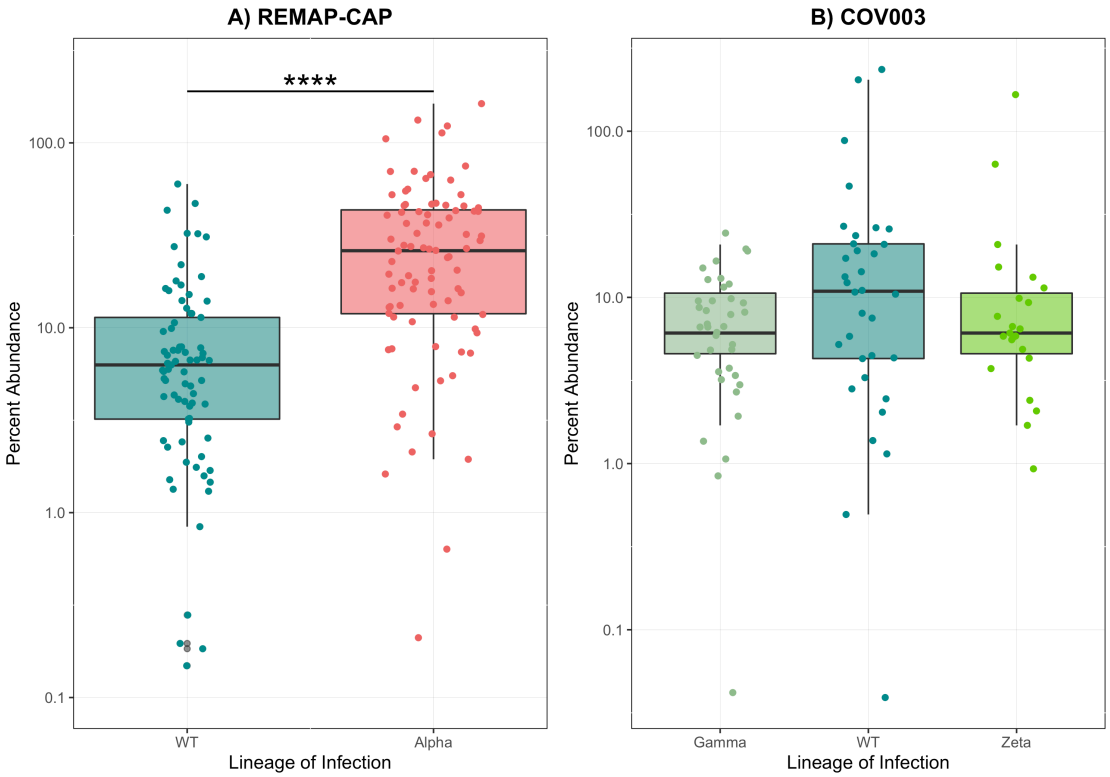


Figure 5.9: N Gene sgRNA Abundance by Lineage of Infection. Percent abundance calculated using $2^{-\Delta Ct}$ analysis with ΔCt calculated as the difference between Ct of the N gene total RNA RT-qPCR and Ct of the N gene sgRNA RT-qPCR. **** = p < 0.00005 and from two-tailed Mann-Whitney U Test

quantified sgRNA and N-gene typing was performed. WT infections with N:RG (n = 52) or with N:KR (n = 9) had no difference in relative abundance values (median values of 6.5% vs 4.8%, respectively, p = 0.33, two-tailed Mann-Whitney U Test). Sequencing of the novel sgRNA in Mears et al. (2022) demonstrated that it is lacking a significant portion of the 5' end of the transcript.[351] Therefore, this transcript is likely to be undetectable by both the CDC N1 gene RT-qPCR and the N gene sgRNA RT-qPCR, although confirmatory tests should be performed. The evidence here weakly suggests no role for the N:KR mutation in increasing relative expression of the canonical N gene sgRNA and that the difference between relative expression in WT and Alpha infections is not due to the N:KR mutation in isolation.

5.2.5 SARS-CoV-2 Viral load dynamics in intensively sampled patients hospitalized with COVID-19

Within the REMAP-CAP trial, 187 patients were enrolled into an intensive arm within the trial and were intended to be swabbed for SARS-CoV-2 repeatedly during their time in ITU up until 28 days post-randomization. 153 of these patients ultimately had >1 sample collected. From these patients, 879 samples were processed and quantified by RT-qPCR, providing a unique population for investigating the intra-host dynamics of SARS-CoV-2 in severely unwell patients. In total, by the difference between the first and last quantified sample for all patients, 1752 days in ITU are accounted for in this analysis. Both the total number of samples per patient and the days between first and last samples were non-normally distributed ($p = 0.00010$ and $1.14*10^{-11}$, respectively, Shapiro-Wilk Normality Test). At the individual level, patients provided a median number of 5 samples and their final sample was given on their ninth day in ITU (Figure 5.10A,B).

Patients were assigned to one of six viral dynamics classifications based on their qPCR results as depicted in Figure 5.10C. 40.5% of samples did not experience dynamic changes in viral load and either all samples had viral loads $\leq 1,000$ IU/mL (Ct ∼35.5 in CDC N1; classified as "cytokine storm") or viral loads ≥ 500 IU/mL (Ct $\sim$ 37.4 in CDC N1; classified as "persistently positive"). Overall classification percentages are included in Figure 5.10C. Example plots for the dynamics classifications are provided in Figure 5.11A-D.

As the intensively sampled group included samples collected after randomization and CP administration within the trial, the effect of the treatment on viral dynamics classification was investigated. As the impact of CP treatment would manifest in its antiviral effects, patients who did not have detectable viral load during their time in ITU (i.e., "cytokine storm" patients) were excluded. Of the remaining patients, 64.2% (77/120) were allocated (and received) CP and 35.8% (43/120) were given standard of care only. There was no association between CP treatment and viral dynamics classification ($p = 0.50$, Fisher's Exact test) or baseline viral load ($p = 0.29$, two-tailed Mann-Whitney U Test). As CP was collected from

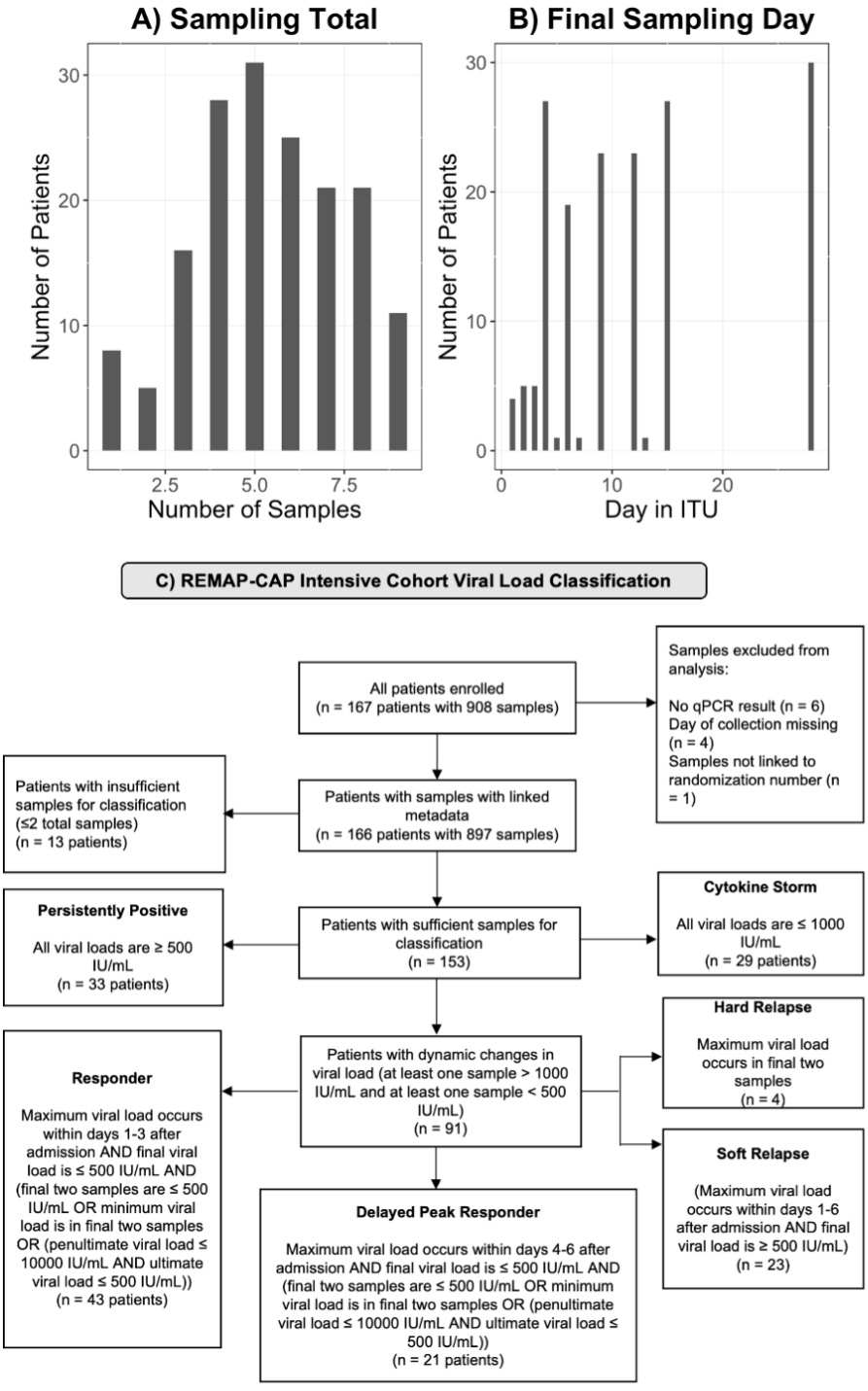


Figure 5.10: Viral Dynamics Classification System. (A) Total number of samples provided by each ITU patient (median = 5). (B) Final day of sampling for each ITU patient (median = 9). (C) Decision tree for Viral Dynamics Classifications.

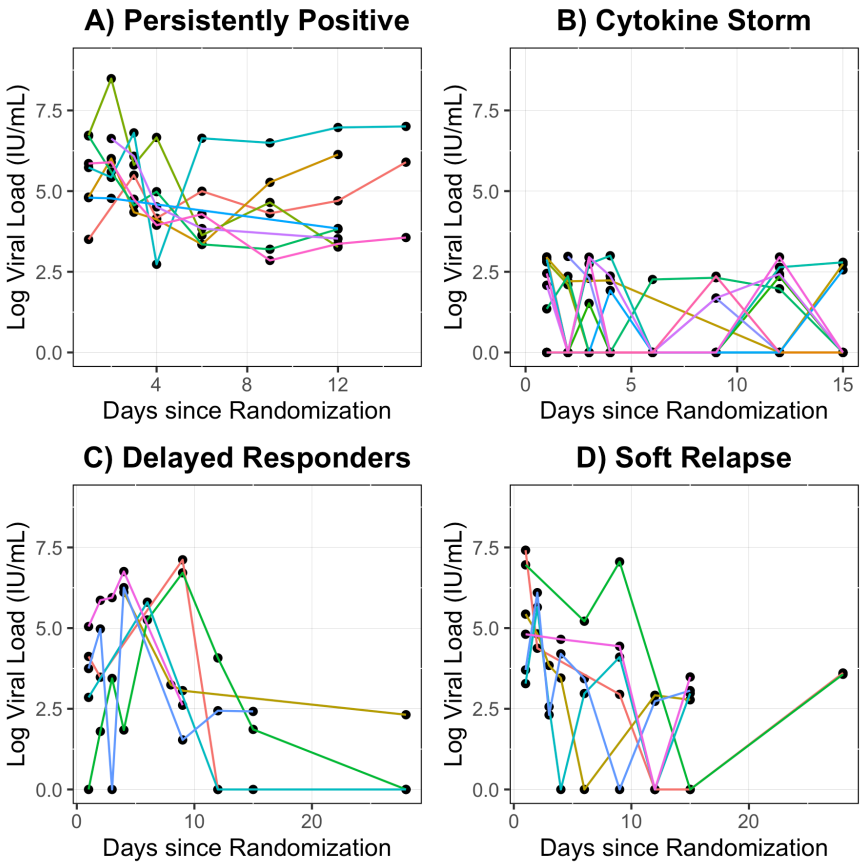


Figure 5.11: Representative Plots of Viral Dynamics Classifications. Patients classified as depicted in Figure 5.10C. Plots only depict a random subset of all patients classified within that group.

patients infected before the emergence of VoCs, CP treated patients infected with Alpha (n = 13) may have experienced an antigenic mismatch that would manifest in delayed viral clearance. Among those that received CP, being antigenically matched vs mismatched did not impact classification (p = 0.71, Fisher's Exact test) but mismatched patients did have a later date of final sampling in ITU than matched (n = 38; median date of 9 days post admission for both groups, p = 0.0097, two-tailed Mann-Whitney U Test; Figure 5.12).

5.2.6 Association of viral dynamics classification with patient and virological factors

Similar as to the baseline viral load, the relationship between patient and virological factors and viral dynamics classification were investigated. Hypotheses were not

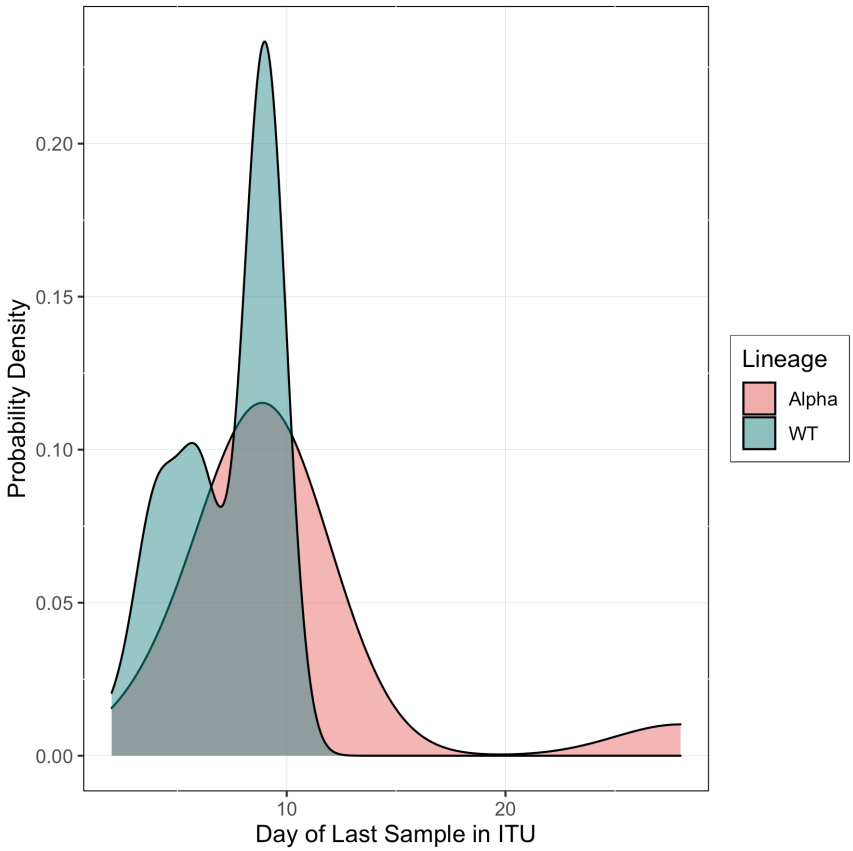


Figure 5.12: Date of Final Sampling in ITU among Convalescent Plasma-treated patients. Difference in date of final sampling among CP-treated patients infected with WT (n = 38) and CP-treated patients infected with Alpha (n = 13). Counts (not displayed) significantly different in two-tailed Mann-Whitney U Test (p = 0.0097).

prespecified in this instance. APACHE II score and invasive ventilation were the only variables that associated with classification (p = 0.046, Kruskal-Wallis Rank Sum Test and p = 0.043, Fisher's Exact Test, respectively), although neither were significant upon a Bonferroni correction for multiple comparisons. For APACHE II scores, where higher numbers indicate greater severity,[352] the highest median scores (17) were seen in both "cytokine storm" and "delayed peak responder" patients while the lowest median scores (12) were seen for "responder" and "persistently positive" patients. The highest percentage of patients requiring invasive ventilation was in the "cytokine storm" classification (53.6%) while the lowest was in the "relapse" classification (23.1%). Given the unclear biological mechanism underlying these relationships and that APACHE II score and invasive ventilation are directly related

and indeed weakly correlated (τ_b = -0.2, p = 0.00062, Kendall's Rank Correlation), it is assumed that they are likely chance associations. Neither the total number of samples nor the days between an individual's first and last sample differed between viral dynamics classifications (Kruskal-Wallis Rank Sum Test), suggesting little to no role for these groupings on length of stay in ITU.

5.3 Discussion

This chapter provides a sweeping overview of the intra-host viral load dynamics of SARS-CoV-2. The intent of this investigation was to better understand the patient and virological factors that influence the SARS-CoV-2 viral load of COVID-19 patients sampled at similar points in disease progression. The studies revealed a remarkable diversity in viral load in both trials, with a 6.5-million fold difference between the minimum and maximum viral loads detected in the COV003 trial and a nearly 6-billion fold difference in REMAP-CAP. As expected given their relative stages in an acute model of infection, viral loads in the COV003 trial were much higher than in REMAP-CAP and patients who had seroconverted in REMAP-CAP had lower viral loads than those who had not. Ct values generated by RT-qPCR are a relevant measurement and there can be minor variation depending on the oligonucleotide set used for amplification. This phenomenon has been widely noted for SARS-CoV-2 assays.[353–355] To mitigate any risk of bias due to varying PCR efficiencies, RT-qPCR on all samples was performed on a single real-time thermocycler with the same oligonucleotides, reaction conditions, and quantified standard for both trials. This approach allows for robust intra- and inter-trial comparisons. For inter-trial comparisons, the differences in viral load seen between COV003 and REMAP-CAP are a combination of a higher percentage of patients in the COV003 trial being earlier in the disease course than in REMAP-CAP and more COV003 patients being infected with high viral load producing variants.

The virological and serological characteristics of COVID-19 patients admitted to intensive care in the REMAP-CAP trial were remarkably diverse. 10.8% of this patient population had successfully cleared the virus prior to their admission to

ITU, suggestive that a substantive element in the disease mechanisms in this group was inflammatory- or cytokine-related rather than being directly virally induced. In contrast, 30.6% of patients had not even seroconverted by the time of their admission. Further analysis of the clinical factors associated with viral load revealed two markers of disease severity (high-flow oxygen and invasive ventilation), patient age, and serostatus as being associated with viral loads at time of admission to ITU. No co-morbidity or prior treatment significantly impacted viral load, including treatment with antivirals (e.g., remdesivir) or immune-modulating drugs (e.g., corticosteroids, IL-6 inhibitors), or evidence of immunosuppression (although this effect may be masked by the analysis being controlled for serostatus).

In general, the viral load association studies for REMAP-CAP presented here are weakened by the unfortunate reality that time since symptom onset was not collected as part of the trial protocol. Any variable that is unknowingly correlated with developing severe symptoms earlier in infection (e.g., BMI)[356] would be associated with higher viral loads. Here, serostatus has been used as a binary indicator of being early vs late in infection, but there is documented heterogeneity in time to seroconversion.[357–359] Further, the factors that would lead a physician to recommend earlier or later admission to ITU may have a role in the results presented. Patients on high flow oxygen therapy or invasive ventilation may have been admitted to ITU due to deterioration during an earlier hospital stay, which would likely be later in the disease course. Elderly patients may have been admitted earlier in the disease course (and thus with higher viral loads) simply due to their greatly increased risk of severe outcome relative to the rest of the population.[360]

Prospectively, measurements of viral loads and serostatus against the SARS-CoV-2 N protein may provide valuable information for patient indication for treatment or as a pre-specified subgroup in clinical trials. Therapies based on antiviral effects, such as CP, remdesivir, or nirmatrelvir (PAXLOVID), may be most suitable in patients where viral loads are still detectable, N gene antibodies are undetectable, and respiratory tract pathology may be due to virus-driven effects. The Randomised Evaluation of COVID-19 Therapy (RECOVERY) trial demonstrated this effect in

their investigation of the casirivimab and imdevimab antibody cocktail.[361] In their study, the antibody cocktail reduced the relative risk of 28-day mortality by 20% in seronegative patients. This reduction was not present when patients were unstratified by baseline antibody status. Conversely, serious pathology in patients who have seroconverted to N (rather than S, which would be detectable following vaccination) may primarily be driven by inflammation-mediated damage and these patients may better respond to immune-modulating drugs (corticosteroids, IL-6 receptor antagonists). The timing of treatment with immune-modulating drugs is key to avoid potential enhancement of viral replication. This is weakly evidenced by the higher viral loads seen in patients treated with IL-6 receptor antagonists prior to admission to intensive care in the REMAP-CAP trial. Similar to age, this comparison is at risk of bias as the factors leading a physician to prescribe IL-6 inhibitors pre-ITU admission are unclear but likely nonrandom.

The difference in viral loads between WT, Zeta, Gamma, and Delta reported for COV003 is consistent with those reported elsewhere, although this is the only analysis encompassing each of these specific lineages in a trial setting.[301, 362–364] In this trial, where patients were sampled following presentation to trial sites after developing symptoms, the differences in viral load likely reflect altered replicative capacity between the variants where VoCs are able to produce more virions in a shorter period of time. However, the differences between lineages may also be impacted by uncontrolled variables such as vaccination status, previous infection, and time since vaccination. While the efficacy for the ancestral-targeting ChAdOx1 nCoV-19 vaccine was not lower for the Zeta variant, this could not be determined for the Gamma or Delta variants.[267] Of note, all Delta infected patients and the vast majority of Gamma infected patients were vaccinated at their point of infection due to the structure of the trial. Vaccination and previous infection may serve to depress detected viral loads as the immune system may be able to more rapidly clear infection, although research on the Delta variant has demonstrated that peak viral load is similar between unvaccinated and vaccinated individuals.[365]

In REMAP-CAP, however, the data presented here suggests that differences in viral load between Alpha and WT are largely due to time of sampling and that these lineages likely reach equivalent or near-equivalent peak viral loads. This result conflicts with many published papers that have reported higher viral loads for Alpha.[362, 366–368] These results are, however, consistent with analysis by the COVID-19 Infection Survey Team of over 11,000 Alpha and non-Alpha test results in the United Kingdon.[369] That the difference in viral load is almost entirely driven by delayed clearance of virus in seropositive individuals is consistent with a less effective cellular or humoral immune response to contain infection; Alpha's apparent IFN-resistance supports this model being driven by evasion of innate rather than adaptive immunity.[346, 347]

The opposite effect of the N:KR mutation on non-Alpha viral loads in REMAP-CAP patients compared to a study of hospitalized patients in Saudi Arabia by Mourier et al. (2022) is peculiar.[304] In that study, the authors were able to control their association studies for time since first positive test which was unavailable metadata for this analysis, but the persistent trend of lower viral loads in seropositive and seronegative REMAP-CAP patients challenges the interpretation that this effect is an artefact of sampling time. Mourier et al. (2022) extend their work characterizing patient samples to *in vitro* characterization of the N:KR mutation. There, the authors show that the double substitution leads to increased N protein oligomerization and viral RNA binding (potentially impacting viral RNA packaging) and increased serine phosphorylation leading to overexpression of interferon-related genes. Polymorphisms determining prenylation of 2'-5'-Oligoadenylate Synthetase 1 (OAS1), an IFN-stimulated antiviral effector molecule, have previously been shown to influence severity of COVID-19.[370] The mutation has strong geographic influence with the highest levels of prenylation in African populations. It is conceivable that the different effects manifesting from the N:KR mutation in these two populations is due to host genetics (highly likely not to be the same polymorphisms underlying the differential phenotype of OAS1 as the dominant allele in African populations

protected against severe COVID-19), but neither study performed whole genome sequencing on their patient population to support this investigation.

The influence of sgRNA expression on clinical viral load has not been explored in the literature. The data presented here suggests inadvertent detection of sgRNA in RT-qPCR assays likely deflate N gene Ct values by a value of ∼0.1 in non-Alpha infections, which is a minimal contribution. The relatively low abundance of sgRNA in diagnostic samples is consistent with quantification of sgRNA and gRNA using *in situ* hybridization of infected Vero E6 cells.[371] There, the ratio of sgRNA/gRNA declined after 24 hours post infection. Given the time delay from infection to symptom onset, the *in vivo* transcription bias is likely to be in the direction of gRNA and virion production.

The sgRNA qPCR was implemented using the same probe and reverse primer as the CDC N1 oligonucleotide set to limit any impact of PCR efficiency on that analysis. This approach does prevent multiplexing of total and sgRNA targets and opens the risk of degradation of extracted RNA from repeated freeze-thaw cycles. In general, PCR-based approaches based on the common leader sequence are more appropriate for sgRNA quantification than short read sequencing technologies, although sgRNA detection NGS methods based on long read technologies with detection of the 5'-m7G cap have been described.[372] Importantly, the N gene sgRNA RT-PCR developed here is only able to detect N gene sgRNAs. In contrast, the CDC N1 gene RT-qPCR would signal for gRNA, N gene sgRNA, and all other canonical sgRNAs (S, 3a, E, M, 6, 7a, 7b, and 8) as they are inclusive of the N gene target sequence.[47] Further, the long read, 5'-m7G cap approach could differentiate the sgRNAs and allow for robust measurement of the relative quantification differences between transcripts.

In a model wherein seroconversion leads to viral inactivation by antibody binding and the persistence of inactivated virions in the extracellular space, sgRNA levels would skew towards decreased $2^{-\Delta Ct}$ values later in infection due to their inability to be produced in infected host cells. In that scenario, sgRNA positivity would represent a stable marker of productive cellular infection and could differentiate patients

testing positive via N gene PCR due to active infection or simply from the presence of inactivated virions with packaged gRNA in the extracellular space. Unfortunately, the data presented here does not support this model as $2^{-\Delta Ct}$ values were strongly linearly correlated to N1 gene Ct in both trials. Further, previous research has suggested that detection of SARS-CoV-2 sgRNAs through NGS-based methods is not indicative of active cellular infection due to the molecule's likely inclusion in double-membrane vesicles and associated resistance to extracellular nucleases.[373] Altogether, deployment of sgRNA-targeting assays are unlikely to provide additional diagnostic benefit beyond those of standard gRNA-targeting assays.

The altered ratio of sgRNA to gRNA in Alpha infections has been previously reported in clinical samples sequenced using the ARTIC protocol by Parker et al. (2021).[374] In their study, the authors found significantly higher expression of canonical sgRNAs in Alpha infections compared to lineage B.1.177. The authors hypothesized a role for the ACE2-affinitiy increasing S:N501Y substitution in leading to enhanced sgRNA expression early in infection, but the lack of a difference in S:N501Y-containing Gamma infections in this trial allows for this hypothesis to be dismissed. Importantly, the authors found increases in normalized sgRNA expression for the S, N, M, E, and ORF6 genes suggesting the altered ratio found for the N gene in this study is not an artefact of the PCR-based screening method used.

The increased relative abundance of sgRNA in clinical samples from Alpha-infected patients may be due to an increased preference for the Alpha variant to spread via syncytia formation rather than through virion budding and release. In this model, the ratio of virions (almost entirely gRNA) to infected cells (mix of sgRNA and gRNA) would skew towards infected cells, resulting in increased relative abundance of sgRNA as seen in this data. This increased fusogenicity for Alpha has been demonstrated *in vitro* in low multiplicity of infection (MOI) studies in Rajah et al. (2021).[375] Other *in vitro* experiments in Vero E6 and A549 cells failed to replicate the altered sgRNA/gRNA ratios.[371] Importantly, these experiments were carried out at an MOI of 1, where approximately 63% of cells would have been

infected by at least one virion (based on Poisson distribution), and the opportunity for cell-cell spread via syncytia formation is low.

The viral load dynamics in patients intensively sampled in the REMAP-CAP cohort revealed substantially more diversity than expected. In particular, the pathology leading to continued intensive care for patients serially negative by RT-qPCR (n = 29) vs serially positive (n = 33) may be remarkably distinct. For the several patients (n = 27) whose viral loads relapsed during their stay, low sequencing performance on samples unfortunately prevented interrogation of whether these resurges in viral load were due to rebound of the initial infection or nosocomial transmission of antigenically distinct lineages. Work by collaborators to immunophenotype and quantify the cytokine release profiles of all the intensively sampled patients is ongoing and will hopefully lead to targeted therapies based on virological and immunological characterization.

Lastly, quantification of SARS-CoV-2 RNA at the population level has been shown to be valuable. As an alternative use of viral load data, some public health agencies are quantifying SARS-CoV-2 RNA levels in wastewater to track infection dynamics within communities.[376, 377] While this approach is unable to identify infected individuals, wastewater quantification has proven to be a leading indicator of rising infection levels within a community when compared to estimating prevalence based on diagnosed cases.[378] Further, the epidemic trajectory of SARS-CoV-2 in a population can be estimated based on the cross-sectional distribution of viral loads; a larger proportion of high viral loads are indicative of a growing epidemic and vice versa.[379]

Viral load measurements in COVID-19 patients are useful on both the individual- and population-level but are influenced by a litany of virological and host factors. Further studies using well-characterized cohorts (especially including data based on time since symptom onset) would further elucidate the contributing characteristics with higher confidence than in this study.

6

Interactions of SARS-CoV-2 Virological Characteristics with COVID-19 Severity and Outcomes

Contents

6.1 Introduction 119
6.2 Results 120
6.2.1 Comparative incidence of infection with Alpha or WT with N:KR in REMAP-CAP vs English population . . . 120
6.2.2 Virological factors associated with disease severity . . . 122
6.2.3 Virological factors associated with severe outcomes . . . 123
6.3 Discussion 127

Abstract

SARS-CoV-2 and the COVID-19 pandemic have been responsible for substantial mortality globally. In the pre-vaccine era, the virus' infection fatality rate was estimated to be approximately 10 times higher than that of influenza A/H1N1. Understanding the mechanistic underpinnings of COVID-19 mortality enables focused development of therapeutics and identification of patients most at risk of severe outcomes. Within the cohort of intensely sick COVID-19 patients in REMAP-CAP, no virological factor (viral load at time of ITU admission, variant infection,

presence of N:KR mutation, viral dynamics classification) influenced disease severity at time of enrollment or risk of death in ITU. Compared to publicly available sequencing data, both the Alpha variant and the N:KR mutation in WT infections emerged in the patient population at a similar rate as in the broader community. This evidence, in concert with that presented in Chapter Five, supports the majority of severe COVID-19 pathology being due to immune-mediated effects rather than viral, at least in the pre-vaccine and pre-Delta VoC era. Therapeutics developed and trialed for late stage COVID-19 should focus on dampening immune-driven pathology rather than antiviral mechanisms of action.

6.1 Introduction

As of June 8, 2022, the COVID-19 pandemic has claimed more than 6.3 million lives, making it one of the deadliest infectious disease outbreaks in history.[380, 381] In the pre-vaccine and pre-corticosteroid therapy era, estimates for median SARS-CoV-2 infection fatality rates (IFR) centered around 0.31% with strong geographic and age associations.[382] This estimate is at least an order of magnitude higher than the 2009 A/H1N1 outbreak in the United Kingdom, on par with estimates for Hepatitis A virus infection in unvaccinated individuals, and roughly 35 times lower than for SARS-CoV.[383–385] Viral infections lead to pathology through a combination of direct virus killing of infected cells (e.g., death of neuronal cells following infection of immunodeficient mice with Venezuelan equine encephalitis virus)[386] and immune-mediated killing of infected cells (e.g., HBV is itself non-cytopathic for human hepatocytes; liver injury is only apparent following activation of the host adaptive immune response)[387]. Despite these mechanistic differences, the majority of treatments for viral infections are still antiviral or based on controlling symptoms, such as the historical use of pegylated-IFN plus ribavirin and the modern use of direct-acting antiviral treatment for HCV infection or the cell-budding inhibitor oseltamivir (TAMIFLU) for IAV and IBV.[388–390] In the 2002-2003 SARS-CoV outbreak, there was some evidence that treatment with corticosteroids led to decreased mortality but the evidence was mixed and potentially biased.[391]

Understanding the relative contributions of the two mechanisms to the elevated SARS-CoV-2 mortality rate can inform therapeutic development priorities and potentially identify patients at greater risk of succumbing to COVID-19.

Much attention has been paid to the role of demographic factors such as age, sex, and comorbidities on COVID-19 severity and risk of death.[392–395] Less attention has been paid to the role of virological factors in COVID-19 morbidity and mortality. This short chapter leverages the REMAP-CAP cohort to disentangle the virological factors associated with increased disease severity and risk of severe outcomes among intensely sick patients.

6.2 Results

6.2.1 Comparative incidence of infection with Alpha or WT with N:KR in REMAP-CAP vs English population

All of the patients enrolled in the REMAP-CAP trial presented with evidence of severe pathology motivating their admission to ITU. As such, virological factors that are overrepresented in this cohort relative to those infected in the general population may themsleves be an indicator of increased severity. To test this hypothesis, the temporal emergence of the Alpha variant in REMAP-CAP was compared to that in the general population based on publicly available GISAID sequences. Due to low numbers of REMAP-CAP recruitment sites in Northern Ireland, Wales, and Scotland and to limit the potential impacts of geographical sampling bias, this was limited to REMAP-CAP patients recruited in England (n = 1145). These samples were compared to all publicly available GISAID sequences collected in England from April 20, 2020 to January 7, 2022 that were published as of April 11, 2022 (n = 129828). Proportions of Alpha per randomization week were generally higher in REMAP-CAP than GISAID but the difference in overall distribution was not significant (p = 0.86, Anderson-Darling k-Sample Test; Figure 6.1A). The large number of samples from the GISAID dataset per randomization week allowed for robust interrogation of proportion of Alpha within individual weeks. Zero of 11 tested weeks (after the stable emergence of Alpha in REMAP-CAP,

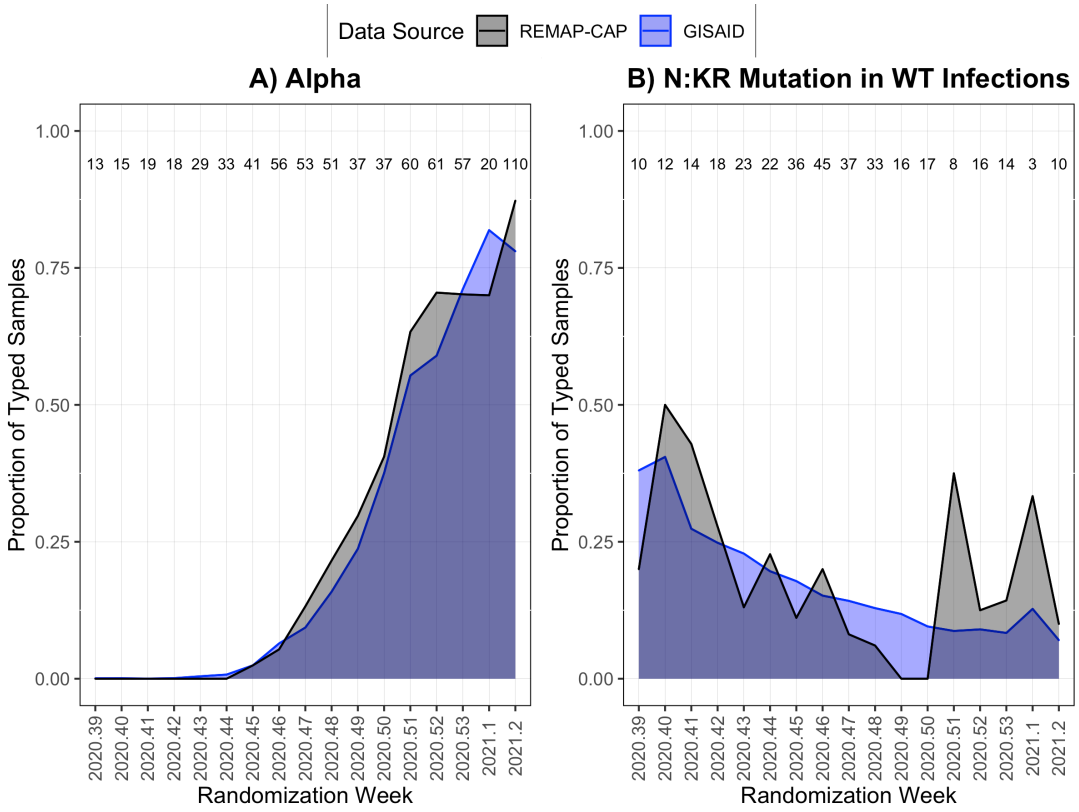


Figure 6.1: Alpha and N:KR in REMAP-CAP versus General Population in England. Numbers above plots indicate number of typed, WT REMAP-CAP samples included in that randomization week. Analysis of GISAID samples includes 129828 full-length sequences deposited as of April 11, 2022.

weeks 45-2020 to 2-2021) were shown to be significantly different using a Fisher's Exact Test with a Bonferroni correction.

To assess whether the N:KR mutation was independently predictive of increased disease severity, the same comparison against the general population was undertaken. The analysis of GISAID sequences was limited to non-Alpha sequences containing full coverage of N:203/204 (n = 101714). From REMAP-CAP, there were 360 WT infections that had been N-gene typed for comparison. No randomization week prior to 2020.39 had >10 WT infections that were N-gene typed, so the analysis was limited to weeks occurring that week or later. The distribution of N:KR in REMAP-CAP was more stochastic than for Alpha. However, there was a significant difference in overall distribution of N:KR containing WT infections in REMAP-CAP vs GISAID (p = 0.004, Anderson-Darling k-Sample Test; Figure 6.1B). Fisher's

Exact Tests for weeks 39-2020 to 2-2021 demonstrated no weeks where there were significant differences in N:KR representation between the two sources after applying a Bonferroni correction. As the deployed Anderson-Darling k-Sample Test uses proportions and not raw counts as inputs (in contrast to Fisher's Exact Test), the observed significant difference may be an artefact of the weeks with few typed samples in REMAP-CAP, particularly toward the end of trial enrollment period when the majority of total samples were Alpha. Of note, a Kolmogorov-Smirnov test, which, similar to an Anderson-Darling k-Sample Test, assesses cumulative distribution functions but is less sensitive to values at the tails of the distribution than an Anderson-Darling k-Sample Test,[396] was still significant ($p = 0.035$). Therefore, it is reasonable to conclude the significance in the Anderson-Darling k-Sample Test is not simply due to increased representation of N:KR in REMAP-CAP compared to GISAID in weeks 50.2020 and later. Neither the Anderson-Darling k-Sample Test nor the Kolmogorov-Smirnov Test assess direction of effect and are simply testing the null hypothesis that two numeric vectors are drawn from the same cumulative distribution function. Given the up-and-down nature of the N:KR proportion in REMAP-CAP relative to GISAID, it cannot be reasonably concluded that there exists a difference in hospitalization risk between those WT infections with and without the double substitution.

6.2.2 Virological factors associated with disease severity

Heterogenous disease severity at time of ITU admission within the cohort enabled investigation of the virological factors associated with severity (other than viral load, which was investigated in Chapter Five). In terms of diversity, 43.6% of patients (n = 601) were invasively ventilated at time of admission and 20.4% (n = 281) were on high-flow oxygen therapy (patients could not be ventilated and on high-flow oxygen therapy concurrently). While the vast majority (76.4%, n = 1055) of patients scored a 15 on the Glasgow Coma Scale (fully alert), 9.2% of patients (n = 127) scored a 3 (completely unresponsive). Association studies of virus type revealed no association with various metrics of disease severity after application of a Bonferroni correction

(Table 6.1A). This lack of association was present upon multivariable binomial regression including patient age, serostatus, and log-transformed viral load as inputs. In seronegative populations, WT infections had significantly higher APACHE II scores, but this value was just above the Bonferroni corrected significance cutoff of 0.004 upon multivariable regression. There were no significant associations in seropositive patients. Both the presence of the N:KR mutation among WT-typed samples and, within the intensively sampled cohort, the viral dynamics classification failed to associate with any marker of severity (Table 6.1B,C). Statistical tests based on multiple categories (i.e. Kruskal-Wallis Rank Sum Test or Fisher's Exact Test) may mask true differences between two categories in isolation. Post-hoc analyses may reveal associations not found in the initial statistical analysis but run an elevated risk of being a type I error. Here, a post-hoc comparison of invasive ventilation rates between delayed peak responders and soft relapsers, which visually appear to be the greatest difference in table 6.1C, was similarly not significant ($p = 0.11$, Fisher's Exact Test). In summary, disease severity at time of admission to ITU in this cohort was not influenced by any of the virological factors measured.

6.2.3 Virological factors associated with severe outcomes

1380 patients with viral load data had known outcomes within the trial, 65.0% (n = 897) of whom were discharged from ITU alive. Both higher baseline viral loads and being seronegative at time of ITU admission were associated with an increased risk of death (Table 6.2). When the Alpha variant emerged there were concerns that it would lead to more severe disease and higher mortality amongst patients compared to WT. In contrast, in this cohort being infected with WT rather than Alpha was associated with a significant increase in ITU mortality (57.8% vs 65.7% discharged alive, respectively, $p = 0.025$, Chi-Squared Test).

Age and patient BMI have been demonstrated to be strongly negatively associated with COVID-19 survival. In the REMAP-CAP cohort, older age was strongly associated with an increased risk of death while increased BMI was surprisingly protective ($p = 1.73*10^{-23}$ and $p = 0.008$, respectively, binomial

6.1A) Association with Virus Type					
Patient Population	Severity Marker	WT (n = 556)	Alpha (n = 303)	p^X	MR p^Y
All patients (n = 859)	APACHE II Score	13 (8 - 20)	12 (8 - 18)	0.028	0.08
	Glasgow Coma Score	15 (14 - 15)	15 (15 - 15)	0.32	0.65
	Invasive ventilation	41.0%	43.3%	0.563	0.38
	High-flow oxygen therapy	17.5%	17.2%	1	0.75
Seronegative (n = 325)	**APACHE II Score**	**15 (10 - 21)**	**13 (7 - 18)**	**0.0014**	0.008
	Glasgow Coma Score	15 (14 - 15)	15 (15 - 15)	0.75	0.61
	Invasive ventilation	41.1%	40.6%	1	0.88
	High-flow oxygen therapy	17.8%	20.8%	0.55	0.48
Seropositive (n = 530)	APACHE II score	12 (7 - 19)	12 (8 - 18)	0.98	0.90
	Glasgow Coma Score	15 (14 - 15)	15 (15 - 15)	0.41	0.52
	Invasive ventilation	41.2%	44.1%	0.52	0.43
	High-flow oxygen therapy	17.0%	15.4%	0.72	0.85

6.1B) Associations with N:KR Mutation in WT-typed Samples			
Severity Marker	N:RG (n = 341)	N:KR (n = 80)	p^X
APACHE II score	13 (9 - 20)	12.5 (8 -18.25)	0.39
Glasgow Coma Score	15 (14 - 15)	15 (14 - 15)	0.45
Invasive ventilation	43.4%	36.3%	0.26
High-flow oxygen therapy	17.0%	22.5%	0.26

6.1C) Associations with Viral Dynamics Classifications							
Severity Marker	Cytokine Storm (n = 27)	Persistently Positive (n = 28)	Responder (n = 43)	Delayed Peak Responder (n = 14)	Soft Relapse (n = 21)	Hard Relapse (n = 3)	p^Z
APACHE II score	17 (10.5 - 22)	13 (9.75 - 20.25)	12 (8 - 19)	18 (10 - 22.75)	15 (8 - 21)	12 (9 - 20)	0.74
Glasgow Coma Score	15 (14.5 - 15)	15 (14 - 15)	15 (15 – 15)	15 (15 – 15)	15 (15 - 15)	15 (15 - 15)	0.35
Invasive ventilation	51.8%	32.1%	37.2%	64.3%	19.0%	66.7%	0.05
High-flow oxygen therapy	25.9%	21.4%	23.2%	21.4%	23.8%	0%	0.99

Table 6.1: Association of disease severity at admission and virological factors. Values in table display median with 25th and 75th percentile values in parentheses for continuous data and percentages for categorical data. X) P value from two-tailed Mann-Whitney U Test for continuous data and Fisher's Exact Test for categorical data. Y) P value from multivariable regression including log-transformed baseline viral loads, serostatus (for all patients category), and patient age. Z) P value from Kruskal-Wallis Rank Sum Test for continuous data and Fisher's Exact Test for categorical data. Significant variables in bold.

Variable		Binomial regression coefficient estimate	p^A	MV p^B
Continuous	**Viral Load (log(IU/mL))**	**-0.14**	**$1.43*10^{-07}$**	0.024
	Age (Years)	**-0.052**	**1.73e-23**	**$1.14*10^{-10}$**
	BMI (Units)	0.012	**0.0079**	0.44
Categorical	**Serostatus (Seropositive)**	**0.55**	**$4.96*10^{-06}$**	0.017
	Virus Type (WT infection)	-0.34	0.020	0.085

Table 6.2: Association of risk of death in ITU and patient virological factors. P values from (A) binomial and (B) multivariable binomial regression. Significant variables in bold

regression). However, patient age and BMI were weakly but significantly negatively associated (τ_b = -0.2, p = $1.21*10^{-28}$, Kendall's Rank Correlation). This result indicated that the protective effect observed for BMI may simply be due to age acting as a confounder. Upon multivariable regression of patient age and BMI with viral load, only patient age remained significantly associated with death in ITU (p = $3.84*10^{-22}$ and p = 0.99, respectively, multivariable binomial regression). This finding confirmed age as the most impactful demographic variable available for this

analyses and the neutral effect of BMI on patient survival here. This finding is important when analyzing ITU discharge rates between WT and Alpha because patients infected with Alpha were significantly younger than those infected with WT (median age of 60 vs 63, p = 0.003, two-tailed Mann-Whitney U Test).

Due to the previously demonstrated association of age with viral load (see Chapter Five), multivariable binomial regression was carried out for virological and serological traits accounting for age and BMI (viral load, serostatus, and virus type). In this analysis, only patient age remained significant upon application of a Bonferroni correction for six hypotheses, with baseline serostatus and viral load just above the significance cutoff of 0.0083 (Table 6.2). The proportion of patients treated by corticosteroid therapy prior to ITU admission was not different for the two virus types (85.4% for WT vs 86.8% for Alpha, p = 1, Chi-Squared Test). Therefore, differences in terms of treatment received between Alpha-infected and WT-infected patients due to being recruited at different periods of the pandemic can largely be discounted. ITU discharge rates between WT-infected patients with N:RG (n = 341) vs N:KR (n = 80) were similarly not significantly different (55.4% vs 56.3%, respectively, p = 0.99, Chi-squared Test).

Lastly, the intensively sampled cohort allowed for the investigation of viral load dynamics with mortality rather than only baseline viral load. Within the intensively sampled cohort, 141 patients had linked outcome data available. Of these patients, 70.9% (n = 100) were discharged from the ITU alive. 136 of 141 had sufficient samples processed for viral dynamics classification. Discharge statistics for each viral load classification are shown in table 6.3. The two groupings with the lowest mortality were delayed peak responders and responders, although a Fisher's Exact test across all groups was not significant (p = 0.62). Further, multivariable binomial regression accounting for patient age showed no association for viral dynamics classification and ITU discharge status. Lastly, a post-hoc Fisher's Exact Test between delayed peak responders (lowest mortality) and those persistently positive (highest mortality) was similarly not significant (p = 0.62).

Viral Dynamics Classification	Percent patients discharged from ITU	Median Age	MR p
Cytokine Storm	66.7% (18/27)	60	N/A
Persistently Positive	60.7% (17/28)	58.5	0.64
Responder	74.4% (32/43)	58	0.44
Delayed Peak Responder	85.7% (12/14)	58	0.23
Soft Relapse	71.4% (15/21)	57	0.65
Hard Relapse	66.7% (2/3)	60	0.85

Table 6.3: Discharge rates and median ages for viral dynamics classifications. P value from multivariable binomial regression incorporating classification and patient age. Null hypothesis is no significantly different ITU discharge rate compared to "Cytokine Storm" patients, which are listed first alphabetically.

6.3 Discussion

None of the SARS-CoV-2 virological factors tested associated with any measure of disease severity or death in ITU. While the use of a Bonferroni correction across all statistical tests may be conservative, that only a specific metric of severity (APACHE II score) in one of the subgroups analyzed (seronegative patients) was close to significance upon multivariable regression suggests that result may be down to chance rather than a biological effect. The failure to observed increased severity when patients were infected with Alpha is at odds with some published literature. Using a matched cohort study of 368 deaths among 109,812 participants in the United Kingdom, Challen et al. (2021) calculated a mortality hazard ratio of 1.64 when comparing risk of death for Alpha versus WT infections.[397] Grint et al. (2021) similarly calculated an adjusted hazard ratio for death within 28 days of testing positive of 1.67 for Alpha versus non-Alpha based on 184,786 infections in the United Kingdom.[398] The null result presented for the REMAP-CAP cohort could either be the result of the multivariable regression not including sufficient variables (e.g., co-morbidities) to properly control for confounding, an improper power (although

this is unlikely given the number of patients enrolled), or there being true differences in the risk of hospitalization but not the risk of mortality once in hospital (which would manifest as increased risk of mortality if measured from the community level as in Challen et al. (2021) and Grint et al. (2021) but not if patients are enrolled in hospital). Given the third explanation, true differences in IFR would not be apparent in REMAP-CAP as all patients were recruited after having presented to ITU. This explanation is bolstered by the findings of Patone et al. (2021) which used linked SGTF results to identify probably Alpha infections among 198,420 cases in the United Kingdom (where 3432 were hospitalized and 822 died).[399] There, Alpha infection was associated with an adjusted hazard ratio of 1.65 for mortality within 28 days of testing positive but there was no difference in mortality risk among patients admitted to critical care, as observed for REMAP-CAP.

As in Chapter Five, the results for the effect of the N:KR mutation on WT samples are in conflict with literature published from Saudi Arabia.[304] There, the authors found increased severity and mortality for infections containing the N:KR mutation. Once Alpha-infected patients are excluded, this study is based on a smaller population than Mourier et al. (2021) and thus may be slightly underpowered to detect differences in mortality. Additionally, Mourier et al. (2021) defines severity based on binary presentation to hospital while this study is focused on an already intensively sick patient population. Similar to the Alpha hypothesis presented above, N:KR may be associated with an altered risk of presenting to hospital but not in severity of symptoms of hospitalized patients. Due to the global dominance of the N:KR-containing Omicron lineages, studies investigating this further would have to be completed using residual samples from previously collected hospitalized cohorts, which are the one of the target populations for the N:KR ASP-PCR assay described in Chapter Four.

Lastly, more rapid emergence and dominance of a variant within a population presenting to hospital than in the general population might be indicative of increased severity for that variant compared to co-circulating lineages. In the pre-VoC and pre-vaccine era, the time delay from infection to hospitalization and the time delay

from infection to diagnosis via community based testing were shown to not be significantly different across multiple age and ethnic groups.[400] Therefore, date of sampling in REMAP-CAP and date of sample collection for GISAID samples should be roughly equivalent for time since infection, allowing for fair comparison of the Alpha variant emergence in both populations. This analysis provided no evidence in favor or a difference in the risk of admission to ITU for individuals infected with Alpha or WT, although time delays from hospitalization to ITU admission would mask increased Alpha severity by 'pushing' Alpha infections in REMAP-CAP to appear later in time.

The totality of the evidence presented here (and partially in Chapter Five) support a model wherein the majority of pathology for patients once they have been admitted to hospital with COVID-19 is due to inflammatory- and immune-mediated effects rather than viral. Indeed, this may explain the general failure of antiviral therapies trialed in hospitalized patients but the relative success of immune-modulating drugs at reducing mortality.[401–403] Future work trialing these immune-modulating drugs in hospitalized patients should continue while antivirals would be best trialed in patients closer to time of infection.

7

SARS-CoV-2 Within-Host Evolution in Convalescent Plasma Treated Patients in the REMAP-CAP Immunological Domain

Contents

7.1 Introduction 131
7.2 Results 134
7.2.1 Cohort characterization 134
7.2.2 Emergence of variant-associated spike substitutions . . . 136
7.2.3 Within-host substitution rates 138
7.2.4 Molecular evolution in immunocompromised patients . . 143
7.2.5 Detection of C>U APOBEC-like substitutions 143
7.3 Discussion 146

Abstract

The factors leading to the emergence of SARS-CoV-2 VoCs are of great interest for both academic research and public health but are complex and mostly uncharacterized. Administration of CP therapy has been widespread despite limited demonstrated efficacy in severely ill patients. Concerns that CP use in critically ill patients may lead to the emergence of VoCs and VoIs have been buoyed by *in vitro* evidence and case reports of emergent spike substitutions in immunocompromised

patients after CP administration. The intensively sampled cohort in the REMAP-CAP CP trial contained 68 patients for whom sequential samples could successfully be sequenced, 43 of whom had been administered CP. Analysis of the within-patient emergent mutations revealed no spike substitutions similar to those characterizing VoCs or in the S RBD in either CP treated or control patients. Overall rates of molecular evolution, as measured by SNPs/site/year, were similar between both groups. No genes displayed elevated substitution rates between the two groups. The six immunocompromised patients in the study with multiple sequenced samples had unremarkable mutation profiles. Overall, the evidence fails to support the claim that widespread use of CP therapy was responsible for the generation of VoCs with antigenically dissimilar spike proteins. Future clinical trials should proceed without concerns for driving evolution in immunocompetent hosts. Lastly, the context distribution for C>U substitutions did not support the model of hypermutation by APOBEC and were better explained by selection against specific dinucleotides.

7.1 Introduction

The consequences of the ongoing adaptive evolution of SARS-CoV-2 to spreading in human populations are clear. Global rises in cases and deaths have generally followed the emergence of novel VoCs and VoIs with increased transmissibility, displacing prior circulating lineages.[294, 404] The factors leading to the emergence of specific mutations are complex.[405] The S:D614G substitution is polyphyletic, arose early in the pandemic before the accumulation of widespread immunity, and has a role in stabilizing the open state of the spike protein.[69, 291, 406, 407] Several spike substitutions, notably at residues S:484, S:501, and S:655, have been described as arising *in vitro* and *in vivo* in the absence of immune pressure and likely increase SARS-CoV-2 infectivity of human cells.[408, 409] Substitutions at the same residues in the spike protein have also been shown to decrease the neutralization titer of monoclonal antibodies and vaccine- and infection-derived antibodies, suggesting an additional role in escape of prior infection- or vaccine-elicited adaptive immune responses.[345, 410–413] While these studies have focused on the impact of infection

of human hosts, the emergence of Omicron (pangolin: BA.*), which contains an unprecedented number of spike mutations compared to the ancestral lineage, has been proposed to be due to reverse spillovers from animal reservoirs.[414] While the evidence in favor of this model does not come from field sampling, the S:Q498R and S:N501Y substitutions have previously appeared concurrently *in vitro* in virus adapted to bind rat ACE2.[415] In all likelihood, a combination of selective pressures to increase transmissibility and evade vaccine-derived immunity have driven and will continue to drive the evolutionary trajectory of the virus in human populations.

SARS-CoV-2 VoCs present an odd conflict with the observed evolutionary patterns of the virus. SARS-CoV-2 and other coronaviruses, due to the action of an encoded 3'-5' exonuclease, demonstrate a remarkably diminished mutation rate compared to other respiratory viruses.[243] Estimates for this rate center around $8*10^{-4}$ SNPs/site/year, while averages estimates for IAV, IBV, and RSV, are roughly two to three times higher.[416–420] In studies of viral evolution, regression of root-to-tip genetic distances to date of sampling allows for estimates of time since divergence for closely related lineages based on a linear accumulation of synonymous substitutions. However, for the major VoCs Alpha, Zeta, Gamma, and Omicron these regressions have been incompatible with emergence via the standard, linear accumulation of synonymous substitutions, strongly suggestive of undetected periods of rapid molecular evolution.[301, 348, 421] The reasons for the explosions of diversity in individual patients generating well-adapted viral lineages are unclear and under debate although a leading theory is prolonged replication in an immunocompromised host. Targeted sequencing of SARS-CoV-2 RBD from fecal shedding in wastewater has revealed circulating cryptic lineages with unique RBD configurations, suggestive of undetected chronic infections within the community leading to substantial intrahost evolution.[422]

It has been well-established that therapeutics can drive the emergence of specific SARS-CoV-2 mutations *in vivo* in immunocompetent and immunocompromised patients. A substitution leading to resistance to remdesivir, a nucleoside analogue

interfering with the activity of the RdRP, has been described once in an immunocompromised patient.[408] Three independent studies have demonstrated the emergence of the S:E484K substitution in a total of eleven high-risk or immunocompromised individuals treated with the monoclonal antibody Bamlanivimab.[423–425] An observational study of 100 Delta-infected patients treated with the monoclonal antibody Sotrovimab in Australia found four who developed S:E340K/A/V substitutions that are associated with monoclonal antibody resistance *in vitro*.[426]

CP is inherently polyclonal and thus the barrier for immune escape is likely higher than for monoclonal therapy. For example, in one of the observational case studies of Bamlanivimab the only patient (of six) that failed to produce the S:E484K substitution was also the only patient additionally treated with CP.[423] However, escape mutations after CP treatment have been described *in vitro* and in immunocompromised patients. Following twelve passages over 80 days of high titer live virus in the presence of potent CP (at the highest dose insufficient to completely inhibit replication), the S:E484K substitution emerged (along with a spike N-terminal domain insertion and deletion).[427] A patient treated with transplant-associated immunosuppressives and undergoing steroid and CP therapy for SARS-CoV-2 infection had the S:E484K substitution and S:del144-147 detected in a tracheal aspirate sample 27 days after treatment.[428] One highly-cited case study examined the evolutionary trajectory of a patient undergoing treatment with B-cell depletion chemotherapy over 102 days of SARS-COV-2 infection and two treatment rounds with CP.[429] This patient developed S:H69/70del (also in the Alpha variant) and S:D796H after the initial CP treatment, and the frequency of these mutations as a proportion of total sequencing reads decreased until the second dose of CP, after which they reemerged. These case studies have raised anxiety regarding the potential role of CP in driving SARS-CoV-2 antigenic evolution.

The intensively sampled domain of the REMAP-CAP CP trial offers a unique opportunity to interrogate the role of CP therapy on a largely immunocompetent population. Leveraging whole genome sequences generated as part of the trial, this research sought to answer three questions regarding CP therapy and SARS-CoV-2

molecular evolution: did CP therapy lead to 1) the generation of immune escape mutations; 2) an increased rate of subsitution; and/or 3) an altered distribution of substitutions towards structural proteins. Lastly, the type and context of the emergent SNPs was analyzed to determine whether this cohort lent support to APOBEC-mediated hypermutation of SARS-CoV-2.

7.2 Results

7.2.1 Cohort characterization

A subset of patients (n = 187 of 1917) in the REMAP-CAP trial were sampled repeatedly during their treatment in intensive care (Figure 7.1A). Of the 187 patients enrolled in the intensive arm, at least one sample was acquired, extracted, and quantified for 163. 269 samples from 94 patients had adequate viral loads for sequencing with veSeq, of which 256 samples returned sequencing coverage (defined as at least one mapped read with ≥two deduplicated bases). A single sample collected on a patient's fourth day in ITU showed strong evidence of contamination (assigned pangolin lineages from samples with 100% coverage cycled from B.1.1.301 to B.1.1.37 back to B.1.1.301 on days three, four, and six, respectively) and was excluded. 26 of the remaining sequenced samples were from patients where sequencing was only performed or successful on a single patient sample. Therefore, a total of 230 samples from 68 total patients were brought forward for analysis. The overall genome coverage as a percent of total bases called was adequate. The median coverage was 99.99% and 68.6% of samples (n = 157) had coverage values exceeding 95%. However, 10.5% of samples (n = 24) had less than 50% coverage, seven of which had less than 10% coverage. All samples were included in the analysis regardless of coverage or read depth. Sites without base calls were masked from the calling of substitutions in the next sequential sample; therefore, the total number of SNPs is likely an underestimate due to the proportionally reduced likelihood of calling a SNP following a low coverage sample.

Days of sampling were highly non-normally distributed ($p = 2.58*10^{-18}$, Shapiro-Wilk Test) with 59.0% of samples with known collection dates being collected on

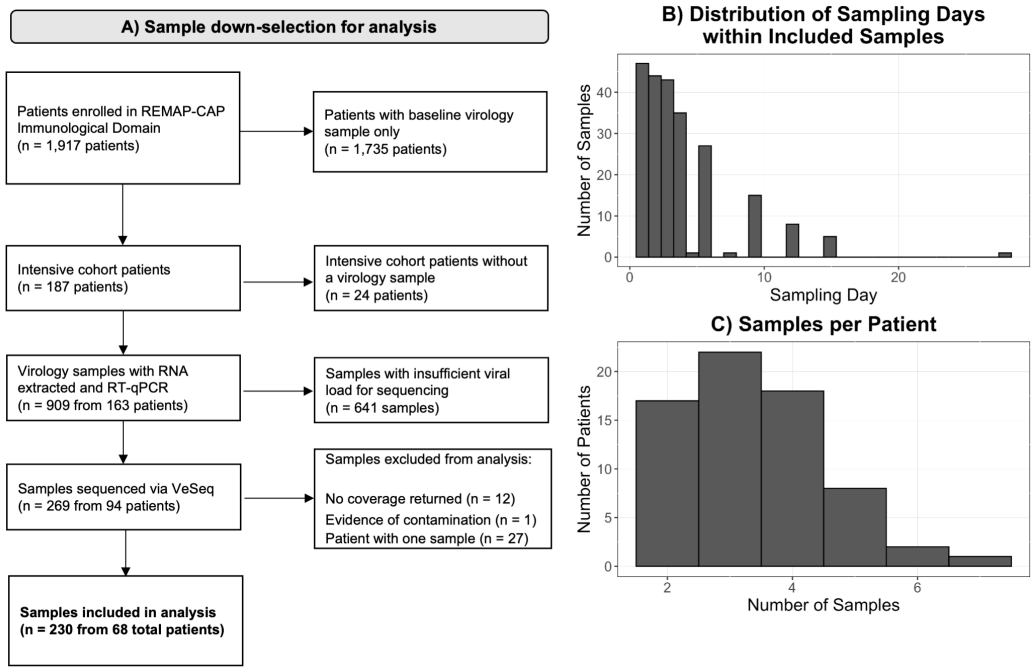


Figure 7.1: Description of Samples Included in Analysis. (A) Description of the sample selection for analysis. (B) Distribution of the days of sampling for the samples included (n = 230). 59.0% of samples were collected on days one, two, or three. (C) Total number of sequenced samples included for each patient (median value of three).

days 1, 2, or 3 in ITU and only 2.6% of samples coming on days 15 or later (Figure 7.1B). In terms of days between first and final sequenced sample, patients had a median number of four days of observation with 33% of patients only contributing samples covering one or two days. Number of included samples per patient was similarly non-normally distributed ($p = 2.05*10^{-5}$, Shapiro-Wilk Test), with patients contributing a median of three samples and a maximum of seven samples (Figure 7.1C). In total, this analysis covers 351 observed days in ITU.

43 of the 66 patients were treated with CP (two patients did not have randomization numbers that mapped to the trial database). 19/66 patients had also been treated with remdesivir prior to admission to ITU (11 of whom received CP). Six of the patients were identified as being immunocompromised, with all six being randomized to receive CP and one also receiving remdesivir. Eight patients that received CP were infected with Alpha.

7.2.2 Emergence of variant-associated spike substitutions

To investigate the concern that treatment with CP would lead to the emergence of spike substitutions characteristic of SARS-CoV-2 VoCs and VoIs, the intra-host emergent substitutions on the spike protein were identified and compared to a list of spike substitutions of concern and interest collated by the European Centre for Disease Control and Prevention.[430] These substitutions and their associated lineages (as of March 31, 2022) are listed in Table 7.1. Further, any emergent mutations in the spike RBD (amino acids 319-541) or S1 domain (amino acids 613-705) were scrutinized for their appearance in other sequenced samples globally as evidence for potential biological significance.

Table 7.1: Specific Spike Mutations of Interest. Adapted from The European Centre for Disease Control and Prevention.[430] Last updated on March 31, 2022. Tabulation does not distinguish BA.1 or BA.2 from B.1.1.529.

Spike Mutations	Characteristic Pangolin Lineage (WHO Classification)
R346K	B.1.621 (Mu)
K417N	B.1.1.529 (Omicron) B.1.351 (Beta)
K417T	P.1 (Gamma)
L452Q	C.37 (Lambda)
L452R	A.27 B.1.427/B.1.429 (Epsilon) B.1.617.1 (Kappa) B.1.617.2 (Delta) B.1.617.3 C.16
S477N	B.1.1.529 (Omicron) B.1.620
T478K	B.1.1.519 B.1.1.529 (Omicron) B.1.617.2 (Delta)
E484K	A.28 AT.1 AV.1 B.1.1.318 B.1.351 (Beta) B.1.525 (Eta) B.1.526 (Iota) B.1.620 B.1.621 (Mu) P.1 (Gamma) P.2 (Zeta) P.3 (Theta)
E484Q	B.1.617.1 (Kappa)
F490S	C.37 (Lambda)
N501T	A.28

Continued on next page

Table 7.1 – continued from previous page

Spike Substitution	Characteristic Pangolin Lineage (WHO Classification)
N501Y	A.27 B.1.1.529 (Omicron) B.1.1.7 (Alpha) B.1.351 (Beta) B.1.621 (Mu) B.1.640 C.1.2 P.1 (Gamma) P.3 (Theta)
T547K	B.1.1.529 (Omicron)
D614G	All circulating lineages
H655Y	A.27 A.28 B.1.1.529 (Omicron) C.1.2 P.1 (Gamma)
N679K	AT.1 B.1.1.529 (Omicron) C.1.2
P681H	AV.1 B.1.1.529 (Omicron) B.1.1.7 (Alpha) B.1.620 B.1.621 (Mu) B.1.640 P.3 (Theta)
P681R	B.1.617.1 (Kappa) B.1.617.2 (Delta) B.1.617.3
Q677H	B.1.525 (Eta)
A701V	B.1.351 (Beta) B.1.526 (Iota)

Across the patient cohort, there were four synonymous SNPs and 16 nonsynonymous substitutions in the spike gene. None of the 16 substitutions matched the substitutions of concern listed in Table 7.1, no mutations occurred within the spike RBD, and only one substitution was in the S1 domain. A CP-treated patient who was sampled on days one, two, and four in ITU and was infected with Alpha had an S:L629P substitution arise on the sample taken on the fourth day. This substitution had only been detected twice in the over 9.5 million sequences deposited on GISAID up to April 5, 2022. The strict purifying selection argues against increased fitness being provided by this substitution. Potentially of note, this patient also had a S:G75C substitution on the day four sample, which is similar to the S:G75V substitution in the Lambda variant.[431] S:G75C has been detected 226 times and does not associate with any particular lineage. This patient had further samples taken on days nine and 12 in ITU, which were not sequenced due to low viral loads (306 and 0 IU/mL, respectively), before later succumbing to COVID-19 in ITU.

The nsp12:E802D mutation has been described as leading to resistance to remdesivir treatment *in vitro* and *in vivo*.[408, 409] This substitution did not emerge within patients regardless of treatment, nor was nsp12:802D present in any patients (whether in the intensive cohort or not) at time of admission to ITU. The number of remdesivir doses received and the time since being treated were not available, limiting the utility of this cohort to interrogate remdesivir's impact on SARS-CoV-2 evolution.

7.2.3 Within-host substitution rates

There were a total of 215 intra-host SNPs detected in the cohort. 153 of these were nonsynonymous substitutions and 64 were synonymous. There were no intrahost insertions or deletions detected, inclusive of the S:D69/70 deletion whose emergence has previously been linked to CP treatment.[429] The C19217T SNP underlying the ORF1ab:P6318L mutation was detected in two individual patients; there were no other instances of homoplastic SNPs. Practically all of the observed SNPs were detected from two sequential samples where one sequence had low genome coverage and the other had high genome coverage. Fewer than 10% of detected SNPs (n = 21) were from two sequential sequences with >95% genome coverage. Only nine of these 21 detected SNPs were nonsynonymous substitutions. Requiring this high quality control threshold for genome coverage also reduces the number of patients eligible for this analysis to 47 with only a total of 201 days under observation (30.8% and 42.7% reductions, respectively). The decision was made to continue the below analyses based on the minimum two deduplicated read depth requirement and not implement genome coverage-based quality controls thresholds.

When examining substitution rates, per patient median values, where all patient rates are effectively treated equally, were considered inappropriate for two reasons: a) if the ground truth substitution rate was lower than 1 SNP/genome/day, then patients with fewer days in ITU would have had fewer opportunities to accumulate SNPs, driving median values slightly lower; b) patients accumulating 0 mutations over two and 27 days of observation would both contribute 0 SNPs/site/day to the

total estimate, despite the latter being indicative of a lower overall substitution rate. To correct for these biases, group level estimates used a weighted average based on total days observed per patient.

Across the whole genome (excluding the 3' UTR, due to VeSeq commonly producing sequencing/alignment errors for the 5' three bases of this genome section), the weighted average SNP/ site/year for all patients was $7.5*10^{-3}$, which is about 1 log higher than estimates for inter-host evolution.[417] There was no difference in overall substitution rates between CP treated and control patients (weighted averages of $6.9*10^{-3}$ vs $8.1*10^{-3}$, respectively, $p = 0.19$, Student's T test modified for weighted averages; Figure 7.2A) nor were the substitution rate densities significantly different ($p = 0.073$, Anderson-Darling k-Sample Test; Figure 7.2B) although the trend was towards slightly higher evolutionary rates in control patients. Individual level substitution rates were highly diverse and non-normally distributed for both patient populations ($p = 4.8*10^{-10}$, Shapiro-Wilk Test). Curiously, there was no linear relationship between days spent under observation and total SNPs detected ($r = 0.028$, $p = 0.83$, univariate linear regression; Figure 7.3).

While CP may not have influenced overall substitution rates, it was hypothesized that it may drive selection towards structural rather than nonstructural genes (structural defined as all S, E, M, and N) due to these being antigenic targets of CP. Gene-specific weighted average mutation rates are displayed in figure 7.4. In summary, there were no observed differences between substitution rates for specific genes between patient groups nor did any genes show elevated mutation rates within either or both patient populations.

One of the more common methods for detecting positive or negative selection is through the calculation of a dN/dS ratio for coding regions. Unfortunately, the derivation of this ratio depends on the occurrence of at least one synonymous substitution in a gene or the ratio fails to converge. Synonymous substitutions only occurred for 35 genes across 29 patients (and only four times in the spike gene), precluding use of this metric to detect CP-driven selection pressures at the individual level. As an alternative method for detecting selection, the ratio of

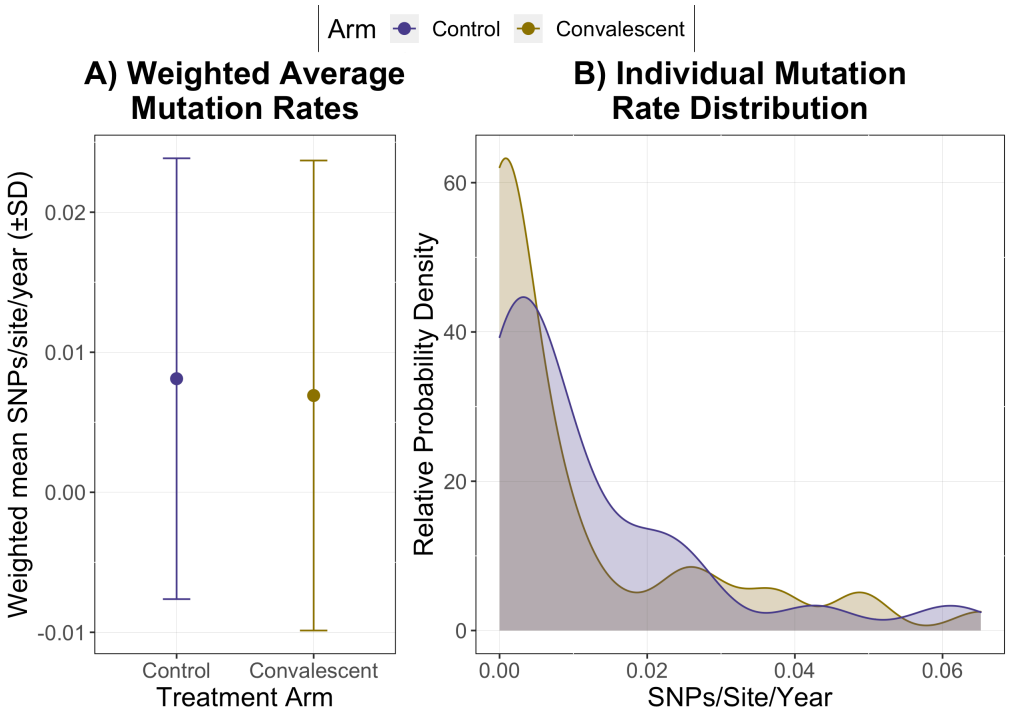


Figure 7.2: Whole Genome Mutation Rates in Control and Convalescent Plasma Treated Patients. (A) Weighted average calculated as described in methods. Averages not significantly different (p = 0.087, Student's T Test). (B) Distribution of individual mutation rates not significantly different (p = 0.073, Anderson-Darling k-Sample Test). Statistical tests on data displayed in plot (B) cannot treat patients with zero mutations and varying days of observation differently, motivating the analysis in plot (A).

total nonsynonymous to synonymous substitutions between groups was calculated. The results are summarized in table 7.2. Neither for the total genome nor for any individual gene was there a statistically significant difference between ratios for CP and control patient groups. The largest ratio of nonsynonymous to synonymous substitutions (6.5) was in the spike gene of the CP-treated group. A potential weakness with this approach is that the total number of synonymous and nonsynonymous sites will differ between lineages in a way that is not accounted for in this calculation, although those would be expected to be randomized between treatment groups. Further, given the low levels of overall diversity between lineages during the period in which this trial enrolled patients, the total number of nonsynonymous and synonymous sites should be roughly comparable, allowing this

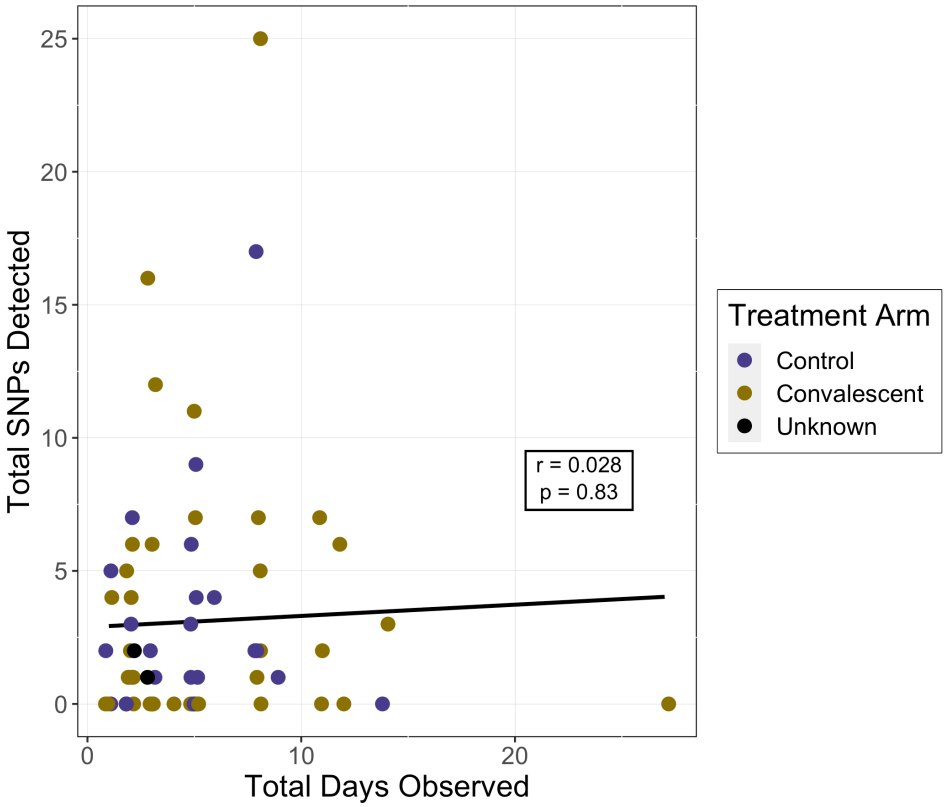


Figure 7.3: Correlation Between Total SNPs Accumulated and Time in ITU. P and r value from simple linear regression. Linear regression includes all samples irrespective of treatment arm.

measure to be an adequate substitute for dN/dS at the consensus level.

Reverting substitutions are those that occurred in sample *t+1* but revert to the original, nonmutated base in sample *t+n* (where n > 1). These reversions complicate analyses as they are scored equivalently to other SNPs but, functionally, are more of a product of sampling density than a reflection of diverse biology between patients (i.e., detection of a reverting SNP strictly depends on three samples being taken which bookend and include the period where the reverting SNP is dominant at the consensus level). Further, as a reversion represents the detection of two mutation events relative to their most recent sequence, they may inappropriately inflate the total. Among the 215 SNPs detected, 21.4% (n = 46) were associated with reversions. Among the three patients with the most hypermutated genomes (25, 17, and 16 SNPS, respectively), reversions were 8%, 47.1%, and 0% of the total SNP burden, respectively, suggesting this mechanism alone was not responsible

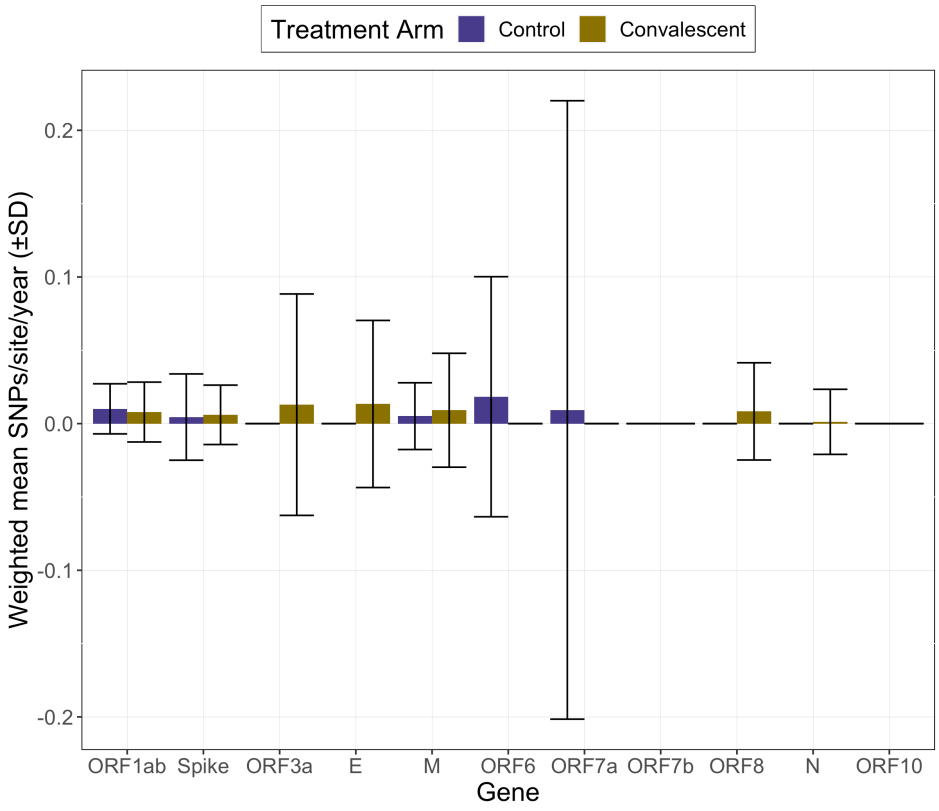


Figure 7.4: Gene-Specific Weighted Substitution Rates. Gene coordinates and lengths based on NCBI annotation of NC_045512. Error bars are ± standard deviation.

Gene	Ratio in Control Group	Ratio in Convalescent Group	P value[A]
Total	2.09 (48/23)	2.62 (102/39)	0.52
ORF1ab	2.32 (44/19)	2.55 (72/31)	0.86
Spike	1.5 (3/2)	6.5 (13/2)	0.25
ORF3a	N/A (0/0)	0.75 (3/4)	1
E	N/A (0/0)	Undefined (2/0)	1
M	0 (0/1)	Undefined (4/0)	0.2
ORF6	0 (0/1)	N/A (0/0)	1
ORF7a	Undefined (1/0)	N/A (0/0)	1
ORF7b	N/A (0/0)	N/A (0/0)	1
ORF8	N/A (0/0)	1 (1/1)	1
N	N/A (0/0)	Undefined (1/0)	1
ORF10	N/A (0/0)	N/A (0/0)	1

Table 7.2: Ratios of Total Nonsynonymous to Synonymous Substitutions Observed. (A) P values derived from Fisher's Exact Test.

for the extreme number of substitutions in some of these patients. The ratio of nonsynonymous/synonymous substitutions within SNPs that reverted (n = 46) vs those that did not (n = 169) were not significantly different (2.3 vs 2.52, p = 0.93, Chi-Squared Test). Therefore, the reversions were also not an outsize source of detected amino acid diversity.

7.2.4 Molecular evolution in immunocompromised patients

The above evidence fails to support a role for CP-driven selection in immunocompetent individuals, but this may be an inappropriate study population for VoC and VoI emergence. Therefore, a focused analysis of the six immunocompromised patients who received CP and had multiple sequenced samples was undertaken. For this trial, immunocompromised individuals either had active treatment with immunosuppressive drugs or had a known disease causing immune deficiency.[266] The patient factors and emergent substitutions of these six patients are summarized in table 7.3. Collectively, these patients were under observation for 34 days and had seven emergent SNPs, all of which were in a single patient, were nonsynonymous, and in ORF1ab. Further, six of these SNPs were reversions, leaving an A21185G/ORF1ab:*6974W substitution the only non-reverting substitution in an immunocompromised patient (* indicates a stop codon). This substitution restored the WT amino acid at that site; the sequenced stop codon may have been due to either sequencing error or a transiently dominant consensus base. In summary, none of the seven immunocompromised patients treated with CP developed any mutations relative to the WT sequence.

7.2.5 Detection of C>U APOBEC-like substitutions

As introduced in Chapter Two, the action of APOBEC enzymes on SARS-CoV-2 RNA would result in C>U substitutions in gRNA. The potential impact of this mutational driver was assessed within this cohort. By far the two most commonly detected substitutions were the transitions C>U (20.9%; 45/215) and U>C (19.1%; 41/215). This contrasts with the Simmonds (2020) analysis that found an excess of

Age	Infecting Lineage	Prior Remdesivir Treatment	Number of Sequenced Samples	Days between first and final sample	ITU Discharge Status	Emergent SNPs; *Substitutions*	
49	Presumptive WT	No	3	2	Alive	Day 2 > Day 3	T3949A; *ORF1ab:D1128E*
							G10112C; *ORF1ab:V3283L*
							G10119T; *ORF1ab:C3285F*
							A21885G; *ORF1ab:*6974W*
						Day 3 > Day 4	A3949T; *ORF1ab:E1228D*
							C10112G; *ORF1ab:L3283V*
							T10119G; *ORF1ab:F3285C*
60	Alpha	No	4	9	Alive	None	
67	WT	No	3	3	Deceased	None	
56	WT	Yes	4	5	Alive	None	
51	WT	No	3	8	Alive	None	
70	WT	No	5	4	Alive	None	

Table 7.3: Unremarkable SARS-CoV-2 Evolution in Immunocompromised Patients Receiving Convalescent Plasma Therapy. Of the seven total SNPs, six were paired reversions. "Presumptive WT" indicates sample not typed but collected prior to the earliest typed Alpha infection in the trial.

C>U relative to all other transitions among SARS-CoV-2 sequences deposited on GISAID.[246] In line with other analyses, however, there was a stark normalized mutational asymmetry for C>U vs U>C (ratio of 1.91) that was closer to the null expectation of one for G>A vs A>G (ratio of 1.22).[220]. Exploring the base context for these four transitions revealed no upstream or downstream context bias for counts of C>U vs U>C or for G>A vs A>G (i.e., frequency of 5'U for C>U vs U>C; p = 1 in all instances, Fisher's Exact Test). Upstream and downstream base context normalized to monodinucleotide frequencies revealed no bias for C>U towards favorable contexts for APOBEC deamination (5'U or 5'C; Figure 7.5A,B). A>G and G>A similarly displayed a stochastic distribution, likely due to the low number of overall purine transitions (n = 36) (Figure 7.5C,D).

All SNPs occur within the context of two dinucleotides where the mutated base is either the 5' base or 3' base. By analyzing the context for observed C>U transitions relative to the frequency of specific dinucleotides in the SARS-CoV-2 reference genome, neither 5'-Up<u>C</u>-3' nor 5'-Cp<u>C</u>-3', the preferred contexts of A3

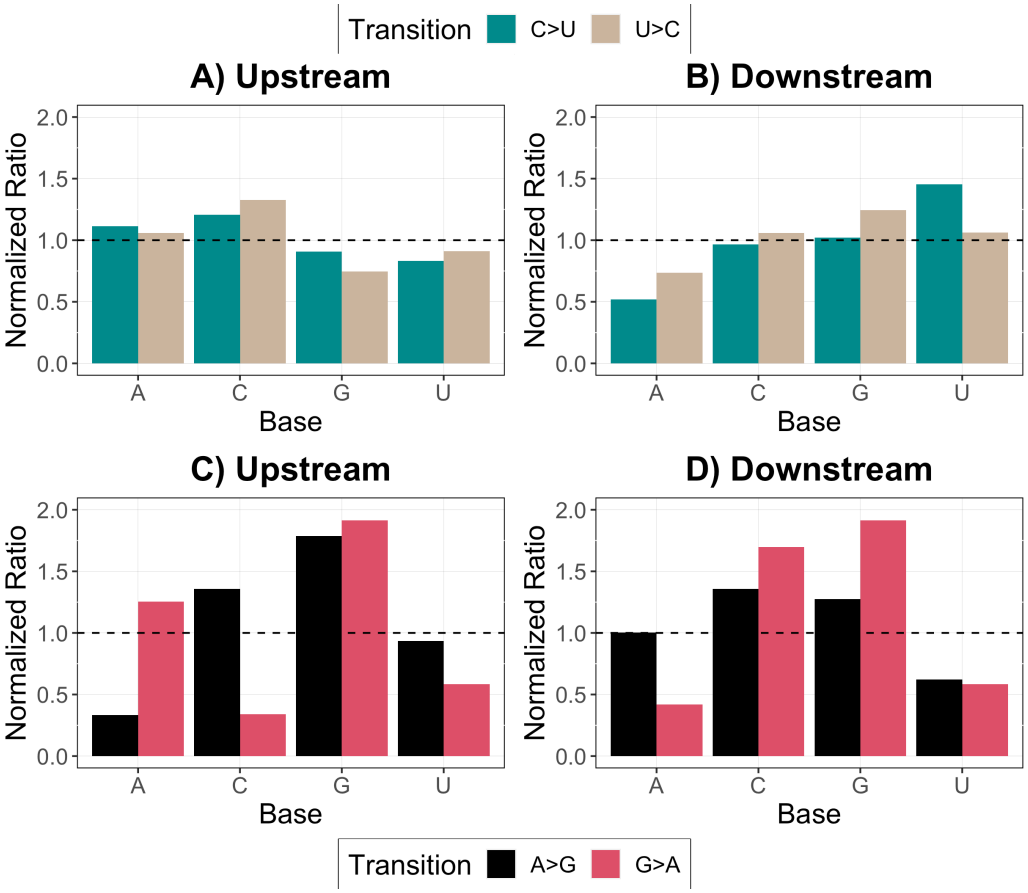


Figure 7.5: Influence of Base Context on Transitions. 56.7% (n = 122) of observed SNPs were transitions. Panels (A) and (B) for pyrimidines and (C) and (D) for purines. Overall distribution of counts for U>C vs C>U or for A>G vs G>A were not significantly different (Fisher's Exact Test). Visualization displays values normalized to mononucleotide frequency of SARS-CoV-2 reference genome NC_045512. Dashed line is null expectation of 1.

proteins, were mutated at a rate different from chance (26.7% observed vs 25.7% expected and 22.2% observed and 16.2% expected, respectively; Figure 7.6A). Interestingly, the distribution of downstream context was significantly different than expected (p = 0.002, Fisher's Exact Test). This impact was driven by decreased mutation of 5'-CpA-3' (15.6% observed vs 37.9% expected) and increased mutation of 5'-CpG-3' (20.0% observed vs 8.0% expected; Figure 7.6B), neither of which have been described as impacting A3 function.

Further, that the only homoplastic substitution detected (C19217T) was C>U is not itself evidence of A3 activity; given the total number of observed SNPs (215), SARS-CoV-2 genome length (29003 bases), and observed C>U substitution

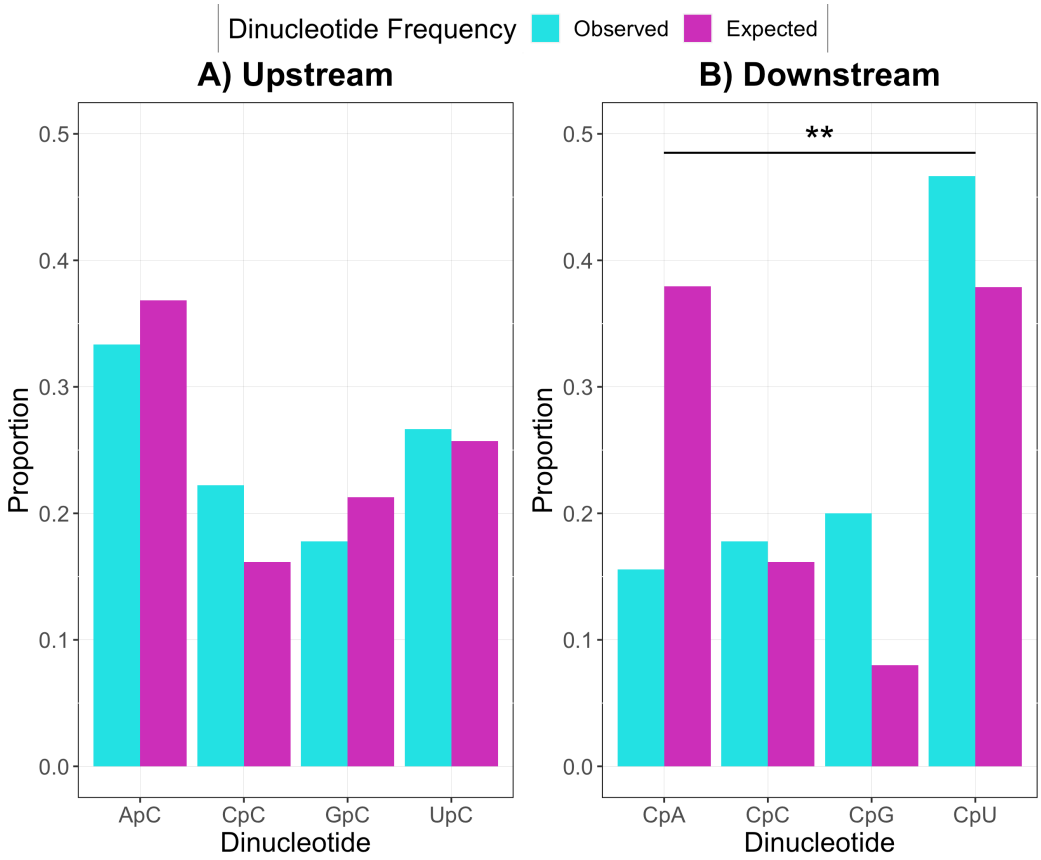


Figure 7.6: Impact of Dinucleotide Context of Observed C>U Substitutions Compared to Expected Dinucleotide Frequencies. Expected dinucleotide frequencies are based on the relative frequency within the SARS-CoV-2 reference genome NC_045512; a random distribution of C>U transitions should match the expected frequencies. Overall distribution of counts of downstream context significantly different by Fisher's Exact Test. ** = $p < 0.005$.

frequencies (20.9%), the probability of a C>U homoplastic substitution occurring in two patients was approximately 11.5% and therefore not significant.

7.3 Discussion

This chapter aimed to interrogate the potential role of CP therapy on the development of immune-escape mutations in severely ill COVID-19 patients. Among immunocompetent individuals, this population may have been the most optimal progenitors for CP-driven immune escape variants – these patients had prolonged infections with viral loads sufficient for sequencing (suggestive of a failure of immune control), several patients were infected with an antigenically mismatched variant,

and the vast majority (75%) were actively being treated with steroid therapy to dampen the immune response. Still, a variety of different approaches to detect CP-driven selection were all unsuccessful.

There are a number of limitations of this study. Principally, the study sequenced virus isolated from nasopharyngeal swabs. While likely an appropriate marker of active cellular infection, these samples may display lower diversity than BALF samples.[432] This analysis was also limited to detecting changes at the consensus level, which prevented the use of dN/dS ratios for detecting selection due to the limited accumulation of synonymous substitutions. Future investigations based off read level data and minor variants rather than consensus changes may have more specific and sensitive detection of selective pressures. The low number of SNPs detected from sequential, high genome coverage samples may also indicate that a large portion of the observed diversity in simply artefactual. Still, the null results for differences between treatments presented in this chapter are similarly null when analyzed using this tighter quality control threshold. Lastly, this study was not designed to power detection of a minor difference in overall mutation rates. Based on the number of individuals who were able to be included in this analysis, only an increase in mutation rate of 25% or greater in the CP group would have been appropriately powered. Larger studies may have been able to detect a subtle influence, although the lack of directionality in favor of CP-driven molecular evolution in this study challenges that assertion.

The evidence presented here does not preclude the scenario wherein a single immunocompromised patient, upon receiving CP therapy, was the progenitor of one or more of the circulating VoCs and VoIs. Indeed, the lack of intermediate lineages for the majority of VoCs despite their considerable branch length is evidence of generation in individual hosts. However, the evidence that the evolution was specifically due to CP therapy is weak to nonexistent. Further, that after over two years of transmission and over 500 million documented infections only five WHO recognized VoCs have arisen (albeit the antigenic distance behind BA.1 and BA.2 is substantial) speaks to the infrequency of sweeping changes in the SARS-CoV-2

genome leading to remarkably increased fitness. While CP has shown limited efficacy in hospitalized patients,[266, 433] its use should not be limited by concerns of instigating the emergence of novel variants.

The observed elevated mutation rate and high rates of reversion being coupled with the low number of sequential high coverage samples may be indicative of sequencing error or the stochastic effects of low read depth. Unpublished data from Prof. Peter Simmonds on the same sample set has indicated that the proportion of reads coinciding to minor variant populations is strongly inversely proportional to viral load. For this hospitalized cohort, this phenomenon is most readily explained by increased diversity as patients progress further along in the disease course. As the true frequency of a minor population variant increases, the likelihood that a random sample of low read depth will lead to that minor variant appearing as consensus will increase. Assuming the minor population variant frequency is static, the likelihood of reaching consensus in two sequential samples is the square of the probability of sampling the minor variant once (and thus always lower than the probability of one event), inevitably leading to some of the detected "reversions". Some of these reversions are likely to be artefacts of low read depth and do not represent true changes in the consensus genome of the virus. An approach with a more stringent read depth requirement would reduce this risk but may also obscure true mutational phenomena. The trade-offs of these two approaches must be balanced for the objective sought. As this study was primarily interested in detecting minor signals of therapeutic-induced selection, it was reasoned that an approach limiting sensitivity would be higher risk and thus less appropriate.

Lastly, the lack of a signal of APOBEC-driven evolution in this population is at odds with much published literature (see citations in Chapter Two). In particular, the lack of an overrepresentation of 5'U in C>U transitions in either the transition context (Figure 7.5A) or dinucleotide context (Figure 7.6A) challenges the potential role of APOBEC deamination in this population. The impacts of downstream base contexts on C>U occurrence are consistent with the dinucleotide selection pressures described in Chapter One. C>U mutation of 5'-CpA-3' dinucleotide would create

the unfavored 5'-UpA-3' dinucleotide; C>U mutation of 5'-CpG-3' would remove these dinucleotides from the genome. However, the low number of CpG dinucleotides in the SARS-CoV-2 genome caution overinterpretation of the results on dinucleotide context. In this study, the observed mutation distribution is better explained by selection against specific dinucleotides rather than APOBEC-driven evolution.

8

Dinucleotide Suppression and Host Species Optimization in Arboviruses

Contents

8.1 Introduction 151
8.2 Results 155
8.2.1 Dinucleotide Suppression Across Human-Infective Arboviruses 155
8.2.2 Dinucleotide Suppression and Transmissibility in Arboviruses 159
8.2.3 Influence of Flavivirus Host Range on CpG Dinucleotide Suppression 161
8.2.4 Shifts in dinucleotide suppression and viral fitness . . . 165
8.3 Discussion 169

Abstract

Viruses optimize their genome composition to replicate within the species they infect, limiting their ability to spillover to distantly related host species. Arboviruses, an informal classification of viruses, are constrained by the need to replicate in the diverse intracellular environments of arthropod and non-arthropod hosts. Suppression of the CpG dinucleotide frequency has previously been shown to be a strong determinant of *in vitro* and *in vivo* viral fitness in mammalian hosts but is not present in viruses exclusively infecting insects. Leveraging a database

cataloging human-infective arboviruses and their transmissibility, strong suppression of the CpG dinucleotide was demonstrated in six of the eight arbovirus families known to infect humans, mimicking the selective pressures seen in viruses infecting mammals exclusively. Analysis of human-infective members of the *Flaviviridae* family (genus: *Flavivirus*) revealed a strong correlation between increased CpG frequency and decreased transmissibility in human populations. Members of a single highly-transmissible arbovirus genus, *Alphavirus*, appear to partially evade CpG suppression through an as-of-yet uncharacterized mechanism. Further analysis across a broader range of members of the *Flaviviridae* family demonstrated very strong association between phylogenetic grouping and CpG suppression; increased time spent replicating in arthropod hosts was associated with increased frequency of CpG. Additionally, replication cycles in bird hosts were further associated with increased CpG content in mosquito-borne flaviviruses. Recent shifts in transmissibility for Zika virus identified during the 2013 outbreak in South America were not associated with shifts in CpG content; levels of CpG suppression were stable in samples collected over 35 years. Altogether, these results suggest that dinucleotide suppression is a strong determinant of arbovirus host range and a barrier to pathogen spillover among arboviruses that are poorly optimized for replication in mammalian hosts. Increased surveillance of members of the genus *Flavivirus* without documented human transmission are warranted over surveillance of poorly transmitted members of the genus.

8.1 Introduction

Spillover of zoonotic pathogens into human populations has had devasting consequences in recent history. The SARS-CoV-2 pandemic resulted in immeasurable economic and educational disruption; WHO estimates place the death toll over 15 million.[52] Over 40 million people are living with chronic HIV-1 or HIV-2 infections, both of which are the result of cross-species transmission of simian-immunodeficiency viruses infecting African primates.[434, 435] The zoonosis of IAV and IBV is unknown and may go as far back as the 5th century BC, but the current

burden, driven by evolution and recombination during circulation in swine and poultry, kills well over 100,000 people per year and puts nearly 10 million people in hospital.[436, 437] Being able to identify potentially zoonotic pathogens prior to spillover would allow for early stage development of vaccines and/or therapeutics in an effort to mitigate the impact of a spillover event.

A wealth of literature dedicated to understanding the risk factors driving the emergence of novel infectious diseases has been published since before the SARS-CoV-2 pandemic, although with mixed and sometimes conflicting conclusions.[106–108, 438–446] There have been large-scale projects funded on the basis that an increased understanding of virus diversity is the best approach to pandemic preparedness. The United States Agency for International Development (USAID) PREDICT project, which ran from 2009 to 2019, aimed to increase zoonotic surveillance of potential wildlife reservoirs and lab capacity in regions deemed 'hotspots' for emerging infectious diseases (such as South-East Asia and Sub-Saharan Africa).[447, 448] Building on PREDICT, the Global Virome Project (GVP), which estimates there remain 600,000-800,000 unknown viruses with zoonotic potential, seeks to vastly increase the number of identified and sequenced viruses globally through virus discovery projects.[449] The Vietnam Initiative on Zoonotic Infections (VIZIONS) enrolls patients presenting to hospital in Vietnam with diseases of unknown origin; pathogen-agnostic sequencing methods performed on patients for whom no pathogen is detected by standard screening methods will hopefully identify novel human-infective viruses.[450] Lastly, UC Davis, in partnership with USAID and GVP, have launched a web tool called SpillOver to calculate a 'risk score' for novel viruses based on a variety of virological factors.[451] The core thrust of these projects is the belief that comparing and contrasting the characteristics of novel viruses with those previously shown to transmit in human populations would allow for generalized and accurate risk assessment.

This chapter aims to apply that characterization framework to a niche of group of well-studied viruses. Arboviruses, as introduced in Chapter One, represent an interesting group through which to study the viral determinants of infectivity and

transmissibility due to their complex and constrained life cycle. Their dedicated replication cycles in distantly related hosts necessitate the conservation of fitness across diverse intracellular environments and cell entry pathways. In combination, these limit the evolutionary space available for exploration through diversification.[70] For instance, the lifecycle of mosquito-borne arboviruses in their arthropod hosts requires infection of mesenteronal epithelial cells, deposition into the hemocoel, and infection of salivary gland acinar cells with infectious hemolymph.[452] Infection of mammalian host cells must also occur alongside infection of these two distinct mosquito cell types. If the virus is unable to replicate to high titer within the mesenteron, secondary amplification in hemocoel-associated cells and tissues may also be necessary, increasing the required number of productively engaged cell receptors. If the capacity to infect these cell types is partially or completely compromised, the virus will be unable to be transmitted to a non-arthropod host. Further, vector-borne transmission pathways have previously been identified as being predictive of lower transmission risk in human populations.[438, 441] Within the transmission levels described by Woolhouse et al. (2016) and introduced in Chapter One, vector-borne viruses are overrepresented within T3 levels (human-human transmissible, $R_0 < 1$) and underrepresented in T4 levels (human-human transmissible, $R_0 > 1$).[108, 453] Analysis of the small subset of these viruses that can infect humans in an attempt to identify the differences between those that transmit poorly versus successfully may provide further clarity of the virological determinants of transmissibility. For arboviruses, human-human transmission proceeds from human host to vector to new human host. Arbovirus transmissibility is therefore the measurement of how rapidly and consistently an arbovirus is able to infect a new human host through its vector step.

The selective pressure against the CpG dinucleotide in viral genomes infecting mammals has been shown to partially determine their *in vitro* and *in vivo* fitness. Artificially increasing the frequency of the CpG (and UpA) dinucleotides through synonymous recoding severely attenuates viral replication and pathogenesis in mammalian cells and mouse models.[86–88, 91, 454, 455] However, CpG suppression

is absent in those viruses exclusively infecting insects, suggesting that these hosts do not contain the intracellular protein (ZAP) responsible for this phenotype.[85] Multiple members of the *Alphavirus* genus have been shown *in vitro* to be restricted by ZAP and members of the *Flavivirus* genus have diverse levels of CpG suppression.[456, 457] This failure to fully adapt for replication in mammalian hosts or escape the selective pressure entirely suggests conflicting evolutionary pressures in insect vs mammalian cells, which has recently been demonstrated *in vitro* within a Zika virus (ZIKV) model.[458] There, increased CpG content was associated with increased infectivity of mosquito cell lines while decreased CpG content was associated with increased infectivity of mammalian cell lines. The role of these conflicting evolutionary pressures on CpG on arbovirus fitness in human populations has not previously been explored.

To investigate the relationship between dinucleotide suppression and human-human transmissibility in arboviruses, publicly available sequence data was combined with the transmission classifications described by Woolhouse and Brierley (2018).[103] Of note, use of the family designations *Flaviviridae*, *Togaviridae*, and others in this chapter refer only the arboviruses within the families and not viruses transmitted by other routes (such as the genus *Hepacivirus* in the family *Flaviviridae*). The results showed strong correlation between transmissibility and CpG suppression that was most pronounced in the *Flavivirus* genus; this prompted more in-depth analysis of the broader role of host range on dinucleotide suppression within *Flaviviridae*. Altogether, the results presented suggest arboviruses optimize their genome composition for the hosts in which they currently replicate, representing a barrier to increased transmissibility in weakly human-transmissible arboviruses and spillover of insect-specific flaviviruses. For the purpose of this thesis, genome composition refers to mononucleotide and dinucleotide frequencies consistent with the definitions in the literature used for synonymously recoded virus clones.[86, 88, 458]

8.2 Results

8.2.1 Dinucleotide Suppression Across Human-Infective Arboviruses

Reference sequences of all human-infective, vector-borne RNA viruses identified in Woolhouse and Brierley (2018) were downloaded from Genbank.[103] Consistent with the dataset designations, the four serotypes of Dengue virus (DENV1-4) and two serotypes of West Nile Virus (WNV1 and WNV2) were included as independent entries. The dataset included 93 viruses across eight different virus families with representative viruses having genomes composed of (+)ssRNA, (-)ssRNA, or dsRNA. 36.6% of the viruses had segmented genomes (all composed of (-)ssRNA or dsRNA). The virus family with the most representatives was *Flaviviridae* (n = 30) while there was only a single member of the family *Orthomyxoviridae*. 71 viruses were classified as L2, 15 as L3, and seven as L4. The distribution of virus families within transmission levels was not significantly different (p = 0.46, Fisher's Exact Test) although all L4 viruses were either members of *Togaviridae* (n = 1) or *Flaviviridae* (n = 6), both of which are (+)ssRNA, nonsegmented families. The human-infective arboviruses included in this project are listed in Table 8.1; GenBank accession numbers are provided in Table 10.2.

Two methods for measuring dinucleotide suppression were deployed. The first method (termed "simple") calculates a ratio of observed vs expected dinucleotide frequencies (O/E) based on observed base composition across the entire genome. The second method (termed "corrected") performs a similar calculation to the simple method but is limited to coding regions and normalizes to the amino acid sequence. Only the positions with degeneracy are scored, eliminating any skew from the inclusion of rare amino acids in the protein product (e.g., tryptophan is only encoded by UGG) or amino acids that encode for generally suppressed dinucleotides (e.g., tyrosine will always encode a UpA). The corrected method was applied to the longest ORF for each virus. O/E values were significantly different for nine of the 16 dinucleotide pairings when comparing estimates given by the simple method versus the corrected method (Table 8.2). In particular, the estimates for

Virus Family	Genome Composition	L2	L3	L4
Flaviviridae	(+)ssRNA	Aroa virus Bagaza virus Banzi virus Cacipacore virus Edge Hill virus Gadgets Gully virus Ilheus virus Kokobera virus Kyasanur Forest Disease virus Langat virus Louping Ill virus Murray Valley Encephalitis virus Ntaya virus Omsk Hemorrhagic Fever virus Powassan virus Saint Louis Encephalitis virus Tembusu virus Tick-borne Encephalitis virus Uganda S virus Wesselsbron virus	Japanese Encephalitis virus West Nile virus 1 West Nile virus 2 Usutu virus	Dengue virus 1 Dengue virus 2 Dengue virus 3 Dengue virus 4 Yellow Fever virus Zika virus
Nairoviridae	(-)ssRNA	Dugbe virus Thiafora virus	Crimean-Congo Hemorrhagic Fever virus	
Orthomyxoviridae	(-)ssRNA	Thogoto Thogotovirus		
Peribunyaviridae	(-)ssRNA	Bunyamwera virus California Encephalitis virus Caraparu virus Catu virus Guama virus Guaroa virus Kairi virus Madrid virus Marituba virus Nyando virus Oriboca virus Patois virus Shuni virus Tacaiuma Orthobunyavirus Wyeomyia virus	Bwamba virus Oropouche virus	
Phenuiviridae	(-)ssRNA	Candiru virus Naples virus Punta Toro virus Uukuniemi virus	Dabie Bandavirus Rift Valley Fever virus	
Reoviridae	dsRNA	Banna virus Cooriparta virus Eyach virus Great Island virus Lebombo virus Orungo virus	Colorado Tick Fever virus	
Rhabdoviridae	(-)ssRNA	Alagoas virus Chandipura virus Cocal virus Indiana virus Isfahan virus Maraba virus New Jersey virus Piry virus		
Togaviridae	(+)ssRNA	Eastern Equine Encephalitis virus Everglades virus Getah virus Highlands J virus Mayaro virus Mosso des Pedras virus Mucambo virus Ndumu virus Pixuna virus Rio Negro virus Sindbis virus Tonate virus Una virus Western Equine Encephalitis virus Whataroa virus	Barmah Forest virus Ross River virus Venezuelan Equine Encephalitis virus Semliki Forest virus O'nyong-nyong virus	Chikungunya virus

Table 8.1: Human-Infective Arbovirus Families and their Members. List of human-infective arboviruses included in this study. Adapted from Woolhouse and Brierley (2018).[103]

Dinucleotide	Median O/E (simple method)	Median O/E (corrected method)	p value
ApA	0.99	1.00	0.13
ApC	0.98	0.99	0.98
ApG	**1.03**	**1.10**	**$1.14*10^{-13}$**
ApU	0.98	0.99	0.51
CpA	1.25	1.25	0.27
CpC	**1.05**	**0.98**	**$1.06*10^{-6}$**
CpG	0.52	0.55	0.59
CpU	1.09	1.07	0.22
GpA	**1.11**	**1.06**	**$1.49*10^{-8}$**
GpC	0.98	0.97	0.60
GpG	**1.03**	**0.94**	**$1.57*10^{-11}$**
GpU	**0.85**	**0.93**	**$6.26*10^{-12}$**
UpA	**0.69**	**0.84**	**$2.08*10^{-7}$**
UpC	1.01	0.96	0.0010
UpG	**1.28**	**1.19**	**$1.25*10^{-9}$**
UpU	**1.05**	**0.97**	**$2.63*10^{-19}$**

Table 8.2: Comparison between Simple and Corrected Methods for Estimating Dinucleotide Observed/Expected Ratios in Human-Infective Arboviruses. Includes median observed/expected value for all human-infective arboviruses (n = 93) calculated using SSE v1.4. p value calculated from two-tailed Mann-Whitney U Test; significant dinucleotides (after Bonferroni correction) in bold.

under- and overrepresentation were largely dampened for UpA and UpA-adjacent dinucleotides (e.g., GpA, UpG, etc.). Based on this result, the additional analyses in this project were performed using the corrected method.

73% of the included viruses were members of the *Flaviviridae*, *Togaviridae*, or *Peribunyaviridae* families, creating the risk that analyses based on collective traits of all individual viruses would be biased towards the traits of those families specifically. To mitigate this risk, analyses generally investigated within-family or within-group trends rather than global. Analyzing dinucleotide frequencies across the eight virus families represented in the dataset demonstrated that the majority of dinucleotides are not present at corrected O/E values expected by chance alone (Figure 8.1). In particular, the CpG dinucleotide was strongly suppressed in six of the eight viral families (mean O/E <0.6) and the UpA dinucleotide was suppressed in all eight (mean O/E < 1.0). Mean CpG O/E across all eight viral families showed higher variance than for UpA (standard deviations of 0.27 vs 0.12, respectively). Members

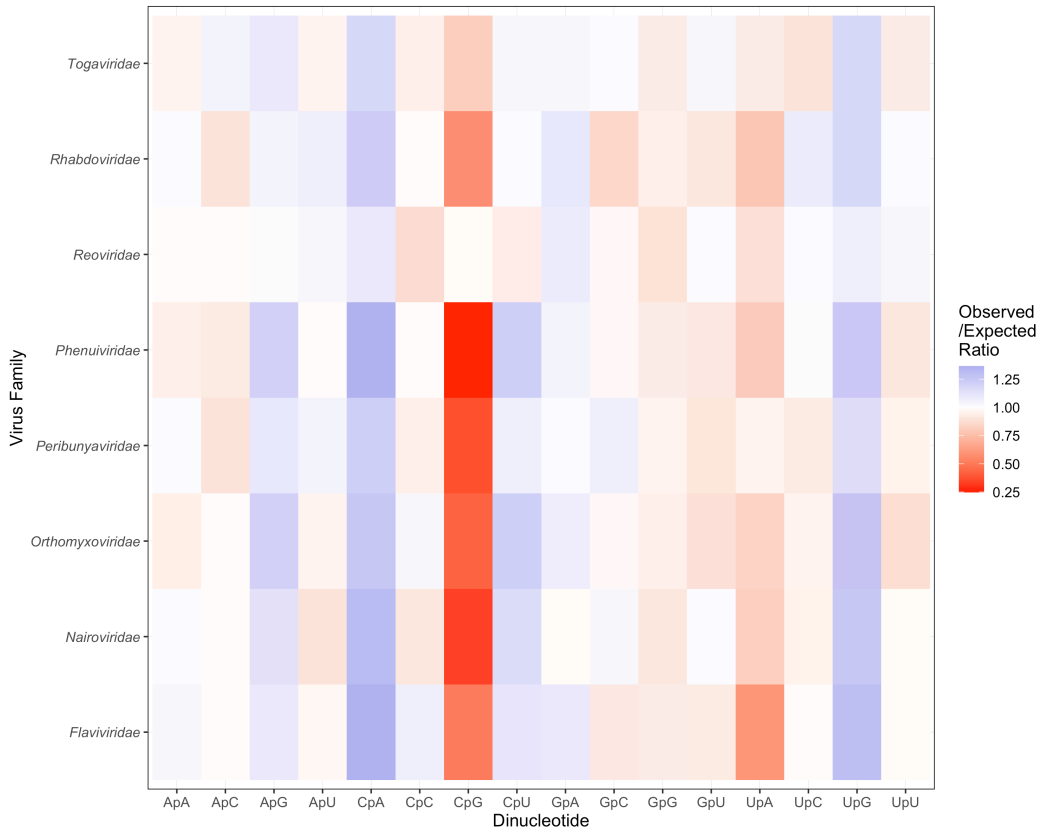


Figure 8.1: Dinucleotide Odds Ratios for Human-Infective Arbovirus Families. Dinucleotide O/E ratios calculated using the "corrected" method for coding regions as described in methods. CpG suppressed in all measured families except for the dsRNA *Reoviridae*.

of the *Reoviridae* and *Togaviridae* families escaped the strong selection against CpG, but *Togaviridae* still demonstrated some mild suppression (mean O/E = 0.81). The dinucleotides adjacent to CpG were commonly overrepresented, with CpA and UpG being present at mean O/E ratios > 1.2 in six and four families, respectively.

GC content significantly positively correlated with corrected CpG O/E across all viruses ($\tau_b = 0.32$, p = $4.72{*}10^{-6}$, Kendall's Rank Correlation; Figure 8.2A) as reported previously.[85] GC content was positively correlated with corrected CpG O/E for every virus family with more than one virus, but the association was only significant for *Flaviviridae* and *Peribunyaviridae* ($\tau_b = 0.34$ and p = 0.0075 and $\tau_b = 0.37$ and p = 0.042, respectively, Kendall's Rank Correlation). The highest median GC content levels were in the *Togaviridae* and *Flaviriviridae* familes (49.8%, and 49.8%) families while the lowest was in the *Peribunyaviridae* family (40.0%); Figure

8.2B). GC content was significantly different when viruses were grouped by genome composition. Virus families with (+)ssRNA genomes (n = 51) had higher median GC content than families with (-)ssRNA genomes (n = 35; 49.8% vs 41.9%, p = $1.48*10^{-12}$, two-tailed Mann-Whitney U Test; Figure 8.2A). GC content further complicates analyses that fail to consider virus family as a confounding variable.

8.2.2 Dinucleotide Suppression and Transmissibility in Arboviruses

The potential influence of dinucleotide frequency on transmissibility in human populations was investigated using the two virus families containing L2, L3, and L4 viruses (*Togaviridae* and *Flaviviridae*). Three hypotheses for the relation between CpG suppression and transmissibility in human populations were posited. First, if transmissibility was directly related to intracellular replicative capacity, CpG frequency would decrease as transmissibility increased in a manner from L2>L3>L4. Second, the binary trait of transmissibility could be determined by factors separate from intracellular replicative capacity while the scale of transmissibility is directly related. This effect would manifest in the highest CpG O/E levels in the poorly optimized genome composition (L3), the lowest in the most well-optimized genome composition (L4), and with a large spread but overall middling values for those without the trait to transmit but with a mixture of intracellular replicative fitness (L2). Third, if individual virus families were resistant to the intracellular selective pressure against CpG or CpG suppression didn't influence fitness, then there would be no relation between CpG suppression and transmissibility.

The *Flaviviridae* family was the most appropriate family in which to study this effect. The family includes 20 L2 viruses, four L3 viruses, and six L4 viruses (four of which are DENV serotypes). Median CpG O/E values across the three transmission levels were 0.55, 0.59, and 0.43, respectively (Figure 8.3). O/E values were not quite significantly lower for L2 viruses compared to L3 viruses (medians of 0.55 and 0.59, respectively, p = 0.068, two-tailed Mann-Whitney U Test) but were significantly higher for L2 viruses compared to L4 viruses (medians of 0.55 and

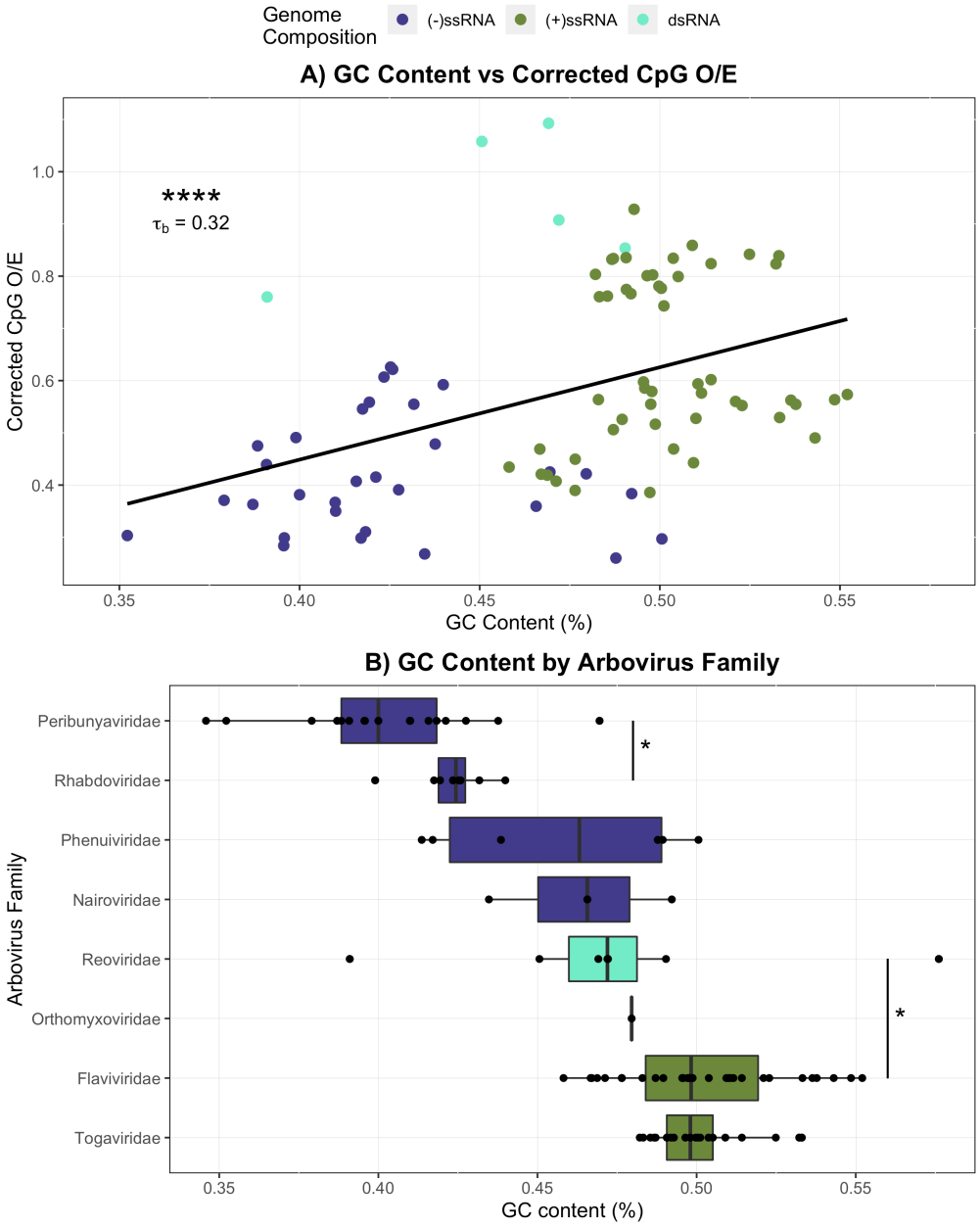


Figure 8.2: Interaction of GC Content, CpG Suppression, and Arbovirus Family. Whole genome GC content percentage vs corrected CpG O/E. * = $p < 0.05$ and **** = $p < 0.00005$ and from Kendall's Rank Correlation Test (A) or two-tailed Mann-Whitney U Test (B).

0.43, respectively, p = 0.0011, two-tailed Mann-Whitney U Test). The *Togaviridae* family has 15 L2 viruses, five L3 viruses, and one L4 virus. Median CpG O/E values across the transmission levels were 0.83, 0.76, and 0.78, respectively (Figure 8.3). CpG O/E levels in L3 viruses were significantly lower than for L2 viruses (p = 0.042, two-tailed Mann-Whitney U Test). There was no difference between median CpG O/E values comparing old and new world *Alphavirus* serotypes (0.80 vs 0.81, p = 0.70, two-tailed Mann-Whitney U Test) despite the lineages having been estimated to have diverged over 1000 years ago.[459, 460] Old world serotypes are slightly overrepresented in the higher transmissibility categories, being 33% of L2, 80% of L3, and 100% of L4 among the *Togaviridae*, but the comparison did not reach significance potentially due to the small number of viruses in the family (p = 0.078, Fisher's Exact Test).

As *Flaviviridae* and *Togaviridae* have similar GC content, genome composition, and spread of transmissibility values, the difference in CpG O/E between the viral families in each transmission level was compared. CpG O/E values were significantly higher for *Togaviridae* in L2 and L3 transmission levels (0.83 vs 0.55 and 0.76 vs 0.59, p = $6.16*10^{-10}$ and p = 0.016, respectively, two-tailed Mann-Whitney U Test; Figure 8.3). For L4 viruses, the difference was not significant (p = 0.29, two-tailed Mann-Whitney U Test) but this is likely due to the test being underpowered due to the single L4 *Togaviridae* representative (CHIKV).

8.2.3 Influence of Flavivirus Host Range on CpG Dinucleotide Suppression

The robust trend in transmissibility and CpG suppression within *Flaviviridae* prompted further investigation of the trends of dinucleotide suppression within the family beyond only those documented to be human-infective. Members of the *Flaviviridae* family have broad host ranges and representative viruses have diverse circulation patterns including via mosquitoes and/or ticks to mammals and/or birds, only within vector populations, and only within mammal populations (no vector). To investigate the influence of host and vector range on dinucleotide suppression, a

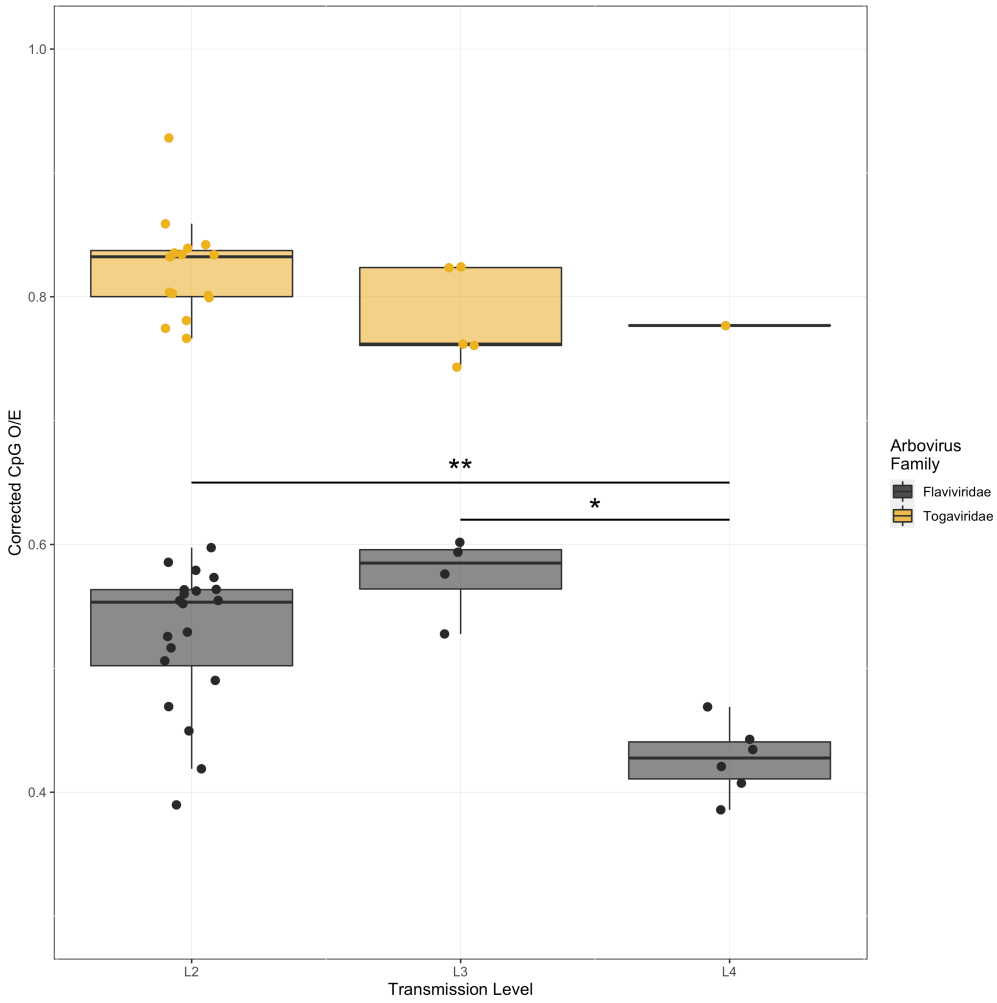


Figure 8.3: CpG Suppression and Transmissibility in *Flaviviridae* and *Togaviridae*. Corrected CpG O/E values versus transmission level in members of the families *Flaviviridae* and *Togaviridae*. Boxplots and dots colored by virus family. * = $p < 0.05$, ** = $p < 0.005$ and from two-tailed Mann-Whitney U Test.

larger database of 88 flaviviruses and their phylogenetic relationships were collected from various review papers and original research (Table 8.3).[286–288] Flaviviruses were classified as: classical insect-specific flaviviruses (cISF; n = 11); divergent insect-specific flaviviruses (dISF; n = 7); tick-borne flaviviruses (TBF; n = 15); mosquito-borne flaviviruses (MBF; n = 38); or no-known-vector flaviviruses (NKVF; n = 17). As their names suggest, cISFs and dISFs circulate only in arthropods and NKVFs, which are not monophyletic, have not had a vector identified.[287] Phylogenetic reconstruction suggests dISFs more recently diverged from MBFs while cISFs diverged prior to the divergence of dual-host flaviviruses.[286]

Table 8.3: Flaviviruses Included and their Phylogenetic Groupings. References for phylogenetic groupings are provided in the "Phylogenetic Group" column. If viruses were identified via other sources than the most common for the group, references are provided in the "Virus Name" column.

Phylogenetic Group	Virus Name
Classical Insect-Specific Flavivirus[286]	Aedes flavivirus Calbertado virus Cell fusing agent virus Culex flavivirus Culex theileri flavivirus Hank virus Kamiti River virus Nakiwogo virus Nienokoue virus Palm Creek virus Quang Binh virus
Divergent Insect-Specific Flavivirus[286]	Barkedji virus Chaoyang virus Donggang virus Ilomantsi virus Lammi virus Nhumirim virus Nounane virus
Tick-Borne Flavivirus[286]	Alkhumra hemorrhagic fever virus Gadgets Gully virus Kadam virus Kama virus Koutango virus[461] Kyasanur forest disease virus[462] Langat virus Louping ill virus Meaban virus Omsk hemorrhagic fever virus Powassan virus Royal Farm virus Suamarex Reef virus Tick-borne encephalitis virus Tyuleniy virus
Mosquito-Borne Flavivirus[286]	Alfuy virus Aroa virus Bagaza virus Banzi virus Bouboui virus Cacipacore virus[463] Dengue virus 1 Dengue virus 2 Dengue virus 3 Dengue virus 4 Edge Hill virus Iguape virus Ilheus virus Japanese encephalitis virus Jugra virus Kedougou virus Kokobera virus Kunjin virus[464] Murray Valley encephalitis virus New Mapoon virus Ntava virus Rocio virus Saboya virus Saint Louis encephalitis virus Sepik virus Spondweni virus Stratford virus Tembusu virus

Continued on next page

Table 8.3 – continued from previous page

Phylogenetic Group	Virus Name
	Torres virus Uganda S virus Usutu virus Wesselsbron virus West Nile virus lineage 1 West Nile virus lineage 2 Yaounde virus Yellow fever virus Zika virus
No-Known-Vector Flavivirus[287]	Apoi virus Batu Cave virus Bukalasa bat virus Carey Island virus Cowbone Ridge virus Dakar bat virus Entebbe bat virus Jutiapa virus Modoc virus Montana myotis leukoencephalitis virus Phnom Penh bat virus Rio Bravo virus Sal Vieja virus San Perlita virus Sokoluk virus Tamana bat virus Yokose virus

Corrected O/E values for each phylogenetic grouping were determined (Figure 8.4A-E). All five groupings demonstrated underrepresentation of the UpA dinucleotide (median O/E ranging from 0.57 to 0.69) and overrepresentation of the CpA dinucleotide (median O/E ranging from 1.17 to 1.37). CpG O/E was the only dinucleotide odds ratio that visually appeared to vary between groupings, although a Kruskal-Wallis Rank Sum Test did not reveal a statistically significant overall association between CpG O/E and phylogenetic group ($p = 0.48$). As CpG suppression is a hallmark of replication in non-vector hosts, it was hypothesized that the relative length of time spent in vector hosts may influence CpG suppression. While both dISF and cISF do not replicate in non-vector hosts, dISF were ranked above cISF due to their more recent divergence from dual-host flaviviruses. Ticks only feed once per life stage, leading to a total of three to four feedings over a two- to three-year lifespan, while mosquitoes can feed multiple times per egg-laying cycle;[465, 466] therefore, TBFs were ranked to spend less time in non-vector hosts than MBFs. By ranking groupings based on proportion of time spent in non-vector hosts (cISF<dISF<TBF<MBF<NKVF) and comparing CpG O/E across adjacent groupings, a stepwise decrease in O/E moving from cISF to NKVF was revealed

with statistically significant differences across each comparison except for TBFs vs MBFs ($p < 0.05$, two-tailed Mann-Whitney U Test; Figure 8.5).

Interestingly, MBFs and NKVFs both demonstrated CpG O/E variance measures 4.2 times and 7.1 times higher than TBFs. NKVFs' variance was likely due to positive skewness from five of their members displaying substantially higher O/E than the rest of the population; the parsimonious explanation is that these viruses represent MBFs or TBFs being misclassified as NKVFs due to insufficient field sampling. The variance in MBFs was investigated as being influenced by the multiple non-vector hosts in which representative viruses from this grouping circulate. Categorizing the MBFs as circulating in ((birds) or (birds and mammals)) vs (mammals only) revealed an association between increased O/E values and replication in bird hosts (Figure 8.6, Table 8.4). MBFs with documented infectivity of birds ($n = 16$) had median corrected CpG O/E values 0.1 points higher than those that infect only mammals (($n = 22$); 0.57 vs 0.47, $p = 0.011$, two-tailed Mann-Whitney U Test).

8.2.4 Shifts in dinucleotide suppression and viral fitness

It was hypothesized that increases in arboviral transmissibility would be mirrored in shifts towards mammalian-optimized (i.e., CpG depleted) genome composition. ZIKV was chosen as the model organism to test this hypothesis for several reasons. As a flavivirus, it is a member of the family that demonstrated the clearest relationship between CpG suppression and transmissibility, the virus was identified in the 1950s but only recently became responsible for a large outbreak of disease, and it is not known to infect birds.[468] In Hay et al. (2013), ZIKV was estimated to have an $R_0 < 1$ based on research conducted in 2010.[469, 470] In the 2014-2016 outbreak in the Americas, estimates ranged from 2.5 to 6.[471, 472] To test if this dramatic increase in transmissibility was associated with a global shift in dinucleotide suppression, 176 ZIKV whole genomes (>10,000 bases) with at most two passages in cell culture were downloaded from Genbank. These viruses were collected from 33 total countries with collection dates ranging from 1984 to 2019

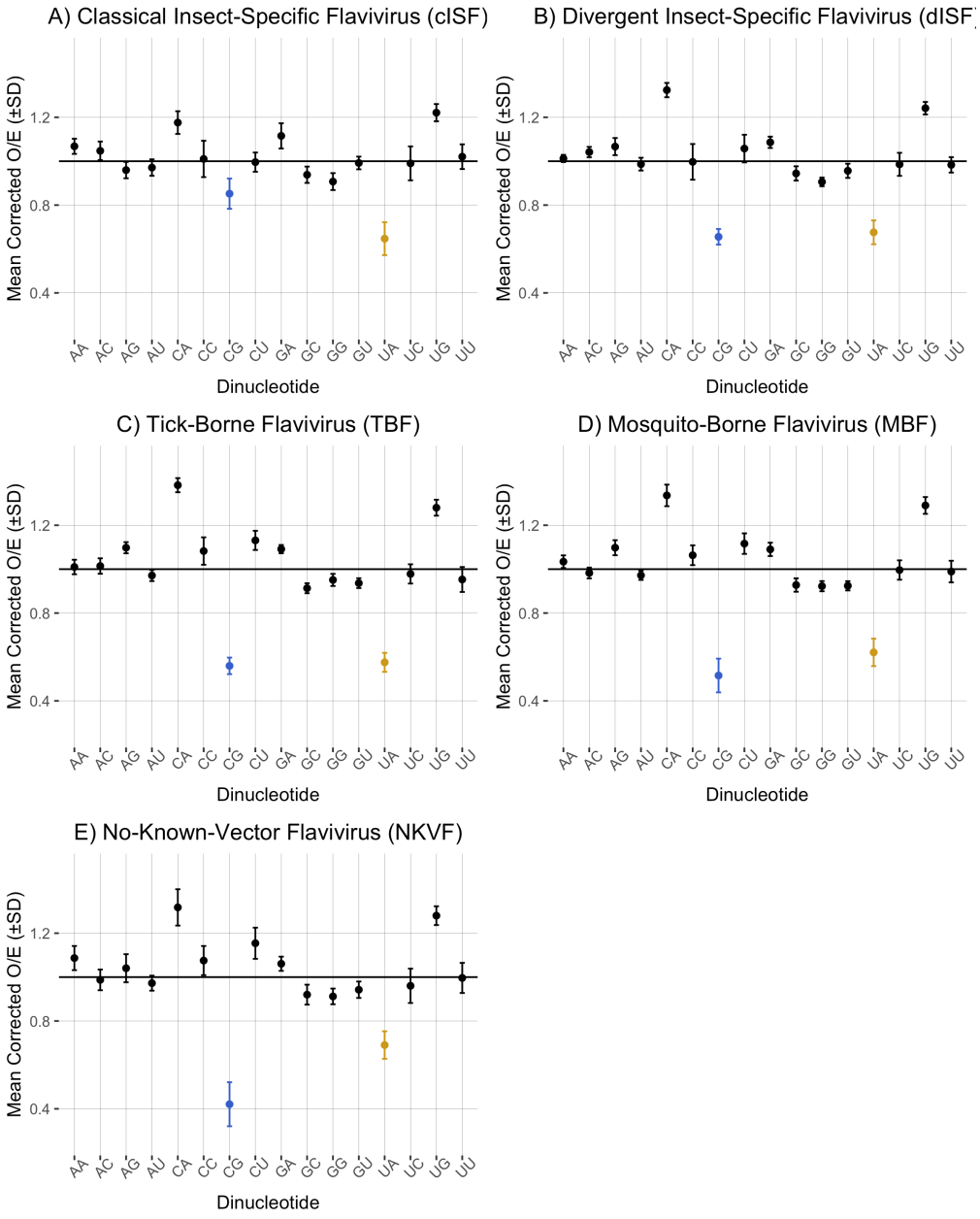


Figure 8.4: Corrected Dinucleotide Odds Ratios for Flavivirus Phylogenetic Groupings. Figure adapted from Lobo et al. (2009).[457] Alterations include use of corrected instead of simple method, larger number of samples included, and inclusion of dISFs as an independent grouping from cISF. Points represent means and error bars signify ± standard deviation. CpG and UpA colored for ease of visualization.

Virus Name	Infection of Birds	Infection of Mammals	Source
Alfuy virus	1	0	US CDC Arbocat
Aroa virus	0	1	Pandit et al. (2018)[288]
Bagaza virus	1	0	Pandit et al. (2018)
Bainyik virus	0	1	May et al. (2013)[467]
Banzi virus	0	1	Pandit et al. (2018)
Bouboui virus	0	1	Pandit et al. (2018)
Cacipacore virus	1	1	Figueiredo et al. (2017)[463]
Dengue virus 1	0	1	Pandit et al. (2018)
Dengue virus 2	0	1	Pandit et al. (2018)
Dengue virus 3	0	1	Pandit et al. (2018)
Dengue virus 4	0	1	Pandit et al. (2018)
Edge Hill virus	0	1	US CDC Arbocat
Iguape virus	0	1	Pandit et al. (2018)
Ilheus virus	1	1	Pandit et al. (2018)
Japanese encephalitis virus	1	1	Pandit et al. (2018)
Jugra virus	0	1	US CDC Arbocat
Kedougou virus	0	1	US CDC Arbocat
Kokobera virus	0	1	US CDC Arbocat
Kunjin virus	1	1	Pandit et al. (2018)
Murray Valley encephalitis virus	1	1	Pandit et al. (2018)
New Mapoon virus	0	1	May et al. (2013)
Ntaya virus	0	1	US CDC Arbocat
Rocio virus	1	1	Pandit et al. (2018)
Saboya virus	1	1	US CDC Arbocat
Saint Louis encephalitis virus	1	1	Pandit et al. (2018)
Sepik virus	0	1	US CDC Arbocat
Spondweni virus	0	1	US CDC Arbocat
Stratford virus	0	1	US CDC Arbocat
Tembusu virus	1	0	Pandit et al. (2018)
Torres virus	0	1	May et al. (2013)
Uganda S virus	1	1	Pandit et al. (2018)
Usutu virus	1	1	Pandit et al. (2018)
Wesselsbron virus	0	1	Pandit et al. (2018)
West Nile virus lineage 1	1	1	Pandit et al. (2018)
West Nile virus lineage 2	1	1	Pandit et al. (2018)
Yaounde virus	1	1	US CDC Arbocat
Yellow fever virus	0	1	Pandit et al. (2018)
Zika virus	0	1	Pandit et al. (2018)

Table 8.4: Host Range of Mosquito-Borne Flaviviruses. Data used for analysis presented in Figure 8.6. Sources chosen by priority of primary literature > secondary literature.

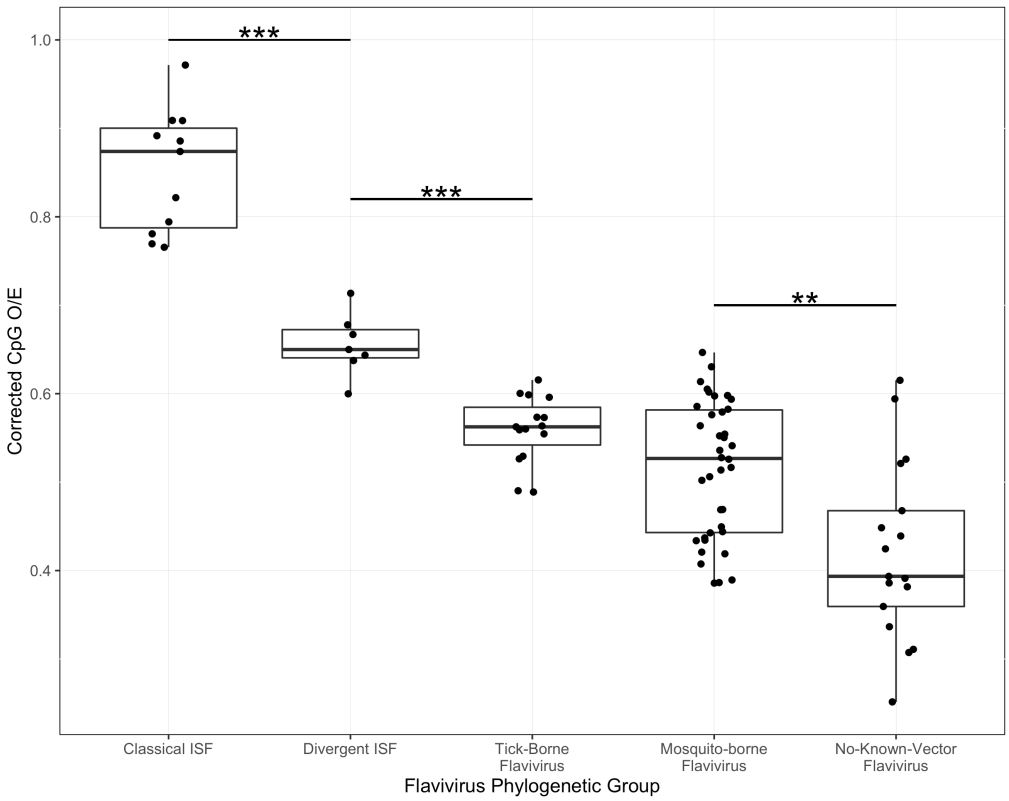


Figure 8.5: Corrected CpG O/E in Flavivirus Phylogenetic Groups. * = p < 0.05, ** = p < 0.005,*** = p <0.0005 and from two-tailed Mann-Whitney U Test.

with only 10 being collected before 2014; their year and country of collection and GenBank accession number are provided in table 10.3.

Corrected CpG O/E values for ZIKV varied from a minimum of 0.4199 (collected in 1984 in Senegal) to a maximum of 0.4806 (collected in 2018 in Cuba) with a mean value of 0.47 and standard deviation of 0.0077. In general, year of collection did not correlate with CpG O/E value (τ_b = 0.053, p = 0.36, Kendall's Rank Correlation). Unexpectantly, median CpG O/E values from samples collected prior to 2014 were significantly lower than those collected during or after 2014 (0.45 vs 0.47, p = 3.21*10^{-5}, two-tailed Mann-Whitney U Test). In this instance, shifts in transmissibility were clearly not associated with a decrease in CpG content.

Sylvatic DENV circulates in non-human primates and is ecologically and evolutionary distinct from non-sylvatic (urban) DENV lineages that circulate in human populations.[473] The four serotypes of DENV (used in this study) represent four independent spillovers of sylvatic DENV into human populations

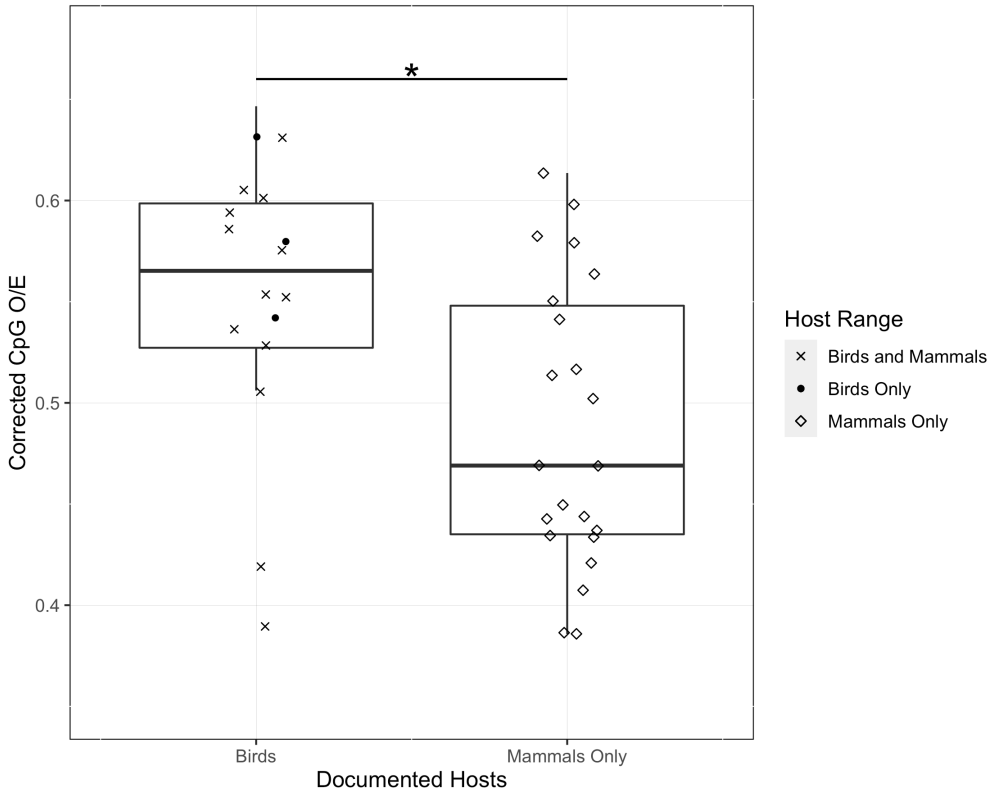


Figure 8.6: Influence of Infection of Birds with Corrected CpG Values in Mosquito-Borne Flaviviruses. Symbol indicates host range. Due to limited numbers of "bird only" MBFs (n = 3), comparison made categorizing viruses passed on evidence of infection of birds. Of note, median CpG O/E value of "bird only" MBFs is higher than "bird and mammals" MBFs (0.60 vs 0.55). * = $p < 0.05$ and from two-tailed Mann-Whitney U Test.

that successfully persisted in human populations. To investigate if CpG suppression creates a barrier to sylvatic DENV spillover, 16 sylvatic DENV and 337 urban DENV genomes were downloaded from GenBank. Accession numbers are provided in table 10.4. Median CpG O/E values were not significantly different between sylvatic and urban DENV, suggestive that genome composition is not a barrier to human-human transmissibility of sylvatic DENV (0.446 vs 0.445, $p = 0.225$, two-tailed Mann-Whitney U Test).

8.3 Discussion

Understanding the virological traits that *determine* rather than *correlate* with virus zoonotic risk would allow for far more robust risk assessment of novel RNA viruses

based on sequence data and minimal phenotypic characterization. The results presented here suggest that CpG O/E correlates strongly with transmissibility within the *Flaviviridae* family but its utility in other arboviral families may be low. CpG O/E further strongly correlates with host range in the *Flaviviridae* family. In the case of ISFs, it may even determine infectivity as viruses from these classifications contain CpG levels that may be completely incompatible with vertebrate replication. Bolstering this claim, wild-type A549 cells (human lung carcinoma epithelial cells) have been shown to be completely resistant to infection with dISFs; however, knocking out ZAP expression made them permissive to infection with live virus at 34°C.[474] For the more distantly related cISFs, inhibition of replication occurs at several steps pre- and post-entry.[475, 476] Experiments using mutant infectious clones of flaviviruses with varied CpG content would further refine this phenotype and its impact on transmissibility, potentially including competition assays using mixed populations of viruses in blood meals and transmission studies from mosquitoes to mice or hamsters. However, the higher mosquito infectivity seen for CpG-high mutants of ZIKV may complicate these experiments.[458]

Other virological traits have been identified as correlating or determining transmissibility, although the majority have little applicability to human-infective arboviruses. Being nonenveloped has repeatedly been shown to increase transmissibility,[438, 441, 445] but *Flaviviridae*, *Peribunyaviridae*, and *Togaviridae*, which collectively contain 86% of all L3 and L4 arboviruses, are enveloped. One of the strongest determinants of cross-species rabies virus transmission is the similarity between body temperature of the index and novel hosts;[442] arboviruses already withstand the ambient temperatures of their arthropod hosts so this is unlikely to apply.[477] Having a dsRNA genome has also been shown to predict transmissibility;[445] in human-infective arboviruses, only 7.5% (7/93) have dsRNA genomes (members of the family *Reoviridae*), of which only one transmits in humans. In summary, prediction of zoonotic risks based on virological traits is likely to be virus-family specific and generalizable heuristics may not exist.

Approaches to predicting zoonotic risk based on virological traits have been fairly criticized.[478, 479] The chief complaints generally center around inadequate and biased data underlying the development of these statistical models and the futility of attempts to catalog an ever shifting population of viruses. These concerns certainly apply to this project but are mitigated for several reasons. First and foremost, rather than performing multivariable regression on traits with and without clear biological mechanisms (as performed elsewhere[451]), this analysis is based on a single continuous variable that has a reproducible impact on virus fitness *in vitro* and *in vivo*. The sequence data used in the calculation of O/E are based on high coverage, RefSeq-curated sequences which have high quality control standards.[480] The use of CpG O/E values based on the corrected method further improves the reliability of these estimates.

The higher risk for inaccurate data in this project is the R_0 estimates used by Woolhouse and Brierley (2018) to determine transmission levels. In their dataset, these values were sourced from a tour de force review of infectious diseases published by Hay et al. in 2013.[469] That review itself includes many references that may have since come out of date (e.g., the reference for Oropouche virus transmissibility was published in 1981).[481] An updated version of these estimates would likely require new epidemiological studies but may update the transmission level categorizations used for the analysis in figure 8.3. Similarly, the catalog of human-infective arboviruses is certainly incomplete as demonstrated by the recently published evidence of human infection by both the Dhori virus and the Bourbon virus of the *Orthomyoxiviridae* family.[482, 483] A recently published seroprevalence study in Missouri, USA found Bourbon virus neutralizing antibodies in almost 1% of a random sample of human sera collected in 2020.[484] The intent of this project was not to review, identify, and catalog newly discovered arboviruses, so the database underlying the analyses was not updated. Further, it is unlikely the conclusions of this study will be impacted from the discovery of new human-infective arboviruses due to the high likelihood that newly discovered arboviruses will be

designated L2 (as noted in Lu et al. (2019)[441]), which would have minimal statistical impact on the group trends.

For a virus to infect a new host, it must successfully overcome various intracellular and extracellular barriers to generate progeny virions. Failure to do so results in the new host being a dead-end, such as what is seen for L2 arboviruses infecting humans. Evidence of infection is indicative of a virus' ability to survive within a subset of that host's cells. In contrast, the ability to transmit within novel host populations is dependent on a much wider range of factors including cell and tissue tropism, environmental factors (e.g., presence of vector in geographic area, shifting climate), and socioeconomic factors.[485] CpG and UpA suppression, which are exclusively due to intracellular selective pressures, should therefore manifest in a virus' ability to successfully infect and replicate within a host and not itself be a determinant of transmissibility but rather a factor that influences the degree of transmissibility. For human-human transmission of arboviruses, release of infectious virions into the human bloodstream is essential for infectivity of subsequent arthropod hosts. Arboviruses that fail to deposit virions into the bloodstream (viremia) will not be able to transmit, as may be the case for the L2 viruses. Further, scale of viremia is likely a strong determinant of arthropod infectivity and thus of human-human transmissibility, with viruses causing higher viral loads likely being more easily transmitted.

The working model developed here is that transmissibility is a trait that can be gained (or lost) and dinucleotide suppression is a proxy value for host species fitness. Here, the highest likelihood trait is presence or absence of viremia. Given this model, L2 viruses from virus families shown to be sensitive to ZAP (e.g., *Flaviviridae*) with genome composition currently optimized for replication in mammals should be prioritized for surveillance over L3 viruses from the same family with poorly-optimized genome composition. For example, Edge Hill Virus (EHV) is a flavivirus with a corrected CpG O/E value of 0.39 (lower than DENV1-4 and on par with Yellow Fever virus (YFV)). EHV was first discovered in Australia in the 1960s and contains antigenic and sequence homology to YFV and other viruses in the YFV

subgroup.[486] The virus primarily infects marsupials (based on serological data) and has only once been suspected of causing clinical disease in a human.[487] Lu et al. (2019) found that vector-borne viruses were overrepresented in L2->L3/L4 transitions versus L1->L3/L4, emphasizing that known human-infective arboviruses are likely more key for surveillance of emerging pathogens than those without documented infection.[441] The association with phylogenetic grouping and effects of host range further refine this model; MBFs that don't circulate in birds are likely the prime virus population to survey. Of note, the effect of phylogenetic grouping on CpG suppression was previously reported by Lobo et al. (2009), but with a more limited dataset (28 viruses vs 88) and more crude phylogeny (that publication did not distinguish dISFs from cISFs).[457]

The factors leading to increased transmissibility in established transmissible arboviruses are unclear and likely both multifactorial and context-specific. The data presented here suggests that the increase in transmissibility observed for ZIKV during the 2014-2016 outbreak was not due to shifts in CpG frequency. Further, the barrier between sylvatic and urban DENV is not due to differences in CpG suppression. This result is not surprising given the presence of ZAP in most mammalian species.[488] One of the more dramatic shifts in arbovirus transmissibility in recent history involved the CHIKV outbreak in the Indian Ocean during 2005-2006, particularly on the French overseas department of La Réunion. Estimates range from one-sixth to over one-third of the island being infected during the seventh months of transmission; an active case-finding system detected 10 times as many clinical CHIKV cases during the outbreak as compared to the same time period the year before.[489, 490] Perhaps more concerningly from a public health standpoint, this was the first CHIKV outbreak that resulted in fatalities. *Aedes aegypti*, the vector often implicated in outbreaks of CHIKV, had been virtually eliminated from the island in the 1950s, and the outbreak was associated with a vector change to the much more common *Aedes albopictus*.[491] In an elegant study, Tsetsarkin et al. (2007) clearly implicated the E1:A226V mutation as dramatically increasing vector competence of *Aedes albopictus* for CHIKV.[492] A single point

mutation was responsible for converting CHIKV from a nuisance to a public health strain on the island of La Réunion. This event emphasizes that transmissibility is a product of many factors and that measuring CpG suppression alone is not sufficient for understanding shits in zoonotic risk.

That two arbovirus families (*Togaviridae* and *Reoviridae*) seemingly escape the intense selective pressure against CpG raises the question as to their avoidance mechanisms. As introduced in Chapter One, the current model of ZAP restriction of CpG-high RNA requires binding to the target and sequestration away from ribosomes (potentially to stress granules). The dsRNA genomes of members of the *Reoviridae* family may shield CpGs from being targeted by ZAP; dsRNA is known to not be a target for ZAP binding and genomic ssRNA from members of *Reoviridae* has been shown to be insulated from the cytoplasm within viral inclusions.[493, 494] The escape for members of the *Togaviridae* family is likely to be more complicated. *In vitro* studies of Sindbis Virus and Semliki Forest Virus have demonstrated that alphaviruses are sensitive to ZAP activity through a mechanism that prevents the accumulation of genomic viral RNA in the cytoplasm (after virus uncoating) and that the degree of sensitivity varies between alphaviruses and cell types.[456, 495] OAS3, which has been implicated in the ZAP pathway, has also been shown to have antiviral activity against CHIKV by blocking viral RNA accumulation and inhibiting viral protein synthesis.[91, 496] This suppression was independent of RNAse L. Still, the analysis presented in figure 8.3 suggest at least partial escape from the suppression seen in other human transmissible arboviruses. The macrodomain present in nsp3 in Alphaviruses may be part of a mechanism of escape. The ADP-ribosylhydrolase activity of this protein in CHIKV has been shown to be crucial for replication in golden hamster kidney cells and *Aedes albopictus* cells.[497] Additionally, the ADP-ribosylhydrolase activity was recently shown to disrupt the formation of stress granules, which may be a key component of the ZAP restriction pathway.[498] However, macrodomains are conserved across many RNA viruses (including those with low CpG O/E levels, such as *Coronaviridae*), so its activity may only be part of the inhibition pathway.[499] Identification of the key viral

proteins and development of inhibitory drugs may be a highly effective antiviral therapy for alphaviral disease; development in this area should be a priority.

Lastly, the analyses presented here are themselves somewhat inappropriate. All the deployed statistical tests assume independence between samples, but this is not an appropriate assumption for tests of within-family samples due to phylogenetic dependence.[500] Further, as introduced in Chapter One, the criteria used to define a species, which are the ultimate statistical object being used in these analyses, are not consistent between virus families. In instances where a greater number of closely related viruses are delineated into separate species in one family than in another (i.e., the antigenically-related *Alphaviruses* have been classically defined by cross-reactivity in immunoassays and not by phylogeny),[501] the estimate for group characteristics for the more taxonomically-refined family may present artefactually high levels of precision due to an inflated sample size. A salient example impacting this analysis is that Woolhouse and Brierley (2018) considers the four serotypes of DENV and two serotypes of WNV as independent species. This likely inflates the effect of CpG suppression on transmissibility for *Flaviviridae* reported in figure 8.3. The observed statistically significant difference is ablated when WNV and DENV are limited to single entries (this effect is mainly a consequence of the reduced sample size; two-tailed Mann-Whitney Tests employed on two numeric vectors each containing only three values cannot return a p value < 0.1 regardless of differences in magnitude). General discussion on the inconsistencies in designating virus species between viral families is available in Simmonds and Aiewsakun (2018).[39] In general, any approach to zoonotic risk prediction that uses species or family characteristics (e.g., as in Grange et al. (2021)[451]) will be susceptible to these biases.

The interplay between CpG suppression and arbovirus host species optimization allows for general predictions of host range for specific arbovirus families from a sequence alone. Caution should be applied when extrapolating these results to other virus families without the conflicting evolutionary pressures of arthropod and non-arthropod hosts. Further functional research is warranted and would benefit greatly from mutant clones of members of the *Flaviviridae* and *Togaviridae* families.

9
General Discussion

Zoonotic viruses will continue to enter the human population, leading to morbidity, mortality and, potentially, to a pandemic similar to those caused by HIV and SARS-CoV-2. In fact, recent modeling has suggested that climate change, through novel interactions between viruses and hosts as geographically-isolated species come into contact, may increase the frequency of spillover events.[502] To provide further evidence of this risk, the largest outbreak outside of Africa of the zoonotic pathogen Monkeypox, a member of the *Poxviridae* family, had just begun while this thesis was being completed. The lessons learned described here develop a framework for rapid characterization of novel viral pathogens. From collection of clinical samples from patients at various stages of disease and the application of relatively standard diagnostic virology techniques, an understanding of the intra-host dynamics, temporal risk of transmission, potential evolutionary pressures, and pathological mechanisms can be obtained. The utility of the approaches described here are dependent on the prior identification of the pathogen (through pathogen-agnostic methods such as metagenomic sequencing), dissemination of high-quality sequences through platforms such as GISAID and GenBank (enabling design of robust PCR-based assays), availability of diagnostics to identify infected individuals, and systems through which to rapidly consent and recruit patients into clinical trials.

The COVID-19 results Chapters (Four through Seven) build on the subsequent work of the previous chapter, motivating their order. The ASP-PCR and PCR-RFLP methods were essential for identifying the infecting variant for a large proportion of patients, particularly given the low SARS-CoV-2 viral loads in the REMAP-CAP cohort and fragmented RNA in COV003. The use of this method is baked into the results presented in Chapters Five, Six, and Seven; any comparison that includes lineage of infection was dependent on the development, validation, and use of those methods. The remarkably diverse viral loads in REMAP-CAP motivated specific interrogation of the role of SARS-CoV-2 virological factors in observed COVID-19 morbidity and mortality. Lastly, the small cohort of intensively sampled COVID-19 patients for whom SARS-CoV-2 was persistently detected inspired sequencing of these samples to detect any acquired SNPs explaining their failure to be cleared. The finding of a limited role for SARS-CoV-2 virological factors or evolution in patient morbidity and mortality is informative for clinical trial design in hospitalized patients; recognition of this phenomenon earlier in the COVID-19 pandemic could have focused trials on corticosteroids (such as dexamethasone) rather than antivirals (such as remdesivir and CP) and saved lives. Importantly, the immunopathological dominance may not hold for novel human-infective RNA viruses. In these future outbreaks of novel pathogens, assessing the contribution of virological factors, particularly relative viral load, in intensely sick individuals must be a research priority. Designing trials in such a way to enable virological assessment, including the collection of potentially relevant metadata such as date of symptom onset, would significantly facilitate this research.

The inception of the SARS-CoV-2 pandemic and the research focus of Chapter Eight are both the spillover of zoonotic RNA viruses. While the bioinformatic research on arboviruses is a project conceived and started prior to COVID-19, it provides a useful framework for understanding the early evolutionary pressures of SARS-CoV-2. Arboviruses and *Coronaviridae* have starkly different host ranges, intracellular selection pressures, cell tropisms, and life cycles so the commonalities between the two classifications are limited. Nevertheless, dinucleotide selection

helped poise SARS-CoV-2 for efficient replication in humans; both SARS-CoV and SARS-CoV-2 exhibit suppression of CpG to a stronger extent than other bat- and human-infecting coronaviruses.[503] That the observed C>U mutational pattern appears to select against the UpA and CpG dinucleotides further supports the role of this mutational pressure on the short-term adaptation of the virus to human hosts. Chapter Eight ends with a call to incorporate considerations of genome composition when assessing virus zoonotic potential. While this may have limited value in zoonotic viruses circulating in mammalian populations, it may reveal currently unrecognized influences of host range on these measures.

Looking to the future, there are several specific innovations from this thesis that can be carried into the field. The ASP-PCR method described in Chapter Four is not specific to SARS-CoV-2 and can be leveraged in any instance where typing a point mutation is informative (such as HIV genotyping, as mentioned in Chapter Four). While the method currently displays markedly improved sensitivity compared to NGS, improvements on the limit-of-detection, portability, and cost-effectiveness of NGS-based methods will be far more useful for general pandemic preparedness and should be a research priority. For ASP-PCR, further development of oligonucleotide modifications would expand the use cases for this technology but will not overcome the need for a large number of high-quality sequences from which to design the assay. The sgRNA RT-qPCR is novel and can be used to assess sgRNA production *in vitro* and *in vivo* for any (+)ssRNA virus that produces sgRNA through discontinuous transcription (which attaches a common leader sequence to the 5' end of the transcript). While primarily described in families within the *Nidovirales* order (such as *Coronaviridae* and *Arteriviridae*), the transcription method is also present in the plant-infecting family *Tombusviridae*.[504, 505] The relative importance (if any) of the altered transcription ratios between SARS-CoV-2 variants is still unclear; expanding this analysis to include hospitalized, Omicron-infected patients would help narrow down the mutations driving this phenotype but that patient population was unfortunately not available for the work presented here. Lastly, the statistical approaches to comparing evolution between groups and assessing

homoplastic mutations in Chapter Seven can be used in other studies attempting to directly compare substitution rates and emergent SNPs between two populations.

The COVID-19 pandemic has demonstrated the failures in the current pandemic preparedness system. As society charts a new path forward, the technological advancements of the past two years coupled with the specific lessons learned from SARS-CoV-2 will lay the foundation of a new era in pandemic preparedness. This research is part of that foundation.

10 Appendix

Contents

10.1 Supplementary Tables 180
10.5 References 192

10.1 Supplementary Tables

Table 10.1: Global Performance of Lineage-Defining SNPs. VoC and VoI sublineages included as their parent lineages. Analysis includes sequences deposited up until 21/01/2022. All global sequences totaled 7,040,262.

[a]Calculated as: Number of lineage with SNP / Total number of lineage
[b]Calculated as: Number of non-lineage without SNP / Total number of non-lineage
[c]Calculated as: Number of lineage with SNP / Total number with SNP
[d]Calculated as: Number of non-lineage without SNP / Total number without SNP

Variant	SNP	PPA[a]	NPA[b]	PPV[c]	NPV[d]
Alpha (n = 1,146,701)	ORF1a:T1001I	99.40%	99.74%	98.65%	99.88%
	ORF1a:A1708D	99.07%	99.84%	99.19%	99.82%
	ORF1a:del3675/3677	96.69%	95.47%	80.60%	99.33%
	ORF1b:P314L	99.03%	1.97%	16.43%	91.26%
	S:del69/70	96.65%	91.81%	69.67%	99.29%
	S:A570D	99.22%	99.87%	99.32%	99.85%
	S:P681H	99.04%	90.86%	67.83%	99.79%
	S:S982A	98.78%	99.87%	99.33%	99.76%
	S:D1118H	99.18%	99.84%	99.20%	99.84%
	N:S235F	98.67%	99.81%	99.01%	99.74%
Gamma (n = 120,458)	ORF1a:K1795Q	98.38%	99.99%	99.17%	99.97%
	ORF1b:P314L	98.69%	1.82%	1.72%	98.76%
	ORF1b:E1264D	98.02%	99.75%	87.30%	99.97%

Continued on next page

Table 10.1 – continued from previous page

Variant	SNP	PPA[a]	NPA[b]	PPV[c]	NPV[d]
	S:L18F	97.37%	98.30%	49.91%	99.95%
	S:P26S	97.12%	99.77%	88.09%	99.95%
	S:K417T	95.07%	100.00%	99.79%	99.91%
	S:H655Y	98.02%	93.42%	20.59%	99.96%
	S:V1176F	97.60%	99.70%	84.94%	99.96%
	ORF8:E92K	97.24%	99.96%	97.44%	99.95%
	N:P80R	98.21%	99.99%	99.27%	99.97%
Zeta (n = 5,497)	ORF1a:L3468V	96.51%	100.00%	97.66%	100.00%
	ORF1a: L3930F	99.44%	99.79%	26.84%	100.00%
	ORF1b:P314L	98.49%	1.81%	0.08%	99.93%
	S:E484K	95.85%	96.68%	2.21%	100.00%
	S:D614G	99.45%	1.20%	0.08%	99.96%
	S:V1176F	98.64%	98.11%	3.92%	100.00%
	N:A119S	99.36%	99.84%	32.87%	100.00%
	N:R203K	98.53%	72.01%	0.27%	100.00%
	N:G204R	98.76%	73.17%	0.29%	100.00%
	N:M234I	98.65%	97.73%	3.28%	100.00%
Beta (n = 40,593)	ORF1a:T265I	97.35%	94.73%	9.68%	99.98%
	ORF1a:K1655N	92.94%	99.85%	78.53%	99.96%
	ORF1a:K3353R	97.10%	98.51%	27.37%	99.98%
	S:D80A	97.13%	99.97%	95.60%	99.98%
	S:D215G	94.63%	99.95%	92.11%	99.97%
	S:K417N	92.91%	97.18%	16.05%	99.96%
	S:A701V	96.40%	98.55%	27.77%	99.98%
	S:D614G	97.77%	1.20%	0.57%	98.93%
	E:P71L	96.85%	99.94%	90.55%	99.98%
	N:T205I	96.01%	98.01%	21.88%	99.98%
Delta (n = 4,044,563)	ORF1b:G662S	97.52%	99.19%	99.39%	96.74%
	ORF1b:P1000L	98.87%	99.06%	99.30%	98.48%
	S:T19R	98.37%	99.62%	99.72%	97.84%
	S:L452R	97.11%	95.99%	97.03%	96.09%
	S:P681R	99.23%	98.67%	99.02%	98.96%
	ORF3a:S26L	99.27%	98.96%	99.23%	99.01%
	M:I82T	98.95%	98.31%	98.75%	98.58%
	ORF7a:T120I	95.29%	99.61%	99.70%	93.99%
	N:R203M	98.65%	99.07%	99.31%	98.20%
	N:D377Y	97.88%	98.14%	98.61%	97.17%
Lambda (C.37) (n = 9,629)	ORF1a:T1246I	98.27%	99.76%	36.18%	100.00%
	ORF1a:P2287S	97.01%	49.55%	0.26%	99.99%
	ORF1a:F2387V	98.77%	100.00%	99.20%	100.00%
	ORF1a:L3201P	98.84%	99.15%	13.74%	100.00%
	ORF1a:T3255I	99.04%	41.21%	0.23%	100.00%
	S:G75V	94.61%	99.88%	51.26%	99.99%
	S:T76I	97.25%	99.94%	70.42%	100.00%
	S:L452Q	97.96%	99.70%	88.77%	99.99%
	S:F490S	97.10%	99.91%	58.80%	100.00%
	N:G214C	97.86%	100.00%	96.76%	100.00%
Mu (B.1.621) (n = 15,084)	ORF1a:T1055A	96.98%	99.99%	93.84%	99.99%
	ORF1a:T1538I	96.67%	99.97%	85.81%	99.99%
	ORF1a:T3255I	96.66%	41.24%	0.35%	99.98%
	ORF1a:Q3729R	95.95%	99.90%	66.49%	99.99%
	ORF1b:P1342S	96.10%	99.97%	86.99%	99.99%
	S:R346K	96.26%	98.39%	11.41%	99.99%
	ORF3a:Q57H	95.92%	92.63%	2.72%	99.99%
	ORF8:T11K	96.50%	99.98%	92.32%	99.99%
	ORF8:P38S	96.35%	99.91%	69.86%	99.99%
	N:T205I	95.88%	97.67%	8.12%	99.99%
R.1 (n = 11,047)	ORF1b:P314L	99.78%	1.81%	0.16%	99.98%
	ORF1b:G1362R	99.76%	99.98%	89.97%	100.00%
	ORF1b:P1936H	99.60%	99.99%	91.77%	100.00%
	S:W152R	99.25%	99.83%	47.34%	100.00%
	S:G769V	98.79%	99.88%	56.93%	100.00%
	M:F28L	99.68%	99.87%	54.36%	100.00%
	N:S187L	99.96%	99.93%	69.79%	100.00%
	N:R203K	99.88%	72.07%	0.56%	100.00%
	N:G204R	99.98%	73.23%	0.58%	100.00%

Continued on next page

Table 10.1 – continued from previous page

Variant	SNP	PPA[a]	NPA[b]	PPV[c]	NPV[d]
	N:Q418H	99.73%	99.88%	55.84%	100.00%
B.1.526 (n = 42,901)	ORF1a:T265I	98.93%	94.77%	10.39%	99.99%
	ORF1a:L3201P	98.70%	99.62%	61.15%	99.99%
	ORF1b:P314L	98.58%	1.81%	0.61%	99.52%
	ORF1b:Q1011H	96.02%	99.98%	96.01%	99.98%
	S:L5F	96.18%	98.26%	25.31%	99.98%
	S:T95I	98.37%	71.11%	2.05%	99.99%
	S:D253G	97.78%	99.90%	86.25%	99.99%
	ORF3a:P42L	97.69%	99.61%	60.68%	99.99%
	ORF3a:Q57H	98.86%	93.00%	7.97%	99.99%
	ORF8:T11I	97.11%	99.74%	69.73%	99.98%
B.1.427/B.1.429 (n = 69,022)	ORF1a:T265I	99.59%	95.13%	16.83%	100.00%
	ORF1b:P314L	95.11%	1.78%	0.95%	97.35%
	ORF1b:D1183Y	98.46%	99.95%	94.83%	99.98%
	S:S13I	94.69%	99.98%	97.95%	99.95%
	S:W152C	94.71%	100.00%	99.48%	99.95%
	S:L452R	96.71%	42.89%	1.65%	99.92%
	ORF3a:Q57H	99.54%	93.35%	12.91%	100.00%
	N:T205I	98.13%	98.42%	38.03%	99.98%
Omicron (n = 447,878)	ORF1a:K856R	96.14%	99.97%	99.61%	99.74%
	ORF1a:T3255I	98.92%	43.88%	10.69%	99.83%
	ORF1a:P3395H	98.66%	99.98%	99.74%	99.91%
	ORF1a:I3758V	97.37%	99.96%	99.33%	99.82%
	S:A67V	96.35%	99.64%	94.79%	99.75%
	S:Q954H	97.63%	99.99%	99.85%	99.84%
	S:N969K	97.97%	99.99%	99.82%	99.86%
	E:T9I	98.70%	99.90%	98.53%	99.91%
	M:Q19E	95.63%	99.99%	99.90%	99.70%
	M:A63T	95.67%	99.99%	99.79%	99.71%
BA.1 (n = 436,781)	ORF1a:K856R	98.08%	99.94%	99.09%	99.87%
	ORF1a:S2083I	97.57%	99.93%	98.91%	99.84%
	S:A67V	98.26%	99.61%	94.27%	99.88%
	S:T547K	99.10%	99.95%	99.22%	99.94%
	S:N856K	97.84%	99.95%	99.30%	99.86%
	S:L981F	97.97%	99.95%	99.26%	99.87%
BA.2 (n = 8,070)	ORF1a:T842I	98.98%	99.98%	83.63%	100.00%
	ORF1a:L3027F	98.46%	99.87%	45.84%	100.00%
	ORF1b:R1315C	99.44%	99.83%	40.32%	100.00%
	ORF1b:T2163I	97.96%	99.84%	41.74%	100.00%
	S:S371F	95.81%	99.99%	95.43%	100.00%
	ORF6:D61L	99.05%	99.86%	44.75%	100.00%

Table 10.2: Accession Numbers for Viruses Included in Chapter 8. List of accession numbers for arboviruses included in analyses presented in chapter 8, exclusive of section 8.2.4. Segmented viruses have multiple accession numbers included within one row.

Virus Name	GenBank Accession Number
Aedes flavivirus	NC_012932
Alagoas virus	NC_025353
Alfuy virus	AY898809
Alkhumra hemorrhagic fever virus	KU884971
Apoi virus	NC_003676
Aroa virus	NC_009026
Bagaza virus	NC_012534
Bainyik virus	KM225264
Banna virus	NC_004198 NC_004200 NC_004201 NC_004202 NC_004203 NC_004204 NC_004211 NC_004217 NC_004218 NC_004219 NC_004220 NC_004221
Banzi virus	NC_043110
Barkedji virus	LC497470
Barmah Forest virus	NC_001786
Batu Cave virus	KJ469370
Bouboui virus	NC_033693
Bukalasa bat virus	NC_043111
Bunyamwera virus	NC_001925 NC_001926 NC_001927
Bwamba virus	NC_034479 NC_034480 NC_034490
Cacipacore virus	NC_026623
Calbertado virus	KX669688
California encephalitis virus	KX817312 KX817313 KX817314
Candiru virus	NC_015373 NC_015374 NC_015375
Caraparu virus	NC_034477 NC_034478 NC_034489
Carey Island virus	NC_043112
Catu virus	NC_038726 NC_038727 NC_038728
Cell fusing agent virus	NC_001564
Chandipura virus	NC_020805
Chaoyang virus	NC_017086
Chikungunya virus	NC_004162
Cocal virus	NC_028255
Colorado Tick Fever virus	NC_004180 NC_004181 NC_004182 NC_004183 NC_004184 NC_004185 NC_004186

Continued on next page

Table 10.2 – continued from previous page

Virus Name	Accession Number
	NC_004187 NC_004188 NC_004189 NC_004190 NC_004191
Corriparta virus	NC_038564 NC_038565 NC_038566 NC_038567 NC_038568 NC_038569 NC_038570 NC_038571 NC_038572 NC_038573
Cowbone Ridge virus	NC_043113
Crimean-Congo hemorrhagic fever virus	KJ682797 KJ682813 KJ682822
Culex flavivirus	NC_008604
Culex theileri flavivirus	NC_040682
Dabie bandavirus	NC_043450
Dabie bandavirus	NC_043451
Dabie bandavirus	NC_043452
Dakar bat virus	NC_043114
Dengue virus 1	NC_001477
Dengue virus 2	NC_001474
Dengue virus 3	NC_001475
Dengue virus 4	NC_002640
Donggang virus	NC_016997
Dugbe virus	NC_004157 NC_004158 NC_004159
Eastern equine encephalitis virus	NC_003899
Edge Hill virus	NC_030289
Entebbe bat virus	NC_008718
Everglades virus	NC_038671
Eyach virus	NC_003696 NC_003697 NC_003698 NC_003699 NC_003700 NC_003701 NC_003702 NC_003703 NC_003704 NC_003705 NC_003706 NC_003707
Gadgets Gully virus	NC_033723
Getah virus	NC_006558
Great Island virus	NC_014522 NC_014523 NC_014524 NC_014525 NC_014526 NC_014527 NC_014528 NC_014529 NC_014530 NC_014531
Guama virus	NC_038735 NC_038736 NC_038737
	NC_034486
Guaroa virus	Continued on next page

Table 10.2 – continued from previous page

Virus Name	Accession Number
	NC_034487 NC_034506
Hank virus	NC_030401
Highlands J virus	NC_012561
Iguape virus	MK332102
Ilheus virus	NC_009028
Ilomantsi virus	NC_024805
Indiana virus	NC_038236
Isfahan virus	NC_020806
Japanese encephalitis virus	NC_001437
Jugra virus	NC_033699
Jutiapa virus	NC_026620
Kadam virus	NC_033724
Kairi virus	NC_038738 NC_038739 NC_038740
Kama virus	NC_023439
Kamiti River virus	NC_005064
Kedougou virus	NC_012533
Kokobera virus	NC_009029
Koutango virus	NC_043116
Kunjin virus	AY274504
Kyasanur forest disease virus	NC_039218
Lammi virus	NC_024806
Langat virus	NC_003690
Lebombo virus	NC_038594 NC_038595 NC_038596 NC_038597 NC_038598 NC_038599 NC_038600 NC_038601 NC_038602 NC_038603
Louping ill virus	NC_001809
Madrid virus	NC_034497 NC_034498 NC_034505
Maraba virus	NC_025255
Marituba virus	NC_034495 NC_034496 NC_034504
Mayaro virus	NC_003417
Meaban virus	NC_033721
Modoc virus	NC_003635
Montana myotis leukoencephalitis virus	NC_004119
Mosso des Pedras virus	NC_038857
Mucambo virus	NC_038672
Murray Valley encephalitis virus	NC_000943
Nakiwogo virus	NC_030400
Naples virus	HM566176 HM566177 HM566178
Ndumu virus	NC_016959
New Jersey virus	NC_024473
New Mapoon virus	NC_032088
Nhumirim virus	NC_024017
Nienokoue virus	NC_024299
Nounane virus	NC_033715
Ntaya virus	NC_018705
Nyando virus	NC_034481 NC_034491 NC_034492

Continued on next page

Table 10.2 – continued from previous page

Virus Name	Accession Number
O'nyong-nyong virus	NC_001512
Omsk Hemorrhagic Fever virus	NC_005062
Oriboca virus	NC_034475 NC_034476 NC_034488
Oropouche virus	NC_005775 NC_005776 NC_005777
Orungo virus	NC_038604 NC_038605 NC_038606 NC_038607 NC_038608 NC_038609 NC_038610 NC_038611 NC_038612 NC_038613
Palm Creek virus	NC_033694
Patois virus	NC_043689 NC_043690 NC_043691
Phnom Penh bat virus	NC_034007
Piry virus	NC_038286
Pixuna virus	NC_038673
Powassan virus	NC_003687
Punta Toro virus	NC_039191 NC_039192 NC_039193
Quang Binh virus	NC_012671
Rift Valley fever virus	NC_014395 NC_014396 NC_014397
Rio Bravo virus	NC_003675
Rio Negro virus	NC_038674
Rocio virus	NC_040776
Ross River virus	NC_001544
Royal Farm virus	NC_039219
Saboya virus	NC_033697
Saint Louis encephalitis virus	NC_007580
Sal Vieja virus	NC_043117
San Perlita virus	NC_043118
Saumarez Reef virus	NC_033726
Semliki Forest virus	NC_003215
Sepik virus	NC_008719
Shuni virus	NC_043697 NC_043698 NC_043699
Sindbis virus	NC_001547
Sokoluk virus	NC_026624
Spondweni virus	NC_029055
Stratford virus	KF917540
Tacaiuma orthobunyavirus	NC_043692 NC_043693 NC_043694
Tamana bat virus	NC_003996
Tembusu virus	NC_015843
Thiafora virus	NC_034254 NC_034255 NC_034256 NC_034261 NC_034262 NC_034263 NC_039220 NC_039221

Continued on next page

Table 10.2 – continued from previous page

Virus Name	Accession Number
	NC_039222
Thogoto thogotovirus	NC_006495 NC_006496 NC_006504 NC_006506 NC_006507 NC_006508
Tick-borne encephalitis virus	NC_001672
Tonate virus	NC_038675
Torres virus	KM225265
Tyuleniy virus	NC_023424
Uganda S virus	NC_033698
Una virus	NC_043403
Usutu virus	NC_006551
Uukuniemi virus	NC_005214 NC_005220 NC_005221
Venezuelan Equine Encephalitis virus	NC_001449
Wesselsbron virus	NC_012735
West Nile virus lineage 1	NC_009942
West Nile virus lineage 2	NC_001563
Western Equine Encephalitis virus	NC_003908
Whataroa virus	NC_016961
Wyeomyia virus	NC_039186 NC_039187 NC_039188
Yaounde virus	NC_034018
Yellow fever virus	NC_002031
Yokose virus	NC_005039
Zika virus	NC_012532

Table 10.3: Accession Numbers and Characteristics for ZIKV Genomes Used in Section 8.2.4.

Year of Collection	Country	GenBank Accession Number
1984	Senegal	HQ234501
1984	Senegal	MK028860
1984	Senegal	MF510857
2006	Thailand	MT377502
2006	Thailand	MT377504
2007	Micronesia	EU545988
2012	Philippines	KU6810823
2013	Thailand	KX051561
2013	Thailand	KX051560
2013	French Polynesia	KJ776791
2014	Thailand	KU6810813
2014	Haiti	OK571913
2014	Haiti	KY415988
2014	Haiti	KY415986
2014	Haiti	KY415987
2014	Haiti	KY415990
2014	Haiti	KY415989
2014	Haiti	KY415991
2015	Thailand	KX051562
2015	Thailand	MT377503
2015	Panama	MK028859
2015	Colombia	KY989971
2015	Mexico	KY631493
2015	Puerto Rico	MK713748
2015	Colombia	MF574561
2015	Cape Verde	MK241416
2015	Brazil	MK566202
2015	Colombia	MF574552
2015	Colombia	MF574562
2015	Colombia	MF574582
2015	Colombia	MF574584
2015	Colombia	MF574563
2015	Colombia	MF574560
2015	Haiti	OL450365
2015	Haiti	OL450364
2015	Cape Verde	MK241415
2015	Brazil	MF352141
2015	Haiti	OL450366
2015	Brazil	MF073359
2015	Brazil	NC_035889
2016	Vietnam	MW915414
2016	Thailand	MT377500
2016	Thailand	MT377499
2016	Nicaragua	OK054394
2016	Panama	MN473454
2016	Thailand	MT377498
2016	Dominican Republic	KY785415
2016	Honduras	MF801387
2016	Nicaragua	OK054415
2016	Nicaragua	OK054391
2016	Singapore	KY241733
2016	United States Virgin Islands	MW165883
2016	Singapore	KY241671
2016	Singapore	KY241686
2016	Singapore	KY241694
2016	Singapore	KY241718
2016	Singapore	KY241728
2016	Singapore	KY241777
2016	Singapore	KY241675
2016	Singapore	KY241676
2016	Singapore	KY241685

Continued on next page

Table 10.3 – continued from previous page

Year of Collection	Country	GenBank Accession Number
2016	Singapore	KY241787
2016	Thailand	MT377493
2016	Nicaragua	MF434522
2016	Singapore	KY241673
2016	Thailand	MT377492
2016	USA	KY785422
2016	Nicaragua	OK054387
2016	Vietnam	MW915411
2016	Nicaragua	OK054414
2016	Nicaragua	OK054416
2016	Colombia	KY785469
2016	Brazil	MH882542
2016	Brazil	MH882527
2016	Brazil	MH882528
2016	Vietnam	MW915413
2016	Honduras	MF801384
2016	Puerto Rico	MW122388
2016	Mexico	MF801402
2016	Nicaragua	OK054434
2016	Brazil	MH882541
2016	Nicaragua	MF434517
2016	Guatemala	MF801378
2016	Nicaragua	OK054436
2016	Nicaragua	OK054396
2016	Nicaragua	OK054433
2016	Puerto Rico	MW122375
2016	Trinidad and Tobago	MW008559
2016	Trinidad and Tobago	MW012427
2016	USA	MK269360
2016	Puerto Rico	MW122382
2016	Mexico	MF801413
2016	Nicaragua	OK054369
2016	Nicaragua	OK054413
2016	Nicaragua	OK054435
2016	Trinidad and Tobago	MW012428
2016	Nicaragua	OK054386
2016	China	MF964216
2016	China	MF964216
2016	South Korea	KY553111
2016	Nicaragua	OK054395
2016	China	KY379148
2016	Guadeloupe	MN185331
2016	Mexico	MF801412
2016	China	MF167360
2016	Nicaragua	OK054390
2016	Nicaragua	OK054412
2016	Puerto Rico	MW122377
2016	French Guiana	MN185326
2016	Colombia	KY785466
2016	Guadeloupe	MN185329
2016	Thailand	MT377497
2016	Puerto Rico	MW122379
2016	Puerto Rico	MH675619
2016	South Korea	KY120353
2016	Colombia	MF574586
2016	Dominican Republic	MF098764
2016	Peru	KY693679
2016	Puerto Rico	MW122385
2016	USA	MH675630
2016	Thailand	MT377491
2016	Dominican Republic	MF098768
2016	Dominican Republic	MF098769
2016	Puerto Rico	MW122383

Continued on next page

Table 10.3 – continued from previous page

Year of Collection	Country	GenBank Accession Number
2016	Mexico	MF801403
2016	Dominican Republic	KY785420
2016	Haiti	MF384325
2016	Guadeloupe	MN185330
2016	Puerto Rico	MW122374
2016	Puerto Rico	MW122381
2016	Puerto Rico	MW122414
2016	Cape Verde	MK241417
2016	Dominican Republic	MF098766
2016	Puerto Rico	MW122378
2016	Jamaica	KY785419
2016	Thailand	MT377496
2016	French Guiana	MN185328
2016	Brazil	MH882543
2016	Thailand	MT377494
2016	Mexico	MF801406
2016	Brazil	MH882548
2016	Brazil	MH882547
2016	Brazil	MT483911
2016	Haiti	MN577544
2016	Mexico	MF801414
2016	Puerto Rico	MW122380
2016	Guadeloupe	MN185327
2016	Suriname	KY348640
2016	Puerto Rico	MW122376
2016	Puerto Rico	MW122373
2016	United States Virgin Islands	MW165882
2016	Jamaica	KY785424
2016	Honduras	KY785414
2016	Puerto Rico	MW122384
2016	United States Virgin Islands	MW165880
2016	Venezuela	KY693680
2016	Vietnam	MW915412
2016	Thailand	MT377495
2016	United States Virgin Islands	MW165884
2016	USA	MK269357
2016	Puerto Rico	MW122386
2017	Cuba	MH063264
2017	Thailand	MT377501
2017	Singapore	MF988734
2017	Cuba	MF159531
2017	Cuba	MK560178
2017	Brazil	MN101548
2017	Brazil	MT078742
2017	Cuba	MH063265
2018	Guinea	MN025403
2018	India	MK238037
2018	India	MK238035
2018	Cuba	MK560179
2019	China	MN611472
2019	China	MN190155
2019	Cambodia	MZ008356

Urban									
KC762633	EU677174	AY732477	GQ868630	FJ410257	GQ199785	FJ898394	FJ410230	GQ199796	JN054255
EU359008	FJ639669	GQ199790	KC762644	EU081257	FJ410238	EU677158	FJ410282	EU482652	JQ955624
FJ744717	EU482658	FM210210	DQ181801	FJ205877	FJ410221	GU131902	EU482782	FM210223	KM279592
FJ882602	EU482783	FM210216	KR779786	FJ547064	EU482677	FM210236	FM210244	JN368476	KM279515
KF955332	EU081206	KJ622199	KC261634	FJ639713	FJ562103	EU081197	GU131937	FJ461338	EU482455
AY858048	JN662391	FJ639723	KC762689	EU081223	GQ868628	FJ898429	JQ045634	JQ048541	LC011948
EU677171	FJ882516	EU482534	FJ461318	JQ045659	HG316481	FJ461340	FJ882520	GQ868606	FJ882540
GQ199799	GU131782	GQ199786	EU482718	GQ868633	FJ882554	JQ045665	FJ024435	FJ461312	GQ199820
FJ898411	FJ687443	FM210246	DQ181804	GQ868591	JQ955623	FJ744716	EU660416	EU482678	FJ205878
JQ045670	GQ868621	EU726776	FM210207	FJ547067	KC964093	EU482778	KJ010186	KM279514	EU482664
GQ398265	GQ868638	EU482774	AF204178	EU081198	FJ639726	GU131914	KF041258	FJ744739	DQ675519
EU081185	FJ744732	KC762688	DQ675527	EU081187	EU482457	EU081201	FJ639728	KF955458	FJ461334
EU081278	EU677176	FJ182024	FJ182028	EU249492	EU482480	EU482516	JQ045631	GQ199852	EU677159
JQ922548	FJ410262	FJ410268	FJ898422	FJ898428	GQ868619	GQ868607	GQ868613	EU482512	AY835999
JN638340	GQ398255	FJ461308	FJ898386	HM181939	KJ734727	FJ205880	JX475906	FJ410241	FM210243
EU482784	KM279532	FJ687445	EU482445	FJ687435	FJ461309	EU482663	JQ922550	KC762677	EU482644
JX470186	GQ398260	EU482661	KF041236	FJ639707	FJ432726	KM279527	EU482788	JQ045688	EU367962
FJ639729	DQ675522	EU482462	EU081199	EU081225	FJ744731	KJ622198	AY858042	GU131904	FJ461337
EU081214	EU081182	GQ466079	GU131935	JQ045638	KC762640	JQ922545	FJ639689	EU677167	FJ410211
FJ898427	EU482507	EU081258	EU677156	EU081230	FJ687430	FJ390383	JQ045642	FJ882546	GQ868614
KF289072	EU081245	AY732478	EU687247	EU081274	FJ024441	GU131721	FJ024425	FJ024429	AF276619
FJ024452	EU482467	EU687250	EU726767	EU569721	GQ398267	AY858036	DQ645555	FJ639706	FJ461305
KM279526	KM279573	EU687248	KM279554	EU482785	FJ467493	DQ181797	KM279564	FJ024458	EU482649
FM210203	FJ639717	GQ868629	EU482454	GU131946	DQ675529	FJ744728	EU081220	AY858041	FJ744730
EU081189	AY496871	GU131940	AY676353	FJ410229	EU081191	EU482456	FJ882572	JQ045637	KC762645
FJ182035	FJ898371	AB608789	EU677155	EU677178	JQ045647	FJ410220	FJ461327	FJ882563	EU482825
EU660393	FJ373296	EU081237	EU081260	FJ024462	FJ461313	FJ882519	GU131778	KC762631	FJ882543
EU482528	EU081246	EU081273	KC964094	KC762657	JQ922552	FJ744718	EU677138	FM210234	GQ398264
GQ398262	KM279530	EU482447	FJ024461	KM279536	FJ898454	DQ645540	KM279565	AY858035	KM279586
KM279545	FJ410193	KM279593	GQ868543	EU482472	KM279578	AY496874	AY858044	DQ675520	FJ562097
AY676352	AY876494	GU131917	FJ639721	AY676349	DQ863638	FJ639720	FJ547066	KC762684	EU081196
KJ737430	EU081190	FJ432725	FJ547065	GQ868632	JQ287661	FJ898392	FJ182021	EU482519	JQ045628
HQ891314	EU081227	GQ868635	FJ882560	EU482820	EU482502	KC762638	EU482520	GQ868602	EU660390
EU677160	FJ898372	EU081232	EU081234	EU081261	EU660419	FJ898425			
Sylvatic									
	EF105378	EF105379	EF105380	EF105381	EF105382	EF105383	EF105384	EF105385	
	EF105386	EF105387	EF105388	EF105389	EF105390	EF457904	EF457905.1	EF457906.1	

Table 10.4: Accession Numbers and Type for DENV Genomes Used in Section 8.2.4

10.5 References

1. Suttle, C. A. Viruses: unlocking the greatest biodiversity on Earth. *Genome* **56,** 542–544 (Oct. 2013).
2. Legendre, M., Arslan, D., Abergel, C. & Claverie, J.-M. Genomics of Megavirus and the elusive fourth domain of Life. *Communicative & Integrative Biology* **5,** 102–106 (Jan. 2012).
3. Louten, J. Virus Structure and Classification. *Essential Human Virology,* 19–29 (2016).
4. Makino, S. *et al.* Molecular cloning and sequencing of a human hepatitis delta virus RNA. *Nature* **329,** 343–346 (Sept. 1987).
5. La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* **455,** 100–104 (Sept. 2008).
6. Choi, K. H. Viral Polymerases. *Advances in experimental medicine and biology* **726,** 267–304 (2012).
7. Payne, S. Introduction to RNA Viruses. *Viruses,* 97–105 (2017).
8. Dall'Ara, M., Ratti, C., Bouzoubaa, S. E. & Gilmer, D. Ins and Outs of Multipartite Positive-Strand RNA Plant Viruses: Packaging versus Systemic Spread. *Viruses* **8,** 228 (Aug. 2016).
9. Ryu, W.-S. Virus Life Cycle. *Molecular Virology of Human Pathogenic Viruses,* 31–45 (2017).
10. Rogers, G. N. & Paulson, J. C. Receptor determinants of human and animal influenza virus isolates: differences in receptor specificity of the H3 hemagglutinin based on species of origin. *Virology* **127,** 361–373 (June 1983).
11. Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 11727–11734 (May 2020).
12. Wilen, C. B., Tilton, J. C. & Doms, R. W. HIV: Cell Binding and Entry. *Cold Spring Harbor Perspectives in Medicine* **2,** a006866 (Aug. 2012).
13. Rampersad, S. & Tennant, P. Replication and Expression Strategies of Viruses. *Viruses,* 55–82 (2018).
14. De Vlugt, C., Sikora, D. & Pelchat, M. Insight into Influenza: A Virus Cap-Snatching. *Viruses* **10,** 641 (Nov. 2018).
15. Reich, S. *et al.* Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516,** 361–366 (Dec. 2014).

16. Ahi, Y. S. & Mittal, S. K. Components of Adenovirus Genome Packaging. *Frontiers in Microbiology* **7** (2016).

17. Veesler, D. & Johnson, J. E. Virus Maturation. *Annual review of biophysics* **41,** 473–496 (2012).

18. Bennett, A. *et al.* Duration and Density of Fecal Rotavirus Shedding in Vaccinated Malawian Children With Rotavirus Gastroenteritis. *The Journal of Infectious Diseases* **222,** 2035–2040 (Nov. 2020).

19. Ip, D. K. M. *et al.* The Dynamic Relationship Between Clinical Symptomatology and Viral Shedding in Naturally Acquired Seasonal and Pandemic Influenza Virus Infections. *Clinical Infectious Diseases* **62,** 431–437 (Feb. 2016).

20. Munywoki, P. K. *et al.* Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiology and Infection* **143,** 804–812 (Mar. 2015).

21. Peltola, V. *et al.* Rhinovirus Transmission within Families with Children: Incidence of Symptomatic and Asymptomatic Infections. *The Journal of Infectious Diseases* **197,** 382–389 (Feb. 2008).

22. Takeyama, A. *et al.* Respiratory syncytial virus shedding by children hospitalized with lower respiratory tract infection. *Journal of Medical Virology* **88,** 938–946 (June 2016).

23. Vetter, P. *et al.* Ebola Virus Shedding and Transmission: Review of Current Evidence. *The Journal of Infectious Diseases* **214,** S177–S184 (Oct. 2016).

24. Wu, Q.-s. *et al.* Norovirus shedding among symptomatic and asymptomatic employees in outbreak settings in Shanghai, China. *BMC Infectious Diseases* **19,** 592 (July 2019).

25. Kane, M. & Golovkina, T. Common Threads in Persistent Viral Infections. *Journal of Virology* **84,** 4116–4123 (May 2010).

26. Decman, V., Freeman, M. L., Kinchington, P. R. & Hendricks, R. L. Immune control of HSV-1 latency. *Viral Immunology* **18,** 466–473 (2005).

27. Efstathiou, S. & Preston, C. M. Towards an understanding of the molecular basis of herpes simplex virus latency. *Virus Research* **111,** 108–119 (Aug. 2005).

28. Kennedy, P. G. E., Rovnak, J., Badani, H. & Cohrs, R. J. A comparison of herpes simplex virus type 1 and varicella-zoster virus latency and reactivation. *The Journal of General Virology* **96,** 1581–1602 (July 2015).

29. Hindmarsh, P. & Leis, J. Retroviral DNA Integration. *Microbiology and Molecular Biology Reviews* **63,** 836–843 (Dec. 1999).

30. Bode, J. G., Brenndörfer, E. D. & Häussinger, D. Hepatitis C virus (HCV) employs multiple strategies to subvert the host innate antiviral response. **389,** 1283–1298 (Oct. 2008).

31. Goodrum, F., Caviness, K. & Zagallo, P. Human Cytomegalovirus Persistence. *Cellular Microbiology* **14,** 644–655 (May 2012).

32. Schindell, B. G., Webb, A. L. & Kindrachuk, J. Persistence and Sexual Transmission of Filoviruses. *Viruses* **10,** 683 (Dec. 2018).

33. Baron, E. J. in *Medical Microbiology* (ed Baron, S.) 4th (University of Texas Medical Branch at Galveston, Galveston (TX), 1996).

34. Van Regenmortel, M. Viruses are real, virus species are man-made, taxonomic constructions. *Archives of Virology* **148,** 2481–2488 (Dec. 2003).

35. Simmonds, P. A clash of ideas – the varying uses of the 'species' term in virology and their utility for classifying viruses in metagenomic datasets. *Journal of General Virology* **99,** 277–287 (Mar. 2018).

36. Baltimore, D. Expression of animal virus genomes. *Bacteriological Reviews* **35,** 235–241 (Sept. 1971).

37. Koonin, E. V., Krupovic, M. & Agol, V. I. The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? *Microbiology and Molecular Biology Reviews* **85,** e00053–21 (June 2021).

38. Burrell, C. J., Howard, C. R. & Murphy, F. A. Classification of Viruses and Phylogenetic Relationships. *Fenner and White's Medical Virology,* 15–25 (2017).

39. Simmonds, P. & Aiewsakun, P. Virus classification – where do you draw the line? *Archives of Virology* **163,** 2037–2046 (Aug. 2018).

40. Aiewsakun, P. & Simmonds, P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome* **6,** 38 (Feb. 2018).

41. Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15,** 161–168 (Mar. 2017).

42. Simmonds, P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology* **96,** 1193–1206 (June 2015).

43. Steward, G. F. *et al.* Are we missing half of the viruses in the ocean? *The ISME journal* **7,** 672–679 (Mar. 2013).

44. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology* **19,** 155–170 (Mar. 2021).

45. Baranov, P. V. *et al.* Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology* **332,** 498–510 (Feb. 2005).

46. Finkel, Y. *et al.* The coding capacity of SARS-CoV-2. *Nature* **589,** 125–130 (Jan. 2021).

47. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181,** 914–921.e10 (May 2020).

48. Sawicki, S. G. & Sawicki, D. L. Coronaviruses use discontinuous extension for synthesis of subgenome-length negative strands. *Advances in Experimental Medicine and Biology* **380,** 499–506 (1995).

49. Masters, P. S. The molecular biology of coronaviruses. *Advances in Virus Research* **66,** 193–292 (2006).

50. Gaunt, E. R., Hardie, A., Claas, E. C. J., Simmonds, P. & Templeton, K. E. Epidemiology and clinical presentations of the four human coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time PCR method. *Journal of Clinical Microbiology* **48,** 2940–2947 (Aug. 2010).

51. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (Mar. 2020).

52. World Health Organization. *The true death toll of COVID-19: estimating global excess mortality*

53. Hamre, D. & Procknow, J. J. A new virus isolated from the human respiratory tract. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine* **121,** 190–193 (Jan. 1966).

54. Van der Hoek, L. *et al.* Identification of a new human coronavirus. *Nature Medicine* **10,** 368–373 (Apr. 2004).

55. Woo, P. C. Y. *et al.* Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *Journal of Virology* **79,** 884–895 (Jan. 2005).

56. McIntosh, K., Becker, W. B. & Chanock, R. M. Growth in suckling-mouse brain of "IBV-like" viruses from patients with upper respiratory tract disease. *Proceedings of the National Academy of Sciences of the United States of America* **58,** 2268–2273 (Dec. 1967).

57. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *The New England Journal of Medicine* **367,** 1814–1820 (Nov. 2012).

58. Ksiazek, T. G. *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *The New England Journal of Medicine* **348,** 1953–1966 (May 2003).

59. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579,** 265–269 (Mar. 2020).

60. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England Journal of Medicine* **382,** 727–733 (Feb. 2020).

61. Kuno, G. & Chang, G.-J. J. Biological Transmission of Arboviruses: Reexamination of and New Insights into Components, Mechanisms, and Unique Traits as Well as Their Evolutionary Trends. *Clinical Microbiology Reviews* **18,** 608–637 (Oct. 2005).

62. Gaudreault, N. N., Madden, D. W., Wilson, W. C., Trujillo, J. D. & Richt, J. A. African Swine Fever Virus: An Emerging DNA Arbovirus. *Frontiers in Veterinary Science* **7,** 215 (May 2020).

63. Scott-Phillips, T. C., Laland, K. N., Shuker, D. M., Dickins, T. E. & West, S. A. The niche constuction perspective: a critical appraisal*. *Evolution; International Journal of Organic Evolution* **68,** 1231–1243 (May 2014).

64. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9,** 267–276 (Apr. 2008).

65. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nature Reviews Microbiology* **9,** 617–626 (Aug. 2011).

66. Steel, J. & Lowen, A. C. Influenza A virus reassortment. *Current Topics in Microbiology and Immunology* **385,** 377–401 (2014).

67. Wright, S. *The roles of mutation, inbreeding, cross-breeding and selection in evolution, p 356–366* in *Proceedings of the Sixth Annual Congress of Genetics* **1** (1932).

68. Trubenová, B., Krejca, M. S., Lehre, P. K. & Kötzing, T. Surfing on the seascape: Adaptation in a changing environment. *Evolution; International Journal of Organic Evolution* **73,** 1356–1374 (July 2019).

69. Wolf, K. A., Kwan, J. C. & Kamil, J. P. Structural Dynamics and Molecular Evolution of the SARS-CoV-2 Spike Protein. *mBio,* e0203021 (Mar. 2022).

70. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war - host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology* **17,** 321–328 (May 2019).

71. Thompson, C. P. *et al.* A naturally protective epitope of limited variability as an influenza vaccine target. *Nature Communications* **9,** 3859 (Sept. 2018).

72. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3,** 418–426 (Sept. 1986).

73. Gillespie, J. H. Natural selection and the molecular clock. *Molecular Biology and Evolution* **3,** 138–155 (Mar. 1986).

74. Holmes, E. C. Molecular Clocks and the Puzzle of RNA Virus Origins. *Journal of Virology* **77,** 3893–3897 (Apr. 2003).

75. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* (Sept. 1994).

76. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2,** 150–174 (Mar. 1985).

77. Yang, Z. & Nielsen, R. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Molecular Biology and Evolution* **17,** 32–43 (Jan. 2000).

78. McDonald, J. & Krietman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351** (June 1991).

79. Smyth, R. P., Negroni, M., Lever, A. M., Mak, J. & Kenyon, J. C. RNA Structure—A Neglected Puppet Master for the Evolution of Virus and Host Immunity. *Frontiers in Immunology* **9** (2018).

80. Mailliot, J. & Martin, F. Viral internal ribosomal entry sites: four classes for one goal. *Wiley interdisciplinary reviews. RNA* **9** (Mar. 2018).

81. Lee, K.-M., Chen, C.-J. & Shih, S.-R. Regulation Mechanisms of Viral IRES-Driven Translation. *Trends in Microbiology* **25,** 546–561 (July 2017).

82. Martinez-Salas, E., Francisco-Velilla, R., Fernandez-Chamorro, J. & Embarek, A. M. Insights into Structural and Mechanistic Features of Viral IRES Elements. *Frontiers in Microbiology* **8** (2018).

83. Schlesinger, S., Makino, S. & Linial, M. L. cis-acting genomic elements and trans-acting proteins involved in the assembly of RNA viruses. *Seminars in Virology* **5,** 39–49 (Feb. 1994).

84. Kelly, J. A. *et al.* Structural and functional conservation of the programmed 1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *Journal of Biological Chemistry* **295,** 10741–10748 (July 2020).

85. Simmonds, P., Xia, W., Baillie, J. K. & McKinnon, K. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics* **14,** 610 (Sept. 2013).

86. Atkinson, N. J., Witteveldt, J., Evans, D. J. & Simmonds, P. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Research* **42,** 4527–4545 (Apr. 2014).

87. Fros, J. J. *et al.* CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. *eLife* **6,** e29112 (Sept. 2017).

88. Gaunt, E. *et al.* Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. *eLife* **5** (ed Lipsitch, M.) e12735 (Feb. 2016).

89. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Human Genetics* **83,** 181–188 (Sept. 1989).

90. Ficarelli, M. *et al.* CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-dependent and -independent mechanisms). *Journal of virology* (Feb. 2020).

91. Odon, V. *et al.* The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. *Nucleic Acids Research* **47,** 8061–8083 (Sept. 2019).

92. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550,** 124–127 (Oct. 2017).

93. Beutler, E., Gelbart, T., Han, J. H., Koziol, J. A. & Beutler, B. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **86,** 192–196 (Jan. 1989).

94. King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science (New York, N.Y.)* **164,** 788–798 (May 1969).

95. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **217,** 624–626 (Feb. 1968).

96. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).

97. Kern, A. D. & Hahn, M. W. The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution* **35,** 1366–1371 (June 2018).

98. Masel, J. Genetic drift. *Current biology* **21,** R837–838 (Oct. 2011).

99. Weaver, S. C., Forrester, N. L., Liu, J. & Vasilakis, N. Population bottlenecks and founder effects: implications for mosquito-borne arboviral emergence. *Nature Reviews. Microbiology* **19,** 184–195 (Mar. 2021).

100. McCrone, J. T. & Lauring, A. S. Genetic bottlenecks in intraspecies virus transmission. *Current Opinion in Virology* **28,** 20–25 (Feb. 2018).

101. Samuel, C. E. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. *Virology* **411,** 180–193 (Mar. 2011).

102. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **356,** 983–989 (July 2001).

103. Woolhouse, M. E. J. & Brierley, L. Epidemiological characteristics of human-infective RNA viruses. *Scientific Data* **5,** 180017 (Feb. 2018).

104. Babkin, I. V. & Babkina, I. N. The origin of the variola virus. *Viruses* **7,** 1100–1112 (Mar. 2015).

105. Ida-Hosonuma, M. *et al.* Host range of poliovirus is restricted to simians because of a rapid sequence change of the poliovirus receptor gene during evolution. *Archives of Virology* **148,** 29–44 (Jan. 2003).

106. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447,** 279–283 (May 2007).

107. Morse, S. S. *et al.* Prediction and prevention of the next pandemic zoonosis. *The Lancet* **380,** 1956–1965 (Dec. 2012).

108. Woolhouse, M. E. J., Brierley, L., McCaffery, C. & Lycett, S. Assessing the Epidemic Potential of RNA and DNA Viruses. *Emerging Infectious Diseases* **22,** 2037–2044 (Dec. 2016).

109. Woolhouse, M. & Gaunt, E. Ecological origins of novel human pathogens. *Critical Reviews in Microbiology* **33,** 231–242 (2007).

110. Chen, S. H. *et al.* Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* **238,** 363–366 (Oct. 1987).

111. Powell, L. M. *et al.* A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50,** 831–840 (Sept. 1987).

112. Teng, B., Burant, C. F. & Davidson, N. O. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* **260,** 1816–1819 (June 1993).

113. Davidson, N. O. *et al.* Proposed nomenclature for the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme: APOBEC-1. *RNA* **1,** 3 (Mar. 1995).

114. Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79,** 285–296 (Mar. 2002).

115. Rogozin, I. B., Basu, M. K., Jordan, I. K., Pavlov, Y. I. & Koonin, E. V. APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle* **4,** 1281–1285 (Sept. 2005).

116. Burns, M. B., Leonard, B. & Harris, R. S. APOBEC3B: pathological consequences of an innate immune DNA mutator. *Biomedical Journal* **38,** 102–110 (Apr. 2015).

117. Green, A. M. *et al.* APOBEC3A damages the cellular genome during DNA replication. *Cell Cycle* **15,** 998–1008 (2016).

118. Henderson, S. & Fenton, T. APOBEC3 genes: retroviral restriction factors to cancer drivers. *Trends in Molecular Medicine* **21,** 274–284 (May 2015).

119. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics* **45,** 970–976 (Sept. 2013).

120. Smith, H. C. RNA binding to APOBEC deaminases; Not simply a substrate for C to U editing. *RNA biology* **14,** 1153–1165 (Sept. 2017).

121. Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* **2,** e00534 (Apr. 2013).

122. Salter, J. D., Bennett, R. P. & Smith, H. C. The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends in Biochemical Sciences* **41,** 578–594 (July 2016).

123. Harari, A., Ooms, M., Mulder, L. C. F. & Simon, V. Polymorphisms and splice variants influence the antiretroviral activity of human APOBEC3H. *Journal of Virology* **83,** 295–303 (Jan. 2009).

124. Lee, R. M., Hirano, K., Anant, S., Baunoch, D. & Davidson, N. O. An alternatively spliced form of apobec-1 messenger RNA is overexpressed in human colon cancer. *Gastroenterology* **115,** 1096–1103 (Nov. 1998).

125. Van Maldegem, F. *et al.* AID splice variants lack deaminase activity. *Blood* **113,** 1862–1864, author reply 1864 (Feb. 2009).

126. Fossat, N. *et al.* C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. *EMBO reports* **15,** 903–910 (Aug. 2014).

127. Mehta, A., Kinter, M. T., Sherman, N. E. & Driscoll, D. M. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Molecular and Cellular Biology* **20,** 1846–1854 (Mar. 2000).

128. Rosenberg, B. R., Hamilton, C. E., Mwangi, M. M., Dewell, S. & Papavasiliou, F. N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nature Structural & Molecular Biology* **18,** 230–236 (Feb. 2011).

129. Skuse, G. R., Cappione, A. J., Sowden, M., Metheny, L. J. & Smith, H. C. The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic Acids Research* **24,** 478–485 (Feb. 1996).

130. Yamanaka, S., Poksay, K. S., Arnold, K. S. & Innerarity, T. L. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes & Development* **11,** 321–333 (Feb. 1997).

131. Backus, J. W. & Smith, H. C. Apolipoprotein B mRNA sequences 3' of the editing site are necessary and sufficient for editing and editosome assembly. *Nucleic Acids Research* **19,** 6781–6786 (Dec. 1991).

132. Blanc, V. *et al.* Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. *Genome Biology* **15,** R79 (June 2014).

133. Wolfe, A. D., Arnold, D. B. & Chen, X. S. Comparison of RNA Editing Activity of APOBEC1-A1CF and APOBEC1-RBM47 Complexes Reconstituted in HEK293T Cells. *Journal of Molecular Biology* **431,** 1506–1517 (Mar. 2019).

134. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics* **11,** 47–59 (Jan. 2010).

135. Rogozin, I. B. *et al.* Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nature Immunology* **8,** 647–656 (June 2007).

136. Chaudhuri, J. *et al.* Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature* **422,** 726–730 (Apr. 2003).

137. Dickerson, S. K., Market, E., Besmer, E. & Papavasiliou, F. N. AID mediates hypermutation by deaminating single stranded DNA. *The Journal of Experimental Medicine* **197,** 1291–1296 (May 2003).

138. Muramatsu, M. *et al.* Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *The Journal of Biological Chemistry* **274,** 18470–18476 (June 1999).

139. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424,** 103–107 (July 2003).

140. Etard, C., Roostalu, U. & Strähle, U. Lack of Apobec2-related proteins causes a dystrophic muscle phenotype in zebrafish embryos. *The Journal of Cell Biology* **189,** 527–539 (May 2010).

141. Liao, W. *et al.* APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. *Biochemical and Biophysical Research Communications* **260,** 398–404 (July 1999).

142. Sato, Y. *et al.* Deficiency in APOBEC2 leads to a shift in muscle fiber type, diminished body mass, and myopathy. *The Journal of Biological Chemistry* **285,** 7111–7118 (Mar. 2010).

143. Vonica, A., Rosa, A., Arduini, B. L. & Brivanlou, A. H. APOBEC2, a selective inhibitor of TGF signaling, regulates left-right axis specification during early embryogenesis. *Developmental Biology* **350,** 13–23 (Feb. 2011).

144. Okuyama, S. *et al.* Excessive activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide 2 (APOBEC2) contributes to liver and lung tumorigenesis. *International Journal of Cancer* **130,** 1294–1301 (Mar. 2012).

145. Yamanaka, S. *et al.* Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proceedings of the National Academy of Sciences of the United States of America* **92,** 8483–8487 (Aug. 1995).

146. Shi, M. *et al.* Characterization and functional analysis of chicken APOBEC4. *Developmental and Comparative Immunology* **106,** 103631 (May 2020).

147. McDaniel, Y. Z. *et al.* Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. *Nucleic Acids Research* **48,** 1353–1371 (Feb. 2020).

148. Sharma, S. *et al.* APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nature Communications* **6,** 6881 (Apr. 2015).

149. Sharma, S., Patnaik, S. K., Taggart, R. T. & Baysal, B. E. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Scientific Reports* **6,** 39100 (Dec. 2016).

150. Sharma, S. *et al.* Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. *Genome Biology* **20,** 37 (Feb. 2019).

151. Bonvin, M. *et al.* Interferon-inducible expression of APOBEC3 editing enzymes in human hepatocytes and inhibition of hepatitis B virus replication. *Hepatology* **43,** 1364–1374 (June 2006).

152. Mohanram, V., Sköld, A. E., Bächle, S. M., Pathak, S. K. & Spetz, A.-L. IFN- induces APOBEC3G, F, and A in immature dendritic cells and limits HIV-1 spread to CD4+ T cells. *Journal of Immunology* **190,** 3346–3353 (Apr. 2013).

153. Peng, G., Lei, K. J., Jin, W., Greenwell-Wild, T. & Wahl, S. M. Induction of APOBEC3 family proteins, a defensive maneuver underlying interferon-induced anti-HIV-1 activity. *The Journal of Experimental Medicine* **203,** 41–46 (Jan. 2006).

154. Weisblum, Y. *et al.* APOBEC3A Is Upregulated by Human Cytomegalovirus (HCMV) in the Maternal-Fetal Interface, Acting as an Innate Anti-HCMV Effector. *Journal of Virology* **91,** e01296–17 (Dec. 2017).

155. Aydin, H., Taylor, M. W. & Lee, J. E. Structure-guided analysis of the human APOBEC3-HIV restrictome. *Structure* **22,** 668–684 (May 2014).

156. Delviks-Frankenberry, K. A., Desimmie, B. A. & Pathak, V. K. Structural Insights into APOBEC3-Mediated Lentiviral Restriction. *Viruses* **12,** E587 (May 2020).

157. Desimmie, B. A. *et al.* Multiple APOBEC3 restriction factors for HIV-1 and one Vif to rule them all. *Journal of Molecular Biology* **426,** 1220–1245 (Mar. 2014).

158. Haché, G., Mansky, L. M. & Harris, R. S. Human APOBEC3 proteins, retrovirus restriction, and HIV drug resistance. *AIDS reviews* **8,** 148–157 (Sept. 2006).

159. Harris, R. S. & Dudley, J. P. APOBECs and virus restriction. *Virology* **479-480,** 131–145 (May 2015).

160. Xu, W. K., Byun, H. & Dudley, J. P. The Role of APOBECs in Viral Replication. *Microorganisms* **8,** E1899 (Nov. 2020).

161. Mangeat, B. *et al.* Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **424,** 99–103 (July 2003).

162. Kaiser, S. M. & Emerman, M. Uracil DNA glycosylase is dispensable for human immunodeficiency virus type 1 replication and does not contribute to the antiviral effects of the cytidine deaminase Apobec3G. *Journal of Virology* **80,** 875–882 (Jan. 2006).

163. Langlois, M.-A. & Neuberger, M. S. Human APOBEC3G can restrict retroviral infection in avian cells and acts independently of both UNG and SMUG1. *Journal of Virology* **82,** 4660–4664 (May 2008).

164. Weil, A. F. *et al.* Uracil DNA glycosylase initiates degradation of HIV-1 cDNA containing misincorporated dUTP and prevents viral integration. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E448–457 (Feb. 2013).

165. Yang, B., Chen, K., Zhang, C., Huang, S. & Zhang, H. Virion-associated uracil DNA glycosylase-2 and apurinic/apyrimidinic endonuclease are involved in the degradation of APOBEC3G-edited nascent HIV-1 DNA. *The Journal of Biological Chemistry* **282,** 11667–11675 (Apr. 2007).

166. Bruner, K. M. *et al.* Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nature Medicine* **22,** 1043–1049 (Sept. 2016).

167. Janini, M., Rogers, M., Birx, D. R. & McCutchan, F. E. Human immunodeficiency virus type 1 DNA sequences genetically damaged by hypermutation are often abundant in patient peripheral blood mononuclear cells and may be generated during near-simultaneous infection and activation of CD4(+) T cells. *Journal of Virology* **75,** 7973–7986 (Sept. 2001).

168. Kieffer, T. L. *et al.* G–>A hypermutation in protease and reverse transcriptase regions of human immunodeficiency virus type 1 residing in resting CD4+ T cells in vivo. *Journal of Virology* **79,** 1975–1980 (Feb. 2005).

169. Conticello, S. G., Harris, R. S. & Neuberger, M. S. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. *Current biology: CB* **13,** 2009–2013 (Nov. 2003).

170. Kao, S. *et al.* The human immunodeficiency virus type 1 Vif protein reduces intracellular expression and inhibits packaging of APOBEC3G (CEM15), a cellular inhibitor of virus infectivity. *Journal of Virology* **77,** 11398–11407 (Nov. 2003).

171. Marin, M., Rose, K. M., Kozak, S. L. & Kabat, D. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nature Medicine* **9,** 1398–1403 (Nov. 2003).

172. Mehle, A., Goncalves, J., Santa-Marta, M., McPike, M. & Gabuzda, D. Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation. *Genes & Development* **18,** 2861–2866 (Dec. 2004).

173. Mehle, A. *et al.* Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. *The Journal of Biological Chemistry* **279,** 7792–7798 (Feb. 2004).

174. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418,** 646–650 (Aug. 2002).

175. Sheehy, A. M., Gaddis, N. C. & Malim, M. H. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nature Medicine* **9,** 1404–1407 (Nov. 2003).

176. Stopak, K., de Noronha, C., Yonemoto, W. & Greene, W. C. HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Molecular Cell* **12,** 591–601 (Sept. 2003).

177. Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302,** 1056–1060 (Nov. 2003).

178. Fribourgh, J. L. *et al.* Core binding factor beta plays a critical role by facilitating the assembly of the Vif-cullin 5 E3 ubiquitin ligase. *Journal of Virology* **88,** 3309–3319 (Mar. 2014).

179. Guo, Y. *et al.* Structural basis for hijacking CBF- and CUL5 E3 ligase complex by HIV-1 Vif. *Nature* **505,** 229–233 (Jan. 2014).

180. Jäger, S. *et al.* Vif hijacks CBF- to degrade APOBEC3G and promote HIV-1 infection. *Nature* **481,** 371–375 (Dec. 2011).

181. Zhang, W., Du, J., Evans, S. L., Yu, Y. & Yu, X.-F. T-cell differentiation factor CBF- regulates HIV-1 Vif-mediated evasion of host restriction. *Nature* **481,** 376–379 (Dec. 2011).

182. Kristbjörnsdóttir, H. B. *et al.* The vif gene of maedi-visna virus is essential for infectivity in vivo and in vitro. *Virology* **318,** 350–359 (Jan. 2004).

183. Nakano, Y. *et al.* A conflict of interest: the evolutionary arms race between mammalian APOBEC3 and lentiviral Vif. *Retrovirology* **14,** 31 (May 2017).

184. Wang, J., Shaban, N. M., Land, A. M., Brown, W. L. & Harris, R. S. Simian Immunodeficiency Virus Vif and Human APOBEC3B Interactions Resemble Those between HIV-1 Vif and Human APOBEC3G. *Journal of Virology* **92,** e00447–18 (June 2018).

185. Zhang, W. *et al.* Cellular requirements for bovine immunodeficiency virus Vif-mediated inactivation of bovine APOBEC3 proteins. *Journal of Virology* **88,** 12528–12540 (Nov. 2014).

186. Zhao, Z. *et al.* CAEV Vif Hijacks ElonginB/C, CYPA and Cullin5 to Assemble the E3 Ubiquitin Ligase Complex Stepwise to Degrade oaA3Z2-Z3. *Frontiers in Microbiology* **10,** 565 (2019).

187. Zielonka, J. *et al.* Vif of feline immunodeficiency virus from domestic cats protects against APOBEC3 restriction factors from many felids. *Journal of Virology* **84,** 7312–7324 (July 2010).

188. Jaguva Vasudevan, A. A. *et al.* Prototype foamy virus Bet impairs the dimerization and cytosolic solubility of human APOBEC3G. *Journal of Virology* **87,** 9030–9040 (Aug. 2013).

189. Löchelt, M. *et al.* The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 7982–7987 (May 2005).

190. Russell, R. A. *et al.* Foamy virus Bet proteins function as novel inhibitors of the APOBEC3 family of innate antiretroviral defense factors. *Journal of Virology* **79,** 8724–8731 (July 2005).

191. Derse, D., Hill, S. A., Princler, G., Lloyd, P. & Heidecker, G. Resistance of human T cell leukemia virus type 1 to APOBEC3G restriction is mediated by elements in nucleocapsid. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 2915–2920 (Feb. 2007).

192. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40,** 491–500 (Mar. 1985).

193. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biology* **19,** 199 (Nov. 2018).

194. Rubin, G. M., Kidwell, M. G. & Bingham, P. M. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell* **29,** 987–994 (July 1982).

195. Biémont, C. & Vieira, C. Genetics: junk DNA as an evolutionary force. *Nature* **443,** 521–524 (Oct. 2006).

196. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).

197. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 5280–5285 (Apr. 2003).

198. Richardson, S. R. *et al.* The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum* **3,** MDNA3–0061–2014 (Apr. 2015).

199. Schumann, G. G. APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition. *Biochemical Society Transactions* **35,** 637–642 (June 2007).

200. Ikeda, T. *et al.* Intrinsic restriction activity by apolipoprotein B mRNA editing enzyme APOBEC1 against the mobility of autonomous retrotransposons. *Nucleic Acids Research* **39,** 5538–5554 (July 2011).

201. MacDuff, D. A., Demorest, Z. L. & Harris, R. S. AID can restrict L1 retrotransposition suggesting a dual role in innate and adaptive immunity. *Nucleic Acids Research* **37,** 1854–1867 (Apr. 2009).

202. Pizarro, J. G. & Cristofari, G. Post-Transcriptional Control of LINE-1 Retrotransposition by Cellular Host Factors in Somatic Cells. *Frontiers in Cell and Developmental Biology* **4,** 14 (2016).

203. Richardson, S. R., Narvaiza, I., Planegger, R. A., Weitzman, M. D. & Moran, J. V. APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. *eLife* **3,** e02008 (Apr. 2014).

204. Ostertag, E. M. & Kazazian, H. H. Biology of mammalian L1 retrotransposons. *Annual Review of Genetics* **35,** 501–538 (2001).

205. Anwar, F., Davenport, M. P. & Ebrahimi, D. Footprint of APOBEC3 on the genome of human retroelements. *Journal of Virology* **87,** 8195–8204 (July 2013).

206. Moris, A., Murray, S. & Cardinaud, S. AID and APOBECs span the gap between innate and adaptive immunity. *Frontiers in Microbiology* **5,** 534 (2014).

207. Suspène, R. *et al.* Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 8321–8326 (June 2005).

208. Gonzalez, M. C. *et al.* Human APOBEC1 cytidine deaminase edits HBV DNA. *Retrovirology* **6,** 96 (Oct. 2009).

209. Liang, G. *et al.* RNA editing of hepatitis B virus transcripts by activation-induced cytidine deaminase. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 2246–2251 (Feb. 2013).

210. Renard, M., Henry, M., Guétard, D., Vartanian, J.-P. & Wain-Hobson, S. APOBEC1 and APOBEC3 cytidine deaminases as restriction factors for hepadnaviral genomes in non-humans in vivo. *Journal of Molecular Biology* **400,** 323–334 (July 2010).

211. Seeger, C. & Mason, W. S. Hepatitis B virus biology. *Microbiology and molecular biology reviews: MMBR* **64,** 51–68 (Mar. 2000).

212. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (Jan. 2015).

213. Warren, C. J. *et al.* APOBEC3A functions as a restriction factor of human papillomavirus. *Journal of Virology* **89,** 688–702 (Jan. 2015).

214. Warren, C. J., Van Doorslaer, K., Pandey, A., Espinosa, J. M. & Pyeon, D. Role of the host restriction factor APOBEC3 on papillomavirus evolution. *Virus Evolution* **1,** vev015 (Jan. 2015).

215. Suspène, R. *et al.* Genetic editing of herpes simplex virus 1 and Epstein-Barr herpesvirus genomes by human APOBEC3 cytidine deaminases in culture and in vivo. *Journal of Virology* **85,** 7594–7602 (Aug. 2011).

216. Tsuge, M. *et al.* G to A hypermutation of TT virus. *Virus Research* **149,** 211–216 (May 2010).

217. Bishop, K. N., Holmes, R. K., Sheehy, A. M. & Malim, M. H. APOBEC-mediated editing of viral RNA. *Science* **305,** 645 (July 2004).

218. Milewska, A. *et al.* APOBEC3-mediated restriction of RNA virus replication. *Scientific Reports* **8,** 5960 (Apr. 2018).

219. Poulain, F., Lejeune, N., Willemart, K. & Gillet, N. A. Footprint of the host restriction factors APOBEC3 on the genome of human viruses. *PLoS pathogens* **16,** e1008718 (Aug. 2020).

220. Simmonds, P. & Ansari, M. A. Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. *PLoS pathogens* **17,** e1009596 (June 2021).

221. Perelygina, L. *et al.* Infectious vaccine-derived rubella viruses emerge, persist, and evolve in cutaneous granulomas of children with primary immunodeficiencies. *PLoS pathogens* **15,** e1008080 (Oct. 2019).

222. Chen, H. *et al.* APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Current biology* **16,** 480–485 (Mar. 2006).

223. Fehrholz, M. *et al.* The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses. *The Journal of General Virology* **93,** 565–576 (Mar. 2012).

224. Narvaiza, I. *et al.* Deaminase-independent inhibition of parvoviruses by the APOBEC3A cytidine deaminase. *PLoS pathogens* **5,** e1000439 (May 2009).

225. Stenglein, M. D. & Harris, R. S. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *The Journal of Biological Chemistry* **281,** 16837–16841 (June 2006).

226. Kremer, M. *et al.* Vaccinia virus replication is not affected by APOBEC3 family members. *Virology Journal* **3,** 86 (Oct. 2006).

227. Pauli, E.-K. *et al.* High level expression of the anti-retroviral protein APOBEC3G is induced by influenza A virus but does not confer antiviral activity. *Retrovirology* **6,** 38 (Apr. 2009).

228. Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K. & Neuberger, M. S. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution* **22,** 367–377 (Feb. 2005).

229. Gerber, A. P. & Keller, W. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends in Biochemical Sciences* **26,** 376–384 (June 2001).

230. Iyer, L. M., Zhang, D., Rogozin, I. B. & Aravind, L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Research* **39,** 9473–9497 (Dec. 2011).

231. Krishnan, A., Iyer, L. M., Holland, S. J., Boehm, T. & Aravind, L. Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. *Proceedings of the National Academy of Sciences of the United States of America* **115,** E3201–E3210 (Apr. 2018).

232. Liu, M.-C. *et al.* AID/APOBEC-like cytidine deaminases are ancient innate immune mediators in invertebrates. *Nature Communications* **9,** 1948 (May 2018).

233. Hsu, E. Assembly and Expression of Shark Ig Genes. *Journal of Immunology* **196,** 3517–3523 (May 2016).

234. Münk, C., Willemsen, A. & Bravo, I. G. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC evolutionary biology* **12,** 71 (May 2012).

235. Ito, J., Gifford, R. J. & Sato, K. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 610–618 (Jan. 2020).

236. Mikkelsen, T. S. *et al.* Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447,** 167–177 (May 2007).

237. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453,** 175–183 (May 2008).

238. LaRue, R. S. *et al.* Guidelines for naming nonprimate APOBEC3 genes and proteins. *Journal of Virology* **83,** 494–497 (Jan. 2009).

239. Hayward, J. A. *et al.* Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. *Molecular Biology and Evolution* **35,** 1626–1637 (July 2018).

240. LaRue, R. S. *et al.* The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. *BMC molecular biology* **9,** 104 (Nov. 2008).

241. Sawyer, S. L., Emerman, M. & Malik, H. S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS biology* **2,** E275 (Sept. 2004).

242. Eckerle, L. D., Lu, X., Sperry, S. M., Choi, L. & Denison, M. R. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *Journal of Virology* **81,** 12135–12144 (Nov. 2007).

243. Eckerle, L. D. *et al.* Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS pathogens* **6,** e1000896 (May 2010).

244. Smith, E. C., Blanc, H., Surdel, M. C., Vignuzzi, M. & Denison, M. R. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. *PLoS pathogens* **9,** e1003565 (Aug. 2013).

245. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22,** 30494 (Mar. 2017).

246. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5,** e00408–20 (June 2020).

247. Klimczak, L. J., Randall, T. A., Saini, N., Li, J.-L. & Gordenin, D. A. Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. *PloS One* **15,** e0237689 (2020).

248. Oude Munnink, B. B. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371,** 172–177 (Jan. 2021).

249. Braun, K. M. *et al.* Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathogens* **17** (ed Weaver, S. C.) e1009849 (Aug. 2021).

250. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372,** eabg0821 (Apr. 2021).

251. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R. & Ramazzotti, D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24,** 102116 (Feb. 2021).

252. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10,** e66857 (Aug. 2021).

253. Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances* **6,** eabb5813 (June 2020).

254. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nature Genetics* **47,** 1067–1072 (Sept. 2015).

255. Simmonds, P. Pervasive RNA Secondary Structure in the Genomes of SARS-CoV-2 and Other Coronaviruses. *mBio* **11,** e01661–20 (Oct. 2020).

256. Wang, S.-M. & Wang, C.-T. APOBEC3G cytidine deaminase association with coronavirus nucleocapsid protein. *Virology* **388,** 112–120 (May 2009).

257. Kim, K. *et al.* APOBEC-mediated Editing of SARS-CoV-2 Genomic RNA Impacts Viral Replication and Fitness. *bioRxiv,* 2021.12.18.473309 (Dec. 2021).

258. Van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **83,** 104351 (Sept. 2020).

259. Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *Virus Evolution* **7,** veaa098 (Jan. 2021).

260. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS biology* **19,** e3001115 (Mar. 2021).

261. Berkhout, B. & van Hemert, F. On the biased nucleotide composition of the human coronavirus RNA genome. *Virus Research* **202,** 41–47 (Apr. 2015).

262. Woo, P. C. Y., Wong, B. H. L., Huang, Y., Lau, S. K. P. & Yuen, K.-Y. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. *Virology* **369,** 431–442 (Dec. 2007).

263. Baker, M. L., Schountz, T. & Wang, L.-F. Antiviral immune responses of bats: a review. *Zoonoses and Public Health* **60,** 104–116 (Feb. 2013).

264. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and Sources of Endemic Human Coronaviruses. *Advances in Virus Research* **100,** 163–188 (2018).

265. Angus, D. C. *et al.* The REMAP-CAP (Randomized Embedded Multifactorial Adaptive Platform for Community-acquired Pneumonia) Study. Rationale and Design. *Annals of the American Thoracic Society* **17,** 879–891 (July 2020).

266. Writing Committee for the REMAP-CAP Investigators. Effect of Convalescent Plasma on Organ Support–Free Days in Critically Ill Patients With COVID-19: A Randomized Clinical Trial. *JAMA* (Oct. 2021).

267. Clemens, S. A. C. *et al.* Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 lineages circulating in Brazil. *Nature Communications* **12,** 5861 (Oct. 2021).

268. Voysey, M. *et al.* Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet* **397,** 99–111 (Jan. 2021).

269. Voysey, M. *et al.* Single-dose administration and the influence of the timing of the booster dose on immunogenicity and efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine: a pooled analysis of four randomised trials. *The Lancet* **397,** 881–891 (Mar. 2021).

270. NIBSC. *Research Reagent for SARS-CoV-2 RNA NIBSC code 19/304 (Version 2)* Jan. 2021.

271. Wymant, C. *et al.* Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evolution* **4,** vey007 (Jan. 2018).

272. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30,** 772–780 (Apr. 2013).

273. Chakrabarti, R. & Schutt, C. E. The enhancement of PCR amplification by low molecular weight amides. *Nucleic Acids Research* **29,** 2377–2381 (June 2001).

274. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution* (July 2021).

275. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32,** 268–274 (Jan. 2015).

276. R Core Team. *R: A language and environment for statistical computing.*

277. Simpson, G. *Confidence intervals for GLMs* Dec. 2018.

278. Choudhuri, J. *et al.* SARS-CoV-2 PCR cycle threshold at hospital admission associated with patient mortality. *PloS One* **15,** e0244777 (2020).

279. Tso, C. F., Garikipati, A., Green-Saxena, A., Mao, Q. & Das, R. Correlation of Population SARS-CoV-2 Cycle Threshold Values to Local Disease Dynamics: Exploratory Observational Study. *JMIR Public Health and Surveillance* **7,** e28265 (Mar. 2021).

280. National SARS-CoV-2 Serology Assay Evaluation Group. Performance characteristics of five immunoassays for SARS-CoV-2: a head-to-head benchmark comparison. *The Lancet Infectious Diseases* **20,** 1390–1400 (Dec. 2020).

281. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20,** 1160–1166 (July 2019).

282. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (June 2009).

283. Madansky, A. Alternative Approaches to Significance Testing with Weighted Means. *The Analytical Group,* 7.

284. Simmonds, P. SSE: a nucleotide and amino acid sequence analysis platform. *BMC Research Notes* **5,** 50 (Jan. 2012).

285. Von Mises, R. Selected Papers of Richard von Mises. *American Mathematical Society* **2,** 313–334 (1964).

286. Blitvich, B. J. & Firth, A. E. Insect-Specific Flaviviruses: A Systematic Review of Their Discovery, Host Range, Mode of Transmission, Superinfection Exclusion Potential and Genomic Organization. *Viruses* **7,** 1927–1959 (Apr. 2015).

287. Blitvich, B. J. & Firth, A. E. A Review of Flaviviruses that Have No Known Arthropod Vector. *Viruses* **9,** 154 (June 2017).

288. Pandit, P. S. *et al.* Predicting wildlife reservoirs and global vulnerability to zoonotic Flaviviruses. *Nature Communications* **9,** 5425 (Dec. 2018).

289. Vasilakis, N. *et al.* Evolutionary Processes among Sylvatic Dengue Type 2 Viruses. *Journal of Virology* **81,** 9591–9595 (Sept. 2007).

290. World Health Organization. *Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health.* Jan. 2021.

291. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182,** 812–827.e19 (Aug. 2020).

292. Bekliz, M. *et al.* Neutralization of ancestral SARS-CoV-2 and variants Alpha, Beta, Gamma, Delta, Zeta and Omicron by mRNA vaccination and infection-derived immunity through homologous and heterologous variants. *medRxiv,* 2021.12.28.21268491 (Jan. 2022).

293. Liu, L. *et al.* Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602,** 676–681 (Feb. 2022).

294. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nature Reviews Genetics,* 1–17 (Sept. 2021).

295. World Health Organization. *Tracking SARS-CoV-2 variants* Mar. 2022.

296. Chiara, M. *et al.* Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Briefings in Bioinformatics* **22,** 616–630 (Mar. 2021).

297. Xiao, M. *et al.* Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome Medicine* **12,** 57 (June 2020).

298. Deng, G. R. A sensitive non-radioactive PCR-RFLP analysis for detecting point mutations at 12th codon of oncogene c-Ha-ras in DNAs of gastric cancer. *Nucleic Acids Research* **16,** 6231 (July 1988).

299. Ota, M., Fukushima, H., Kulski, J. K. & Inoko, H. Single nucleotide polymorphism detection by polymerase chain reaction-restriction fragment length polymorphism. *Nature Protocols* **2,** 2857–2864 (Nov. 2007).

300. England, P. H. *Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01* Dec. 2020.

301. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372,** 815–821 (May 2021).

302. Voloch, C. M. *et al.* Genomic Characterization of a Novel SARS-CoV-2 Lineage from Rio de Janeiro, Brazil. *Journal of Virology* **95,** e00119–21 (Apr. 2021).

303. Johnson, B. A. *et al.* Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *bioRxiv,* 2021.10.14.464390 (Oct. 2021).

304. Mourier, T. *et al.* SARS-CoV-2 genomes from Saudi Arabia implicate nucleocapsid mutations in host response and increased viral load. *Nature Communications* **13,** 601 (Feb. 2022).

305. Wu, H. *et al.* Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host & Microbe* **29,** 1788–1801.e6 (Dec. 2021).

306. Gangavarapu, K. *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *MedRxiv,* 2022.01.27.22269965 (Jan. 2022).

307. Weissman, J. L., Fagan, W. F. & Johnson, P. L. F. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLOS Genetics* **15,** e1008493 (Nov. 2019).

308. Cavazzoni, P. *Coronavirus (COVID-19) Update: FDA Limits Use of Certain Monoclonal Antibodies to Treat COVID-19 Due to the Omicron Variant* Jan. 2022.

309. Brown, K. A. *et al.* S-Gene Target Failure as a Marker of Variant B.1.1.7 Among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021. *JAMA* **325,** 2115–2116 (May 2021).

310. Lee, W. L. *et al.* Quantitative SARS-CoV-2 Alpha Variant B.1.1.7 Tracking in Wastewater by Allele-Specific RT-qPCR. *Environmental Science & Technology Letters* **8,** 675–682 (Aug. 2021).

311. Babiker, A. *et al.* Single-amplicon, multiplex real-time RT-PCR with tiled probes to detect SARS-CoV-2 spike mutations associated with variants of concern. *Journal of Clinical Microbiology,* JCM0144621 (Aug. 2021).

312. Vogels, C. B. F. *et al.* Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLOS Biology* **19,** e3001236 (May 2021).

313. Harper, H. *et al.* Detecting SARS-CoV-2 variants with SNP genotyping. *PLOS ONE* **16,** e0243185 (Feb. 2021).

314. Stadhouders, R. *et al.* The Effect of Primer-Template Mismatches on the Detection and Quantification of Nucleic Acids Using the 5 Nuclease Assay. *The Journal of Molecular Diagnostics* **12,** 109–117 (Jan. 2010).

315. Quick, J. *nCoV-2019 sequencing protocol* Jan. 2020.

316. Lambisia, A. W. *et al.* Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol. *Frontiers in Medicine* **9** (2022).

317. Varadaraj, K. & Skinner, D. M. Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene* **140,** 1–5 (Mar. 1994).

318. Boden, D. *et al.* HIV-1 Drug Resistance in Newly Infected Individuals. *JAMA* **282,** 1135–1141 (Sept. 1999).

319. Cassedy, A., Parle-McDermott, A. & O'Kennedy, R. Virus Detection: A Review of the Current and Emerging Molecular and Immunological Methods. *Frontiers in Molecular Biosciences* **8** (2021).

320. Larremore, D. B. *et al.* Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science Advances* **7,** eabd5393 (Jan. 2021).

321. Tom, M. R. & Mina, M. J. To Interpret the SARS-CoV-2 Test, Consider the Cycle Threshold Value. *Clinical Infectious Diseases,* ciaa619 (May 2020).

322. Arons, M. M. *et al.* Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. *New England Journal of Medicine* **382,** 2081–2090 (May 2020).

323. Cevik, M. *et al.* SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet Microbe* **2,** e13–e22 (Jan. 2021).

324. Challenger, J. D. *et al.* Modelling upper respiratory viral load dynamics of SARS-CoV-2. *BMC Medicine* **20,** 25 (Jan. 2022).

325. Lennon, N. J. *et al.* Cross-Sectional Assessment of SARS-CoV-2 Viral Load by Symptom Status in Massachusetts Congregate Living Facilities. *The Journal of Infectious Diseases* **224,** 1658–1663 (Nov. 2021).

326. Walsh, K. A. *et al.* SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *Journal of Infection* **81,** 357–371 (Sept. 2020).

327. Fontana, L. M., Villamagna, A. H., Sikka, M. K. & McGregor, J. C. Understanding viral shedding of severe acute respiratory coronavirus virus 2 (SARS-CoV-2): Review of current literature. *Infection Control and Hospital Epidemiology* **42,** 659–668 (June 2021).

328. Simmonds, P., Williams, S. & Harvala, H. Understanding the outcomes of COVID-19 – does the current model of an acute respiratory infection really fit? *The Journal of General Virology* **102,** 001545 (Dec. 2020).

329. Liu, Y. *et al.* Viral dynamics in mild and severe cases of COVID-19. *The Lancet Infectious Diseases* **20,** 656–657 (June 2020).

330. Wang, J. *et al.* Persistent SARS-COV-2 RNA positivity in a patient for 92 days after disease onset: A case report. *Medicine* **99,** e21865 (Aug. 2020).

331. Zhang, L., Li, C., Zhou, Y., Wang, B. & Zhang, J. Persistent viral shedding lasting over 60 days in a mild COVID-19 patient with ongoing positive SARS-CoV-2. *Quantitative Imaging in Medicine and Surgery* **10,** 1141–1144 (May 2020).

332. Byington, C. L. *et al.* Community Surveillance of Respiratory Viruses Among Families in the Utah Better Identification of Germs-Longitudinal Viral Epidemiology (BIG-LoVE) Study. *Clinical Infectious Diseases* **61,** 1217–1224 (Oct. 2015).

333. Jones, T. C. *et al.* Estimating infectiousness throughout SARS-CoV-2 infection course. *Science* **373,** eabi5273 (July 2021).

334. Marc, A. *et al.* Quantifying the relationship between SARS-CoV-2 viral load and infectiousness. *eLife* **10** (eds Cobey, S. E. & Van der Meer, J. W.) e69302 (Sept. 2021).

335. B., S. R. *et al.* Virus load and clinical features during the acute phase of Chikungunya infection in children. *PLoS ONE* **14,** e0211036 (Feb. 2019).

336. Li, C.-C. *et al.* Correlation of Pandemic (H1N1) 2009 Viral Load with Disease Severity and Prolonged Viral Shedding in Children. *Emerging Infectious Diseases* **16,** 1265–1272 (Aug. 2010).

337. Vos, L. M. *et al.* Lower respiratory tract infection in the community: associations between viral aetiology and illness course. *Clinical Microbiology and Infection* **27,** 96–104 (Jan. 2021).

338. Lavezzo, E. *et al.* Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. *Nature* **584,** 425–429 (Aug. 2020).

339. Lee, S. *et al.* Clinical Course and Molecular Viral Shedding Among Asymptomatic and Symptomatic Patients With SARS-CoV-2 Infection in a Community Treatment Center in the Republic of Korea. *JAMA Internal Medicine* **180,** 1447–1452 (Nov. 2020).

340. Yang, Q. *et al.* Just 2% of SARS-CoV-2positive individuals carry 90% of the virus circulating in communities. *Proceedings of the National Academy of Sciences* **118,** e2104547118 (May 2021).

341. Zuin, M., Gentili, V., Cervellati, C., Rizzo, R. & Zuliani, G. Viral Load Difference between Symptomatic and Asymptomatic COVID-19 Patients: Systematic Review and Meta-Analysis. *Infectious Disease Reports* **13,** 645–653 (July 2021).

342. Pujadas, E. *et al.* SARS-CoV-2 viral load predicts COVID-19 mortality. *The Lancet Respiratory Medicine* **8,** e70 (Sept. 2020).

343. Waudby-West, R. *et al.* The association between SARS-CoV-2 RT-PCR cycle threshold and mortality in a community cohort. *European Respiratory Journal* **58** (July 2021).

344. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182,** 1295–1310.e20 (Sept. 2020).

345. Zhou, D. *et al.* Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184,** 2348–2361.e6 (Apr. 2021).

346. Guo, K. *et al.* Interferon resistance of emerging SARS-CoV-2 variants. *Proceedings of the National Academy of Sciences* **119.** Publisher: Proceedings of the National Academy of Sciences, e2203760119 (Aug. 2022).

347. Thorne, L. G. *et al.* Evolution of enhanced innate immune evasion by the SARS-CoV-2 B.1.1.7 UK variant. *bioRxiv,* 2021.06.06.446826 (June 2021).

348. Rambaut, A. *et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology* Dec. 2020.

349. Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581,** 465–469 (May 2020).

350. Leary, S. *et al.* Generation of a novel SARS-CoV-2 sub-genomic RNA due to the R203K/G204R variant in nucleocapsid: homologous recombination has potential to change SARS-CoV-2 at both protein and RNA level. *bioRxiv,* 2020.04.10.029454 (Aug. 2021).

351. Mears, H. V. *et al.* Emergence of new subgenomic mRNAs in SARS-CoV-2. *bioRxiv.* Section: New Results Type: article, 2022.04.20.488895 (Apr. 2022).

352. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: a severity of disease classification system. *Critical Care Medicine* **13,** 818–829 (Oct. 1985).

353. Dahdouh, E., Lázaro-Perona, F., Romero-Gómez, M. P., Mingorance, J. & García-Rodriguez, J. Ct values from SARS-CoV-2 diagnostic PCR assays should not be used as direct estimates of viral load. *The Journal of Infection* **82,** 414–451 (Mar. 2021).

354. Markewitz, R. *et al.* Analysis of SARS-CoV-2 reverse transcription-quantitative polymerase chain reaction cycle threshold values vis-à-vis anti-SARS-CoV-2 antibodies from a high incidence region. *International Journal of Infectious Diseases* **110,** 114–122 (Sept. 2021).

355. Ratcliff, J., Nguyen, D., Andersson, M. & Simmonds, P. Evaluation of Different PCR Assay Formats for Sensitive and Specific Detection of SARS-CoV-2 RNA. *bioRxiv,* 2020.06.24.168013 (July 2020).

356. Gao, M. *et al.* Associations between body-mass index and COVID-19 severity in $6 \cdot 9$ million people in England: a prospective, community-based, cohort study. *The Lancet Diabetes & Endocrinology* **9,** 350–359 (June 2021).

357. Staines, H. M. *et al.* IgG Seroconversion and Pathophysiology in Severe Acute Respiratory Syndrome Coronavirus 2 Infection. *Emerging Infectious Diseases* **27,** 85–91 (Jan. 2021).

358. Yadav, A. K. *et al.* Seroconversion among COVID-19 patients admitted in a dedicated COVID hospital: A longitudinal prospective study of 1000 patients. *Medical Journal, Armed Forces India* **77,** S379–S384 (July 2021).

359. Zhao, J. *et al.* Antibody responses to SARS-CoV-2 in patients of novel coronavirus disease 2019. *Clinical Infectious Diseases,* ciaa344 (Mar. 2020).

360. Yanez, N. D., Weiss, N. S., Romand, J.-A. & Treggiari, M. M. COVID-19 mortality risk for older men and women. *BMC Public Health* **20,** 1742 (Nov. 2020).

361. RECOVERY Collaborative Group. Casirivimab and imdevimab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet* **399,** 665–676 (Feb. 2022).

362. King, K. L. *et al.* SARS-CoV-2 variants of concern Alpha and Delta show increased viral load in saliva. *MedRxiv,* 2022.02.10.22270797 (Mar. 2022).

363. Levi, J. E. *et al.* Dynamics of SARS-CoV-2 Variants of Concern in Brazil, Early 2021. *Frontiers in Public Health* **9** (2021).

364. Naveca, F. G. *et al.* COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nature Medicine* **27,** 1230–1238 (July 2021).

365. Singanayagam, A. *et al.* Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *The Lancet Infectious Diseases* **22.** Publisher: Elsevier, 183–195 (Feb. 2022).

366. Couzens, A. *et al.* Higher Relative Viral Load Excretion Determined by Normalised Threshold Crossing Value in Acute Cases infected with the B.1.1.7 Lineage VOC 202012/01 (Using S gene target failure as a Proxy) When Compared to other Circulating Lineages in Wales. *MedRxiv,* 2021.04.02.21254832 (Apr. 2021).

367. Frampton, D. *et al.* Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *The Lancet Infectious Diseases* **21,** 1246–1256 (Sept. 2021).

368. Kidd, M. *et al.* S-Variant SARS-CoV-2 Lineage B1.1.7 Is Associated With Significantly Higher Viral Load in Samples Tested by TaqPath Polymerase Chain Reaction. *The Journal of Infectious Diseases* **223,** 1666–1670 (May 2021).

369. Walker, A. S. *et al.* Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. *eLife* **10** (eds Teare, M. D. & Davenport, M. P.) e64683 (July 2021).

370. Wickenhagen, A. *et al.* A prenylated dsRNA sensor protects against severe COVID-19. *Science* **374,** eabj3624 (Oct. 2021).

371. Lee, J. Y. *et al.* Absolute quantitation of individual SARS-CoV-2 RNA molecules provides a new paradigm for infection dynamics and variant differences. *eLife* **11** (ed Davenport, M. P.) e74153 (Jan. 2022).

372. Ugolini, C. *et al.* Nanopore ReCappable sequencing maps SARS-CoV-2 5 capping sites and provides new insights into the structure of sgRNAs. *Nucleic Acids Research* **50,** 3475–3489 (Apr. 2022).

373. Alexandersen, S., Chamings, A. & Bhatta, T. R. SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication. *Nature Communications* **11,** 6059 (Nov. 2020).

374. Parker, M. D. *et al.* Altered subgenomic RNA abundance provides unique insight into SARS-CoV-2 B.1.1.7/Alpha variant infections. *Communications Biology* **5.** Number: 1 Publisher: Nature Publishing Group, 1–10 (July 2022).

375. Rajah, M. M. *et al.* SARS-CoV-2 Alpha, Beta, and Delta variants display enhanced Spike-mediated syncytia formation. *The EMBO journal* **40,** e108944 (Dec. 2021).

376. Larsen, D. A. & Wigginton, K. R. Tracking COVID-19 with wastewater. *Nature Biotechnology* **38,** 1151–1153 (Oct. 2020).

377. Peccia, J. *et al.* Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nature Biotechnology* **38,** 1164–1167 (Oct. 2020).

378. Olesen, S. W., Imakaev, M. & Duvallet, C. Making waves: Defining the lead time of wastewater-based epidemiology for COVID-19. *Water Research* **202,** 117433 (Sept. 2021).

379. Hay, J. A. *et al.* Estimating epidemiologic dynamics from cross-sectional viral load distributions. *Science* **373,** eabh0635 (July 2021).

380. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20,** 533–534 (May 2020).

381. Prabhu, M. & Gergen, J. *History's Seven Deadliest Plagues* Nov. 2021.

382. COVID-19 Forecasting Team. Variation in the COVID-19 infection–fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. *The Lancet* **0** (Feb. 2022).

383. Chan-Yeung, M. & XU, R.-H. SARS: epidemiology. *Respirology* **8,** S9–S14 (Nov. 2003).

384. Foster, M. A., Haber, P. & Nelson, N. P. *Pinkbook: Hepatitis A* Aug. 2021.

385. Presanis, A. M. *et al.* Changes in severity of 2009 pandemic A/H1N1 influenza in England: a Bayesian evidence synthesis. *BMJ* **343,** d5408 (Sept. 2011).

386. Charles, P. C., Trgovcich, J., Davis, N. L. & Johnston, R. E. Immunopathogenesis and Immune Modulation of Venezuelan Equine Encephalitis Virus-Induced Disease in the Mouse. *Virology* **284,** 190–202 (June 2001).

387. Chisari, F. V., Isogawa, M. & Wieland, S. F. Pathogenesis of Hepatitis B Virus Infection. *Pathologie-biologie* **58,** 258–266 (Aug. 2010).

388. Foster, G. R. Past, present, and future hepatitis C treatments. *Seminars in Liver Disease* **24 Suppl 2,** 97–104 (2004).

389. Holmes, J. A., Rutledge, S. M. & Chung, R. T. Direct-acting antiviral treatment for hepatitis C. *The Lancet* **393,** 1392–1394 (Apr. 2019).

390. Davies, B. E. Pharmacokinetics of oseltamivir: an oral antiviral for the treatment and prophylaxis of influenza in diverse populations. *Journal of Antimicrobial Chemotherapy* **65,** ii5–ii10 (Apr. 2010).

391. Chen, R.-C. *et al.* Treatment of severe acute respiratory syndrome with glucosteroids: the Guangzhou experience. *Chest* **129,** 1441–1452 (June 2006).

392. Dessie, Z. G. & Zewotir, T. Mortality-related risk factors of COVID-19: a systematic review and meta-analysis of 42 studies and 423,117 patients. *BMC Infectious Diseases* **21,** 855 (Aug. 2021).

393. Jin, J.-M. *et al.* Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. *Frontiers in Public Health* **8** (2020).

394. Jordan, R. E., Adab, P. & Cheng, K. K. Covid-19: risk factors for severe disease and death. *BMJ* **368,** m1198 (Mar. 2020).

395. Pijls, B. G. *et al.* Demographic risk factors for COVID-19 infection, severity, ICU admission and death: a meta-analysis of 59 studies. *BMJ Open* **11,** e044640 (Jan. 2021).

396. Engmann, S. & Cousineau, D. Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitative Methods* **6,** 1–18 (Sept. 2011).

397. Challen, R. *et al.* Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ* **372,** n579 (Mar. 2021).

398. Grint, D. J. *et al.* Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England, 16 November to 5 February. *Eurosurveillance* **26,** 2100256 (Mar. 2021).

399. Patone, M. *et al.* Mortality and critical care unit admission associated with the SARS-CoV-2 lineage B.1.1.7 in England: an observational cohort study. *The Lancet Infectious Diseases* **21,** 1518–1528 (Nov. 2021).

400. Faes, C. *et al.* Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients. *International Journal of Environmental Research and Public Health* **17,** 7560 (Oct. 2020).

401. Ohl, M. E. *et al.* Association of Remdesivir Treatment With Survival and Length of Hospital Stay Among US Veterans Hospitalized With COVID-19. *JAMA Network Open* **4,** e2114741 (July 2021).

402. RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine* **384,** 693–704 (Feb. 2021).

403. RECOVERY Collaborative Group. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet* **397,** 1637–1645 (May 2021).

404. Abdool Karim, S. S. & de Oliveira, T. New SARS-CoV-2 Variants — Clinical, Public Health, and Vaccine Implications. *New England Journal of Medicine* **384,** 1866–1868 (May 2021).

405. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19,** 409–424 (July 2021).

406. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184,** 64–75.e11 (Jan. 2021).

407. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183,** 739–751.e8 (Oct. 2020).

408. Gandhi, S. *et al.* De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an immunocompromised patient: a case report. *Nature Communications* **13,** 1547 (Mar. 2022).

409. Szemiel, A. M. *et al.* In vitro selection of Remdesivir resistance suggests evolutionary predictability of SARS-CoV-2. *PLOS Pathogens* **17,** e1009929 (Sept. 2021).

410. Hoffmann, M. *et al.* SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell* **184,** 2384–2393.e12 (Apr. 2021).

411. Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184,** 4220–4236.e13 (Aug. 2021).

412. Nguyen, D. *et al.* Effects of Severe Acute Respiratory Syndrome Coronavirus 2 Strain Variation on Virus Neutralization Titers: Therapeutic Use of Convalescent Plasma. *The Journal of Infectious Diseases* **225,** 971–976 (Mar. 2022).

413. Planas, D. *et al.* Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature* **602,** 671–675 (Feb. 2022).

414. Mallapaty, S. Where did Omicron come from? Three key theories. *Nature* **602,** 26–28 (Jan. 2022).

415. Bate, N. *et al.* In vitro evolution predicts emerging CoV-2 mutations with high affinity for ACE2 and cross-species binding. *bioRxiv,* 2021.12.23.473975 (Dec. 2021).

416. Bose, M. E. *et al.* Sequencing and analysis of globally obtained human parainfluenza viruses 1 and 3 genomes. *PLOS ONE* **14,** e0220057 (July 2019).

417. MacLean, O. A., Orton, R. J., Singer, J. B. & Robertson, D. L. No evidence for distinct types in the evolution of SARS-CoV-2. *Virus Evolution* **6,** veaa034 (Jan. 2020).

418. Nobusawa, E. & Sato, K. Comparison of the Mutation Rates of Human Influenza A and B Viruses. *Journal of Virology* **80,** 3675–3678 (Apr. 2006).

419. Virk, R. K. *et al.* Divergent evolutionary trajectories of influenza B viruses underlie their contemporaneous epidemic activity. *Proceedings of the National Academy of Sciences* **117,** 619–628 (Jan. 2020).

420. Yu, J.-M., Fu, Y.-H., Peng, X.-L., Zheng, Y.-P. & He, J.-S. Genetic diversity and molecular evolution of human respiratory syncytial virus A and B. *Scientific Reports* **11,** 12941 (June 2021).

421. Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *bioRxiv,* 2022.03.08.481609 (Mar. 2022).

422. Gregory, D. A. *et al. Genetic Diversity and Evolutionary Convergence of Cryptic SARS-CoV-2 Lineages Detected Via Wastewater Sequencing* Pages: 2022.06.03.22275961. June 2022.

423. Jensen, B. *et al.* Emergence of the E484K mutation in SARS-COV-2-infected immunocompromised patients treated with bamlanivimab in Germany. *The Lancet Regional Health – Europe* **8** (Sept. 2021).

424. Lohr, B., Niemann, D. & Verheyen, J. Bamlanivimab Treatment Leads to Rapid Selection of Immune Escape Variant Carrying the E484K Mutation in a B.1.1.7-Infected and Immunosuppressed Patient. *Clinical Infectious Diseases* **73,** 2144–2145 (Dec. 2021).

425. Peiffer-Smadja, N. *et al.* Emergence of E484K Mutation Following Bamlanivimab Monotherapy among High-Risk Patients Infected with the Alpha Variant of SARS-CoV-2. *Viruses* **13,** 1642 (Aug. 2021).

426. Rockett, R. *et al.* Resistance Mutations in SARS-CoV-2 Delta Variant after Sotrovimab Use. *New England Journal of Medicine* **0,** null (Mar. 2022).

427. Andreano, E. *et al.* SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proceedings of the National Academy of Sciences* **118,** e2103154118 (Sept. 2021).

428. Chen, L. *et al.* Emergence of Multiple SARS-CoV-2 Antibody Escape Variants in an Immunocompromised Host Undergoing Convalescent Plasma Treatment. *mSphere* **6,** e0048021 (Aug. 2021).

429. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592,** 277–282 (Apr. 2021).

430. European Centre for Disease Control and Prevention. *SARS-CoV-2 variants of concern as of 31 March 2022* Mar. 2022.

431. Romero, P. E. *et al.* The Emergence of Sars-CoV-2 Variant Lambda (C.37) in South America. *Microbiology Spectrum* **9,** e0078921 (Oct. 2021).

432. Rueca, M. *et al.* Compartmentalized Replication of SARS-Cov-2 in Upper vs. Lower Respiratory Tract Assessed by Whole Genome Quasispecies Analysis. *Microorganisms* **8,** 1302 (Aug. 2020).

433. RECOVERY Collaborative Group. Convalescent plasma in patients admitted to hospital with COVID-19 (RECOVERY): a randomised controlled, open-label, platform trial. *The Lancet* **397,** 2049–2059 (May 2021).

434. Frank, T. D. *et al.* Global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2017, and forecasts to 2030, for 195 countries and territories: a systematic analysis for the Global Burden of Diseases, Injuries, and Risk Factors Study 2017. *The Lancet HIV* **6,** e831–e859 (Dec. 2019).

435. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine:* **1,** a006841 (Sept. 2011).

436. Reperant, L. A., Moesker, F. M. & Osterhaus, A. D. Influenza: from zoonosis to pandemic. *ERJ Open Research* **2,** 00013–2016 (Mar. 2016).

437. Troeger, C. E. *et al.* Mortality, morbidity, and hospitalisations due to influenza lower respiratory tract infections, 2017: an analysis for the Global Burden of Disease Study 2017. *The Lancet Respiratory Medicine* **7,** 69–89 (Jan. 2019).

438. Geoghegan, J. L., Senior, A. M., Di Giallonardo, F. & Holmes, E. C. Virological factors that increase the transmissibility of emerging human viruses. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 4170–4175 (Apr. 2016).

439. Kilpatrick, A. M. & Randolph, S. E. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *The Lancet* **380,** 1946–1955 (Dec. 2012).

440. Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J. & Jiggins, F. M. The evolution and genetics of virus host shifts. *PLoS pathogens* **10,** e1004395 (Nov. 2014).

441. Lu, L. *et al.* Evolutionary origins of epidemic potential among human RNA viruses. *bioRxiv,* 771394 (Sept. 2019).

442. Mollentze, N., Streicker, D. G., Murcia, P. R., Hampson, K. & Biek, R. Virulence mismatches in index hosts shape the outcomes of cross-species transmission. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 28859–28866 (Nov. 2020).

443. Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* **546,** 646–650 (June 2017).

444. Plowright, R. K. *et al.* Pathways to zoonotic spillover. *Nature Reviews. Microbiology* **15,** 502–510 (Aug. 2017).

445. Walker, J. W., Han, B. A., Ott, I. M. & Drake, J. M. Transmissibility of emerging viral zoonoses. *PloS One* **13,** e0206926 (2018).

446. Woolhouse, M. E., Taylor, L. H. & Haydon, D. T. Population biology of multihost pathogens. *Science* **292,** 1109–1112 (May 2001).

447. Kelly, T. R. *et al.* Implementing One Health approaches to confront emerging and re-emerging zoonotic disease threats: lessons from PREDICT. *One Health Outlook* **2,** 1 (Jan. 2020).

448. PREDICT Consortium. Reducing Pandemic Risk, Promoting Global Health (2014).

449. Carroll, D. *et al.* The Global Virome Project. *Science* **359,** 872–874 (Feb. 2018).

450. Rabaa, M. A. *et al.* The Vietnam Initiative on Zoonotic Infections (VIZIONS): A Strategic Approach to Studying Emerging Zoonotic Infectious Diseases. *EcoHealth* **12,** 726–735 (Dec. 2015).

451. Grange, Z. L. *et al.* Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proceedings of the National Academy of Sciences* **118.** interim analysis of, e2002324118 (Apr. 2021).

452. Hardy, J. L., Houk, E. J., Kramer, L. D. & Reeves, W. C. Intrinsic factors affecting vector competence of mosquitoes for arboviruses. *Annual Review of Entomology* **28,** 229–262 (1983).

453. Woolhouse, M. E. J., Adair, K. & Brierley, L. RNA Viruses: A Case Study of the Biology of Emerging Infectious Diseases. *Microbiology Spectrum* **1** (Oct. 2013).

454. Antzin-Anduetza, I., Mahiet, C., Granger, L. A., Odendall, C. & Swanson, C. M. Increasing the CpG dinucleotide abundance in the HIV-1 genomic RNA inhibits viral replication. *Retrovirology* **14,** 49 (Nov. 2017).

455. Witteveldt, J., Martin-Gans, M. & Simmonds, P. Enhancement of the Replication of Hepatitis C Virus Replicons of Genotypes 1 to 4 by Manipulation of CpG and UpA Dinucleotide Frequencies and Use of Cell Lines Expressing SECL14L2 for Antiviral Resistance Testing. *Antimicrobial Agents and Chemotherapy* **60,** 2981–2992 (May 2016).

456. Bick, M. J. *et al.* Expression of the Zinc-Finger Antiviral Protein Inhibits Alphavirus Replication. *Journal of Virology* **77,** 11555–11562 (Nov. 2003).

457. Lobo, F. P. *et al.* Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PloS One* **4,** e6282 (July 2009).

458. Fros, J. J. *et al.* The dinucleotide composition of the Zika virus genome is shaped by conflicting evolutionary pressures in mammalian hosts and mosquito vectors. *PLOS Biology* **19,** e3001201 (Apr. 2021).

459. Levinson, R. S., Strauss, J. H. & Strauss, E. G. Complete sequence of the genomic RNA of O'Nyong-Nyong virus and its use in the construction of alphavirus phylogenetic trees. *Virology* **175,** 110–123 (Mar. 1990).

460. Weaver, S. C. *et al.* A Comparison of the Nucleotide Sequences of Eastern and Western Equine Encephalomyelitis Viruses with Those of Other Alphaviruses and Related RNA Viruses. *Virology* **197,** 375–390 (Nov. 1993).

461. Fall, G., Diallo, M., Loucoubar, C., Faye, O. & Sall, A. A. Vector Competence of Culex neavei and Culex quinquefasciatus (Diptera: Culicidae) from Senegal for Lineages 1, 2, Koutango and a Putative New Lineage of West Nile virus. *The American Journal of Tropical Medicine and Hygiene* **90,** 747–754 (Apr. 2014).

462. Bhat, H. R., Sreenivasan, M. A., Goverdhan, M. K. & Naik, S. V. Transmission of Kyasanur Forest Disease virus by Haemaphysalis kyasanurensis trapido, Hoogstraal and Rajagopalan, 1964 (Acarina: Ixodidae). *The Indian Journal of Medical Research* **63,** 879–887 (June 1975).

463. Figueiredo, M. L. G. d. *et al.* Cacipacore virus as an emergent mosquito-borne Flavivirus. *Revista da Sociedade Brasileira de Medicina Tropical* **50,** 539–542 (Aug. 2017).

464. Prow, N. A. The Changing Epidemiology of Kunjin Virus in Australia. *International Journal of Environmental Research and Public Health* **10,** 6255–6272 (Dec. 2013).

465. Centers for Disease Control and Prevention. *How ticks spread disease | CDC* Sept. 2020.

466. Scott, T. W. & Takken, W. Feeding strategies of anthropophilic mosquitoes result in increased risk of pathogen transmission. *Trends in Parasitology* **28,** 114–121 (Mar. 2012).

467. May, F. J. *et al.* Genetic divergence among members of the Kokobera group of flaviviruses supports their separation into distinct species. *The Journal of General Virology* **94,** 1462–1467 (July 2013).

468. Dick, G. W. A., Kitchen, S. F. & Haddow, A. J. Zika virus. I. Isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **46,** 509–520 (Sept. 1952).

469. Hay, S. I. *et al.* Global mapping of infectious disease. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368,** 20120250 (Mar. 2013).

470. Weissenböck, H., Hubálek, Z., Bakonyi, T. & Nowotny, N. Zoonotic mosquito-borne flaviviruses: Worldwide presence of agents with proven pathogenicity and potential candidates of future emerging diseases. *Veterinary Microbiology. Zoonoses: Advances and Perspectives* **140,** 271–280 (Jan. 2010).

471. Shutt, D. P., Manore, C. A., Pankavich, S., Porter, A. T. & Del Valle, S. Y. Estimating the reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America. *Epidemics* **21,** 63–79 (Dec. 2017).

472. Wang, L., Zhao, H., Oliva, S. M. & Zhu, H. Modeling the transmission and control of Zika in Brazil. *Scientific Reports* **7,** 7721 (Aug. 2017).

473. Vasilakis, N., Cardosa, J., Hanley, K. A., Holmes, E. C. & Weaver, S. C. Fever from the forest: prospects for the continued emergence of sylvatic dengue virus and its impact on public health. *Nature Reviews. Microbiology* **9,** 532–541 (June 2011).

474. Colmant, A. M. G. *et al.* Insect-Specific Flavivirus Replication in Mammalian Cells Is Inhibited by Physiological Temperature and the Zinc-Finger Antiviral Protein. *Viruses* **13,** 573 (Apr. 2021).

475. Junglen, S. *et al.* Host Range Restriction of Insect-Specific Flaviviruses Occurs at Several Levels of the Viral Life Cycle. *mSphere* **2,** e00375–16 (Jan. 2017).

476. Piyasena, T. B. H. *et al.* Infectious DNAs derived from insect-specific flavivirus genomes enable identification of pre- and post-entry host restrictions in vertebrate cells. *Scientific Reports* **7,** 2940 (June 2017).

477. Bellone, R. & Failloux, A.-B. The Role of Temperature in Shaping Mosquito-Borne Viruses Transmission. *Frontiers in Microbiology* **11,** 584846 (2020).

478. Holmes, E. C., Rambaut, A. & Andersen, K. G. Pandemics: spend on surveillance, not prediction. *Nature* **558,** 180–182 (June 2018).

479. Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLOS Biology* **19,** e3001135 (Apr. 2021).

480. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745 (Jan. 2016).

481. Pinheiro, F. P. *et al.* Oropouche virus. I. A review of clinical, epidemiological, and ecological findings. *The American Journal of Tropical Medicine and Hygiene* **30,** 149–160 (Jan. 1981).

482. Kosoy, O. I. *et al.* Novel thogotovirus associated with febrile illness and death, United States, 2014. *Emerging Infectious Diseases* **21,** 760–764 (May 2015).

483. Lledó, L., Giménez-Pardo, C. & Gegúndez, M. I. Epidemiological Study of Thogoto and Dhori Virus Infection in People Bitten by Ticks, and in Sheep, in an Area of Northern Spain. *International Journal of Environmental Research and Public Health* **17,** 2254 (Apr. 2020).

484. Bamunuarachchi, G. *et al.* Detection of Bourbon Virus-Specific Serum Neutralizing Antibodies in Human Serum in Missouri, USA. *mSphere* **0,** e00164–22 (May 2022).

485. Teurlai, M. *et al.* Socio-economic and Climate Factors Associated with Dengue Fever Spatial Heterogeneity: A Worked Example in New Caledonia. *PLOS Neglected Tropical Diseases* **9,** e0004211 (Dec. 2015).

486. Macdonald, J. *et al.* Molecular Phylogeny of Edge Hill Virus Supports its Position in the Yellow Fever Virus Group and Identifies a New Genetic Variant. *Evolutionary Bioinformatics Online* **6,** 91–96 (June 2010).

487. Aaskov, J., Phillips, D. & Wiemers, M. Possible clinical infection with Edge Hill virus. *Transactions of The Royal Society of Tropical Medicine and Hygiene* **87,** 452–453 (July 1993).

488. Gonçalves-Carneiro, D., Takata, M. A., Ong, H., Shilton, A. & Bieniasz, P. D. Origin and evolution of the zinc finger antiviral protein. *PLoS Pathogens* **17,** e1009545 (Apr. 2021).

489. Gérardin, P. *et al.* Estimating Chikungunya prevalence in La Réunion Island outbreak by serosurveys: Two methods for two critical times of the epidemic. *BMC Infectious Diseases* **8,** 99 (July 2008).

490. Renault, P. *et al.* A major epidemic of chikungunya virus infection on Reunion Island, France, 2005-2006. *The American Journal of Tropical Medicine and Hygiene* **77,** 727–731 (Oct. 2007).

491. Reiter, P., Fontenille, D. & Paupy, C. Aedes albopictus as an epidemic vector of chikungunya virus: another emerging problem? *The Lancet Infectious Diseases* **6,** 463–464 (Aug. 2006).

492. Tsetsarkin, K. A., Vanlandingham, D. L., McGee, C. E. & Higgs, S. A Single Mutation in Chikungunya Virus Affects Vector Specificity and Epidemic Potential. *PLOS Pathogens* **3.** Publisher: Public Library of Science, e201 (Dec. 2007).

493. Luo, X. *et al.* Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein. *Cell Reports* **30,** 46–52.e4 (Jan. 2020).

494. Tenorio, R. *et al.* Function, Architecture, and Biogenesis of Reovirus Replication Neoorganelles. *Viruses* **11,** 288 (Mar. 2019).

495. MacDonald, M. R., Machlin, E. S., Albin, O. R. & Levy, D. E. The Zinc Finger Antiviral Protein Acts Synergistically with an Interferon-Induced Factor for Maximal Activity against Alphaviruses. *Journal of Virology* **81,** 13509–13518 (Dec. 2007).

496. Bréhin, A.-C. *et al.* The large form of human 2',5'-Oligoadenylate Synthetase (OAS3) exerts antiviral effect against Chikungunya virus. *Virology* **384,** 216–222 (Feb. 2009).

497. McPherson, R. L. *et al.* ADP-ribosylhydrolase activity of Chikungunya virus macrodomain is critical for virus replication and virulence. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 1666–1671 (Feb. 2017).

498. Jayabalan, A. K., Griffin, D. E. & Leung, A. K. L. Alphavirus nsP3 ADP-ribosylhydrolase Activity Disrupts Stress Granule Formation, 629881 (June 2019).

499. Fehr, A. R., Jankevicius, G., Ahel, I. & Perlman, S. Viral Macrodomains: Unique Mediators of Viral Replication and Pathogenesis. *Trends in Microbiology* **26,** 598–610 (July 2018).

500. Revell, L. J. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* **1,** 319–329 (2010).

501. Gould, E. *et al.* Understanding the alphaviruses: Recent research on important emerging pathogens and progress towards their control. *Antiviral Research* **87,** 111–124 (Aug. 2010).

502. Carlson, C. J. *et al.* Climate change increases cross-species viral transmission risk. *Nature,* 1–1 (Apr. 2022).

503. Digard, P., Lee, H. M., Sharp, C., Grey, F. & Gaunt, E. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. *Virus Evolution* **6,** veaa057 (July 2020).

504. Van Marle, G. *et al.* Arterivirus discontinuous mRNA transcription is guided by base pairing between sense and antisense transcription-regulating sequences. *Proceedings of the National Academy of Sciences* **96,** 12056–12061 (Oct. 1999).

505. Zhang, G., Slowinski, V. & White, K. A. Subgenomic mRNA regulation by a distal RNA element in a (+)-strand RNA virus. *RNA* **5,** 550–561 (Apr. 1999).