Molecular characteristics and transmission dynamics of human respiratory syncytial virus



Gu-Lung Lin

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Abstract

Human respiratory syncytial virus (RSV) is the leading cause of lower respiratory tract infection in young children and an increasingly recognised cause of illness in older adults. To date, the role of viral genetic diversity and infection dynamics in RSV have not been fully understood, and viral genetic studies of clinical samples are critical for developing and deploying preventative measures (e.g., vaccines, monoclonal antibodies) against RSV infection. In this thesis my aim is to identify viral and microbial factors associated with severe RSV disease, to explore RSV diversity during seasonal outbreaks both at the population level and within an individual infection, and to reconstruct the global epidemic history of the recently dominant RSV genotype ON1. RSV genomes and transcriptomes were reconstructed from target-enriched quantitative deep sequencing of >800 RSV-positive nasopharyngeal swabs collected from infants <1 year of age and adults ≥60 years of age in the United Kingdom, Spain, and the Netherland during 2017–2020. In addition to RSV, the sequencing targets included a panel of >100 common bacterial and viral pathogens, enabling simultaneous genotyping of RSV with that of any co-infecting organisms. Two subgroups of RSV, A and B, co-circulated in each RSV season, with RSV-B (genotype BA) dominating the 2017/18 and 2018/19 seasons and RSV-A (genotype ON1) dominating the 2019/20 season. Differential gene expression analysis showed that RSV-A had a lower expression of the P gene (encoding the phosphoprotein, a cofactor of the viral RNA-dependent RNA polymerase) and a higher expression of the open reading frame M2-1 (encoding the M2-1 protein, facilitating transcription of full-length mRNA) than RSV-B. Analysis of the deep sequencing data identified two RSV-B consensus strains and one minor variant carrying amino acid alterations potentially associated with resistance to RSV monoclonal antibodies. RSV-B had greater within-host diversity than RSV-A, while RSV-A exhibited greater population diversity than RSV-B. Further analysis combining metagenomic sequencing data with clinical data showed that RSV-A and RSV-B caused disease of similar severity and exhibited similar viral load dynamics in the infant cohort. Young age and co-infection with *Haemophilus* spp. were significant risk factors for severe RSV disease in infants. Delayed viral clearance was found in infants requiring mechanical ventilation. Transmission dynamics of RSV were examined by reconstructing the epidemic history of genotype ON1 from sequences

generated in this thesis along with publicly available sequences downloaded from GenBank. ON1 was estimated to emerge in the Oceania region between October 2008 and May 2009. After its emergence, it rapidly spread to the Americas and was first detected in Mexico in late 2009. It reached every region of the world within 2 years and gained dominance within 3 years. This thesis provides new insights into the molecular characteristics and transmission dynamics of RSV and emphasises the importance of continued genomic surveillance of RSV to promptly detect emerging variants that may confer resistance to forthcoming RSV preventative measures.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor Professor Andrew Pollard, who gave me the amazing opportunity to work on the multicentre RSV project, and has provided unwavering support and considerable encouragement throughout the various phases of my DPhil study. I would also like to express my heartfelt appreciation to my co-supervisors, Dr Simon Drysdale and Dr Tanya Golubchik. Dr Drysdale assisted me in going through the first year of the DPhil study when I, as an international student, was struggling to keep pace with my studies. The presentation of my studies has benefited enormously from his ongoing support. Dr Golubchik has acquainted me with viral bioinformatics and always offered invaluable insights into my work. Special thanks to Dr Rory Bowden and Professor Nuno Faria, whose help in virus genomics and phylogenetics cannot be overestimated. I would also like to extend my sincere thanks to Dr Daniel O'Connor, who has provided constant support in many details of my work.

The work presented in this thesis would not have been achieved without the extensive support of many of my colleagues at the Oxford Vaccine Group and University of Oxford, and many other colleagues in Europe and the United States working on the REspiratory Syncytial virus Consortium in EUrope (RESCEU) project, especially Amy Beveridge, Elizabeth Clutterbuck, Hannah Robinson, Katja Pfafferott, Kimberley Jefferies, and Matthew Snape at the Oxford Vaccine Group, and Deniz Öner (Janssen Pharmaceutica NV), Koos Korsten (UMC Utrecht), Ryan Thwaites (Imperial College London), and Alberto Gómez Carballa (GENVIP). I have had the great pleasure of working with Joseph McGinley on the RESCEU project, who is always there and willing to help. I would also like to recognise the assistance that I have received from Anthony Brown, George MacIntyre-Cockett, Esther Mellado-Gomez, Mariateresa de Cesare, David Bonsall, and M Azim Ansari. Without their effort, I could not obtain the high-quality sequencing data that were essential to my work.

The completion of this thesis and my DPhil study would not have been possible without the unfaltering support and nurturing of my dearest parents, Ching-Chuan and Chiou-Mei, and my family. I am equally deeply indebted to Thomas for his care, support, and understanding. Last but not least, Kuen-Wei, Mi Jun, Tzu-Hao, Yun-Kai, Ta-Chun, Ming-Shan, Mu-Huan, and Chee Kin have been instrumental in making my life in Oxford more joyful and memorable.

Publications and presentations

Published papers

Lin GL, Golubchik T, Drysdale S, O'Connor D, Jefferies K, Brown A, de Cesare M, Bonsall D, Ansari MA, Aerssens J, Bont L, Openshaw P, Martinón-Torres F, Bowden R, Pollard AJ, and RESCEU Investigators.

Simultaneous viral whole-genome sequencing and differential expression profiling in respiratory syncytial virus infection of infants

Journal of Infectious Diseases 2020;222:S666-S671[1]

Abstract

Targeted metagenomics using strand-specific libraries with target enrichment is a sensitive, generalized approach to pathogen sequencing and transcriptome profiling. Using this method, we recovered 13 (76%) complete human respiratory syncytial virus (RSV) genomes from 17 clinical respiratory samples, reconstructed the phylogeny of the infecting viruses, and detected differential gene expression between 2 RSV subgroups, specifically, a lower expression of the P gene and a higher expression of the M2 gene in RSV-A than in RSV-B. This methodology can help to relate viral genetics to clinical phenotype and facilitate ongoing population-level RSV surveillance and vaccine development.

Lin GL, McGinley JP, Drysdale SB, and Pollard AJ.

Epidemiology and immune pathogenesis of viral sepsis

Frontiers in Immunology 2018;9:2147[2]

Abstract

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection. Sepsis can be caused by a broad range of pathogens; however, bacterial infections represent the majority of sepsis cases. Up to 42% of sepsis presentations are culture negative, suggesting a non-bacterial cause. Despite this, diagnosis of viral sepsis remains very rare. Almost any virus can cause sepsis in vulnerable patients (e.g., neonates, infants, and other immunosuppressed groups). The prevalence of viral sepsis is not known, nor is there enough information to make an accurate estimate. The initial standard of care for all cases of sepsis, even those that are subsequently proven to be culture negative, is the immediate use of broad-spectrum antibiotics. In the absence of definite diagnostic criteria for viral sepsis, or at least to exclude bacterial sepsis, this inevitably leads to unnecessary antimicrobial use, with associated consequences for antimicrobial resistance, effects on the host microbiome and excess healthcare costs. It is important to understand non-bacterial causes of sepsis so that inappropriate treatment can be minimised, and appropriate treatments can be developed to improve outcomes. In this review, we summarise what is known about viral sepsis, its most common causes, and how the immune responses to severe viral infections can contribute to sepsis. We also discuss strategies to improve our understanding of viral sepsis, and ways we can integrate this new information into effective treatment.

Öner D, Drysdale SB, McPherson C, **Lin GL**, Janet S, Broad J, Pollard AJ, Aerssens J, and RESCEU Investigators.

Biomarkers for disease severity in children infected with respiratory syncytial virus: a systematic literature review

Journal of Infectious Diseases 2020;222:S648-S657[3]

Abstract

Background: Clinical manifestations of respiratory syncytial virus (RSV) infection vary widely from mild, self-limiting illness to severe life-threatening disease. There are gaps in knowledge of biomarkers to objectively define severe disease and predict clinical outcomes.

Methods: A systematic search was performed, 1945-March 2019 in databases Ovid Medline, Embase, Global health, Scopus, and Web of Science. Risk of bias was assessed using the Cochrane tool.

Results: A total of 25,132 abstracts were screened and studies were assessed for quality, risk of bias, and extracted data; 111 studies met the inclusion criteria. RSV severity was correlated with antibody titers, reduced T and B cells, dysregulated innate immunity, neutrophil mobilization to the lungs and blood, decreased Th1 response, and Th2 weighted shift. Microbial exposures in respiratory tract may contribute to neutrophil mobilization to the lungs of the infants with severe RSV compared with mild RSV disease.

Conclusions: Although a wide range of biomarkers have been associated with RSV disease severity, robust validated biomarkers are lacking. This review illustrates the broad heterogeneity of study designs and high variability in the definition of severe RSV disease. Prospective studies are required to validate biomarkers. Additional research investigating epigenetics, metabolomics, and microbiome holds promise for novel biomarkers.

Jefferies K, Drysdale SB, Robinson H, Clutterbuck EA, Blackwell L, McGinley J, **Lin GL**, Galal U, Nair H, Aerssens J, Öner D, Langedijk A, Bont L, Wildenbeest JG, Martinón-Torres F, Rodriguez-Tenreiro Sanchez C, Nadel S, Openshaw P, Thwaites R, Widjojoatmodjo M, Zhang L, Nguyen TL, Giaquinto C, Giordano G, Baraldi E, Pollard AJ, and RESCEU Investigators.

Presumed risk factors and biomarkers for severe respiratory syncytial virus disease and related sequelae: protocol for an observational multicenter, case-control study from the respiratory syncytial virus consortium in Europe (RESCEU)

Journal of Infectious Diseases 2020;222:S658-S665[4]

Abstract

Respiratory syncytial virus (RSV) is the leading viral pathogen associated with acute lower respiratory tract infection and hospitalization in children < 5 years of age worldwide. While there are known clinical risk factors for severe RSV infection, the majority of those hospitalized are previously healthy infants. There is consequently an unmet need to identify biomarkers that predict host response, disease severity, and sequelae. The primary objective is to identify biomarkers of severe RSV acute respiratory tract infection (ARTI) in infants. Secondary objectives include establishing biomarkers associated with respiratory sequelae following RSV infection and characterizing the viral load, RSV whole-genome sequencing, host immune response, and transcriptomic, proteomic, metabolomic and epigenetic signatures associated with RSV disease severity. Six hundred thirty infants will be recruited across 3 European countries: the Netherlands, Spain, and the United Kingdom. Participants will be recruited into 2 groups: (1) infants with confirmed RSV ARTI (includes upper and lower respiratory tract infections), 500 without and 50 with comorbidities; and (2) 80 healthy controls. At baseline, participants will have nasopharyngeal, blood, buccal, stool, and urine samples collected, plus complete a questionnaire and 14-day symptom diary. At convalescence (7 weeks $\pm$ 1 week post-ARTI), specimen collection will be repeated. Laboratory measures will be correlated with symptom severity scores to identify corresponding biomarkers of disease severity.

Submitted article

Lin GL, Drysdale S, Snape MD, O'Connor D, Brown A, MacIntyre-Cockett G, Mellado-Gomez E, de Cesare M, Bonsall D, Ansari MA, Öner D, Aerssens J, Butler C, Bont L, Openshaw P, Martinó-Torres F, Nair H, Bowden R, RESCEU Investigators, Golubchik T, and Pollard AJ.

Distinct patterns of within-host virus populations between two subgroups of human respiratory syncytial virus

Nature Communications (accepted in July 2021, under typesetting and production)

Abstract

Human respiratory syncytial virus (RSV) is a major cause of lower respiratory tract infection in young children globally, but little is known about within-host RSV diversity. Here, we characterised within-host RSV populations using deep-sequencing data from 322 nasopharyngeal swabs collected during 2017–2020. RSV-B had lower consensus diversity than RSV-A at the population level, while exhibiting greater within-host diversity. Two RSV-B consensus sequences had an amino acid alteration (K68N) in the fusion (F) protein, which has been associated with reduced susceptibility to nirsevimab (MEDI8897), a novel RSV monoclonal antibody under development. In addition, several minor variants were identified in the antigenic sites of the F protein, one of which may confer resistance to palivizumab, the only licensed RSV monoclonal antibody. The differences in within-host virus populations emphasise the importance of monitoring for vaccine efficacy and may help to explain the variable prevalence of monoclonal antibody-escape mutants between the two subgroups.

Articles in preparation

Lin GL, et al. Host and viral determinants and nasopharyngeal microbiota associated with respiratory syncytial virus disease severity.

Lin GL, et al. Emergence, transmission, and evolutionary dynamics of respiratory syncytial virus A genotype ON1.

Presentations

Lin GL, et al. Targeted metagenomic sequencing of respiratory syncytial virus: a comprehensive method for studying viral genomics and beyond. Oral presentation at the 37th Annual Meeting of the European Society for Paediatric Infectious Diseases (ESPID 2019), Ljubljana, Slovenia, 6–11 May 2019.

Lin GL, et al. Whole-genome reconstruction and transcriptional profiling of human respiratory syncytial virus by targeted metagenomic sequencing. Poster presentation at the 7th Applied Bioinformatics and Public Health Microbiology Conference (ABPHM 2019), Cambridge, United Kingdom, 5–7 June 2019.

Lin GL, et al. Genome-wide evolutionary dynamics of respiratory syncytial virus. Poster presentation at the Global Conference on Novel RSV Preventive and Therapeutic Interventions (RSVVW 2019), Accra, Ghana, 12–14 November 2019.

Lin GL, et al. Molecular epidemiology of respiratory syncytial virus in European countries between 2017 and 2019. Poster presentation at the 38th Annual Meeting of the European Society for Paediatric Infectious Diseases (ESPID 2020), virtual meeting, 26–29 October 2020.

Statement of responsibility

This thesis is based on the results from three sequencing data studies associated with the REspiratory Syncytial virus Consortium in EUrope (RESCEU) project, and a systematic analysis of a collection of global respiratory syncytial virus sequences. The work presented in this thesis is my own work, though it could not have been possible without the considerable support of both the academic and industry partners and the participation of the public in the RESCEU project. In what follows I provide a breakdown of the contributors to the clinical studies and laboratory work. All of the data analyses and interpretations thereof presented in this thesis are my own. Any content that is not solely my own is accordingly indicated.



Dr Gu-Lung Lin
1st August 2021

Clinical studies of the RESCEU project (chapter 2)

Chief investigators: Matthew Snape (birth cohort study), Andrew J Pollard (infant case–control study), Christopher Butler (older adult cohort study)
Investigators: Andrew Pollard, Christopher Butler, Matthew Snape, Simon Drysdale, **Gu-Lung Lin**, Daniel O'Connor, Elizabeth Clutterbuck, Joseph McGinley (University of Oxford, Oxford, United Kingdom); Harish Nair, Harry Campbell, Steve Cunningham, Debby Bogaert (University of Edinburgh, Edinburgh, United Kingdom); Philippe Beutels (Universiteit Antwerpen, Antwerp, Belgium); Louis Bont, Joanne Wildenbeest (University Medical Centre Utrecht, Utrecht, Netherlands); Peter Openshaw, Ryan Thwaites, Dexter Wiseman (Imperial College London, London, United Kingdom); Federico Martinó-Torres, Alberto Gómez-Carballa, Carmen Rodriguez-Tenreiro, Irene Rivero-Calle, Ana Dacosta-Urbieta (Servicio Galego de Saude, Santiago de Compostela, Spain); Terho Heikkinen (University of Turku, Turku, Finland); Adam Meijer (National Institute for Public Health and the Environment, Bilthoven, Netherlands); Thea Kølsen Fischer (Statens Serum Institut, Copenhagen, Denmark); Maarten van den Berge (University of Groningen, Groningen, Netherlands); Carlo Giaquinto (PENTA Foundation, Padua, Italy); Michael Abram (AstraZeneca, Gaithersburg, MD, United States); Philip Dormitzer (Pfizer, Pearl River, NY, United States); Sonia Stoszek (GlaxoSmithKline, Potomac, MD, United States); Scott Gallichan (Sanofi Pasteur, Toronto, Ontario, Canada); Jeroen Aerssens, Deniz Öner (Janssen, Beerse, Belgium); Brian Rosen (Novavax, Potomac, MD, United States); Eva Molero, Nuria Machin, Martina Spadetto (Team-It Research, Barcelona Spain)
Conduction and supervision of the clinical studies at the Oxford site: Hannah Robinson, Katja Pfafferott, Laura Walker, Carrie White, Simon Drysdale, Kimberley Jefferies, Sarah Rhead, Helen Ratcliffe, Matthew Snape, Christopher Butler, Andrew J Pollard
Sample processing at the Oxford site: Gu-Lung Lin, Joseph McGinley, Elizabeth Clutterbuck, Amy Beveridge
Funding: Innovative Medicines Initiative 2 Joint Undertaking (grant number 116019). This Joint Undertaking receives support from the European Union Horizon 2020 Research and Innovation Program and European Federation of Pharmaceutical Industries and Associations.

Targeted metagenomic sequencing (chapter 2)

Protocol design: Mariateresa de Cesare, David Bonsall, Rory Bowden
Probe design: M Azim Ansari
Nucleic acid extraction: Anthony Brown, George MacIntyre-Cockett

Library construction: George MacIntyre-Cockett, Esther Mellado-Gomez, Mariateresa de Cesare
Polymerase chain reaction: Gu-Lung Lin
Funding: National Institute for Health Research (NIHR) Oxford Biomedical Research Centre, the NIHR Thames Valley and South Midlands Clinical Research Network, British Research Council, and Innovative Medicines Initiative 2 Joint Undertaking (grant number 116019). This Joint Undertaking receives support from the European Union Horizon 2020 Research and Innovation Program and European Federation of Pharmaceutical Industries and Associations.

Pilot sequencing data analysis to validate targeted metagenomics (chapter 3)

Study design: Gu-Lung Lin, Tanya Golubchik, Rory Bowden, Andrew J Pollard
Data analysis and interpretation: Gu-Lung Lin
Manuscript preparation: Gu-Lung Lin

Metagenomic analysis on factors associated with severe RSV infection (chapter 4)

Study design: Gu-Lung Lin, Tanya Golubchik, Simon Drysdale, Andrew J Pollard
Data analysis and interpretation: Gu-Lung Lin
Manuscript preparation: Gu-Lung Lin

Within-host virus diversity analysis (chapter 5)

Study design: Gu-Lung Lin, Tanya Golubchik, Andrew J Pollard
Data analysis and interpretation: Gu-Lung Lin
Manuscript preparation: Gu-Lung Lin

Phylogeographic analysis on global respiratory syncytial virus subgroup A sequences (chapter 6)

Study design: Gu-Lung Lin, Nuno Faria
Sequence data compilation: Gu-Lung Lin
Data analysis and interpretation: Gu-Lung Lin
Manuscript preparation: Gu-Lung Lin

Contents

List of Figures **xviii**

List of Tables **xx**

List of Abbreviations **xxii**

1 Respiratory syncytial virus **1**

1.1 Virology . . . 2
1.1.1 Replication and transcription . . . 2
1.1.2 Proteins . . . 4
1.2 Epidemiology and evolution . . . 5
1.2.1 Epidemiology . . . 5
1.2.2 Molecular epidemiology . . . 7
1.2.3 Evolutionary dynamics . . . 8
1.3 Clinical features and treatment . . . 9
1.3.1 Clinical features and diagnosis . . . 9
1.3.2 Risk factors for severe disease . . . 10
1.3.3 Treatment and prognosis . . . 11
1.4 Immune responses to respiratory syncytial virus infection . . . 12
1.4.1 Innate immune cell responses . . . 12
1.4.2 Cellular immunity . . . 15
1.4.3 Humoral immunity . . . 16
1.4.4 T helper type 1 and type 2 responses . . . 17
1.5 Prevention . . . 18
1.5.1 Monoclonal antibodies . . . 18
1.5.2 Vaccines . . . 19
1.6 Thesis outline . . . 23

2 General materials and methods **25**

2.1 REspiratory Syncytial virus Consortium in EUrope (RESCEU) . . . 26
2.1.1 Birth cohort study . . . 26
2.1.2 Infant case–control study . . . 28
2.1.3 Older adult cohort study . . . 30

2.1.4 Definition of disease severity . . . 32
2.1.5 Ethical approval and study conduction . . . 34
2.2 Nucleic acid extraction . . . 34
2.3 Targeted metagenomics . . . 35
2.3.1 Library construction . . . 35
2.3.2 Probe design . . . 36
2.3.3 Sequencing . . . 38
2.4 Genome reconstruction . . . 38
2.5 Polymerase chain reaction . . . 40

3 RSV genome and transcriptome sequencing 41
3.1 Introduction . . . 42
3.2 Methods . . . 43
3.2.1 Sample source and sequencing . . . 43
3.2.2 Viral genome and transcriptome reconstruction . . . 43
3.2.3 Phylogenetic analysis . . . 44
3.2.4 Differential gene expression analysis . . . 45
3.2.5 Statistical analysis . . . 46
3.3 Results . . . 46
3.3.1 Samples and participants . . . 46
3.3.2 RSV genome and transcriptome reconstruction . . . 46
3.3.3 Coverages of genomic and positive-strand reads . . . 52
3.3.4 Relative transcript abundance of RSV genes . . . 53
3.3.5 Differential gene expression analysis . . . 55
3.4 Discussion . . . 59

4 Factors associated with respiratory syncytial virus disease severity 63
4.1 Introduction . . . 63
4.2 Methods . . . 64
4.2.1 Sample source and sequencing . . . 64
4.2.2 Phylogenetic analysis . . . 65
4.2.3 Respiratory virome reconstruction . . . 65
4.2.4 Respiratory bacterial microbiome reconstruction . . . 66
4.2.5 Statistical analyses . . . 67
4.3 Results . . . 68
4.3.1 Study population . . . 68
4.3.2 Sequencing yield of RSV . . . 70
4.3.3 Molecular epidemiology of RSV . . . 73
4.3.4 Respiratory virome composition . . . 78
4.3.5 Respiratory bacterial microbiota . . . 79
4.4 Discussion . . . 83

5 Within-host diversity of respiratory syncytial virus 86
5.1 Introduction . . . 87
5.2 Methods . . . 88
5.2.1 Sample source and sequencing . . . 88
5.2.2 Genome reconstruction . . . 88
5.2.3 Within-host virus diversity analysis . . . 89
5.2.4 Phylogeny reconstruction . . . 90
5.2.5 Statistical analysis . . . 90
5.3 Results . . . 91
5.3.1 Sample population . . . 91
5.3.2 Minor allele frequencies and minor variants . . . 92
5.3.3 Potential antigenic variants . . . 95
5.3.4 Pairwise nucleotide diversity . . . 98
5.3.5 Genetic distance . . . 100
5.3.6 Temporal change of intrahost virus population . . . 101
5.4 Discussion . . . 102

6 The global epidemic history of genotype ON1 108
6.1 Introduction . . . 108
6.2 Methods . . . 110
6.2.1 Sequence compilation . . . 110
6.2.2 Maximum-likelihood phylogeny . . . 113
6.2.3 Molecular clock analysis . . . 114
6.2.4 Phylogeographic analysis . . . 114
6.3 Results . . . 115
6.3.1 Sequence compilation . . . 115
6.3.2 Phylogenetic reconstruction of RSV-A . . . 115
6.3.3 Phylogeographic analysis of ON1 . . . 117
6.4 Discussion . . . 119

7 Summary of the insights and contributions of this thesis 123
7.1 Summary of the novel and important findings resulting from this thesis 124
7.1.1 Transcriptome profiling of respiratory syncytial virus . . . 124
7.1.2 Molecular epidemiology of respiratory syncytial virus in Europe 125
7.1.3 Factors associated with severe RSV disease . . . 126
7.1.4 Patterns of within-host virus diversity . . . 127
7.1.5 Emergence and global spread of genotype ON1 . . . 127
7.2 Limitations . . . 127
7.3 Implications and future directions . . . 130
7.4 Closing remarks . . . 133

Appendices

A Supplementary tables **135**
A.1 Supplementary table for RSV genome and transcriptome sequencing (chapter 3) . . . 136
A.2 Supplementary tables for factors associated with RSV disease severity (chapter 4) . . . 137
A.3 Supplementary table for RSV within-host diversity (chapter 5) . . . 142
A.4 Supplementary tables for global epidemic history of genotype ON1 (chapter 6) . . . 143

B Supplementary figures **151**
B.1 Supplementary figures for RSV genome and transcriptome sequencing (chapter 3) . . . 152
B.2 Supplementary figures for factors associated with RSV disease severity (chapter 4) . . . 158
B.3 Supplementary figures for RSV within-host diversity (chapter 5) . . . 160
B.4 Supplementary figures for global epidemic history of genotype ON1 (chapter 6) . . . 166

C Supplementary methods and results **169**
C.1 Supplementary methods for RSV transcriptomics . . . 169
C.2 Preliminary results for RSV transcriptomics . . . 170

References **174**

List of Figures

1.1 RSV genome and structural proteins 3
1.2 Fusion protein and antigenic sites 6
1.3 Human immune responses to RSV infection 13
1.4 RSV vaccine and monoclonal antibody snapshot 22

2.1 RSV genome reconstruction 39

3.1 Mean percent identity of mapped reads for each depth of coverage . 48
3.2 Sequencing yield and virus load 49
3.3 Maximum-likelihood phylogenies of RSV-A and RSV-B 52
3.4 Coverage plots of RSV genomes and transcriptomes 54
3.5 Positive-strand read coverage across the M2 gene 55
3.6 Relative transcript abundance of RSV genes 56
3.7 Evaluation of differential gene expression 57
3.8 Differential gene expression 58

4.1 Correlation between the number of RSV reads and RSV viral load . 72
4.2 Predicted probability of fever 73
4.3 Temporal change of RSV read counts 74
4.4 Maximum-likelihood phylogenies of RSV-A and RSV-B 77
4.5 Composition of bacteria identified in the nasopharynx of the infants 81

5.1 Adjusting for batch effects on cumulative minor allele frequency . . 94
5.2 Minor variants in the coding regions of the F and G genes 96
5.3 Standardised pairwise nucleotide diversity and pairwise genetic distances 100
5.4 Temporal change of pairwise nucleotide diversity 102
5.5 Temporal change of minor alleles 103

6.1 Geographical distribution of the RSV-A sequences 112
6.2 Maximum-likelihood phylogeny of Dataset 1 116
6.3 Maximum clade credibility tree of Dataset 1 117
6.4 Estimated evolutionary rate and time of the common ancestor . . . 118
6.5 Frequency plots since the emergence of genotype ON1 119

B.1 Coverages of genomic reads and positive-strand reads 152
B.1 Coverages of genomic reads and positive-strand reads (continued) . 153
B.2 Relative transcript abundance of RSV genes 154
B.3 Relative transcript abundance for each sample 155
B.3 Relative transcript abundance for each sample (continued) 156
B.4 Gene expression profiles between samples with complete transcriptome assembly 157
B.5 Phylogenetic reconstruction of human herpesvirus 6 158
B.6 Venn diagram of the three main bacterial genera 159
B.7 Percentage of the samples passing read depth threshold 160
B.8 Correlation between number of minor variants and read count . . . 161
B.9 Correlation between pairwise nucleotide diversity and other variables 162
B.10 Gene-wise comparisons of standardised pairwise nucleotide diversity 163
B.11 Maximum-likelihood phylogenies of RSV-A and RSV-B 164
B.12 Pairwise Manhattan distances 165
B.13 Flowchart of the sequence compilation 166
B.14 Geographical and temporal distribution of the sequences 167
B.15 Regression of root-to-tip divergence on sampling time 168

C.1 Multidimensional scaling plots 171
C.2 Differential gene expression 172

List of Tables

2.1 Primary and secondary objectives of the birth cohort study. 27
2.2 Primary and secondary objectives of the infant case–control study. . 30
2.3 Primary and secondary objectives of the older adult cohort study. . 31
2.4 ReSVinet scale. 33
2.5 Target bacterial and viral species. 36

3.1 Demographic and clinical characteristics of the study participants. . 47
3.2 Viral load, sequencing yield, and characteristics of RSV for each sample. 51

4.1 Characteristics of the RSV-infected infants by study group. 69
4.2 Clinical characteristics of the RSV-infected infants, stratified by age group. 71
4.3 Associations between clinical variables and peak viral burden. . . . 72
4.4 Associations between clinical variables and daily percentage decrease in viral burden. 75
4.5 Characteristics of the RSV-infected infants by RSV subgroup. . . . 76
4.6 The presence and absence of other viruses in the RSV-infected infants (N = 433). 78
4.7 Demographic and clinical features of the RSV-infected infants with and without co-detection of any other virus. 80
4.8 Demographic and clinical features of the RSV-infected infants with and without co-detection of any *Haemophilus* sp. 82

5.1 Characteristics of RSV samples by subgroup. 93
5.2 Characteristics of minor variants within the antigenic sites of the fusion protein. 97
5.3 A multiple linear regression model to evaluate the association between pairwise nucleotide diversity and other variables. 99

A.1 Accession numbers of sequences included in phylogenetic analysis. . 136
A.2 Criteria for the presence status of each target virus. 137
A.3 Number of RSV-infected infants from each country and each season. 138

A.4 Demographic characteristics of the RSV-infected infants, stratified by age group. 139
A.5 Incidences of the two RSV subgroups in each country and season. . 140
A.6 The presence and absence of each identified bacterial genus in the RSV-infected infants (N = 433). 141
A.7 Characteristics of RSV samples by sequencing batch. 142
A.8 GenBank accession numbers of the 1,171 complete RSV-A sequences. 143
A.9 Model selection using marginal likelihood estimation. 149
A.10 Time and location of the emergence of genotype ON1. 150

C.1 Differential expression of RSV genes and open reading frames (N = 406). 173

List of Abbreviations

ADCC . . . . . Antibody-dependent cell-mediated cytotoxicity

AFR . . . . . African region

ANCOVA . . . Analysis of covariance

API . . . . . Application programming interface

APRIL . . . . A proliferation-inducing ligand

ARTI . . . . . Acute respiratory tract infection

BAFF . . . . . B cell-activating factor

BAL . . . . . Bronchoalveolar lavage

BAM . . . . . Binary alignment/map

BSSVS . . . . Bayesian stochastic search variable selection

CCL-5 . . . . . C-C motif chemokine ligand 5

cDC . . . . . Conventional dendritic cell

CDR . . . . . Complementarity-determining region

CXCL . . . . . C-X-C motif chemokine ligand

DC . . . . . Dendritic cell

dN/dS . . . . . Ratio of nonsynonymous to synonymous substitutions

EAR . . . . . East Asian region

EMR . . . . . Eastern Mediterranean region

ECP . . . . . Eosinophil cationic protein

EUR . . . . . European region

Fc . . . . . Fragment crystallisable

GE . . . . . Gene end

GP . . . . . General practitioner

GS . . . . . Gene start

HHV-6 . . . . Human herpesvirus 6

HIV . . . . . Human immunodeficiency virus

HKY . . . . . Hasegawa-Kishino-Yano

HPD . . . . . Highest posterior density

HRQoL . . . . Health-related quality of life

HRV . . . . . Human rhinovirus

GTR . . . . . General time reversible

IFN . . . . . Interferon

IL . . . . . Interleukin

IQR . . . . . Interquartile range

LAC . . . . . Latin American and the Caribbean region

LOD . . . . . Limit of detection

mAb . . . . . Monoclonal antibody

MAF . . . . . Minor allele frequency

MCMC . . . . Markov Chain Monte Carlo

MDS . . . . . Multidimensional scaling

MRCA . . . . Most recent common ancestor

mTOR . . . . Mammalian target of rapamycin

NAR . . . . . North American region

NET . . . . . Neutrophil extracellular trap

NK . . . . . Natural killer

NNI . . . . . Nearest neighbour interchange

nt . . . . . Nucleotide

OCE . . . . . Oceania region

OR . . . . . Odds ratio

ORF . . . . . Open reading frame

PCR . . . . . Polymerase chain reaction

PD-1 . . . . . Programmed cell death 1

pDC . . . . . Plasmacytoid dendritic cell

PS . . . . . Path sampling

R . . . . . Correlation coefficient

$\mathbf{R}^2$ . . . . . Coefficient of determination

RADT . . . . . Rapid antigen detection test

RANTES . . . Regulated on activation, normal T cell expressed and secreted

RESCEU . . . REspiratory Syncytial virus Consortium in EUrope

RdRp . . . . . RNA-dependent RNA polymerase

rMLST . . . . Ribosomal multilocus sequence typing

RNA-seq . . . Transcriptome sequencing

RSV . . . . . Human respiratory syncytial virus

RT-PCR . . . Reverse transcription polymerase chain reaction

SARS-CoV-2 . Severe acute respiratory syndrome coronavirus 2

SD . . . . . Standard deviation

SEA . . . . . South-East Asian region

SPR . . . . . Subtree pruning and regrafting

SS . . . . . Stepping-stone sampling

Th1 . . . . . T helper type 1

Th2 . . . . . T helper type 2

TLR . . . . . Toll-like receptor

TMM . . . . . Trimmed mean of M values

TMRCA . . . Time of the most recent common ancestor

TNF . . . . . Tumour necrosis factor

TPM . . . . . Transcripts per kilobase million

TRAIL . . . . TNF-related apoptosis-inducing ligand

UCLN . . . . . Uncorrelated lognormal

UDI . . . . . Unique dual indexing

Γ . . . . . Gamma-distributed rate heterogeneity among sites

1

Respiratory syncytial virus

Contents

1.1 Virology 2
1.1.1 Replication and transcription 2
1.1.2 Proteins 4
1.2 Epidemiology and evolution 5
1.2.1 Epidemiology 5
1.2.2 Molecular epidemiology 7
1.2.3 Evolutionary dynamics 8
1.3 Clinical features and treatment 9
1.3.1 Clinical features and diagnosis 9
1.3.2 Risk factors for severe disease 10
1.3.3 Treatment and prognosis 11
1.4 Immune responses to respiratory syncytial virus infection 12
1.4.1 Innate immune cell responses 12
1.4.2 Cellular immunity 15
1.4.3 Humoral immunity 16
1.4.4 T helper type 1 and type 2 responses 17
1.5 Prevention 18
1.5.1 Monoclonal antibodies 18
1.5.2 Vaccines 19
1.6 Thesis outline 23

1.1 Virology

Human respiratory syncytial virus (RSV) is an enveloped virus with a non-segmented, single-stranded, negative-sense RNA genome of approximately 15.2 kilobases. A new virus taxonomy was proposed for RSV in 2017[5]. RSV was formerly a member in subfamily *Pneumovirinae* within the family *Paramyxoviridae*. The *Pneumovirinae* has now been reclassified as a family, named *Pneumoviridae*. The new scientific name of RSV is *Human orthopneumovirus*, a member of genus *Orthopneumovirus*, family *Pneumoviridae*, and order *Mononegavirales*. *Bovine orthopneumovirus* and *Murine orthopneumovirus* are the other two species in genus *Orthopneumovirus*. Other viruses in the family *Pneumoviridae* include *Human metapneumovirus* and *Avian metapneumovirus*.

The RSV genome contains 10 genes: NS1, NS2, N, P, M, SH, G, F, M2, and L from the 3' end to the 5' end of the genome (Fig. 1.1, panel A). Each gene is flanked by a conserved gene-start (GS) and gene-end (GE) signals[6]. Additionally, a conserved leader and trailer sequences are located at the 3' and the 5' end of genome, respectively. Neighbouring genes are separated by an intergenic region of variable length, except for an overlap between the M2 and L genes by 68 nucleotides (nt)[7]. Each gene has an open reading frame (ORF) and encodes a single protein. An exception is the M2 gene, which contains two overlapping ORFs, M2-1 and M2-2, by 11–32 nt, and thus encodes two proteins (Fig. 1.1, panel B)[8].

1.1.1 Replication and transcription

RNA replication starts with synthesising a low abundance of the positive-sense replicative intermediate, also called the antigenome, from the 3' end to the 5' end of the genome. This process is followed by synthesis of genome replicates using the antigenome as a template. Both processes require the RNA-dependent RNA polymerase (RdRp) complex, constituted by the L protein (large polymerase), P protein (phosphoprotein, as a cofactor), and N protein (nucleoprotein)[10]. The M2-2 protein has been known as an RNA synthesis regulatory protein, which promotes viral RNA replication[11].

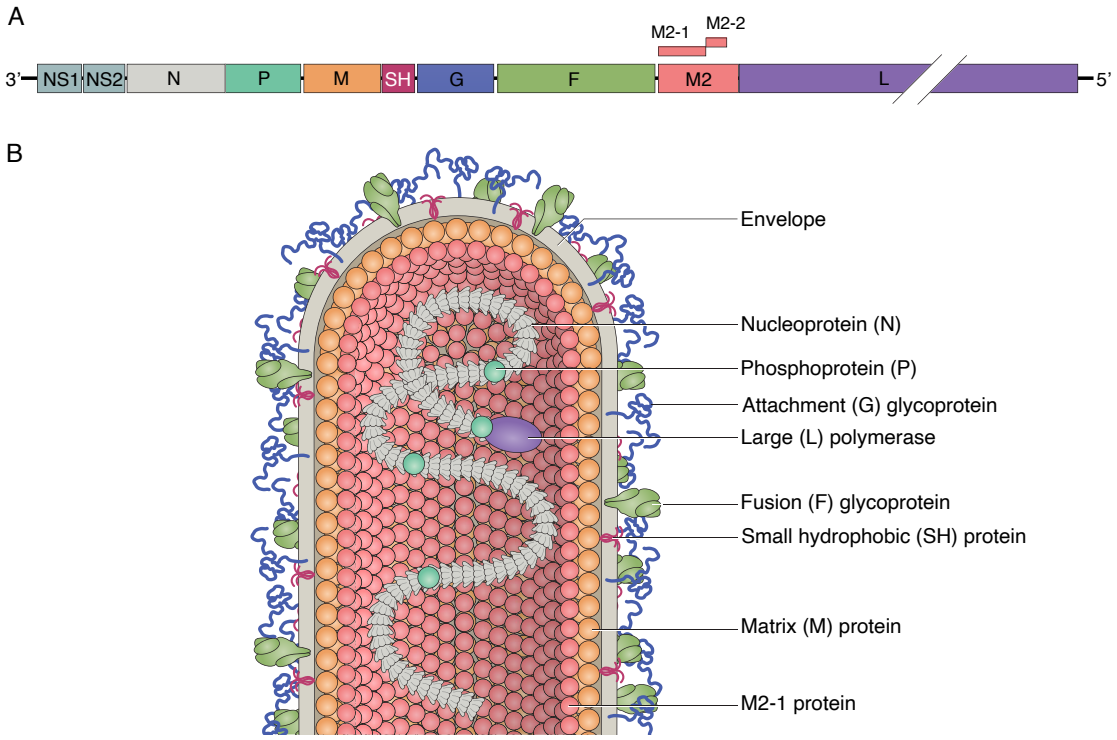


Figure 1.1: RSV genome and structural proteins. **A.** The single-stranded, negative-sense RNA genome contains 10 genes. The M2 gene has two overlapping open reading frames, M2-1 and M2-2. **B.** The structural proteins of RSV. The gene that encodes the protein is in brackets. Adapted from Battles MB, et al. Nat Rev Microbiol 2019;17:233–45[9], with permission (licence number 5066640782501).

The RdRp is also responsible for RNA transcription, but requires an additional protein, the M2-1 protein[12]. The RdRp starts from the 3' end and transcribes each gene sequentially till the 5' end of the genome. It initiates transcription in the response to the GS signal of each gene, and directs termination and polyadenylation at the GE signal[6]. The RdRp continues to move across the intergenic region, and reinitiates transcription at the next GS signal. The M2-1 protein acts as an anti-terminator, which prevents the RdRp from disengagement in the middle of each gene and facilitates the production of full-length mRNAs (i.e., intragenic anti-termination)[13]. The M2-1 protein also facilitates the production of read-through mRNAs (i.e., intergenic anti-termination), where the RdRp fails to respond to the GE signal, continuing to transcribe the intergenic sequence and the next gene[14]. In some cases, the RdRp disengages from the template in an intergenic region, but

the transcription cannot be reinitiated in the middle of the genome[14]. This results in a transcription gradient—genes located nearer the 3' end are being transcribed more than genes located closer to the 5' end of the genome[15]. Therefore, the mRNAs of the 3'-most NS1 gene are the most abundant whereas the mRNAs of the 5'-most L gene are the least abundant. The transcription gradient has been well described in the cell free system[15].

1.1.2 Proteins

The non-structural proteins, NS1 and NS2, inhibits type I and III interferon (IFN) signalling and apoptosis of infected cells[16], thus facilitating virus proliferation. The M (matrix) protein is located beneath the viral envelope and provides structural support. RSV has three transmembrane surface glycoproteins: G (attachment), F (fusion), and SH (small hydrophobic) proteins. The G protein is a type II glycoprotein, responsible for viral attachment[17], whereas the F protein is a type I glycoprotein, mediating the fusion of host cell and viral membranes. The G and F proteins are the only two targets for human neutralising antibodies on the viral surface[18].

Among the 11 RSV proteins, the G protein is the most variable[19]. The extracellular portion (ectodomain) of the G protein contains a conserved central domain, flanked by two variable mucin-like regions (also called hypervariable regions)[9]. The characteristic feature of the mucin-like regions is a high content of serine, threonine, and proline, together accounting for ~40% of the total amino acid composition[20]. Due to this composition, the G protein is heavily O-glycosylated, which potentially contributes to viral evasion of host immunity[18]. The antigenic sites in the G protein can be classified into three groups: (1) conserved epitopes that are present in all strains; (2) group-specific epitopes that are shared by viruses within the same subgroup; (3) strain-specific or variable epitopes that are uniquely found in some strains only[19]. The conserved and group-specific epitopes have been mapped to the conserved central domain by linear epitope mapping using escape

mutants and synthetic peptides. The strain-specific epitopes are mainly mapped to the C-terminal third (the second hypervariable region) of the G protein[19].

Crystal structural analysis of the conserved central domain of the G protein discovered two antigenic sites $\gamma 1$ and $\gamma 2$, bound by broadly neutralising monoclonal antibodies (mAbs) 3D3 and 2D10 respectively[21]. Studies have shown that the conserved epitopes can induce robust immunity and protection against RSV in mice[22,23]. A recombinant subunit RSV vaccine (BBG2Na, comprising the central fragment of the G protein conjugated to part of the streptococcal protein G) showed promising results in a phase I trial[24]. In contrast, the surface structure outside the conserved central domain has yet to be well defined, but studies have shown that multiple neutralising epitopes are also present in the mucin-like regions and could be potential targets for RSV vaccines and mAbs[25,26].

In contrast to high variability of the G protein, the F protein is rather conserved and has thus attracted more interest for the development of vaccines and mAbs. We also have greater understanding of the structural biology of the F protein than that of the G protein. The F protein can rearrange from a metastable prefusion conformation to a highly stable postfusion conformation[27]. Several neutralising epitopes have been located through studies on the crystal structures of the F protein in complex with different mAbs (Fig. 1.2). Antigenic sites II and IV can be found in both prefusion and postfusion conformations, whereas antigenic sites ∅ and V are prefusion-specific[28,29]. Antibodies that bind exclusively to prefusion-specific antigenic sites have been shown to have stronger neutralising activity than antibodies that bind to antigenic sites shared on both confirmations[30]. Recently, stabilised prefusion F subunits used as a vaccine antigen[31,32] and prefusion-specific mAbs[33] have shown promising results in clinical trials (see section 1.5).

1.2 Epidemiology and evolution

1.2.1 Epidemiology

RSV is the leading cause of acute lower respiratory tract infection in young children worldwide. Nearly all children are infected with RSV at least once by the age

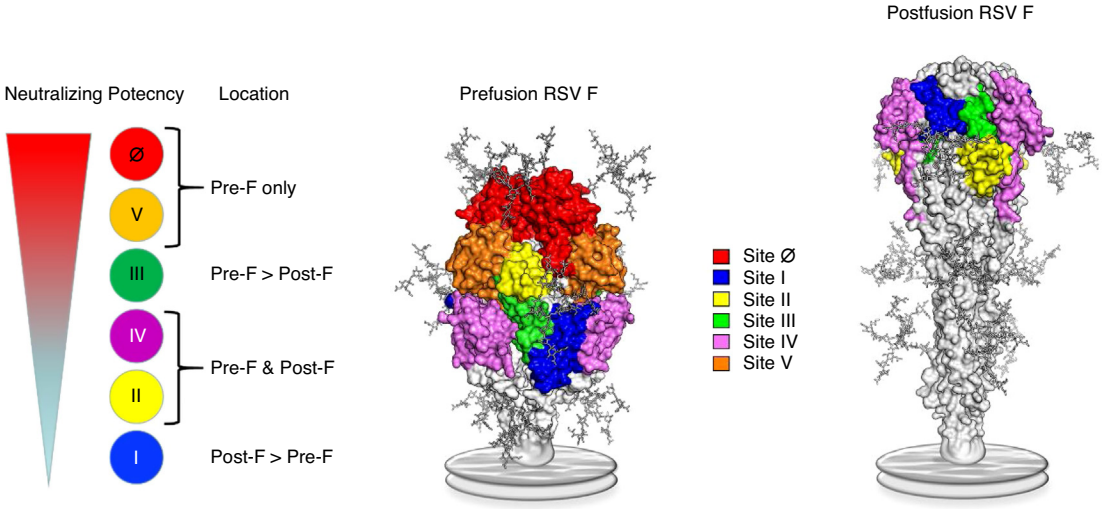


Figure 1.2: Two conformations of the RSV fusion (F) protein and antigenic sites. Antigenic sites ∅ and V have the greatest neutralisation potency and can only be seen in the prefusion conformation. Reproduced from Graham BS, et al. Curr Opin Virol 2017;23:107–112[34], with permission (licence number 5072100249726).

of 2 years[35]. The annual RSV epidemic causes an estimated 33 million episodes of acute lower respiratory tract infection, 3 million hospitalisations, and up to 75,000 in-hospital deaths in children under 5 years of age globally, with mortalities predominantly in low- and middle-income countries[36]. The Global Burden of Disease Study estimated that RSV is responsible for 6.7% of post-neonatal infant deaths, the second most common cause following malaria, which accounts for 11.8% of the deaths[37].

Natural RSV infections only confer transient immune protection against subsequent RSV infections, so repeated infection with RSV is common beyond childhood[38]. School-aged children have been shown to be a frequent introducer of this virus into households[39]. Although repeated RSV infection generally results in self-limited upper respiratory infections, the health impact on adult populations has been increasingly recognised. Elderly adults ≥65 years old and adults who have chronic heart or lung disease are at substantial risk for severe RSV infection[40]. Globally, RSV causes an estimated 336,000 hospital admissions for lower respiratory tract infection and up to 50,000 in-hospital deaths in adults ≥65 years old every year[41].

The burden of RSV infection is similar to that of influenza in older populations where the prevalence of influenza vaccination is high[40,42].

RSV epidemics generally occur once a year[43,44]. A biannual[45,46] or biennial[47,48] pattern of RSV epidemics can be seen in some regions, and persistent RSV transmission with multiple peaks all year round can also be found in other countries, such as Taiwan[49]. In temperate regions, the peak RSV transmission typically coincides with the winter season when temperatures are low and relative humidity is high[50], and occurs before the influenza virus epidemic[43]. In tropical areas, RSV epidemics are more variable[44,50], with an average onset in late summer months[43]. In the United Kingdom, RSV epidemics started in October, peaked in December, and finished in the beginning of February during 2017–2020[51]. In the Netherlands and Spain, RSV epidemics usually start and finish 1–2 months later than that in the United Kingdom[52,53]. Nonpharmaceutical interventions (e.g., social distancing, face masks) imposed to control the current Covid-19 pandemic have substantially reduced RSV transmission in the past year[54,55], with potential RSV outbreaks likely in future years due to increased population susceptibility[55].

1.2.2 Molecular epidemiology

RSV can be classified as two subgroups, A and B, on the basis of their reaction with a panel of mAbs[19]. Each subgroup can be further divided into multiple genotypes by phylogenetic analysis, traditionally using the second hypervariable region of the G gene[56,57]. At least 11 RSV-A genotypes (GA1 to GA7[56,58], SAA1[57], NA1, NA2[59], and ON1[60]) and 17 RSV-B genotypes (GB1 to GB4[56], SAB1 to SAB3[57], and BA1 to BA10[61,62]) have been reported. Nonetheless, there is no standardised definition or nomenclature for RSV genotyping, which complicates data comparison between studies and hinders viral evolutionary studies. Recent studies have proposed new classification schemes based on phylogenies of the G ectodomain or whole-genome sequences[63,64], which provide a higher resolution than that of the second hypervariable region of the G gene[65,66].

The two subgroups, RSV-A and RSV-B, co-circulate in annual epidemics with a biennial, triennial, or quadrennial alternating pattern of subgroup dominance[67]. RSV-A is usually more prevalent than RSV-B[67]. In Belgium, for example, two-year dominance of RSV-A is followed by one-year dominance of RSV-B[68]. In Finland, subgroup dominance alternates every 2 years[47]. The interplay between host immune response and virus genetic variability is believed to play a role in the subgroup shifts. Hypotheses of the dominance of RSV-A over RSV-B include greater genetic diversity[69] and higher transmissibility[70] of RSV-A than that of RSV-B, and more transient host immunity against RSV-A than against RSV-B[68].

Several genotypes within RSV-A and RSV-B also co-circulate in annual epidemics, with new genotypes emerging and old ones disappearing constantly[60,71,72]. Each epidemic usually has one or two dominant genotypes or strains, with the dominant ones being replaced the next year[73]. Short-term herd immunity has been proposed to be the cause of yearly changes in genotype or strain prevalence[19,73]. However, statistical analysis suggests that adaptive host immunity plays a limited role in annual genotype replacement, and strain replacement follows nonselective epidemiological processes[74]. Strains collected in the same year tend to be placed in several phylogenetic clades and have a closer phylogenetic relationship than those collected in different years regardless of sampling locations[19,75,76], suggesting concurrent circulation of multiple strains and rapid global spread and evolution of RSV.

In recent years, genotypes ON1 and BA have become the dominant genotypes for RSV-A and RSV-B, respectively[77–79]. Both genotypes have a distinctive sequence duplication in the second hypervariable region of the G gene, specifically 72-nt for ON1 and 60-nt for BA[60,72]. The emergence and establishment of these two genotypes again exemplify the genetic variability of the G gene and the constant evolution of RSV.

1.2.3 Evolutionary dynamics

As with other RNA viruses, RSV exhibits rapid evolution[80]. Host immunity drives RSV genetic changes, with the G gene experiencing the greatest immune pressure

and being the primary target for diversifying selection[74]. This has been demonstrated by selective pressure analysis and evolutionary analysis. The G gene has the highest ratio of nonsynonymous to synonymous substitutions (dN/dS)[73], followed by the SH and M2-2 genes[81], indicating positive selection acting on these genes. Consistent with the selection pressure analysis, these three genes also share the least amino acid identity between the two subgroups[82]. Sites undergoing positive selection in the G gene include potential N- and O-glycosylation sites[74]. The evolutionary rate of the G gene is estimated at 2.22×10^{-3} substitutions/site/year and 2.78×10^{-3} substitutions/site/year for RSV-A and RSV-B respectively, highest among all RSV genes[82]. The high rate of genetic changes in the G gene has the potential to trigger antigenic drifts[83]. One mechanism could be changes of glycosylation in the G protein resulting in immune evasion of the virus[84].

RSV-B has been shown to have a higher genome-wide evolutionary rate than RSV-A, with $5.68–6.47 \times 10^{-4}$ substitutions/site/year for RSV-A[82,85,86] and $7.47–7.76 \times 10^{-4}$ substitutions/site/year for RSV-B[82,86]. However, one study showed a similar rate between the two subgroups[87]. The most recent common ancestors (MRCA) of RSV-A and RSV-B were dated to 1937–1964 and 1964–1993 depending on the time frames of the datasets analysed[82,86,87]. Although the estimated times of the MRCA vary between studies, the MRCA of RSV-A is consistently older than that of RSV-B, indicating RSV-B carrying lower genomic diversity than RSV-A at the population level[82].

1.3 Clinical features and treatment

1.3.1 Clinical features and diagnosis

RSV causes respiratory tract infection with clinical presentations ranging from asymptomatic or mild upper respiratory infection (e.g., cough and coryza) to severe lower respiratory tract infection (e.g., bronchiolitis and pneumonia), respiratory distress, and even death[88]. Clinical suspicion of RSV infection is usually based on compatible epidemiological and clinical features (e.g., age, known circulation of RSV, and lower respiratory tract disease). Rales and wheezing on auscultation are

typical findings of a physical examination in young children with lower respiratory tract infection. High fever is not very common in RSV infection[88].

Confirmation of RSV infection requires laboratory tests, including virus culture, immunofluorescence, rapid antigen detection test (RADT), and reverse transcriptase polymerase chain reaction (PCR). RADTs are widely used in clinical settings with an overall sensitivity and specificity of 80% and 97% respectively[89]. However, some studies showed a sensitivity below 30%[90,91], and immunoprophylaxis with palivizumab (anti-RSV mAb) may interfere with the performance of RADTs and cause false-negative results[92]. PCR tests have the highest sensitivity and specificity, usually above 90%[93]; however, they are not widely available in resource-limited regions. Multiplex PCR systems for multi-pathogen detection have been increasingly used in clinical laboratories, and RSV is usually included in the panels[94].

1.3.2 Risk factors for severe disease

Most children with RSV infection are previously healthy[95]. However, children under 2 years of age tend to have more severe disease, especially those under 6 months of age and those with a history of prematurity[95]. In addition to young age and prematurity, infants who are at substantial risk for severe RSV disease include those who have haemodynamically significant congenital heart disease or bronchopulmonary dysplasia[96]. As mentioned in section 1.2.1, elderly adults and adults with chronic heart or lung disease are also subject to severe RSV disease.

Other risk factors related to viral types, viral load, or host genetic predisposition are less well defined. A majority of studies have demonstrated that RSV-A caused more severe infection than RSV-B[97,98], but others showed the opposite[99] or similar disease severity in both subgroups[100,101]. Differences in genotype-specific virulence have also been reported. For example, in one study genotype GA3 caused more severe disease than genotypes GA2 and GA4[101], and in another study NA1 was more virulent than genotype ON1[100]. Consistent associations have yet to be established. High RSV load has been shown to be associated with increased length of hospital stay, requirement of intensive care and mechanical ventilation, and/or a high clinical

severity score in young children[102–104]. Other studies, however, found that children with low RSV load are more likely to require hospitalisation or have respiratory failure than those with high viral load[105,106].

Host genetic variations have been linked with an increased risk of severe RSV disease. The majority of these variations are associated with the regulation of host immune responses, particularly chemokine, cytokine, and pattern recognition receptor pathways[107]. For example, −2459G and −2554T variants of the *CCR5* gene[108], T280M polymorphism of the *CX3CR1* gene[109], and −1112C/T variation of the *IL13* gene[110] have been shown to be associated with severe RSV bronchiolitis. Thus, severity of RSV infection relates to viral determinants, host demographic characteristics, host genetic predisposition, and the interplay between these factors. I will evaluate the host demographic and viral determinants for severe RSV disease among an infant cohort in chapter 4.

1.3.3 Treatment and prognosis

Ribavirin is the only licensed antiviral for RSV infection. However, its use is limited and controversial due to uncertain clinical efficacy, potential teratogenicity, and high cost[111,112]. The standard of care for RSV infection remains supportive care with fluid support and supplemental oxygen, as recommended by the American Academy of Pediatrics[113] and the National Institute for Care Excellence[114]. Randomised trials, systematic reviews, and meta-analyses do not support the routine use of bronchodilators (oral or inhaled)[115,116], glucocorticoids (systemic or inhaled)[113,117,118], inhaled hypertonic saline[119], or leukotriene inhibitors[120] in patients with acute bronchiolitis. Many novel therapeutics are currently in development with at least eight antivirals in clinical trials[121]. Preventative measures, including mAbs and vaccines, are discussed in detail in section 1.5. Most patients with RSV infection fully recover without sequelae, but there is ample evidence of the association between RSV-associated lower respiratory tract infection in early childhood and the development of recurrent wheeze and asthma in later childhood up to early adulthood[122–124].

Several studies have evaluated the causal relationship between RSV infection during infancy and subsequent wheeze and asthma. A double-blind, placebo-controlled, randomised trial showed that palivizumab prophylaxis in otherwise healthy preterm infants born at a gestational age of 33 to 35 weeks reduced the number of recurrent wheeze in the first year of life[125]. A follow-up study of the same cohort, however, did not find a significant difference in the incidence of asthma and lung function at the age of 6 years[126]. One case-control study showed that palivizumab prophylaxis reduced the incidence of recurrent wheeze, but not of atopic asthma, in preterm infants born at 33 to 35 weeks' gestational age up to 6 years[127]. Another double-blind, placebo-controlled, randomised trial showed no effect of prophylaxis with motavizumab (another anti-RSV mAb) in healthy term infants on rates of wheeze during the first 3 years[128]. Further studies are required to establish the causal link between RSV infection and subsequent wheeze and asthma.

1.4 Immune responses to respiratory syncytial virus infection

The following sections will discuss the innate immune cell responses, cellular and humoral immune responses, and T helper type 1 (Th1) and type 2 (Th2) responses to RSV infection. Figure 1.3 provides a summary of human immune responses.

1.4.1 Innate immune cell responses

Strong systemic and respiratory tract neutrophil responses predominate the initial phase of RSV infection[130,131]. These neutrophils express myeloid activation markers, such as CD11b and CD18[132], and produce neutrophil elastase, contributing to airway obstruction[133]. The peak neutrophil responses coincide with maximum disease severity and viral load[134]. Substantial pulmonary neutrophil infiltration is evident in severe cases of RSV infection[130,135]. Neutrophil extracellular trap (NET) formation (NETosis) and neutrophil apoptosis, both of which impede the spread of the virus, can be seen in nasopharyngeal and bronchoalveolar lavage (BAL) fluid during acute infection[132,136]. However, excessive NETosis can also cause

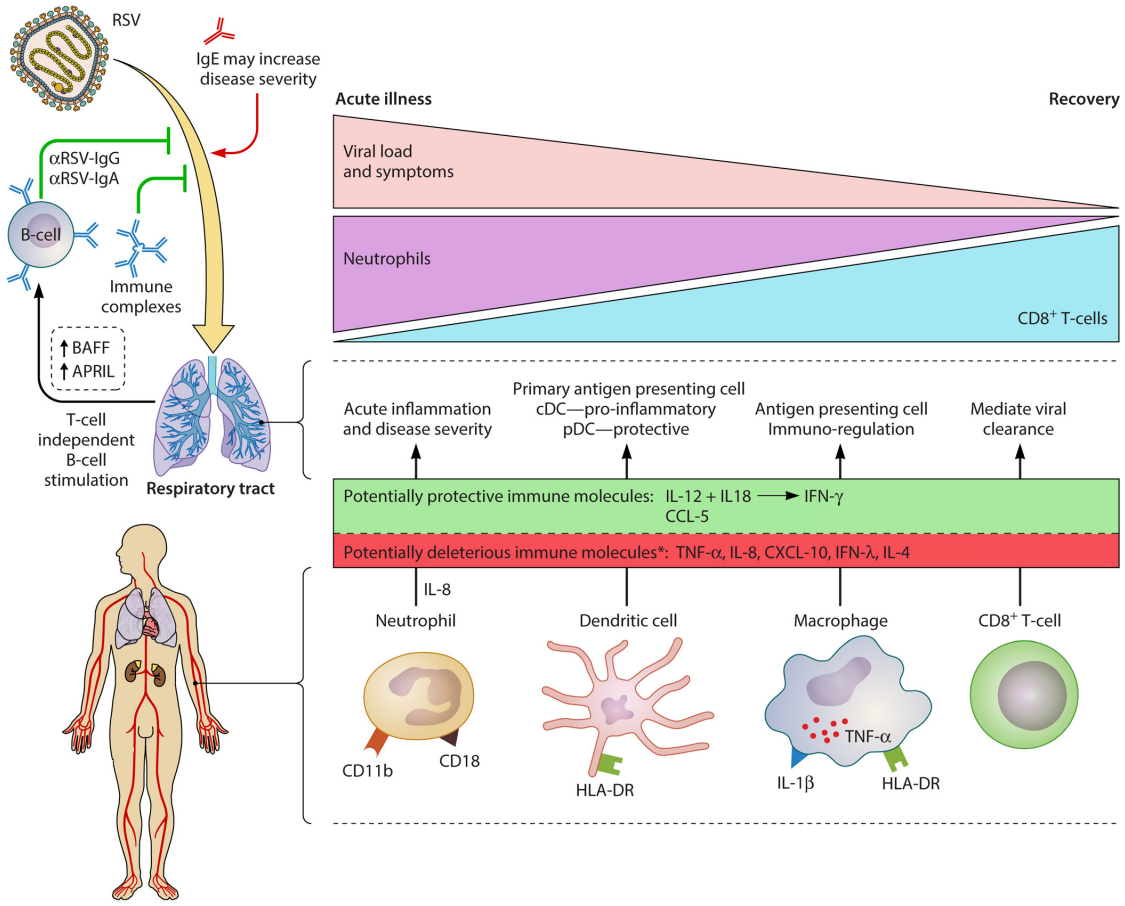


Figure 1.3: Human immune responses to respiratory syncytial virus (RSV) infection. APRIL denotes a proliferation-inducing ligand; BAFF, B cell-activating factor; CCL-5, C-C motif chemokine ligand 5; cDC, conventional dendritic cell; CXCL-10, C-X-C motif chemokine ligand 10; IFN, interferon; IL, interleukin; pDC, plasmacytoid dendritic cell; and TNF, tumour necrosis factor. Adapted from Russell CD, et al. Clin Microbiol Rev 2017;30:481–502[129], with permission.

airway occlusion[136]. Systemic and respiratory tract responses of IL-8, a neutrophil chemoattractant, also increase during RSV lower respiratory tract infection and correlate with disease severity[137–139].

Dendritic cells (DCs) are the primary antigen-presenting cells during RSV infection since conventional DCs (cDCs) and plasmacytoid DCs (pDCs), but not monocytes, migrate to the nasal mucosa from the circulation early in infection[140,141]. In addition, the RSV F protein can be seen within HLA-DR$^+$ DCs in the nasal mucosa[140]. Both DCs have also been found in the lower airways of infants with severe RSV infection requiring mechanical ventilation[142]. The number of pDCs

in blood inversely correlates with disease severity, suggesting a protective role of pDCs[143]. In contrast, cDC counts correlate with levels of proinflammatory cytokines (interleukin-6 [IL-6], tumour necrosis factor alpha [TNF-α], and IL-8) and T cell-derived cytokines (IFN-γ, IL-13, IL-10, and IL-2) in BAL fluid[142]. Upregulated expression of the activation marker CD83 and the co-stimulatory molecule CD40 can also be seen in blood cDCs, suggesting that cDCs contribute to RSV-induced inflammation[142].

Alveolar macrophages are another antigen-presenting cells. Co-expression of RSV surface glycoproteins, HLA-DR molecules, IL-1β, and cytoplasmic TNF-α on alveolar macrophages indicate their immunoregulatory role[144]. Circulating monocytes have reduced expression of Toll-like receptor 8 (TLR8) and TNF-α[145], but elevated synthesis of TLR4 during acute RSV infection[146].

Eosinophils may play a role in the resolution of an RSV infection. Concentrations of the eosinophil cationic protein (ECP, a cytotoxic protein in the granules of eosinophils) are higher in samples collected from convalescent children than in those from children with acute illness[147]. Higher eosinophil counts are also seen in the recovery phase of RSV infection than in the recovery phase of adenovirus or influenza virus infection[148]. Increased expression of the myeloid activation marker CD11b on peripheral eosinophils is associated with a shorter duration of supplemental oxygen[149]. Levels of CCL-5 (C-C motif chemokine ligand 5; also known as RANTES, regulated on activation, normal T cell expressed and secreted; an eosinophil chemoattractant), ECP, and eotaxin progressively increase throughout the course of RSV infection and correlate with respiratory tract eosinophil counts, further supporting the role of eosinophils in recovery[147,148,150]. However, nasal and systemic CCL-5 concentrations inversely correlate with disease severity[151,152].

Systemic natural killer (NK) cell counts are reduced during RSV infection[153], possibly resulting from migration to the lungs[154]. Low numbers of systemic NK cells have been associated with severe RSV disease[151,153]. Type III interferons, IFN-λ1–3, are expressed by epithelial cells in particular[155] and associated with IFN-stimulated

gene expression (e.g., *MxA* and *ISG56*), which exerts an antiviral effect[156]. However, nasal IFN-λ1 levels correlate with disease severity[156].

1.4.2 Cellular immunity

An initial transient T cell lymphopenia (all subsets, including $CD8^+$, $CD4^+$, $CD3^+$, and γδ T cells) can be seen in blood during acute RSV infection[157,158], followed by an increase in $CD8^+$ T cell counts in both peripheral blood and the lungs[157,159], coinciding with the reduction in viral load[159]. In contrast to neutrophils, $CD8^+$ T cell counts are low at the time of maximum disease severity and viral load, and the peak responses coincide with convalescence[134]. Circulating T cell counts and RSV-specific cell-mediated cytotoxic immune responses inversely correlate with disease severity[160–162]. Defective systemic $CD4^+$ and $CD8^+$ T cell responses in the elderly may explain their increased susceptibility to severe RSV infection[163]. Decreased $FOXP3^+$ $CD4^+$ regulatory T cell counts can also be seen during acute infection and persist for at least 3 weeks after acute illness[164,165]. Numbers of circulating regulatory T cells are not associated with disease severity[129]. $CD4^+$ and, to a lesser extent, $CD8^+$ T cells are present in BAL fluid from infants with RSV lower respiratory tract infection[131,166]. The $CD8^+$ T cells express granzyme B, displaying an effector phenotype (HLA-DR^+ granzyme B^+ $CD38^+$)[131]. Elevated levels of pre-existing pulmonary RSV-specific $CD8^+$ T cells (exhibiting a $CD69^+$ $CD103^+$ Trm phenotype) are associated with low viral load and mild disease[159].

Several mechanisms relating to T cells have been proposed for immune evasion of RSV. Circulating $CD4^+$ and $CD8^+$ T cells have higher expression of Fas and TNF-related apoptosis-inducing ligand (TRAIL) receptor, potentially resulting in T cell apoptosis and thus systemic lymphopenia[158]. Upregulated expression of the programmed cell death 1 (PD-1) protein in pulmonary $CD8^+$ T cells is associated with the inhibition of activated T cells and thus reduced cytotoxic T-cell responses[167]. Furthermore, activation of mammalian target of rapamycin (mTOR) is also noted during RSV infection, blunting the differentiation of memory $CD8^+$ T cells[168].

1.4.3 Humoral immunity

RSV infection elicits B cell responses and antibody production. B cell responses have been shown to be T cell independent and driven by type I IFN-induced proteins (myxovirus resistance protein A and 2',5'-oligoadenylate synthetase 1) and B cell tropic factors (a proliferation-inducing ligand [APRIL] and B cell-activating factor [BAFF]) present in airway epithelial cells[169]. A subset of perialveolar plasma cells express the APRIL and BAFF receptors[169]. Both mature ($CD19^+$ $CD5^+$) and precursor ($CD19^+$ $CD10^+$) B cells increase in blood[170] and IgG, IgM, and IgA are present in blood[171], the nasopharynx[172,173], and the lungs[169] of infants with RSV infection. These antibodies target the RSV F and G proteins[171]. Higher levels of APRIL and RSV-directed IgA and IgM are associated with a lower degree of hypoxia[169].

RSV-specific nasal IgA, serum IgG, and serum neutralising titres correlate with protection from RSV (re)infection in adults, with nasal IgA conferring more protection than the others[174–176]. IgA responses, rather than IgG, are also associated with recovery from primary RSV infection in infants and young children[177]. These serum neutralising antibodies and nasal IgG and IgA responses to the RSV F protein are cross-reactive between RSV-A and RSV-B[178], but they are subgroup-specific to the RSV G protein[173,179]. IgG-producing, but not IgA-producing, memory B cells are present following RSV infection[176]. The lack of IgA memory may contribute to recurrent RSV infections throughout life[176,177]. In contrast, elderly adults are subject to severe RSV infection probably due to a deficit in serum neutralising antibodies[175].

Maternally derived circulating RSV-specific IgG levels are highest in infants <1 month of age and decrease after 3 months[180]. RSV-specific IgG levels increase again after 2 years of age, consistent with active antibody production during infection[180]. IgG avidity correlates with protection against severe RSV infection in children[180]. High maternally derived pre-existing RSV-specific mucosal IgG concentrations are more associated with low RSV viral load and low mucosal C-X-C motif chemokine ligand 10 (CXCL-10; an inflammatory chemokine) levels than high plasma IgG concentrations in infants <3 months of age[181]. It has also

been shown that pre-existing serum IgG levels may suppress the development of nasal IgA following infection[182].

RSV-specific immune complexes and antibody-dependent cell-mediated cytotoxicity (ADCC) can be detected in nasopharyngeal secretions as early as 3 days after symptom onset and persist for up to 30 days[183,184]. The detection of immune complexes is associated with temporary disappearance of RSV antigens from the airway epithelium[184]. The ADCC activity has been shown to closely correlate with RSV IgG levels[183]. These immune complexes and ADCC responses may contribute to eradication of RSV.

RSV infection also elicits deleterious IgE responses[185]. In RSV-infected infants, higher serum IgE titres and eosinophil counts have been shown to be associated with prolonged fever, severe disease, and development of wheeze[186,187].

1.4.4 T helper type 1 and type 2 responses

Th1 responses, characterised by production of IFN-γ, IL-1, IL-2, IL-12, IL-18, and TNF-α, are mounted against RSV infection[129]. They are proinflammatory and essential for the generation of cell-mediated immunity to control RSV infection[129]. Reduced circulating and respiratory tract IFN-γ levels and IFN-γ/IL-10 ratios have been associated with severe RSV disease, including hypoxia and wheeze, suggesting a protective role of IFN-γ in RSV infection[138,150,160,161,188–190]. Infants under 6 months of age have low IFN-γ responses during RSV infection, possibly contributing to the increased susceptibility to severe disease[186]. An inverse correlation between circulating IL-12 concentrations and disease severity has also been shown[161]. However, elevated nasal IL-1α and IL-18 levels are associated with the requirement for mechanical ventilation[191] and the development of bronchiolitis[192], respectively.

Th2 responses, characterised by production of IL-4, IL-5, IL-6, IL-9, IL-10, and IL-13, are involved in IgE and eosinophilic responses in atopy and may counteract Th1-mediated inflammation[129]. RSV infection also elicit significant Th2 responses. Systemic IL-6 and IL-10 levels correlate with severe RSV disease[188,193], whereas low respiratory tract IL-10 levels are associated with severe RSV lower respiratory tract

infection, suggesting a protective role in RSV infection[152,194]. However, the effect of respiratory tract IL-10 is reverse in infants <3 months of age[193]. There is conflicting evidence on the correlation between nasal IL-6 responses and disease severity[194,195]. Although inconsistent data exist[196], the majority of studies have linked Th2-biased responses with severe RSV infection, demonstrated by high systemic and nasal IL-4/IFN-γ ratios and a high nasal IL-10/IL-12 ratio[197–199].

1.5 Prevention

1.5.1 Monoclonal antibodies

The only licensed immunoprophylactic agent against RSV is palivizumab, a humanised murine IgG1 mAb. Palivizumab binds to antigenic site II, present in both prefusion and postfusion conformations of the F protein. Immunoprophylaxis with palivizumab in infants with prematurity, congenital heart disease, or bronchopulmonary dysplasia resulted in a 39–78% reduction in hospitalisation for RSV infection[200,201]. However, its use is limited to high risk infants (e.g., those who were born extremely preterm or have congenital heart disease or bronchopulmonary dysplasia) due to its prohibitive cost and the requirement of monthly administrations over the RSV season[202].

Due to the drawbacks of palivizumab, attempts have been made to develop mAbs that are more effective, affordable, and long-lasting. Motavizumab, derived from palivizumab, also targets antigenic site II in the F protein. Due to 13 amino acid alterations, it has 70-fold higher affinity and 20-fold greater neutralising activity *in vitro* than palivizumab[203]. A phase 3 clinical trial showed an 87% relative reduction in the proportion of term infants hospitalised for RSV infection[128]. However, its licensure was declined by the United States Food and Drug Administration due to safety concerns (hypersensitivity and allergic reactions)[128,204]. Suptavumab (REGN2222), another RSV-specific mAb, targets prefusion-specific antigenic site V. It is 5–10 times more potent in RSV neutralisation than palivizumab, but failed to meet its primary end point in a phase 3 clinical trial[205]. This was found to be due to all of the circulating RSV-B strains detected in the trial carrying

two amino acid alternations (L172Q and S173L) in the F protein, conferring resistance to suptavumab. I will discuss in detail the monoclonal antibody–escape mutants in chapter 5.

A promising candidate, nirsevimab (MEDI8897), is derived from a prefusion-specific human IgG1 mAb, D25[206]. Additional germline reversions and amino acid substitutions in the complementarity-determining (CDR) and fragment crystallisable (Fc) regions of D25 extend the serum half-life and increase the neutralising activity of nirsevimab[206]. Hence, the serum level of nirsevimab after a single intramuscular dose can sustain above the target level for at least 5 months, enabling once-per-season dosing[206]. A phase 2b trial in healthy preterm infants born <35 weeks of gestation showed that a single dose of nirsevimab lowered the incidence of medically attended RSV-associated lower respiratory tract infection by 70% and the rate of hospitalisations for this condition by 78% through 150 days[33]. A phase 3 trial in healthy late preterm and term infants (born ≥35 weeks of gestation) is currently ongoing, and the preliminary analysis has shown that nirsevimab significantly reduced the incidence of medically attended RSV-associated lower respiratory tract infection, reaching its primary end point (ClinicalTrials.gov Identifier: NCT03979313)[207].

1.5.2 Vaccines

The attempt to develop a vaccine against RSV dates back to the 1960s, soon after the discovery of this virus. However, a licensed vaccine has yet to become available. The first RSV vaccine, developed in the 1960s, was a formalin-inactivated vaccine. However, this vaccine caused enhanced respiratory disease with subsequent RSV infection in vaccinees[208]. This failure, along with limited knowledge of correlates of protection against RSV disease, hampered the development of RSV vaccine in the following decades.

With advances in the understanding of RSV protein structures and the awareness of the impact of RSV infection, the last decade has seen growing interest in the development of preventative measures against RSV infection. According to PATH,

as of April 7, 2021, there were nine vaccine candidates in phase 1 clinical trials, and nine vaccine candidates and two mAbs have entered into late stage clinical trials (Fig. 1.4)[209]. Different RSV vaccine candidates are designed for different populations, and the three target populations are (1) infants and young children <2 years of age; (2) pregnant women (to protect the newborn infants through placental transfer of maternal antibodies); and (3) the elderly. There are wide variations in vaccine designs, including (1) particle-based, (2) live-attenuated or chimeric, (3) subunit, and (4) vector-based vaccines[210]. In what follows, I will discuss vaccines that have had results from clinical trials in each category.

One of the particle-based vaccines has completed its phase 3 clinical trial. This vaccine consists of postfusion F protein nanoparticles and targets pregnant women at 28 weeks through 36 weeks of gestation. Maternal vaccination with this vaccine reduced RSV-associated, medically significant lower respiratory tract infection in infants by 39.4% during the first 90 days of life, narrowly failing to meet the trial's primary end point[211]. However, subgroup analysis showed that low- and middle-income countries saw a greater vaccine efficacy than high-income countries. In addition, the results relating to other end points of all-cause or RSV-associated respiratory infections also showed the potential benefits of maternal vaccination, warranting further study.

Most live-attenuated vaccines that are being or have been evaluated in phase 1 trials target paediatric populations[209]. Results of phase 1 trials for three of these vaccines have been recently published: D46/NS2/N/ΔM2-2-HindIII[212], LID/ΔM2-2/1030s[213], and RSV/ΔNS2/Δ1313/I1314L[214]. These vaccines were attenuated by deleting either the M2-2 protein (RNA synthesis regulatory protein) or the NS2 protein (IFN antagonist), and were stabilised by mutations in the NS2, N, or L proteins. A single intranasal dose of any of the three vaccines showed great infectivity and immunogenicity, and primed for anamnestic responses to natural RSV infection in RSV-seronegative children aged 6–24 months[212,213] or 12–59 months[214].

DS-Cav1 is an engineered soluble variant of the F protein with stably exposed prefusion-specific antigenic site ∅, which induces potent neutralising antibodies

against RSV[32]. This subunit vaccine has been evaluated in healthy adults 18–50 years of age in a phase 1 trial. The results showed a robust boost in neutralising activity and a substantial increase in levels of RSV F-binding IgG and IgA up to 44 weeks without any serious adverse events[31].

Vector-based vaccines include PanAd3-RSV, MVA-RSV, and ChAd155-RSV among others. These three vaccines use replication-incompetent chimpanzee adenovirus 3, modified vaccinia virus Ankara, and chimpanzee adenovirus 155, respectively, as a viral vector to deliver the F, N, and M2-1 proteins. Phase 1 trials of a prime-boost combination of PanAd3-RSV and MVA-RSV[215,216] and two doses of ChAd155-RSV[217] all showed an acceptable safety profile with a robust boost in humoral and cellular immune responses in healthy adults with pre-existing RSV immunity. A phase 2 trial of ChAd155-RSV in RSV-seropositive children aged 12–24 months has recently completed, and the results are awaited (ClinicalTrials.gov: NCT02927873). Another vector-based vaccine is Ad26.RSV.preF, which is a replication-defective adenovirus 26 vector encoding the F protein stabilised in the prefusion conformation. A phase 2a trial of co-administration of this vaccine and a seasonal influenza vaccine (Fluarix) in adults $\geq$60 years old showed robust neutralising and binding antibody responses[218].

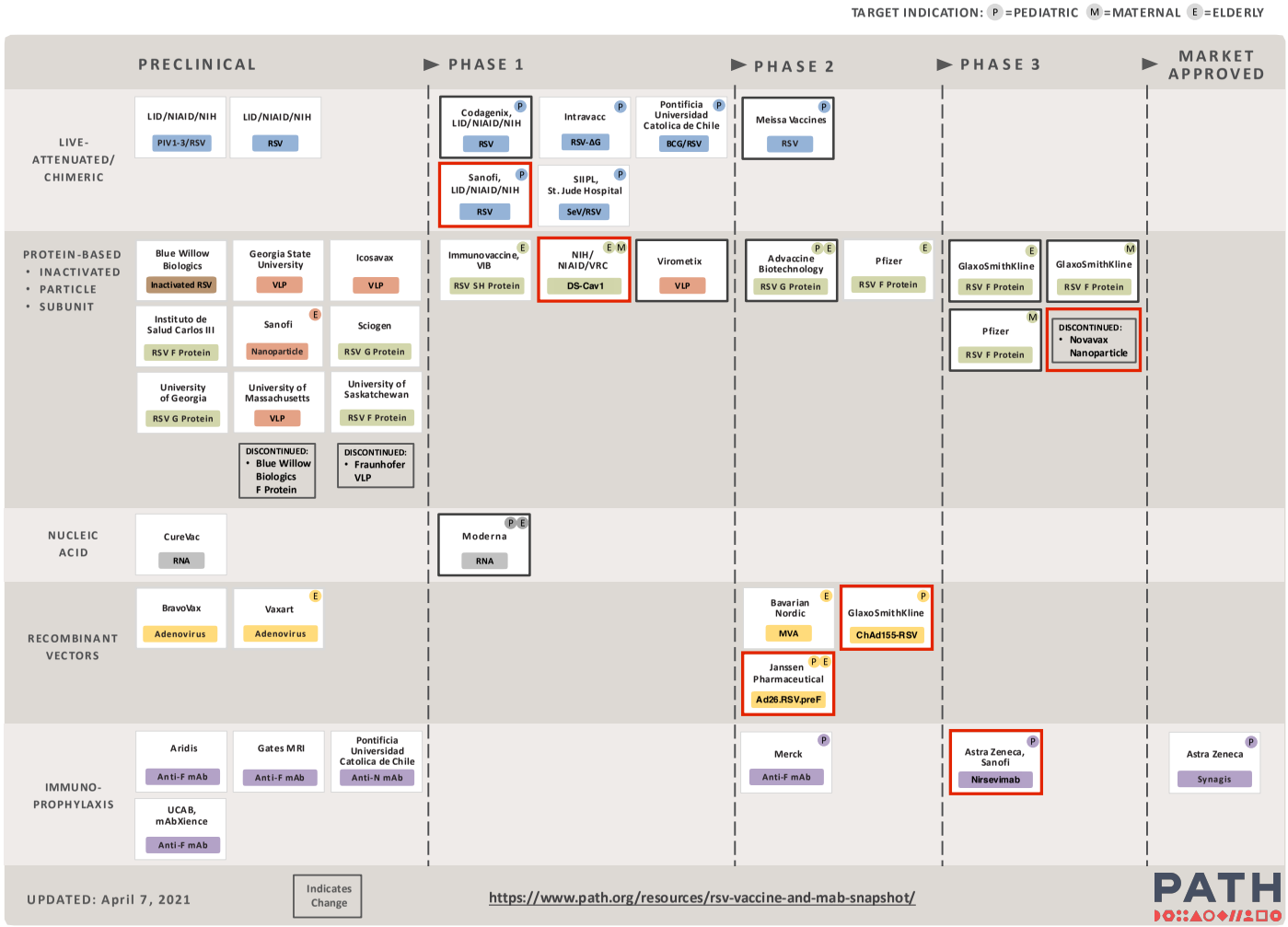


Figure 1.4: RSV vaccine and monoclonal antibody snapshot. Adapted from the PATH website at `www.path.org`, on May 12, 2021. Candidates in red squares are discussed in the main text.

1.6 Thesis outline

In this thesis I intend to examine the molecular characteristics and transmission dynamics of RSV. The objectives of this thesis are:

1. To evaluate the performance of a revised sequencing protocol on RSV samples, and to examine its ability to quantify viral load and gene expression

2. To establish the molecular epidemiology of RSV in three European countries during 2017–2020

3. To identify host, viral, and microbial factors associated with severe RSV disease

4. To analyse RSV populations within each infected individual, and to compare diversity differences between RSV-A and RSV-B at the individual and population levels

5. To study the evolutionary dynamics of RSV-A, and to reconstruct the global epidemic history of genotype ON1

Studies presented in this thesis analysed RSV samples prospectively collected in the REspiratory Syncytial virus Consortium in EUrope (RESCEU) project during 2017–2020. In the next chapter, I will introduce the clinical studies in the RESCEU project and the methodologies of targeted metagenomic sequencing and genome reconstruction applied to the samples (chapter 2). I conducted a pilot study to evaluate the application of targeted metagenomic sequencing in a subset of samples. I successfully recovered RSV genomic sequences and discovered differential expression of viral genes between RSV-A and RSV-B (chapter 3). Applying the same methods to over 800 samples collected from three RSV seasons, I examined viral determinants for disease severity. I also expanded the analysis to evaluate the effect of co-detection of other respiratory viral and bacterial pathogens on RSV disease severity (chapter 4). I then investigated RSV diversity within each infected individual and found distinct features between the two subgroups (chapter

5). Next, I put the RSV-A isolates from the RESCEU project in a global context and uncovered the emergence and epidemic history of genotype ON1 (chapter 6). In the last chapter, I summarise the findings from the whole thesis, and discuss the implications and limitations of my studies (chapter 7).

2

General materials and methods

Contents

2.1 REspiratory Syncytial virus Consortium in EUrope (RESCEU) 26
2.1.1 Birth cohort study 26
2.1.2 Infant case–control study 28
2.1.3 Older adult cohort study 30
2.1.4 Definition of disease severity 32
2.1.5 Ethical approval and study conduction 34
2.2 Nucleic acid extraction 34
2.3 Targeted metagenomics 35
2.3.1 Library construction 35
2.3.2 Probe design 36
2.3.3 Sequencing 38
2.4 Genome reconstruction 38
2.5 Polymerase chain reaction 40

The samples analysed in this thesis were collected from three clinical studies. This chapter details these studies, and the methodologies for nucleic acid extraction, targeted metagenomic sequencing, genome reconstruction, and quantitative reverse transcription PCR (RT-PCR). All of these methods were used in later chapters.

2.1 REspiratory Syncytial virus Consortium in EUrope (RESCEU)

Nasopharyngeal swabs were prospectively collected from participants enrolled in the clinical studies of the RESCEU project. RESCEU is a European multicentre study, investigating the epidemiological, virological, and immunological features of RSV infection. RSV-positive swabs collected from three European countries (United Kingdom, Netherlands, and Spain) across three continuous RSV seasons during 2017–2020 were included in the analyses presented in this thesis. These samples were from three clinical studies: the birth cohort, infant case–control, and older adult cohort studies.

Study visits and sample collection were conducted by research nurses and doctors at each participating site. I was responsible for handling the samples once received in the Oxford Vaccine Group Laboratory.

2.1.1 Birth cohort study

The birth cohort study[219] is a prospective epidemiological and observational cohort study, conducted in the following five centres: Spaarne Gasthuis, Haarlem, the Netherlands; Hospital Clínico Universitario de Santiago, Santiago de Compostela, Spain; participating hospitals in the Thames Valley and South Midlands Clinical Research Network (Oxford), United Kingdom; Royal Hospital for Sick Children, Edinburgh, United Kingdom; and Turku University Hospital, Turku, Finland. This study established a birth cohort of 10,000 healthy term infants, consisting of 1,000 in the active surveillance cohort and 9,000 in the passive surveillance cohort. Participants in the active cohort were enrolled only in the United Kingdom, Spain, and the Netherlands. The primary and secondary objectives of this study are listed in Table 2.1. Continuous recruitment expanded between July 2017 and December 2019 (active cohort) or April 2020 (passive cohort). Follow-up of participants in this study is still ongoing. This study enrolled term newborn babies ($\geq$37 weeks of gestational age) at birth. Neonates with major congenital defects or serious chronic illness (i.e., congenital heart and/or lung disease, genetic, immunological and/or

metabolic disorders) or acute severe medical conditions (e.g., sepsis or asphyxia) were excluded. Neonates who had received immunoglobulin, monoclonal antibodies (mAbs), or an investigational vaccine or medication against RSV were also excluded.

Table 2.1: Primary and secondary objectives of the birth cohort study.

Primary objectives
1. To determine the incidence of RSV-associated acute respiratory tract infection (ARTI) and RSV-associated medically attended ARTI during the first year of life (active cohort) 2. To determine the incidence of RSV-related hospitalisation during the first year of life (active and passive cohorts)
Secondary objectives
1. To estimate how RSV infection of different severity relates to wheeze up to age 3 years (active and passive cohorts) 2. To determine the rate of all-cause medically attended (inpatient or outpatient) ARTI (active cohort) 3. To determine mortality (RSV-associated and all-cause) through all RSV seasons of follow-up (passive and active cohorts) 4. To determine health care costs, health care resource use, interruption of normal activities, and health-related quality of life (HRQoL) in RSV-associated and all-cause medically attended (inpatient or outpatient) ARTI patients and their families (active cohort) 5. To determine the incidence of RSV-related secondary bacterial respiratory tract infections within 21 days after the onset of RSV infection and their association with antibiotic use in hospitalised RSV ARTI patients (active and passive cohorts) and non-hospitalised RSV ARTI patients (active cohort) 6. To collect clinical samples (blood, nasopharyngeal swab, stool, and urine) for biomarker analysis (active cohort) 7. To determine the incidence rate of other respiratory pathogens (influenza, rhinovirus, human metapneumovirus, parainfluenza virus, etc.) associated with all medically attended (inpatient or outpatient) ARTI (active cohort) 8. To determine important risk factors for RSV infection (by severity and healthcare utilisation) (active and passive cohort) 9. To determine the proportion of viral ARTI attributable to RSV (active cohort)

Parents or guardians of the participants in the passive cohort were asked to fill in a questionnaire at enrolment and at the age of 1 year. The baseline questionnaire collected demographic data, medical history, and family history. The one-year questionnaire collected frequency and reason of health care resource use, duration of

symptoms, treatment received, health care costs, interruption of normal activities, and health-related quality of life (HRQoL). Participants who had been hospitalised for acute respiratory tract infection (ARTI) during the first year of life were followed up to the age of maximum 3 years by yearly questionnaires to determine the incidence and severity of wheeze.

Parents or guardians of all participants in the active cohort were asked to complete baseline and yearly questionnaires until the age of maximum 3 years. In addition, blood samples, a nasopharyngeal swab, and stool and urine samples were collected from participants in the active cohort around the fifth day of life as baseline samples. When participants in the active cohort had acute respiratory symptoms during RSV seasons (between October 1st and April 30th) before their first birthday, a nasopharyngeal swab was taken by a research nurse or doctor within 3 days of symptom onset. RSV infection was diagnosed using point-of-care testing on the Alere™ i RSV assay (Abbott, Illinois, United States) for qualitative molecular detection of RSV. If a participant was positive for RSV, informed consent was asked for participation in a biomarker sub-study (n = 480 maximum). Additional blood samples, a nasopharyngeal swab, and stool and urine samples were collected at the time of infection and 6–8 weeks after infection from participants included in the biomarker sub-study.

Regardless of the test results, nasopharyngeal swabs collected when participants had acute respiratory symptoms were immersed in 3 mL of M4RT® viral transport medium (Thermo Fisher Scientific, Massachusetts, United States), aliquoted, and frozen at −80 °C until use. As of March 2020, all available RSV-positive swabs were sequenced using the methods described in sections 2.2 and 2.3 without further sample selection.

2.1.2 Infant case–control study

The infant case–control study is conducted in the following four sites: the University Medical Center Utrecht, the Netherlands; Hospital Clínico Universitario de Santiago, Spain; Imperial College National Health Service Trust, London; and participating

hospitals in the Thames Valley and South Midlands Clinical Research Network (Oxford), United Kingdom[4]. Participant recruitment was between October 2017 and March 2020. Follow-up of participants in this study is still ongoing. This study enrolled infants under 1 year of age with acute RSV infection (RSV cases) from communities within 4 days of symptom onset or from hospitals within 2 days of admission during RSV seasons (between October 1st and April 30th). This study planned to recruit 275 RSV-infected non-hospitalised infants (250 without comorbidities and 25 with comorbidities) and 275 RSV-infected hospitalised infants (250 without comorbidities and 25 with comorbidities) during RSV seasons. Comorbidities included, but were not limited to, prematurity (<37 weeks of gestational age), bronchopulmonary dysplasia, and other congenital abnormalities. This study also planned to enrol 80 healthy term infants (≥37 weeks of gestational age) under 1 year of age without acute RSV infection or any respiratory illness in the past 7 days (healthy controls) outside RSV seasons. Infants who had had any concurrent significant medical illness or received any vaccine in the past 7 days were also not eligible to be included in the control group. Infants with a history of RSV infection, receipt of immunoglobulin or mAbs, or exposure to an RSV investigational vaccine or medication were excluded from this study. Table 2.2 lists the primary and secondary objectives of this study.

RSV infection was confirmed by the Alere™ i RSV assay in a community setting or routine tests (e.g., antigen tests or PCR) at a central laboratory in a hospital setting. Blood samples, nasopharyngeal swabs, urine and stool samples were collected at enrolment (RSV cases and healthy controls) and 6–8 weeks after infection (RSV cases). For RSV-infected hospitalised participants, daily nasopharyngeal swabs were also collected where possible until hospital discharge. A 14-day symptom diary during illness was collected from all RSV cases. Yearly parental questionnaires were completed to quantify number of wheezing episodes, number of hospital admissions, number of days admitted, other healthcare advice sought and how often, health care costs, and HRQoL from both RSV cases and healthy controls until the age of 3 years.

As of March 2020, all available nasopharyngeal swabs collected from RSV-infected participants (at enrolment and daily during hospitalisation) were processed as described in section 2.1.1 and sequenced using the methods described in sections 2.2 and 2.3 without further sample selection.

Table 2.2: Primary and secondary objectives of the infant case–control study.

Primary objective
To establish biomarkers predictive of, or associated with, lower respiratory tract RSV infection and disease severity in infants
Secondary objectives
1. To establish biomarkers associated with sequelae following RSV infection in infants 2. To characterise the viral load and genetic sequence of mild and severe RSV disease 3. To characterise the immune responses to mild and severe RSV disease 4. To describe the transcriptomic, proteomic, metabolomic, and epigenetic signatures associated with RSV disease severity and/or respiratory sequelae 5. To biobank specimens for later analysis of respiratory/stool microbiome, and genetic associations with severe RSV disease and/or respiratory sequelae

2.1.3 Older adult cohort study

The older adult cohort study was a prospective epidemiological and observational cohort study, conducted in the United Kingdom (Oxford) and the Netherlands (Utrecht)[220]. Participant recruitment, follow-up, and sample collection were between August 2017 and June 2019. It enrolled 500 adults aged 60 to <75 years and 500 adults aged ≥75 years from general practitioner (GP) networks before the onset of each RSV season (August and September). Those who had concurrent alcohol or drug abuse, dementia, any known or suspected immunosuppressive condition, chronic use of immunosuppressants or other immune-modifying drugs (defined as >14 continuous days), or had received any investigational vaccine or mAb were excluded. Each participant was followed up for one year. A short questionnaire, blood samples, and a nasopharyngeal and an oropharyngeal swab were collected at

enrolment and after the RSV season (May and June). Table 2.3 lists the primary and secondary objectives of this study.

Table 2.3: Primary and secondary objectives of the older adult cohort study.

Primary objective
To estimate the incidence of RSV-associated ARTI, RSV-associated medically attended ARTI, and RSV hospitalisation in older patients
Secondary objectives
1. To estimate the rate of all-cause medically attended (inpatient or outpatient) ARTI and related medical complications (exacerbations of chronic conditions, acute cardiovascular events) 2. To estimate the RSV-associated and all-cause mortality 3. To estimate health care costs, health care resource use, interruption of normal activities, and health-related quality of life (HRQoL) in RSV-associated and all-cause medically attended (inpatient or outpatient) ARTI patients 4. To estimate the incidence of RSV-related secondary bacterial pneumonia events and their association with antibiotic use within 21 days after the onset of RSV infection 5. To collect clinical samples for biomarker analysis 6. To examine the incidence of other respiratory pathogens associated with all medically attended ARTI 7. To estimate the proportion of viral ARTI attributable to RSV 8. To estimate important risk factors for RSV infections (by severity and healthcare utilisations) 9. To determine change in frailty over the course of the study

In the event of an ARTI during the RSV season (between October 1st and April 30th), an oropharyngeal and two nasopharyngeal swabs were collected from the participant by a research nurse within 3 days after the participant notified the research team. RSV infection was diagnosed using molecular point-of-care testing on the GeneXpert® influenza/RSV system (Cepheid, California, United States). If the participant was positive for RSV, blood samples were collected at the time of the infection, and an additional nasopharyngeal swab was collected 2 weeks after the infection. A short daily diary was completed by the participant to score respiratory symptoms and the quality of life.

As of March 2020, all available nasopharyngeal swabs collected from RSV-infected participants at the time of an RSV infection were processed as described

in section 2.1.1 and sequenced using the methods described in sections 2.2 and 2.3 without further sample selection.

2.1.4 Definition of disease severity

Demographic and clinical information was gathered in these three studies. Clinical severity in infants was determined using the ReSVinet scale (Table 2.4)[221]. This scale accounts for seven clinical variables: feeding intolerance, medical intervention, respiratory difficulty, respiratory frequency, apnoea, general condition, and fever. The score of each variable ranges from 0 to 2 or 3, and the total score is 20. A score of 0–7 was defined as mild, a score of 8–13 as moderate, and a score of 14–20 as severe. In older adults, those who did not require any treatment or medical attendance were defined as having a mild disease, those requiring hospitalisation were defined as having a severe disease, and the rest were defined as having a moderate RSV disease.

A fever was defined as at least one episode of axillary temperature $\geq$37.5 °C or tympanic temperature $\geq$38 °C during the course of acute infection in all of the above-mentioned clinical studies.

Table 2.4: ReSVinet scale.[a]

Variable	0 point	1 point	2 points	3 points
Feeding intolerance	No	Decreased appetite and/or isolated vomits with cough.	Frequent vomits with cough, rejected feed but able to tolerate fluids sufficiently to ensure hydration.	Oral intolerance, not able to guarantee adequate hydration orally. Required nasogastric and/or intravenous fluids.
Medical intervention	No	Nasal secretion aspiration, physical examination, trial of nebulised bronchodilators, antipyretics.	Oxygen therapy required. Complementary exams were needed (chest X-rays, blood test, etc.). Maintained nebulised therapy with bronchodilators.	Required respiratory support with positive pressure (either non-invasive or invasive).
Respiratory difficulty	No	Not in basal situation but did not appear severe. Wheezing only audible with stethoscope. Good air entrance.	Some respiratory effort. Expiratory wheezing audible without stethoscope. Decreased air entry in localised areas.	Obvious respiratory effort. Inspiratory and expiratory wheezing and/or clearly decreased air entry.
Respiratory frequency	Normal	Presented episodes of tachypnea, limited in time by self-resolution or response to secretion aspiration or nebulisation.	Tachypnea persisted or recurred despite secretion aspiration and nebulisation with bronchodilators.	Severe and sustained tachypnea. Very superficial and quick breath rate. Mental status affected.
Apnoea	No			At least one episode of apnoea.
General condition	Normal	Not in basal situation. Child was mildly uncomfortable. Parents were not alarmed. Could wait in the waiting room or even stay at home.	Patient looked ill and needed further medical exam and therapy. Parents were concerned. Could not wait in the waiting room.	Agitated, apathetic, lethargic. No need of medical training to realise severity. Immediate medical evaluation and intervention were required.
Fever	No	$<38.5\,^{\circ}C$	$\geq 38.5\,^{\circ}C$	

[a] Adapted from Justicia-Grande AJ, et al. PLoS One 2016;11:e0157665[221], published under CC BY 4.0.

2.1.5 Ethical approval and study conduction

The three studies have been registered in ClinicalTrials.gov with the identifiers of NCT03627572, NCT03756766, and NCT03621930 for the birth cohort, infant case–control, and older adult cohort studies, respectively. The protocols were reviewed and approved by the relevant ethics committees at each site: the University of Oxford, the Health Research Authority (224156 and 231136), the NHS National Research Ethics Service Oxfordshire Committee A (15/SC/0335), South Central and Hampshire A Research Ethics Committee (17/SC/0522), and London—Central Research Ethics Committee (17/LO/1210) in the United Kingdom; Hospital Clínico Universitario de Santiago de Compostela, and Comité de Ética de la Investigación de Santiago-Lugo (2017/395) in Spain; the Medical Ethical Committee, University Medical Center Utrecht (17/563), and the Ethical Review Authority (NL60910.041.17) in the Netherlands.

These studies were conducted in accordance with the provisions of the Declaration of Helsinki. All adult participants and the parents or guardians of all infant participants provided written, informed consent.

2.2 Nucleic acid extraction

Nucleic acid extraction was carried out by my colleagues, Anthony Brown and George MacIntyre-Cockett, at the Peter Medawar Building for Pathogen Research, University of Oxford. The extraction was performed in four different batches due to the large sample size and the plan to acquire sequencing data at certain times. The same batching was also applied to targeted metagenomic sequencing (section 2.3). To minimise the risk of RNA degradation, nucleic acid was extracted from primary samples, and the extractions were scheduled as close as practical to the time of sequencing.

Automated extraction of total nucleic acid was performed directly from nasopharyngeal swabs using the NucliSENS® easyMAG® system (BioMérieux, Marcy-l'Étoile, France), following the manufacturer's instructions. The duration of lysis

and incubation with silica was 10 minutes each. The volume of each sample used for extraction was 500 µL, and the final eluate was 25 µL in the first and fourth batches and 35 µL in the second and third batches.

2.3 Targeted metagenomics

Library construction and sequencing were performed by my colleagues, George MacIntyre-Cockett, Esther Mellado-Gomez, and Mariateresa de Cesare, at the Wellcome Centre for Human Genetics, University of Oxford.

2.3.1 Library construction

Nucleic acid extracts were plated out in 96-well plates for sequencing library construction. Each plate contained at least one buffer control and one RSV-negative control, extracted with the same method. Buffer controls were the pure M4RT® medium. RSV-negative controls were samples collected from study participants with acute respiratory symptoms, which tested negative by point-of-care molecular testing. These negative controls (i.e., buffer and RSV-negative controls) were processed together with RSV-positive samples.

Sequencing libraries were constructed using the methodology of targeted metagenomics[1], a modification of the veSEQ-HIV protocol[222]. A 12-µL aliquot of each nucleic acid sample, including negative controls, was first concentrated to 3 µL with RNAClean XP magnetic beads (Beckman Coulter, California, United States) and denatured at 72 °C. Dual-indexed libraries for Illumina sequencing were then constructed using the SMARTer Stranded Total RNA-Seq Kit v2—Pico Input Mammalian (Takara Bio USA, California, United States), where first-strand reverse transcription was primed with tagged random hexamers and double-stranded cDNA was synthesised with sets of i5 and i7 index primers, as previously described elsewhere[223]. These primers gave unique dual indexing (UDI) for the samples, thus minimising the risk of index misassignment during sequencing. No fragmentation of the RNA was performed during library preparation. After reduced (i.e., 12) cycles of PCR amplification of the cDNA, 10 µL of each library was pooled, purified using

AMPure XP (Beckman Coulter), and quantified using the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific). A 750-ng aliquot was taken from the pool for capture by a predesigned SureSelect RNA Target Enrichment multi-pathogen probe set using the SureSelectXT Reagent Kit, HSQ (Agilent, California, United States). The post-capture libraries were amplified with 16 cycles of PCR, and the final product was purified by AMPure XP.

2.3.2 Probe design

This multi-pathogen probe set, *Castanet*[224], consisting of 120-mer oligonucleotides, was designed using the algorithm devised in the veSEQ protocol[225]. It targeted more than 100 pathogenic bacteria and viruses, including both RSV-A and RSV-B (Table 2.5). For the target viruses, the probes were designed to cover the variation of whole-genome sequences of each virus with the exception of herpesvirus, only about 10% of whose genome was covered. For the target bacteria, the probes were designed to cover the variation of 53 *rps* genes, encoding the bacterial ribosome protein subunits. The allele characterisation of the 53 *rps* genes has been shown to be able to classify bacteria into groups at all taxonomic and most typing levels[226]. A database containing catalogued *rps* gene variation has been developed for ribosomal multilocus sequence typing (rMLST). The probe set has been validated and benchmarked on a standard reference containing a mixture of pre-quantified viruses[224], and the efficiency of capture has been shown to be consistent at up to 20% sequence divergence between the probes and the target species[225]. Therefore, the target enrichment was unlikely to be biased towards any specific genomic region or strain of the target species.

Table 2.5: Target bacterial and viral species.[a]

Bacteria	
Acinetobacter baumannii	*Acinetobacter calcoaceticus*
Bartonella henselae	*Bordetella pertussis*
Borrelia burgdorferi	*Brucella* spp.
Chlamydia psittaci	*Chlamydophila pneumoniae*
Coxiella burnetiid	*Enterobacter aerogenes*

Continued on next page

Table 2.5 – continued from previous page

Viruses	
Enterobacter cloacae	*Escherichia coli*
Haemophilus influenzae	*Haemophilus parainfluenzae*
Klebsiella pneumoniae	*Klebsiella oxytoca*
Legionella pneumophila	*Leptospira* spp.
Listeria monocytogenes	*Moraxella catarrhalis*
Mycobacterium tuberculosis	*Mycoplasma pneumoniae*
Neisseria meningitidis	*Pseudomonas aeruginosa*
Serratia marcescens	*Staphylococcus aureus*
Stenotrophomonas maltophilia	*Streptococcus agalactiae*
Streptococcus pneumoniae	*Streptococcus pyogenes*
Treponema pallidum	
Viruses	
BK polyomavirus	California encephalitis virus
Cardiovirus A & B	HCoV-229E[b]
HCoV-HKU1 genotypes A, B, C[b]	HCoV-NL63[b]
HCoV-OC43 genotypes A–E[b]	MERS-CoV[c]
SARS-CoV[d]	Dengue virus 1–4
Eastern equine encephalitis virus	Enterovirus A & B
Hendra virus	Henipavirus B & M
Hepatitis A virus	HHV-1—Herpes simplex virus type 1[e]
HHV-2—Herpes simplex virus type 2[e]	HHV-3—Varicella-zoster virus[e]
HHV-4—Epstein-Barr virus[e]	HHV-5—Human cytomegalovirus[e]
HHV-6A—Human herpesvirus 6A[e]	HHV-6B—Human herpesvirus 6B[e]
HHV-7—Human herpesvirus 7[e]	HHV-8—Kaposi's sarcoma herpesvirus[e]
Human bocavirus 1	Human metapneumovirus
Human parainfluenza virus 1, 2, 3, 4a, 4b, 5	Human parechovirus 1–8
Influenza A virus H1N1, H1N2, H2N2, H3N2, H5N1, H7N3, H7N7, H7N9, H9N2	Influenza B virus
Influenza C virus	Japanese encephalitis virus
JC polyomavirus	Lassa mammarenavirus
Lymphocytic chriomeningitis mammarenavirus	Mastadenovirus A–G
Measles virus A, B1–B3, C1, C2, D1–D11, E, F, G1–G3, H1, H2	Mumps virus A–D, F–L, N
Murray Valley encephalitis virus	Parechovirus B
Primate erythroparvovirus 1	Primate tetraparvovirus 1
Respiratory syncytial virus A & B	Rabies virus
Lagos bat virus	Mokola virus
Duvenhage virus	European bat lyssavirus 1

Continued on next page

Table 2.5 – continued from previous page

Viruses	
European bat lyssavirus 2	Australian bat lyssaviruses
Rhinovirus A–C	Rift Valley fever virus
Rosavirus 2	Rotavirus A–C
Rubella virus	Salivirus A & FHB
Sandfly fever Naples virus	Sandfly fever Sicilian virus
Sosuga virus	St. Louis encephalitis virus
Tick-borne encephalitis virus	Venezuelan equine encephalitis virus
West Nile virus	Western equine encephalitis virus
Yellow fever virus	

[a] Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).
[b] HCoV denotes human coronavirus.
[c] MERS-CoV denotes Middle East respiratory syndrome coronavirus.
[d] SARS-CoV denotes severe acute respiratory syndrome coronavirus.
[e] HHV denotes human herpesvirus; the probe set only targets partial genomes of HHVs.

2.3.3 Sequencing

The first and third batches of samples were sequenced on the Illumina MiSeq system (Illumina, California, United States) with the MiSeq Reagent Kit v3 (600-cycle), generating 265-bp and 300-bp paired-end reads respectively. The second and fourth batches were sequenced on the Illumina NovaSeq 6000 system with the NovaSeq 6000 SP Reagent Kit v1.5 (300-cycle), generating 151-bp paired-end reads. The choice of a sequencing platform was determined by sample size: the MiSeq platform was used when there were $\leq$96 samples, and the NovaSeq 6000 system was used for a batch of 384 samples.

2.4 Genome reconstruction

The RSV genome was reconstructed using the pipeline implemented in *shiver*[227] (Fig. 2.1). After obtaining raw sequencing reads, the first six bases of read 1 and the first three bases of read 2 were clipped off using Trimmomatic (v0.39)[228] to remove random hexamer primers and the SMARTer adapter sequences, respectively. An extra two bases at the 5' end of MiSeq-generated read 2 were also cut off due

to reduced quality of the bases. Trimmomatic was then used to trimmed off the Illumina adapter sequences and low-quality bases with a Phred score below 20 (commands: `Adapters:2:10:7:1:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:50`) (Fig. 2.1, step A). *De novo* assembly of contig sequences from the trimmed reads was performed using both IVA (v1.0.8)[229] and metaSPAdes (v3.14.1)[230] (Fig. 2.1, step B), in each sample selecting the contig sequences with a higher N50 for genome reconstruction.

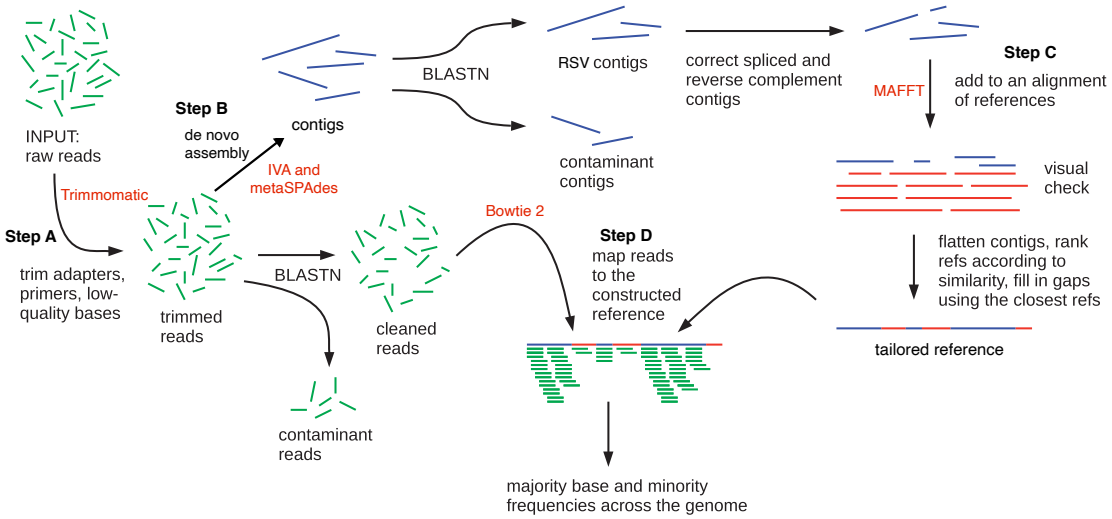


Figure 2.1: RSV genome reconstruction. See main text for details. BLASTN (v2.7.1+)[231] was used to classify the trimmed reads and contig sequences as RSV or contaminants according to the alignment and similarity to the closest reference. Adapted from Wymant C, et al. Virus Evol 2018;4:vey00716[227], published under CC BY 4.0.

A collection of 1,269 RSV whole-genome sequences were downloaded from GenBank[232], aligned using MAFFT (v7.471)[233], and used as references. MAFFT was used to align RSV contigs with the references, and *shiver* generated a tailored reference by filling gaps between contigs with the closest reference (Fig. 2.1, step C). Reads were then aligned to the tailored reference with Bowtie 2 (v2.4.1)[234] with the commands `--very-sensitive-local --maxins 2000 --no-discordant --no-unal` (Fig. 2.1, step D). A minimum base quality of 5 was required for a base to be counted as mapped. No mapping quality threshold was applied. Only reads mapped in proper pair were retained. After mapping, reads were deduplicated with

Picard MarkDuplicates (v2.18.14)[235], retaining the one with the largest sum of base-quality scores. Duplicates were defined as read pairs that shared the same mapping coordinates on the reference. A minimum of two unique reads were required to support base calls in the consensus sequence; the choice of this cut-off value was based on the analysis of a pilot sequencing study, which I will discuss in detail in chapter 3. A second round of mapping to the consensus sequence was repeated to improve sensitivity.

2.5 Polymerase chain reaction

RSV viral load was determined using quantitative RT-PCR on the same nucleic acid samples used for sequencing. PCR assays were prepared using oasig lyophilised OneStep 2X RT-qPCR Master Mix (Primerdesign, Southampton, United Kingdom) and the genesig RSV Standard Kit (Primerdesign), following the manufacturer's instructions. The exception was that only 80% of the recommended volumes of the samples and the reagents were used due to limited sample volume. Specifically, 4 µL of nucleic acid, 8 µL of master mix, 0.8 µL of primer/probe mix, and 3.2 µL of nuclease-free water were pipetted into each well. In addition, six 10-fold dilutions of RSV RNA templates with a known viral load (from 200,000 to 2 copies/mL) were used as calibration standards in each PCR plate. Nuclease-free water was used as a negative control. The PCR primers targeted the same region of the N gene for both RSV-A and RSV-B (sequence proprietary). PCR assays were carried out on the Applied Biosystems StepOnePlus Real-Time PCR System (Thermo Fisher Scientific) with amplification conditions of 55 °C for 10 min, 95 °C for 2 min, followed by 50 cycles of 95 °C for 10 sec and 60 °C for 60 sec.

All assays were run in duplicate for each sample and control, and the results are shown as the mean value in copies per millilitre (copies/mL).

3

RSV genome and transcriptome sequencing

Contents

3.1 Introduction 42
3.2 Methods 43
3.2.1 Sample source and sequencing 43
3.2.2 Viral genome and transcriptome reconstruction 43
3.2.3 Phylogenetic analysis 44
3.2.4 Differential gene expression analysis 45
3.2.5 Statistical analysis 46
3.3 Results 46
3.3.1 Samples and participants 46
3.3.2 RSV genome and transcriptome reconstruction 46
3.3.3 Coverages of genomic and positive-strand reads 52
3.3.4 Relative transcript abundance of RSV genes 53
3.3.5 Differential gene expression analysis 55
3.4 Discussion 59

This chapter is based on my publish paper: Lin GL, et al. Simultaneous viral whole-genome sequencing and differential expression profiling in respiratory syncytial virus infection of infants. Journal of Infectious Diseases 2020;222:S666-S671[1]. The work represents my own contribution to the publication and not that of co-authors.

3.1 Introduction

In the last decade, transcriptome sequencing (RNA-seq) has been extensively used to decipher the functions of a genome and gene expression levels; however, the transcriptional profile of many viral pathogens during active clinical infection, including RSV, remains poorly understood. Sequencing viruses at scale is complicated by marked viral genetic variation, especially for RNA viruses[80], and low abundance of viral genetic material compared with host nucleic acids in clinical samples. Methods to provide high-throughput, whole-genome sequences for specific viruses without a bespoke assay would have substantial value in epidemiological and evolutionary studies. For example, considering the global impact of RSV infection, the lack of a simple, cost-effective assay means that relatively few complete sequences have been determined, and little is known about potential links between viral genetics and pathogenicity.

Different strategies have been employed to overcome the difficulties, including pre-amplification with specific or random PCR primers[65,236,237], but these methods still suffer from limitations on variable PCR primer efficiencies, PCR errors, and sensitivity[237]. In contrast, probe capture methodologies, such as veSEQ[225], can generate complete sequences of all target species at up to 20% sequence divergence between the probes and the target species. A recent adaptation of veSEQ to human immunodeficiency virus (HIV) sequencing (veSEQ-HIV) has demonstrated that this method produces quantitative estimates of genome copy number, enabling viral load to be estimated directly from sequencing output[222]. Since this sequencing method relies only on library construction and enrichment (two processes that do not depend on the particular target pathogen), it can be applied to new viruses without the requirement for optimisation.

A complexity unique to RNA-seq of RNA viruses is the necessity of discriminating between viral genomic material and transcripts, both of which are RNA molecules. The development of strand-specific RNA-seq protocols[238], such as SMART (Switching Mechanism At the 5' end of the RNA Transcript) technology[239], where distinct

adapters are used to distinguish the two ends of cDNA molecules, enables the preservation of directionality of sequenced fragments.

A detailed understanding of RSV clinical infection at the molecular level could aid evaluation of viral responses to immune pressure from new vaccines or monoclonal antibodies that are under development. In this pilot study I evaluate the application of a generalised targeted metagenomic sequencing method, modified from veSEQ-HIV[222] to RSV-positive nasopharyngeal swabs collected from the RESCEU studies. My aim is to recover complete RSV genomes and to quantify RSV gene expression directly from these clinical samples. I adapt existing computational pipelines for use in RSV genomics and transcriptomics. This approach enabled whole-genome reconstruction and virus transcriptome profiling to detect genotype-specific patterns of transcription of potential phenotypic significance.

3.2 Methods

3.2.1 Sample source and sequencing

This pilot study was conducted in June 2018, after the 2017–2018 RSV season. All available nasopharyngeal swabs collected from the birth cohort and infant case–control studies (section 2.1 RESCEU studies) from Santiago de Compostela, Spain, and Oxford, United Kingdom, during the 2017–18 RSV season, were included in this pilot study. The Dutch samples were unavailable at the time of this pilot study due to logistic issues. Total nucleic acid extraction, sequencing library construction, and target enrichment were processed as described in sections 2.2 and 2.3. Sequencing was performed on the Illumina MiSeq platform, generating 265-bp paired-end reads. All samples in this pilot study were in batch 1 (see sections 2.2 and 2.3 for details). Viral load was measured using quantitative RT-PCR on the same nucleic acid samples used for sequencing, as described in section 2.5.

3.2.2 Viral genome and transcriptome reconstruction

RSV genome was reconstructed using the *shiver* pipeline as describe in section 2.4. A script, `LinkIdentityToCoverage.py`, bundled with *shiver* was used to determine an

appropriate threshold for consensus calling. This script calculated the identity value of each read (i.e., the percent identity between a mapped read and the reference) and averaged out the identity values of mapped reads in positions with the same depth of coverage. The minimal number of reads required for consensus calling was the minimal depth of coverage above which the mean percent identity became stable.

After genome reconstruction, each pre-deduplicated binary alignment/map (BAM) file was divided using samtools (v1.10) into two files on the basis of SAM[240] flags: one containing read pairs mapped in the antisense orientation (genomic reads, SAM flags: 83 or 163), and the other containing read pairs mapped in the sense orientation (positive-strand reads, SAM flags: 99 or 147). The two files were then deduplicated and processed using `process_bam.sh`, a script bundled with *shiver*, to generate consensus sequences and per-base coverage and composition for genomic and positive-strand reads respectively.

3.2.3 Phylogenetic analysis

Sixty RSV whole-genome sequences (30 RSV-A and 30 RSV-B) collected from different countries at various times were downloaded from GenBank (Supplementary Table A.1) and included for phylogenetic analysis, along with those complete RSV sequences generated in this study (accession numbers: LR699315, LR699726, LR699734, and LR699736–44). Multiple sequence alignment was carried out using MAFFT with the FFT-NS-i method[241] (commands: `--nwildcard --adjustdirection --maxiterate 1000`). Maximum-likelihood phylogenetic analysis was performed using RAxML (v8.2.12)[242] with the general time reversible (GTR) model[243] and gamma-distributed rate heterogeneity among sites (Γ)[244]. Phylogenies were rooted to the oldest RSV-A (accession number KU316171) and RSV-B (accession number KU316116) sequences respectively. R (v3.6.0)[245] packages—ape (v5.3)[246] and ggtree (v1.16.3)[247,248]—were used to demonstrate the phylogenies.

3.2.4 Differential gene expression analysis

Samples from which complete RSV genomes were reconstructed were included in the differential gene expression analysis. `htseq-count`, a script developed with HTSeq (v0.11.2)[249], was used with the union mode to calculate the number of positive-strand reads mapped to each RSV gene from the pre-deduplicated BAM files created by *shiver*. With the union mode, read pairs partially mapped to a gene (i.e., at the border of this gene) were counted for this feature, while those mapped to more than one gene were counted as ambiguous.

To compare the expression levels of different genes within a sample, the count data were transformed to transcripts per kilobase million RSV-mapped reads (TPM), i.e., divided by the length of each gene in kilobases and then scaled by the total number of positive-strand reads assigned exclusively to a single RSV gene. The gene expression level of each sample was then visualised using box and scatter plots.

Differential gene expression analysis between different samples was performed using the count metrics, following the RNAseq123 workflow with the R packages—limma (v3.40.2)[250], Glimma (v1.12.0)[251], and edgeR (v3.26.5)[252]. The count data were normalised by trimmed mean of M values (TMM)[253] and transformed to the base-2 logarithmic scale ($\log_2$). The normalisation was to ensure that the expression distributions of each sample were similar across the entire dataset, to remove systematic technical effects, and to minimise the effect of technical bias. Multidimensional scaling analysis was carried out on the TMM-normalised and $\log_2$-transformed count data to inspect the sample clustering. Fitted linear models were constructed to compare gene expression levels between two RSV subgroups (RSV-A and RSV-B), three severity groups (mild, moderate, and severe; see section 2.1.4), and two host sex groups (male and female). Finally, empirical Bayes moderated t-statistics were applied to compute the significance of differential gene expression with a minimum $\log_2$ fold change of $\log_2 1.2$. P values were corrected using the Benjamini–Hochberg procedure[254].

3.2.5 Statistical analysis

Pearson correlation was used to evaluate the association between the number of mapped RSV reads and the viral load. A two-tailed Williams' t-test was used to test the difference between two dependent correlations sharing one variable. All statistical analyses were carried out using R (v3.5.1)[245]. P values or adjusted P values less than 0.05 were considered significant. An R package, ggplot2 (v3.2.0)[255], was used to generate figures in this work.

3.3 Results

3.3.1 Samples and participants

A total of 17 RSV-positive nasopharyngeal swabs were available at the time of this pilot study, including 6 from the infant case–control study in Oxford, United Kingdom and 11 from the birth cohort study in Santiago de Compostela, Spain. All samples came from participants younger than 1 year of age (median: 12 weeks, range: 6–26 weeks). Among them, two samples (ES-09 and ES-10) were collected 26 days apart from the same participant in Spain. Demographic and clinical information from the participants is shown in Table 3.1. Except for two participants born prematurely (at 30 and 36 weeks of gestation, respectively), none had comorbidities.

3.3.2 RSV genome and transcriptome reconstruction

All 17 samples were extracted, enriched for RSV nucleic acids, and sequenced on the Illumina MiSeq system. A median of 0.7 million read pairs per sample were generated (range 0.2–5.4 million), where read 2 corresponded to the original template strand. RSV genomes were reconstructed with *shiver*[227]. In this dataset, the saturated percent identity was 98–100%, and there was no decrease in mean percent identity when the depth of coverage became lower (Fig. 3.1). With a depth of coverage of two or more, the percent identity was above 97% in general, so a minimal depth of two reads was used for consensus calling.

Table 3.1: Demographic and clinical characteristics of the study participants.[a]

	UK (N = 6)	**Spain (N = 11)**	**Total (N = 17)**
Median age, wk (range)	13 (7–26)	12 (6–16)	12 (6–26)
Male sex, no. (%)	0	4 (36)	4 (24)
Gestational age			
Median, wk (range)	39 (30–41)	40 (38–42)	40 (30–42)
Prematurity, no. (%)[b]	2 (33)	0	2 (12)
Median birth weight, g (range)	3,302 (1,520–3,498)	3,230 (2,250–4,310)	3,231 (1,520–4,310)
Median days between symptom onset and sampling (range)	5 (3–9)	4 (1–9)	4 (1–9)
ReSVinet score, no. (%)[c]			
0–7	0	9 (82)	9 (53)
8–13	4 (67)	1 (9)	5 (29)
14–20	2 (33)	1 (9)	3 (18)
Hospital admission, no. (%)	6 (100)	1 (9)	7 (41)
HDU or ICU admission,no. (%)[d]	1 (17)	1 (9)	2 (12)

[a] Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

[b] Prematurity was defined as birth before 37 weeks of gestation.

[c] The ReSVinet score was determined on the basis of seven clinical parameters, including feeding intolerance, medical intervention, respiratory difficulty, respiratory frequency, apnoea, general condition, and fever. Each parameter ranges from 0 to 2 or 3 points, with the sum of 20 points. The higher the score, the more severe the RSV disease.

[d] HDU denotes high dependency unit, and ICU denotes intensive care unit.

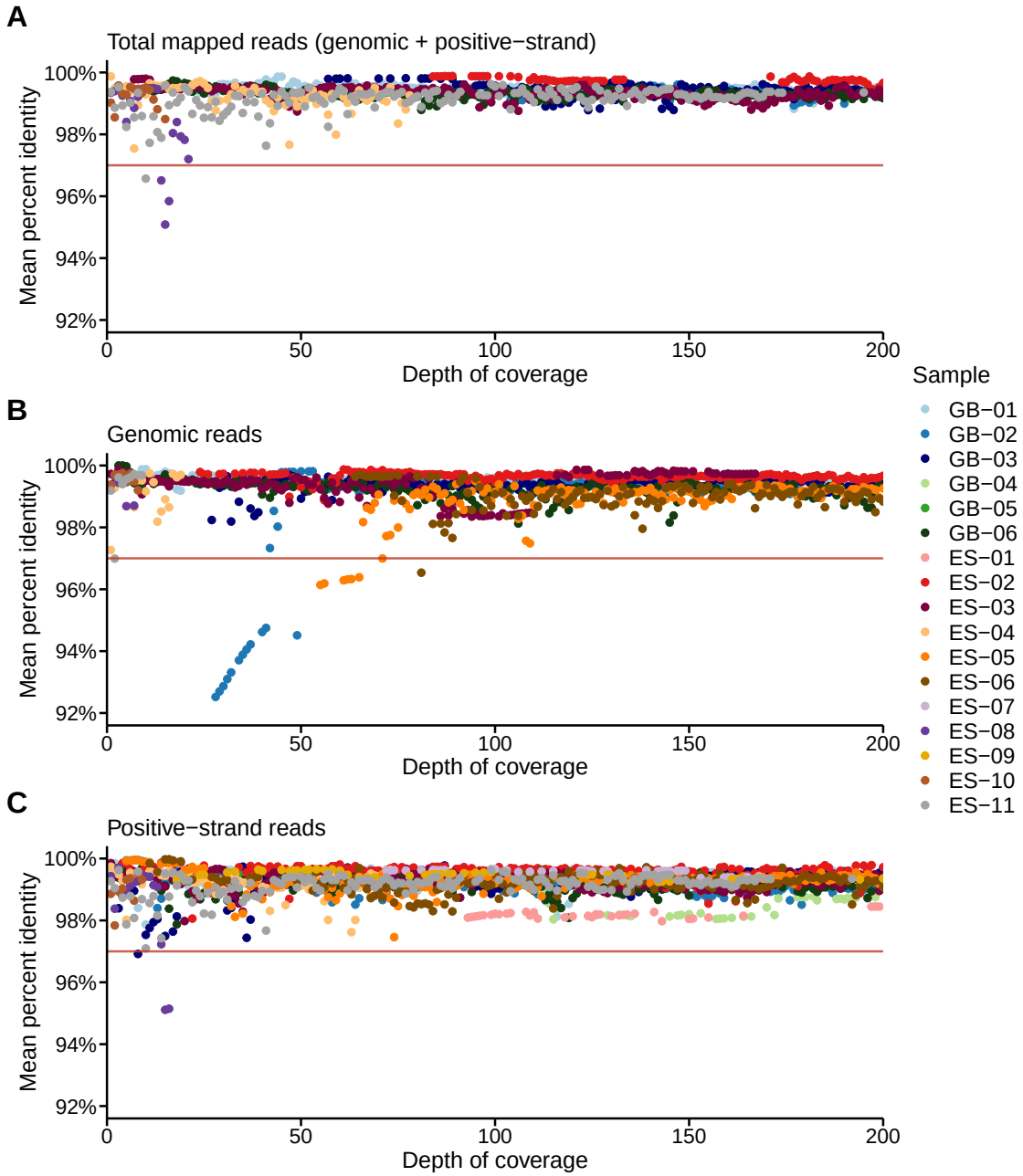


Figure 3.1: Mean percent identity of mapped reads for each depth of coverage. This analysis was done using a script, `LinkIdentityToCoverage.py`, bundled with *shiver*. The mean percent identities of the total, genomic, and positive-strand reads for depth of coverage between 1 and 200 are shown in **A**, **B**, and **C**, respectively. The horizontal line in each panel shows the mean percent identity of 97%.

A median of 30% of total sequencing reads originated from RSV (range 0.02%–75% among samples), and complete RSV genomes were recovered from 13 (76%) of the samples, including all samples that generated >10,000 unique (deduplicated) RSV reads. RSV viral load positively correlated with the number of unique RSV reads [$r(15) = 0.86$, $P < 10^{-5}$] (Fig. 3.2). Since the RSV quantitative RT-PCR primers targeted the N gene, the correlation between the number of unique reads mapped to the N gene and RSV viral load was also investigated. The correlation coefficient [$r(15) = 0.88$, $P < 10^{-5}$] was similar to that using the total mapped reads across the RSV genome [two-tailed Williams' t-test, $t(14) = -1.79$, $P = 0.09$].

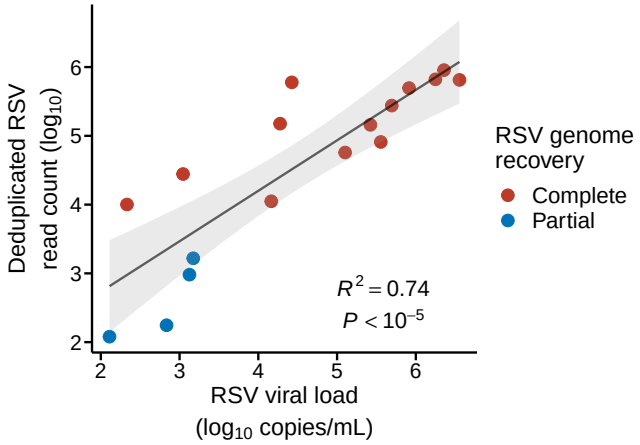


Figure 3.2: Sequencing yield and virus load. The number of uniquely mapped RSV reads (i.e., deduplicated) across the whole RSV genome positively correlated with RSV viral load. The shaded area around the regression line represents the 95% confidence interval. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

Among the 13 samples with complete RSV genome assembly, sequences of the F gene were conserved in the palivizumab-binding site (i.e., antigenic site II in the F protein, residuals 254–277)[28]; no mutation that has been previously associated with resistance to palivizumab was observed.

On average, mapped RSV reads originating from positive-strand templates, corresponding to transcripts and antigenomic replicative intermediates (hereafter called positive-strand reads) were 11 times more than mapped RSV reads originating from negative-strand templates, corresponding to genomic material (hereafter called

genomic reads) before removing read pairs that shared the same mapping coordinates on the reference (i.e., duplicates). After deduplication, the ratio of positive-strand reads to genomic reads decreased to 5. Table 3.2 shows the numbers of RSV read pairs either by mapping with total reads or by mapping with only genomic or positive-strand reads.

Maximum-likelihood phylogenetic analysis identified four RSV-A and nine RSV-B samples among the 13 completely sequenced genomes (Fig. 3.3). All RSV-A samples were assigned to genotype ON1, while all RSV-B samples were located in the clade of genotype BA. In all but one case, the nearest neighbour sequence for each new genome was another sample from this study, although not necessarily from the same country, reflecting the ongoing evolution and limited number of co-circulating RSV lineages.

Table 3.2: Viral load, sequencing yield, and characteristics of RSV for each sample.[a]

ID[b]	RSV viral load[c]	Total read pair	Mapped RSV read pair				RSV genome recovered	RSV Subgroup
			Pre-deduplication		Deduplication			
			Total[d]	Genomic[e]	Total[d]	Genomic[f]		
GB-01	1,108	320,553	30,161 (9)	16%	13,917 (4)	24%	100%	A
GB-02	126,552	688,518	57,472 (8)	21%	28,605 (4)	29%	100%	B
GB-03	18,899	994,215	407,076 (41)	3%	75,146 (8)	10%	100%	A
GB-04	820,113	2,194,241	1,157,676 (53)	15%	247,677 (11)	31%	100%	B
GB-05	26,609	5,368,901	3,448,805 (64)	6%	299,926 (6)	19%	100%	A
GB-06	215	2,226,316	10,837 (<1)	32%	5,017 (<1)	38%	100%	A
ES-01	2,280,533	4,680,563	3,461,692 (74)	19%	451,693 (10)	40%	100%	B
ES-02	358,940	323,314	98,596 (30)	15%	40,726 (13)	24%	100%	B
ES-03	14,599	571,463	10,550 (2)	10%	5,588 (1)	14%	100%	B
ES-04	1,336	479,279	644 (<1)	27%	479 (<1)	28%	86%	B
ES-05	265,222	656,179	240,795 (37)	8%	72,255 (11)	16%	100%	B
ES-06	494,971	1,278,140	557,200 (44)	4%	137,764 (11)	10%	100%	B
ES-07	1,774,886	2,961,629	2,087,121 (70)	8%	331,414 (11)	22%	100%	B
ES-08	684	416,569	101 (<1)	28%	88 (<1)	28%	31%	B
ES-09	3,584,263	3,503,558	2,638,892 (75)	6%	325,491 (9)	20%	100%	B
ES-10	129	393,796	81 (<1)	49%	60 (<1)	52%	23%	B
ES-11	1,492	223,796	1,245 (1)	3%	829 (<1)	4%	65%	B

[a] Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

[b] GB represents samples from the United Kingdom, and ES represents samples from Spain.

[c] Viral load is shown as copies/mL. The median viral loads of the RSV-A and RSV-B samples were 10,004 copies/mL and 265,222 copies/mL, respectively.

[d] Data are shown as number (percentage of total sequencing read pairs).

[e] Data are shown as percentage of mapped RSV read pairs (pre-deduplicated).

[f] Data are shown as percentage of mapped RSV read pairs (deduplicated).

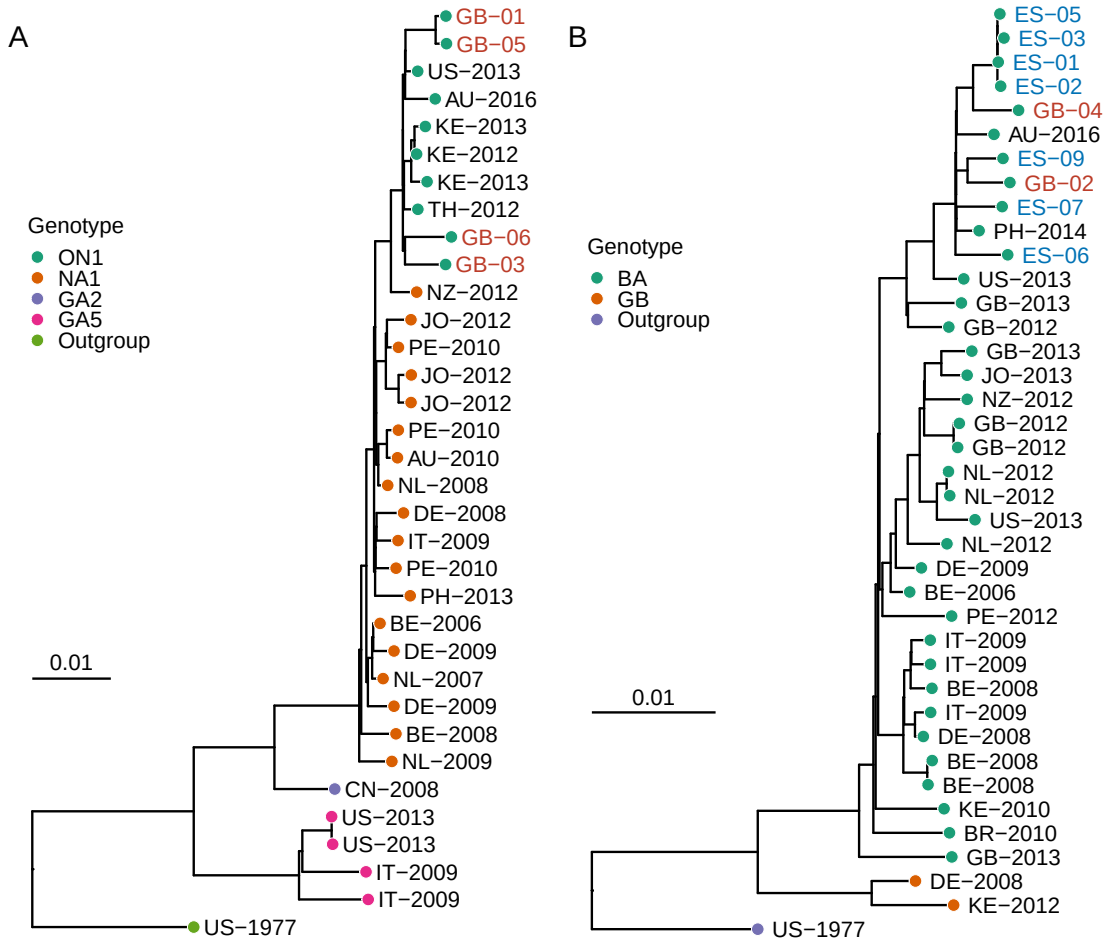


Figure 3.3: Maximum-likelihood phylogenies of RSV-A (**A**) and RSV-B (**B**). Phylogenies were reconstructed using RAxML from consensus sequences of the 13 complete genomes, augmented with 30 reference genomes representing sampled diversity of each of RSV-A and RSV-B. Sequences from this study are coloured in red (United Kingdom) or blue (Spain) and genotypes are coloured according to the legend. Reference taxa are labelled with country (AU, Australia; BE, Belgium; BR, Brazil; CN, China; DE, Germany; ES, Spain; GB, United Kingdom; IT, Italy; JO, Jordan; KE, Kenya; NL, Netherlands; NZ, New Zealand; PE, Peru; PH, Philippines; TH, Thailand; US, United States of America) and year of sampling. Scale bar shows nucleotide substitutions per site. The phylogenies were rooted to the oldest sampled RSV sequences, dating from 1977. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

3.3.3 Coverages of genomic and positive-strand reads

Read coverage plots of the genomic and positive-strand reads from the 13 samples with complete RSV genome assembly revealed characteristic, strand-specific patterns

of read depth along the genome and antigenome (Fig. 3.4 and Supplementary Fig. B.1). Read depth of genomic reads varied locally by up to 100-fold but revealed no trend across the genome (Fig. 3.4, panels A and B). In contrast, positive-strand read coverage was marked by read depths in intergenic regions ~1–10% of those in coding regions (Fig. 3.4, panels C and D), recapitulating the known mechanism of RSV transcription, with a single initiation point at the 3' end of the genome and polyadenylation-termination at each GE signal before reinitiating at the next GS signal[14]. The small number of positive-strand reads spanning intergenic regions presumably represent the detection of read-through transcripts and antigenomes.

The depth of positive-strand reads also decreased progressively along the genome from the 3' end to the 5' end, reflecting the possibility of subsequent dissociation of a fraction of the RNA-dependent RNA polymerase (RdRp) complex before reinitiation[14], leading to progressively weaker transcription. Under this model, due to an overlap between the M2 and L genes, expression of the 5'-most L gene requires the RdRp to terminate at the GE signal of the preceding M2 gene, and then to move backwards to the L GS signal before reinitiation[14]. This inference is supported by a jump in positive-strand read coverage at the start of the L gene (Fig. 3.5).

3.3.4 Relative transcript abundance of RSV genes

For the 13 samples where complete RSV genomes had been recovered, the number of positive-strand reads mapped to each RSV gene (before deduplication) was calculated using `htseq-count`[249]. Raw count data were converted to TPM, representing the transcript abundance. Consistent with the polarity paradigm of decreasing transcript levels along the RSV genome, a marked decrease was observed in the transcript level of the L gene, the median of which was less than 10% that of the other genes (Fig. 3.6, panel A; and Supplementary Fig. B.2). The overall level of transcription declined along the genome for all samples (median gradient −0.12, range −0.14 to −0.08), consistent with the gradient of −0.15 derived from an *in vitro* transcription system[15] [one-way analysis of covariance, $F(13,112) = 0.15$, $P > 0.99$] (Fig. 3.6, panel B; and Supplementary Fig. B.3). However, in these clinical

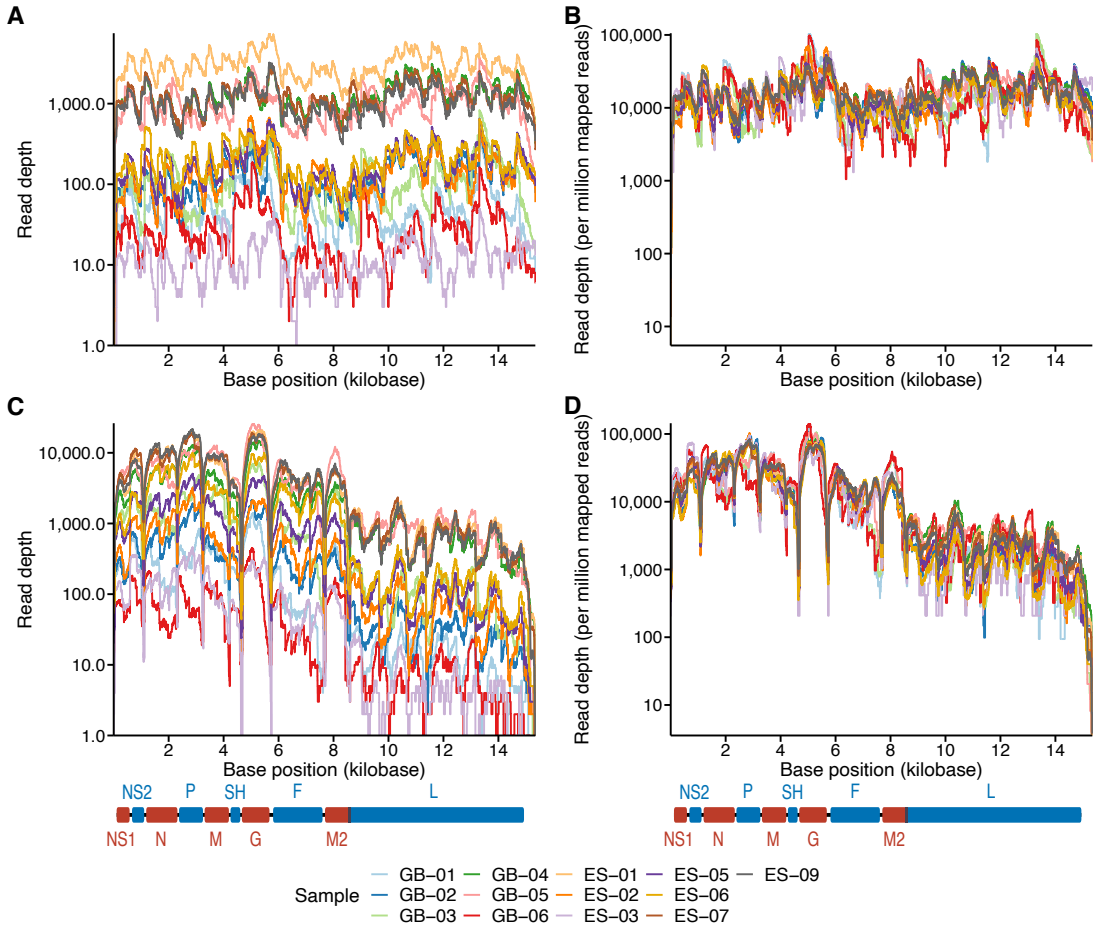


Figure 3.4: Coverage plots of the 13 complete RSV genomes and transcriptomes. **A.** Read coverage of complete RSV genomes. Mapped RSV reads derived from negative-sense templates (i.e., genomic reads) were used to generate the coverage plot. Data are shown as absolute read depth after deduplication. **B.** Read depth was scaled by the total number of deduplicated genomic reads for each sample. Data are shown as normalised read depth per million mapped genomic reads. **C.** Read coverage of RSV transcriptomes. As in **A**, but mapped RSV reads were derived from positive-sense templates (i.e., positive-strand reads). **D.** As in **B**, but read depth was scaled by the total number of deduplicated positive-strand reads for each sample. Legends below the plots represent the position and length of each RSV gene to scale. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

samples, it can be seen that in most samples transcripts from the NS2, P, and G genes were more abundant than those from the genes preceding them (Fig. 3.6, panel A), in contrast with the classical transcriptional gradient in cell-free systems[15].

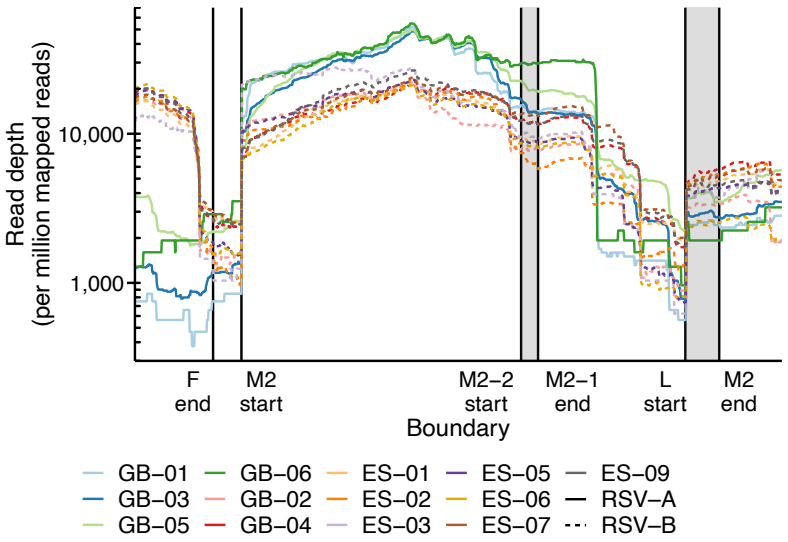


Figure 3.5: Positive-strand read coverage across the M2 gene. There is little evidence for read depth dropping at the end of the M2 gene; instead coverage drops within the open reading frame (ORF) M2-2, which is inconsistent with a single transcript covering both ORFs M2-1 and M2-2. Shaded areas represent gene or open reading frame overlaps. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

3.3.5 Differential gene expression analysis

The proportion of total expression associated with each gene varied among the 13 samples with complete RSV genome assembly (Fig. 3.7, panel A). Noting marked variation among isolates in the proportions of positive-strand reads assigned to each gene, a formal analysis of differential gene expression was undertaken following the RNAseq123 workflow[256] using the count data generated from `htseq-count`. Multidimensional scaling (MDS) analysis was used to contrast the transcriptomic profiles using different covariates: RSV subgroup, disease severity, and host sex. RSV subgroup, but not host sex or disease severity, was significantly associated with differences in specific transcript levels (Fig. 3.7, panel B; and Supplementary Fig. B.4). Only RSV subgroup correlated with differential expression for any genes: the mean level of M2 transcripts in RSV-A (samples GB-01, -03, -05, and -06) was 240% of that in RSV-B (GB-02, -04, and all Spanish isolates), while the expression of the P gene in RSV-A was 57% of that in RSV-B (Fig. 3.8, panels A and B).

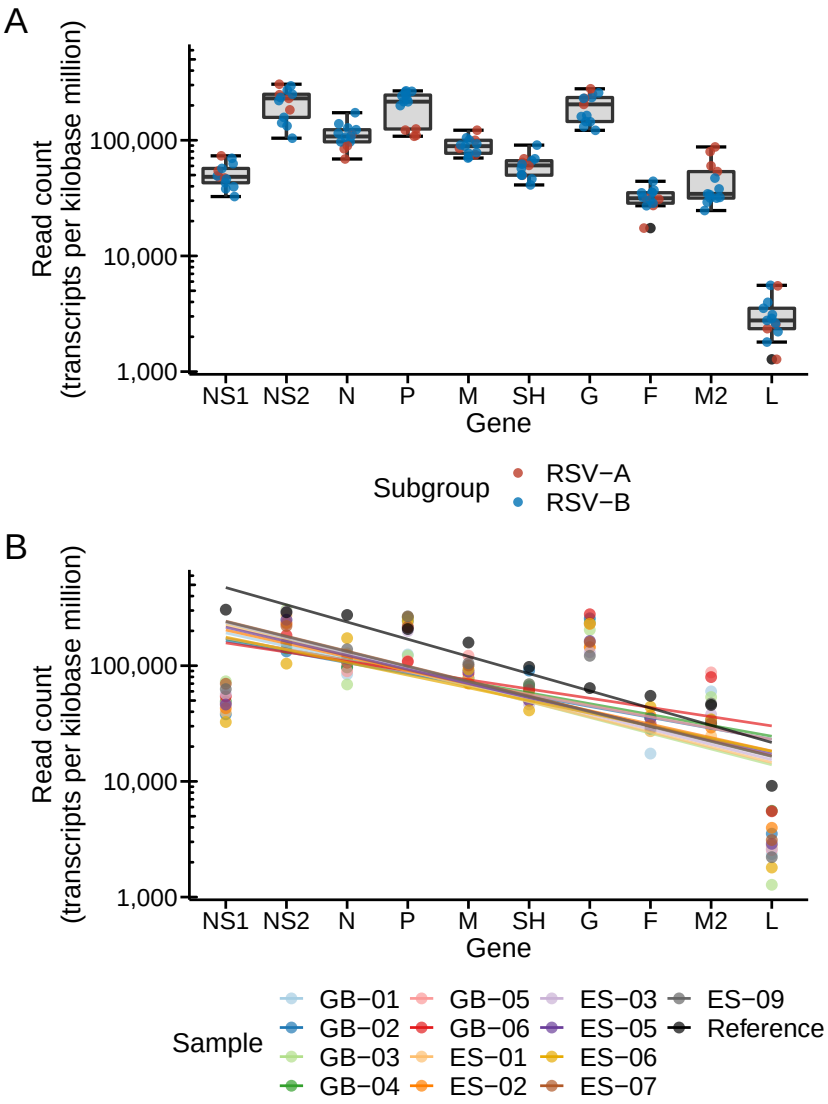


Figure 3.6: Relative transcript abundance of RSV genes. Total transcript counts for each gene in the 13 samples with complete RSV genome assembly were computed using `htseq-count` and transformed to transcripts per kilobase million RSV-mapped reads (TPM). RSV genes are displayed from left to right in 3' to 5' genomic order. Each point represents a single sample, coloured by RSV subgroup (**A**) or sample (**B**). **A.** The box plots show the pattern of gene expression levels across the RSV genome. The centre line of a box denotes the median of the dataset; box limits, first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338). **B.** The regression lines show the overall trend of transcript levels for each sample, together with the expected transcriptional gradient (black) derived using an *in vitro* transcription system[15]. The highest expression level observed from that system was scaled to the maximum TPM in our dataset.

While inter-sample variation was marked, perhaps reflecting clinical heterogeneity between samples, other genes were not significantly differentially expressed.

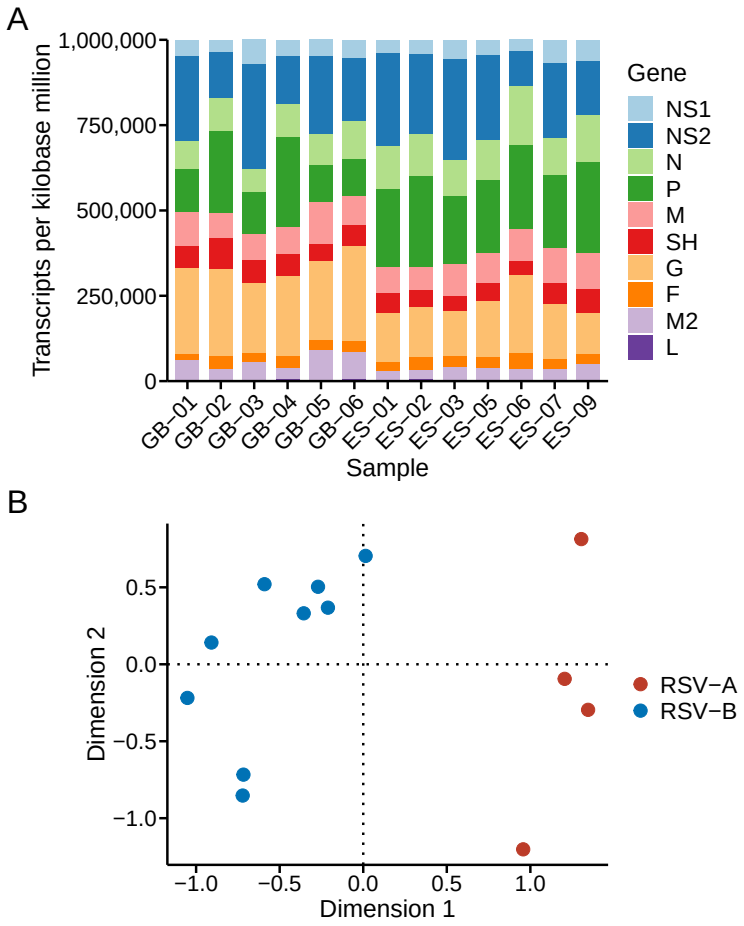


Figure 3.7: Evaluation of differential gene expression among the 13 complete transcriptomes. **A.** The stacked bar chart shows the number of transcripts per kilobase million mapped RSV reads for each gene across the 13 samples. **B.** Multidimensional scaling plot demonstrates the relationship between samples on a 2-dimensional scale, where more similar samples tend to cluster together. Each point represents a single sample coloured by subgroup. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

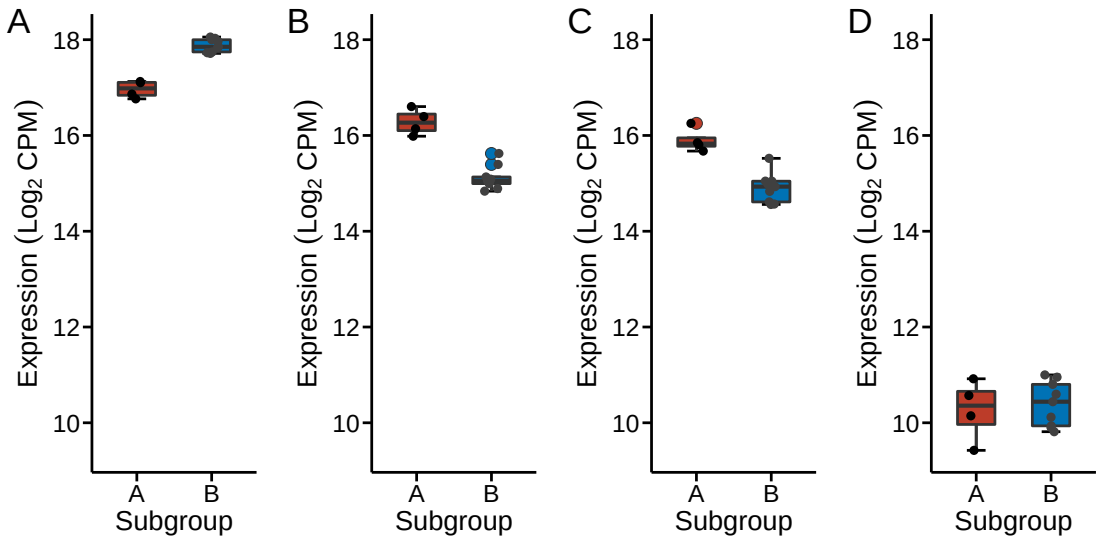


Figure 3.8: Differential expression of the P gene (**A**), M2 gene (**B**), open reading frames M2-1 (**C**), and M2-2 (**D**). RSV-A had lower expression of the P gene [fold change = −1.75, 95% confidence interval (CI) = −1.42 to −2.15, adjusted P = 0.009] and higher expression of the M2 gene (fold change = 2.38, 95% CI = 1.44 to 3.95, adjusted P = 0.032) than RSV-B. Within the M2 gene, RSV-A had higher expression of M2-1 than RSV-B (fold change = 2.29, 95% CI = 1.49 to 3.53, adjusted P = 0.017), while the expression of M2-2 was comparable between RSV-A and RSV-B (adjusted P = 0.703). Centre line of each box denotes the median; box limits, first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers. CPM denotes counts per million. Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

The M2 gene is known to contain two overlapping open reading frames (ORFs): M2-1, whose product interacts with the P gene product (phosphoprotein) to induce antitermination (promoting transcription of full-length subgenomic mRNAs), and M2-2, whose protein conversely acts to favour production of antigenomes over full-length transcripts[14]. Under the accepted model of RSV molecular biology, both ORFs belong to a single transcriptional unit. Given their contrasting roles, the transcription data were scrutinised more carefully to see whether the change in M2 transcription affected both ORFs. In this dataset (Fig. 3.5), (i) read depth corresponding to transcription drops by an aggregate 0.5–1 log-fold downstream of M2-1 but within M2-2; (ii) levels of transcription in the remainder of M2-2 are similar to intergenic levels in the rest of the genome; (iii) the signal for

enhanced transcription of the M2 gene in RSV-A is confined to the M2-1 region (Fig. 3.8, panels C and D).

3.4 Discussion

In this study, I demonstrate the simultaneous recovery of RSV genomes and transcriptomes directly from clinical samples, without a prerequisite for culture or purification of viral particles. Using strand-specific libraries with target enrichment, I was able to examine separately the RSV genome and transcriptome sequences and to relate viral gene expression to the phylogenetic subgroup of the infecting virus. I identified intriguing differences between RSV-A and RSV-B transcriptional profiles, which could provide novel pathogenetic and therapeutic insights.

The quantitative and high-resolution nature of this sequencing protocol would be useful in diagnostics and public health surveillance—one was able to obtain both viral load estimates and RSV sequences from a single assay on clinical samples. A linear relationship was observed between RSV viral load and the yield of uniquely mapped RSV reads in log space ($R^2 = 0.74$), indicating that the method is quantitative enough to be used as an estimate of clinical viral load, as has been previously noted for HIV[222]. Furthermore, high-resolution sequence data generated by this method would also facilitate monitoring of RSV sequence changes[257] and early detection of escape mutants, such as those with point mutations in the monoclonal antibody (e.g., palivizumab or motavizumab) binding sites of the F gene[258,259], in the face of immune pressure from up-and-coming vaccines and monoclonal antibodies. I will exploit these high-resolution sequence data to explore potential monoclonal antibody-resistant variants and RSV diversity within each RSV-infected patients enrolled in the RESCEU studies in chapter 5.

World Health Organization has launched a phase 2 global RSV surveillance pilot programme, which includes genetic characterisation and genome sequencing[260]. The great sensitivity and efficient workflow of this protocol, modified from veSEQ-HIV[222], makes it a potential sequencing strategy for the surveillance pilot. In addition, complete RSV genomes were reconstructed in all samples generating

>10,000 uniquely mapped RSV reads. This high sensitivity, achieved without further optimisation, makes the methodology of targeted metagenomic sequencing suitable for studying a wide variety of infections.

RSV transcript abundance was quantified using strand-specific sequencing. The relative expression level of each gene uncovered in this study was similar to that previously reported from RNA-seq of poly(A)-selected mRNAs from cultured samples[261], with G and P transcripts being the most and L transcripts being the least abundant. Interestingly, both poly(A)-selected RNA-seq[261] and this study do not wholly replicate the classical view of an RSV transcriptional gradient, i.e., uniformly decreasing transcript levels along the genome from the 3' end to the 5' end. This discrepancy may be due to different experimental methods, for example in the difficulty of recovering reads at transcript edges. It may also provide indirect support that during RSV infection there exists a more complex, post-transcriptional control of transcript abundance in addition to the well-described RdRp drop-out[261].

Differential expression analysis showed that the expression of the M2 gene was higher in RSV-A (genotype ON1) than in RSV-B (genotype BA), while the expression of the P gene was the opposite. RSV RNA replication involves the nucleoprotein (N), large polymerase (L), and its cofactor, phosphoprotein (P), all of which constitute the RdRp complex[13]. RNA transcription requires an extra protein, the M2-1 protein, encoded by the ORF M2-1 of the M2 gene, which switches the RNA synthesis from replication to transcription[13,262]. The M2-1 protein is also called the transcription processivity or antitermination factor, preventing the RdRp from premature termination at intergenic junctions[13,263]. The M2-2 protein, encoded by the ORF M2-2, is a regulatory factor of RNA synthesis in favour of RNA replication and viral packaging[14].

The nonconcordant expressions of the ORFs M2-1 and M2-2 lead to a hypothesis that the two ORFs may represent separate transcriptional units, and that the M2-2 protein may be translated from a low-abundance RNA of unknown length (including possibly read-through of M2-1 mRNA or genomic intermediate) and to conclude that the M2 differential expression signal relates specifically to M2-1. Since the M2-1

protein facilitates the synthesis of polycistronic read-through mRNAs[14], a testable hypothesis consistent with the transcriptomics result is that the RSV-A samples should have less marked declines in positive-strand read depth within intergenic regions than RSV-B due to more read-through mRNAs.

Differential expression of the P gene between subgroups is also intriguing because the direct interaction between the M2-1 and P proteins[264] means that the contrasting changes in gene expression could combine in complex ways. For example, a skew towards transcription (and thus antigen synthesis[265]) in RSV-A could foreseeably affect infection at both the cellular and whole-host levels, and may lead to generally higher prevalence of RSV-A in annual epidemics and proposed greater transmissibility of RSV-A than RSV-B[70]. However, during the study period of 2017–2018, RSV-B was predominant over RSV-A, accounting for 74% of the RSV samples collected in this multi-centre study. Although no study has found differences in transcriptional machinery that would potentially cause the differential expressions between subgroups, sequence differences or polymerase dissociation in intergenic regions may play a role.

Limitations of this study include varied duration from symptom onset to sampling time among the clinical samples. However, the median duration for the RSV-A and RSV-B samples was similar, 3.5 days (range 3–6 days) and 4 days (1–9 days), respectively. Moreover, each RSV sample could consist of multiple RSV species at various stages of their life cycle regardless of sampling time. Therefore, this analysis showing the differentially expressed M2 and P genes between RSV-A and RSV-B was robust to variable sampling time. Positive-sense genetic sequences produced by RSV include some antigenomic molecules alongside the sub-genomic transcripts. Removing reads originating from antigenomes is not achievable in this protocol, and the analysis of relative transcript abundance and differential gene expression must of necessity include some antigenomic molecules. However, positive-strand reads in the intergenic regions were only 1–10% as abundant as those in the coding regions (Fig. 3.4, panels C and D), as has been previously reported[14]. Since reads

originating from antigenomes only accounted for a small proportion of total reads, including them would have minimal effect on the transcriptomic analysis.

Another limitation is that I did not perform quantitative PCR to support the findings of differential expression. Nevertheless, RNA-seq has been shown to be highly accurate for quantifying transcriptomes, providing estimates of gene expression levels closely correlated with measurements using quantitative PCR[266].

To the best of my knowledge this is the first observation of relative transcript abundance and differential transcription of RSV genes using RNA-seq directly on clinical samples. The method can also be applied to genomic and transcriptomic studies of other RNA viruses. More generally, this methodology facilitates population-level RSV surveillance in the context of current efforts in global RSV surveillance[260] and vaccine development. With this approach, it is straightforward to relate viral load, sequence variation, and transcriptional profiles from sequencing data in order to investigate the epidemiological and phenotypic consequences (e.g. host biomarkers) of virus evolution. Targeted metagenomics provides a powerful methodology for characterising sets of pathogens that may co-occur in a single sample type.

Since the methodology of targeted metagenomic sequencing worked well on the clinical samples in this pilot study, this method was used to sequence all RSV-positive nasopharyngeal swabs collected from the RESCEU studies during 2017–2020. In the following chapters I will present the results derived from different analyses of these sequencing data.

4

Factors associated with respiratory syncytial virus disease severity

Contents

4.1 Introduction 63
4.2 Methods 64
4.2.1 Sample source and sequencing 64
4.2.2 Phylogenetic analysis 65
4.2.3 Respiratory virome reconstruction 65
4.2.4 Respiratory bacterial microbiome reconstruction 66
4.2.5 Statistical analyses 67
4.3 Results 68
4.3.1 Study population 68
4.3.2 Sequencing yield of RSV 70
4.3.3 Molecular epidemiology of RSV 73
4.3.4 Respiratory virome composition 78
4.3.5 Respiratory bacterial microbiota 79
4.4 Discussion 83

4.1 Introduction

Co-infection and the respiratory microbiome in various viral respiratory infections have attracted great interest recently. Simultaneous detection of multiple viruses is common in children with respiratory infections[267–271], and virus-virus interactions

can be either synergistic or antagonistic. For example, human rhinovirus (HRV), a common cause of acute respiratory infection, has been shown to interfere with influenza virus, reducing the chance of co-detection and co-circulation of both viruses at the individual and population levels[272]. During the first year of life, there are constant changes in the respiratory microbiome[273] with increasing diversity[274]. Colonisation with *Moraxella* spp. and, to a lesser extent, *Haemophilus* and *Streptococcus* spp. has been shown to be associated with viral respiratory infections[273].

Despite high prevalence of colonisation and infection, the impact of other potential pathogens on the progression of RSV infection remains to be understood. A consensus regarding associations between the presence of other pathogens and RSV disease severity has yet to be well established. Multiplex molecular testing and typing of common respiratory viral and bacterial pathogens have been increasingly used in clinical setting. Elucidating these pathogen interactions in RSV infection would provide valuable insight into the pathogenic mechanism of RSV disease.

In this chapter, I evaluate the role of other common bacterial and viral pathogens in RSV infection in infants using a multi-pathogen enrichment panel, *Castanet*[224], to simultaneously identify and sequence targets from over 100 common respiratory pathogens, including RSV. I also seek to explore the host and viral factors associated with severe RSV infection. A better understanding of the host, viral, and microbial factors associated with RSV disease severity will help advance the development of therapeutic and preventative measures.

4.2 Methods

4.2.1 Sample source and sequencing

All RSV-positive nasopharyngeal swabs collected from infants enrolled in the RESCEU birth cohort and infant case–control studies in the Netherlands, Spain, and the United Kingdom during the 2017–2020 RSV seasons were included in this study. Swabs collected at the time of acute infection in both studies and daily swabs collected during hospitalisation from hospitalised participants in the case–control study were both included. The clinical studies and sample collection are described

in detail in section 2.1. Nucleic acid isolation, library construction, sequencing, RSV genome reconstruction, and quantitative RT-PCR were performed as described in sections 2.2, 2.3, 2.4, and 2.5 respectively.

Viral load of a small subset of samples was determined using a quantitative RT-PCR assay performed at GlaxoSmithKline as previously described elsewhere[220]. The primers of this duplex RT-PCR assay targeted the N gene for both RSV-A and RSV-B. The limit of detection (LOD) was 304 copies/mL of swab for RSV-A and 475 copies/mL for RSV-B.

4.2.2 Phylogenetic analysis

RSV consensus sequences covering at least 70% of the coding sequences were used to reconstruct phylogenetic trees. For participants with multiple samples collected, only the earliest sample with the highest coverage was included. To put the study strains in a global context, I also included other RSV strains downloaded from GenBank with at least 70% coverage of the coding sequence collected in different regions of the world from 2015 onwards. RAxML (v8.2.12)[242] was used to reconstruct the maximum-likelihood phylogenies with the general time reversible model[243] and gamma-distributed rate heterogeneity among sites (GTR+Γ). The R package ggtree (v2.2.4)[247] was used for tree visualisation. Patristic distances, used to measure phylogenetic distances between sample pairs, were computed using the `cophenetic` function in R (v4.0.2)[245].

4.2.3 Respiratory virome reconstruction

The *Castanet* pipeline (Tanya Golubchik, available at `https://github.com/tgolubch/castanet`) was used to reconstruct viral genomes for each sample[224]. Reads originating from human cells were removed using BBMap (Brian Bushnell, available at `https://sourceforge.net/projects/bbmap/`). The remaining (host-depleted) reads were mapped using bowtie 2 to the references that covered known sequence variability of the target species. Read depth, genome coverage, and consensus sequences were generated for each target species from BAM files.

Species with more than 20–50 uniquely mapped reads and 30% genome recovery were considered present, and those with <10 unique reads were considered absent. Species fulfilling neither criteria were considered equivocal. Further criteria varied between species due to different efficiencies of the target enrichment and sequence specificities; detailed criteria for each species are shown in Supplementary Table A.2. For patients with multiple samples collected, species that were detected in at least one sample were considered present; species that were not detected in any sample were considered absent; and the rest were considered equivocal. Patients with an equivocal presence of a species were excluded from statistical tests for associations between clinical features and the presence of this species.

Enterovirus genotyping was performed using a web-based typing tool based on phylogenetic analysis of the VP1 gene at `https://www.rivm.nl/mpf/typingtool/enterovirus/`[275]. The genotypes of human herpesvirus 6 (HHV-6) were determined by phylogenetic analysis of the HHV-6 partial genome consensus sequences (approximately from the U23 to the U37 genes) and reference strains. Reference strains included HHV-6A (strains: U1102, GS, and AJ; accession numbers: NC_001664, KC465951, and KP257584), HHV-6B (strains: Z29 and HST; accession numbers: NC_000898 and AB021506), chromosomally integrated HHV-6A (accession numbers: MG894371 and KY315540), and chromosomally integrated HHV-6B (accession numbers: KY316051 and MG894376).

4.2.4 Respiratory bacterial microbiome reconstruction

The *Castanet* gene targets contained rMLST loci[226] for all common respiratory bacterial pathogens. Reconstruction of the bacterial microbiome was based on the rMLST scheme. Contig sequences assembled by metaSPAdes were matched against all known alleles for each *rps* gene to find exact matches. The PubMLST multi-species isolate database, integrating curated allelic and species information, was then searched to identify the matching species (algorithm available at `https://pubmlst.org/bigsdb?db=pubmlst_rmlst_seqdef_kiosk`)[226]. The PubMLST

RESTful application programming interface (API) was applied to efficiently analyse all samples, using the command:

```
(echo -n '{"base64":true,"details":true,"sequence": "'; base64
contigs.fasta; echo '"}') | curl -s -H "Content-Type: application/json"
-X POST "http://rest.pubmlst.org/db/pubmlst_rmlst_seqdef_kiosk/
schemes/1/sequence" -d @- > output.json
```

Species supported by at least two exact and unique allelic matches were considered present. A 'unique' allelic match was an allelic pattern that was linked to exactly one taxon at its lowest taxonomic rank (e.g., *Moraxella catarrhalis* or *Enterobacteriaceae*) or to two species with one being an unspecified species of the same genus as the other specified species (e.g., *Moraxella* sp. and *Moraxella catarrhalis*, where *Moraxella* sp. was ignored). Species supported by only one exact and unique match were considered equivocal. The rest were considered absent. For patients with multiple samples collected, species that were detected in at least one sample were considered present; species that were not detected in any sample were considered absent; and the rest were considered equivocal. Patients with an equivocal presence of a species were excluded from statistical tests for associations between clinical features and the presence of this species. Krona charts were used to visualise the composition of identified bacteria[276].

4.2.5 Statistical analyses

Continuous data were summarised using minimum, mean, median, maximum, interquartile range (IQR), and standard deviation. Count data were summarised using number and percentage. Linear regression and Pearson correlation were used to assess the relationship between two continuous variables. Mann–Whitney U tests were used to compare continuous variables between two groups. Kruskal–Wallis tests were used when there were more than two groups, and pairwise comparisons were conducted using Mann–Whitney U tests with the Benjamini–Hochberg procedure. Chi-square tests using Yates' correction were employed for contingency analyses, and Fisher's exact tests were used when the expected value for a cell was less than five.

Multiple linear regression (for continuous outcome variables), multivariable logistic regression (for dichotomous outcome variables), and proportional odds ordered logistic regression (for ordered outcome variables) were used to adjust for covariates (e.g., gestational age, participant age, sex, comorbidity, and RSV subgroup). Analysis of covariance (ANCOVA) tests were performed to compare the ReSVinet score between three age groups while adjusting for gestational age, and to compare the rates of viral load decline. A post hoc adjustment for multiple comparisons with the Benjamini–Hochberg procedure was applied to determine false discovery rate–corrected Q values in all association analyses between the demographic/clinical features and the presence/absence of any pathogen. The probability of fever was predicted on the basis of a multivariable logistic regression model with age (three groups) and the number of peak RSV reads as explanatory (independent) variables. Each age group had 100 simulated data points with the number of peak RSV reads distributed evenly between the lowest and the highest counts in the dataset. All statistical analyses were performed using R. A P or Q value of less than 0.05 was considered to indicate statistical significance.

4.3 Results

4.3.1 Study population

A total of 440 infants, enrolled from the United Kingdom, Spain, and the Netherlands during 2017–2020, tested positive for RSV using point-of-care molecular testing, antigen tests, or PCR, depending on the clinical setting (Supplementary Table A.3). A single nasopharyngeal swab was obtained from each infant, except for 125 hospitalised infants in the case–control study who had a mean (standard deviation, SD) of 4.2 (2.5) daily swabs collected. In total, 839 nasopharyngeal swabs were collected and sequenced. One infant's sample failed to be sequenced, and six infants' samples failed to generate any RSV reads. These seven infants and their samples (n = 8) were excluded from this study.

Of the remaining 433 infants, 143 (33%) were enrolled in the birth cohort study, and 290 (67%) were enrolled in the infant case–control study. The median age

Table 4.1: Characteristics of the RSV-infected infants by study group.[a]

	Birth cohort study (N = 143)	**Case–control study (N = 290)**	**P value**
Age			
Median (IQR)—mo	5.7 (3.6–8.9)	3.0 (1.4–6.3)	1.7×10^{-10}
Distribution[b]			5.3×10^{-10}
<3 mo	25/142 (18)	145/289 (50)	
3 to <6 mo	50/142 (35)	67/289 (23)	
≥6 mo	67/142 (47)	77/289 (27)	
Gestational age			
Median (IQR)—wk	39.9 (39.0–40.9)	39.4 (38.0–40.3)	7.4×10^{-5}
Distribution[b]			0.0003
<32 wk	0/139 (0)	8/287 (3)	
32 to <37 wk	0/139 (0)	17/287 (6)	
≥37 wk	139/139 (100)	262/287 (91)	
Female sex[b]	67/142 (47)	123/289 (43)	0.42
Comorbidity[b]	0/143 (0)	38/289 (13)	1.9×10^{-7}
ReSVinet score			
Mean ± SD	4.7 ± 2.6	8.9 ± 4.7	3.5×10^{-10}[c]
Distribution[b]			3.9×10^{-10}[c]
0–7	120/136 (88)	120/283 (42)	
8–13	15/136 (11)	107/283 (38)	
14–20	1/136 (1)	56/283 (20)	
Fever[b]	53/136 (39)	90/284 (32)	0.70[c]
Hospitalisation[b]	7/115 (6)	213/289 (74)	2.2×10^{-15}[c]
PICU admission[b]	1/115 (1)	76/289 (26)	0.0032[c]
Any respiratory support[b]	3/112 (3)	175/287 (61)	2.5×10^{-9}[c]
Mechanical ventilation[b]	0/112 (0)	66/287 (23)	0.99[c]

[a] IQR denotes interquartile range; SD, standard deviation; PICU, paediatric intensive care unit.

[b] Data are shown in no./total no. (%). Percentages may not total 100 due to rounding.

[c] Multiple linear regression, ordered logistic regression, and multivariable logistic regression were used to adjust for age, gestational age, sex, and comorbidity.

of the 433 infants at the time of RSV infection was 4.1 months (IQR, 1.8–7.4 months). On average, RSV-infected infants in the birth cohort study were older at the time of RSV infection and had less severe RSV disease than those in the infant case–control study (Table 4.1). This was because infants in the birth cohort study did not have any comorbidity and were actively followed up during the RSV

seasons and tested whenever they developed acute respiratory symptoms, whereas those in the infant case–control study were enrolled from either communities or hospitals and some of them had comorbidities (see sections 2.1.1 and 2.1.2 for details). These comorbidities included prematurity (<37 weeks of gestational age) with or without bronchopulmonary dysplasia, congenital hypothyroidism, wheeze, and other congenital abnormalities. The demographic characteristics of these infants stratified by age are shown in Supplementary Table A.4.

After combining RSV-infected infants enrolled in either studies, the associations between clinical severity and age were evaluated using different clinical outcomes (Table 4.2). Despite multiple clinical parameters suggesting a less severe RSV infection, older infants were more likely to have a fever than younger infants.

4.3.2 Sequencing yield of RSV

Of the 831 samples where sequencing was possible, 597 (72%) and 631 (76%) had sufficient viral load to reconstruct at least 50% and 70% of the RSV genome, respectively. Among samples collected within 7 days of symptom onset, 89% (540/604) were able to generate at least 50% of the genome. The proportion decreased to 43% (80/188) for samples collected between 8 and 14 days, and to 0% for samples collected after 14 days of symptom onset.

Consistent with the pilot study (chapter 3)[1], the number of uniquely mapped RSV reads highly correlated with RSV viral load ($R^2 = 0.9$, $P < 0.001$; Fig. 4.1). However, separation of the regression lines between RSV-A and RSV-B can be seen from viral load generated using the genesig RSV Standard assay (Fig. 4.1, panel A), which may result from different efficacies or sensitivities of the PCR primers against the two subgroups. No separation between the two subgroups was found from viral load generated by GlaxoSmithKline (Fig. 4.1, panel B). Given that a significant correlation was found between read counts and viral load, numbers of unique RSV reads (hereafter referred to as viral burden) were used to represent conventional viral load for the analyses that followed.

Table 4.2: Clinical characteristics of the RSV-infected infants, stratified by age group.[a]

	<3 mo (N = 170)	**3 to <6 mo (N = 117)**	**≥6 mo (N = 144)**	**P value**
ReSVinet score				
Mean ± SD	9.6 ± 4.9	6.7 ± 4.3	5.8 ± 3.3	1.0×10^{-12}[b]
Distribution[c]				3.0×10^{-11}[d]
0–7	63/166 (38)	74/113 (66)	102/139 (73)	
8–13	60/166 (36)	29/113 (26)	33/139 (24)	
14–20	43/166 (26)	10/113 (9)	4/139 (3)	
Fever[c]	33/167 (20)	39/113 (35)	71/139 (51)	6.5×10^{-8}[d]
Hospitalisation[c]	131/168 (78)	47/104 (45)	42/131 (32)	2.1×10^{-15}[d]
PICU admission[c]	58/168 (35)	15/104 (14)	4/131 (3)	2.1×10^{-11}[d]
Any respiratory support[c]	116/167 (69)	36/103 (35)	26/128 (20)	$< 2.2 \times 10^{-16}$[d]
Mechanical ventilation[c]	54/167 (32)	9/103 (9)	3/128 (2)	2.7×10^{-12}[d]

[a] SD denotes standard deviation; PICU, paediatric intensive care unit.

[b] Analysis of covariance (ANCOVA) test was used to adjust for gestational age. Pairwise comparisons using Mann–Whitney U tests with the Benjamini–Hochberg procedure showed the youngest group (<3 mo) had a significantly lower ReSVinet score than the older groups (3 to <6 mo and ≥6 mo) ($P = 5.8 \times 10^{-7}$ and $P = 1.2 \times 10^{-11}$ respectively), but no significant difference in ReSVinet score between the older groups ($P = 0.4$).

[c] Data are shown in no./total no. (%). Percentages may not total 100 due to rounding.

[d] Chi-square tests with Yates' correction. When using proportional odds ordered logistic regression or multivariable logistic regression to adjust for gestational age, the results were still significant.

After adjusting for age and days between symptom onset and sampling, both of which affected clinical outcomes and/or viral burden, neither the peak nor the initial viral burden correlated with the ReSVinet score, the need for hospitalisation or intensive care, or the requirement of any respiratory support or mechanical ventilation (Table 4.3). However, infants with a high peak or initial viral burden were more likely to have a febrile RSV infection than those with a low viral burden after adjusting for age and days between symptom onset and sampling ($P = 0.002$ and $P = 0.004$ for peak and initial viral burden, respectively). Figure 4.2 shows that the probability of fever increased with age and peak viral burden.

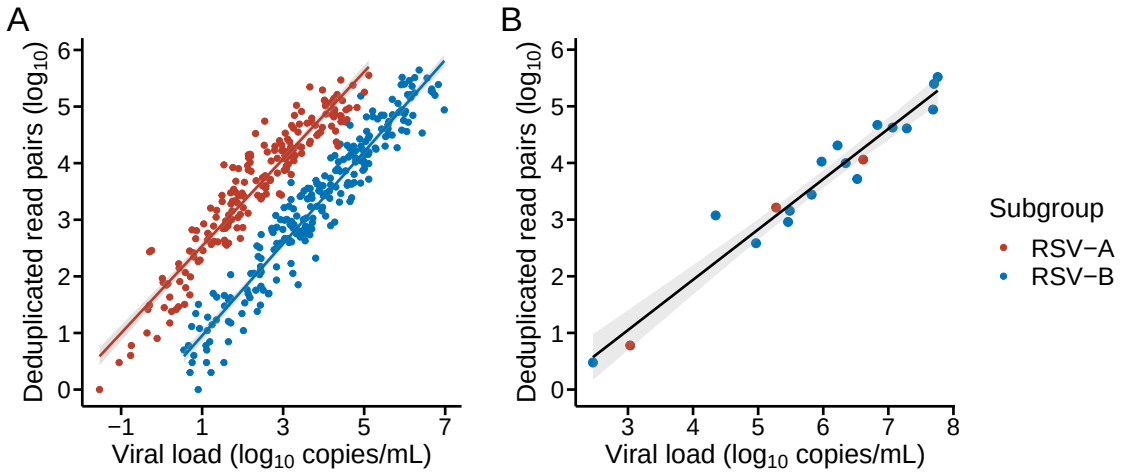


Figure 4.1: Correlation between the number of RSV reads and RSV viral load. **A.** The genesig RSV Standard assay (Primerdesign, Southampton, United Kingdom) performed by me. The coefficient of determination (R^2) of RSV-A was 0.89, and of RSV-B was 0.92. **B.** Quantitative RT-PCR assay performed by GlaxoSmithKline. R^2 of the correlation was 0.95. P values of the three correlations were all < 0.001.

Table 4.3: Associations between clinical variables and peak viral burden.[a]

	Test	No.	T/Z[b]	P
Subgroup	Ordered logistic regression	412	−1.49	0.136
ReSVinet score	Multiple linear regression	415	1.90	0.058
Severity[c]	Ordered logistic regression	415	1.62	0.105
Fever	Multiple logistic regression	416	3.06	0.002
Hospitalisation	Multiple logistic regression	393	1.23	0.217
PICU admission[d]	Multiple logistic regression	393	1.00	0.320
Any respiratory support	Multiple logistic regression	388	−0.12	0.902
Mechanical ventilation	Multiple logistic regression	388	1.55	0.121

[a] For subgroup, an ordered logistic regression model was used to adjust for days between symptom onset and sampling. For other variables, regression models were used to adjust for both age and days between symptom onset and sampling. Participants with missing information on the confounding or outcome variables were excluded from the regression analyses.

[b] For multiple linear regression and ordered logistic regression, the test statistics are T values. For multiple logistic regression, the test statistics are Z values.

[c] Severity was defined using the ReSVinet score: 0–7, mild; 8–13, moderate; and 14–20, severe.

[d] PICU denotes paediatric intensive care unit.

Temporal change of the viral burden was analysed using serial samples collected from hospitalised infants. RSV burden decreased by 53% daily (95% confidence

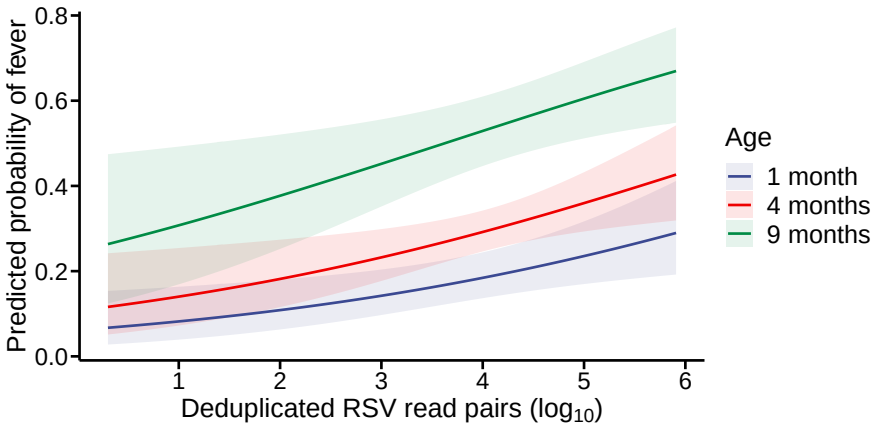


Figure 4.2: Predicted probability of fever for different ages and peak RSV read counts (see section 4.2.5 for details). Simulated data points were distributed evenly between 0.3 $\log_{10}$ and 5.9 $\log_{10}$, the lowest and the highest counts in this dataset.

interval [CI], 49% to 57%) over the 12 days from symptom onset (Fig. 4.3, panel A). Viral clearance was significantly slower in infants requiring mechanical ventilation (daily decrease 49%; 95% CI, 40% to 56%) than in those who did not (daily decrease 60%; 95% CI, 54% to 66%; P = 0.020) (Fig. 4.3 panel B; Table 4.4). However, patients in different age groups or other clinical outcome groups had similar rates of change in viral burden (Table 4.4).

4.3.3 Molecular epidemiology of RSV

Phylogenetic analyses were used to classify RSV into subgroups and genotypes. In total, 216 infants had an RSV-A infection, 207 had an RSV-B infection, nine had a mixed subgroup infection (or cross-sample contamination during library construction or sequencing), and one infant's sample failed to yield enough reads to determine the subgroup. During the 2017–18 and 2018–19 RSV seasons, RSV-B was the predominant subgroup, accounting for 74% (52/70) and 55% (89/161) of the infections, respectively. In contrast, RSV-A dominated the 2019–20 season, accounting for 65% (131/202) of the infections. The incidences of the two subgroups in each country and season are shown in Supplementary Table A.5. RSV-A-infected infants and RSV-B-infected infants had similar demographic profiles and clinical

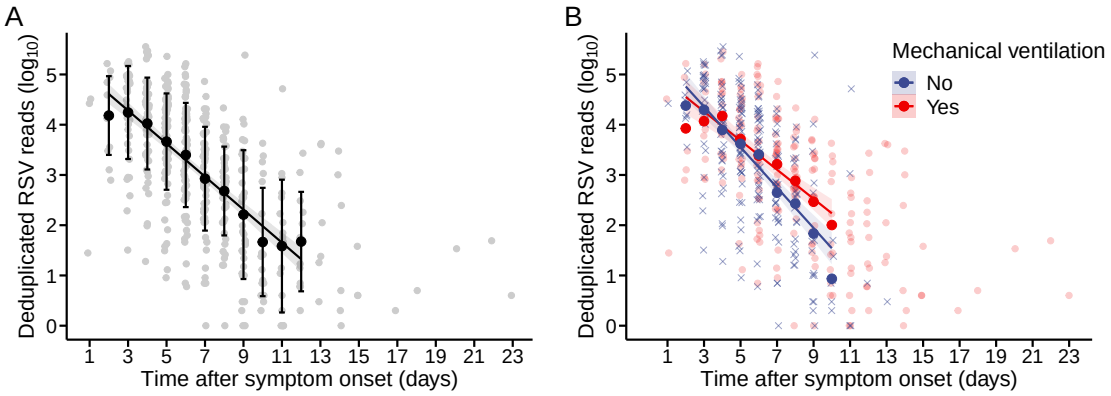


Figure 4.3: Temporal change of RSV read counts. **A.** The change of RSV read counts in all serial samples (N = 463). Each grey dot is a single sample. Black dots and error bars denote the mean value and standard deviation of all data at each time point, respectively. The line across different days is the regression line (R^2 = 0.38), and the shaded area around this line represents the 95% confidence interval. The summary statistics and linear regression were only performed on days with at least 10 samples, which were between day 2 and day 12. **B.** The difference in viral clearance between samples collected from patients requiring mechanical ventilation and those from patients who did not need mechanical ventilation. Similar to **A**, the data were summarised by mean values and regression lines. The summary statistics and linear regression were only performed on days with at least five samples for each group, which were between day 2 and day 10.

severity in the dataset (Table 4.5). RSV burden and its temporal change were also similar between the two groups of infants (Tables 4.3 and 4.4).

All RSV-A and RSV-B strains in this study were classified as genotypes ON1 and BA, respectively. Both genotypes had a distinctive nucleotide duplication in the second hypervariable region of the G gene—72-nt duplication for ON1 and 60-nt duplication for BA. Figure 4.4 shows the phylogenetic relationships between the samples in a global context. ON1 strains had greater patristic distances on average than BA [median (IQR), 0.0109 (0.0089–0.0119) vs. 0.0058 (0.0044–0.0075); Mann–Whitney U test, $P < 0.001$], whereas BA strains showed stronger temporal clustering than ON1.

Table 4.4: Associations between clinical variables and daily percentage decrease in viral burden.[a]

	Daily decrease (95% CI)	**P value**[b]
Age		0.713
<3 mo	55% (47%–61%)	
3 to <6 mo	51% (31%–65%)	
≥6 mo	49% (32%–61%)	
Subgroup		0.309
A	58% (52%–64%)	
B	54% (45%–61%)	
Severity[c]		0.589
Mild	44% (8%–66%)	
Moderate	57% (44%–68%)	
Severe	50% (29%–65%)	
Fever		0.696
No	53% (47%–59%)	
Yes	55% (48%–62%)	
PICU admission		0.059
No	60% (53%–65%)	
Yes	50% (42%–57%)	
Any respiratory support		0.628
No	57% (26%–76%)	
Yes	51% (39%–60%)	
Mechanical ventilation		0.020
No	60% (54%–66%)	
Yes	49% (40%–56%)	

[a] Only serial samples collected from hospitalised infants in the case–control study were included in this analysis. CI denotes confidence interval, and PICU denotes paediatric intensive care unit.

[b] Analysis of covariance (ANCOVA) tests were performed to adjust for days between symptom onset and sampling.

[c] Severity was defined using the ReSVinet score: 0–7, mild; 8–13, moderate; and 14–20, severe.

Table 4.5: Characteristics of the RSV-infected infants by RSV subgroup.[a]

	RSV-A (N = 216)	**RSV-B (N = 207)**	**P value**
Age			
Median (IQR)—mo	4.5 (2.0–7.4)	3.8 (1.7–7.5)	0.29
Distribution[b]			0.28
<3 mo	77/215 (36)	88/206 (43)	
3 to <6 mo	64/215 (30)	50/206 (24)	
≥6 mo	74/215 (34)	68/206 (33)	
Gestational age			
Median (IQR)—wk	39.8 (38.5–40.6)	39.6 (38.6–40.3)	0.53
Distribution[b]			0.54
<32 wk	3/210 (1)	5/206 (2)	
32 to <37 wk	10/210 (5)	6/206 (3)	
≥37 wk	197/210 (94)	195/206 (95)	
Female sex[b]	94/215 (44)	92/206 (45)	0.92
Weight—total no.[c]	141	131	
Mean ± SD—kg	6.2 ± 1.9	6.1 ± 2.1	0.53
Comorbidity[b]	16/216 (7)	21/206 (10)	0.40
ReSVinet score			
Mean ± SD	7.4 ± 4.3	7.6 ± 4.9	0.95
Distribution[b]			0.57
0–7	126/212 (59)	110/197 (56)	
8–13	61/212 (29)	57/197 (29)	
14–20	25/212 (12)	30/197 (15)	
Fever[b]	79/212 (37)	60/198 (30)	0.17
Hospitalisation[b]	112/201 (56)	101/193 (52)	0.57
PICU admission[b]	31/201 (15)	42/193 (29)	0.14
Any respiratory support[b]	85/197 (43)	87/192 (45)	0.74
Mechanical ventilation[b]	27/197 (14)	36/192 (19)	0.23

[a] IQR denotes interquartile range; SD, standard deviation; PICU, paediatric intensive care unit.

[b] Data are shown in no./total no. (%). Percentages may not total 100 due to rounding.

[c] Only infants in the case–control study had information about current body weight.

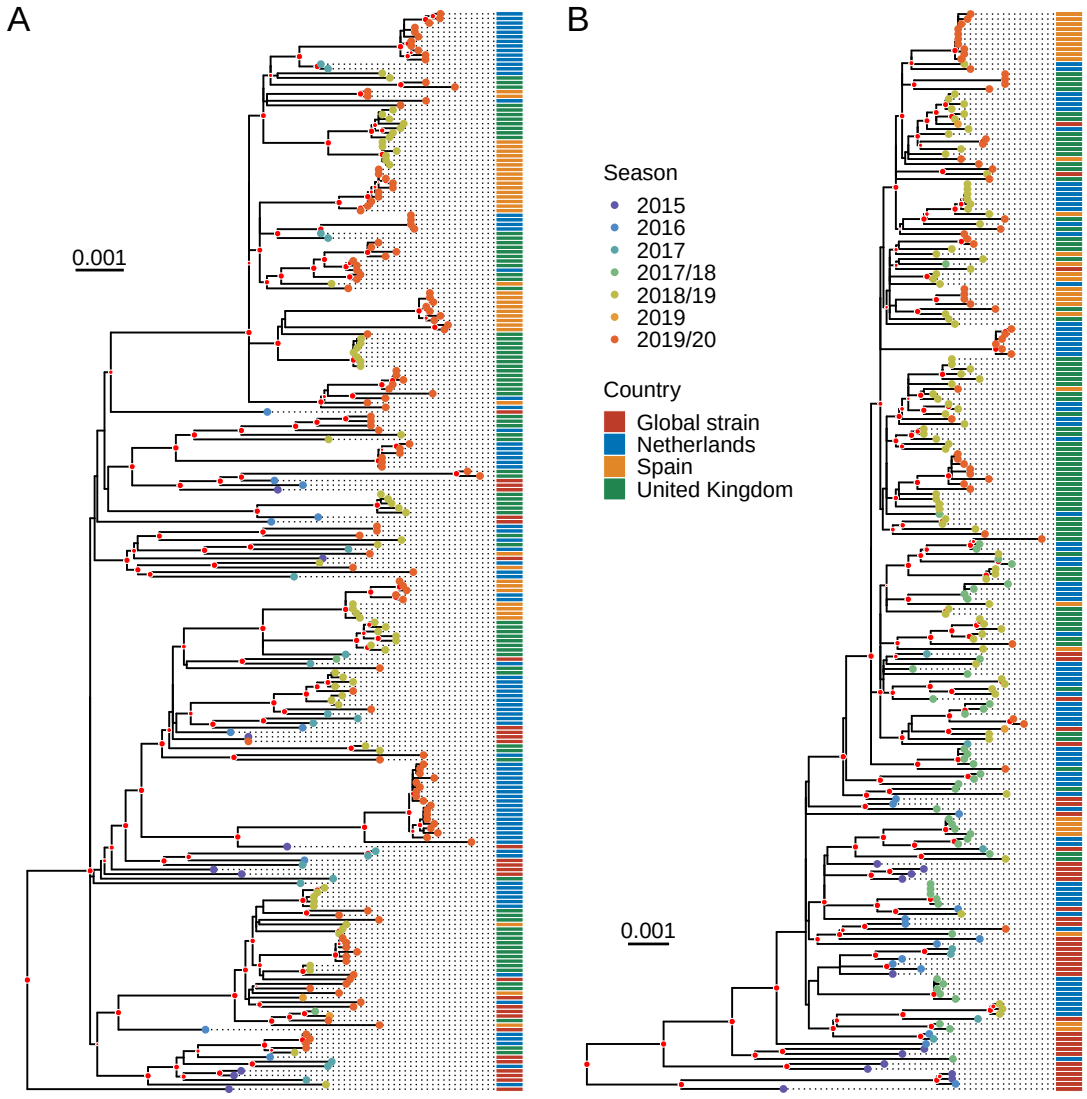


Figure 4.4: Maximum-likelihood phylogenies. **A.** RSV-A phylogeny. All RSV-A strains are genotype ON1, except for the one at the very bottom, which is genotype GA2 downloaded from GenBank. **B.** RSV-B phylogeny. All RSV-B strains are genotype BA. The phylogenies were inferred using RAxML with the general time reversible nucleotide substitution model and gamma-distributed rate heterogeneity among sites. The sampling country and season of strains are illustrated by bar and tip colours. Strains labelled as Netherlands, Spain, and United Kingdom were generated from this study, and the remainder (global strains) were downloaded from GenBank. No RSV whole-genome sequence collected in 2020 was available in GenBank as of 7th November 2020, when the data were downloaded. Red dots show well-supported nodes with a bootstrap value over 70% and are sized in proportion to bootstrap values. The trees were rooted to the oldest strains sampled during 1962–1977, KU316149 and KU316166 for RSV-A, and KU316116 and MG813995 for RSV-B. The oldest strains are not shown here. The scale bars represent the number of nucleotide substitutions per site.

4.3.4 Respiratory virome composition

The *Castanet* enrichment panel included a wide range of human bacterial and viral pathogens in addition to RSV. At least one other virus was recovered in 24% (106/433) of the RSV-infected infants, with ten of these 106 infants having two other viruses identified. HRV was the most frequently co-identified virus, found in 15% (63/433) of the infants, followed by human coronaviruses (4%), human adenoviruses (2%), and HHV-6 (2%) (Table 4.6). No severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was found in any of the samples collected in early 2020. VP1 genotyping of HRV showed that HRV-A, HRV-B, and HRV-C accounted for 41% (26/63), 29% (18/63), and 30% (19/63), respectively. Phylogenetic analysis showed that two infants had HHV-6A and five had HHV-6B (Supplementary Fig. B.5). Phylogenetic segregation of HHV-6A was strongly suggestive of chromosomally integrated HHV-6A in the samples[277].

Table 4.6: The presence and absence of other viruses in the RSV-infected infants (N = 433).[a]

	Presence	**Absence**	**Equivocal**
Any virus	106 (24)	299 (69)	28 (6)
Enterovirus	69 (16)	355 (82)	9 (2)
Rhinovirus	63 (15)	362 (84)	8 (2)
Enterovirus D	3 (1)	429 (99)	1 (<1)
Enterovirus A	2 (<1)	431 (100)	0
Enterovirus B	1 (<1)	431 (100)	1 (<1)
Human coronavirus	16 (4)	415 (96)	2 (<1)
OC43	7 (2)	425 (98)	1 (<1)
HKU1	6 (1)	427 (99)	0
NL63	2 (<1)	430 (99)	1 (<1)
229E	1 (<1)	432 (100)	0
Human adenovirus	10 (2)	419 (97)	4 (1)
Human herpesvirus 6	7 (2)	414 (96)	12 (2)
Human bocavirus	4 (1)	428 (99)	1 (<1)
Human cytomegalovirus	3 (1)	426 (98)	4 (1)
Human parechovirus	3 (1)	425 (98)	5 (1)
Human parainfluenza virus	3 (1)	429 (99)	1 (<1)
Influenza C virus	1 (<1)	432 (100)	0

[a] Data are shown as number (percentage). Percentages may not total 100 due to rounding.

RSV-infected infants with any other virus found in the nasopharynx tended to be older than those without any other virus found [median (IQR), 4.3 (2.0–8.0) months vs. 3.7 (1.7–7.0) months; P = 0.048; Q = 0.175; Table 4.7]. They were also more likely to require intensive care than those without co-detection of any other virus after adjusting for age, sex, comorbidity, and RSV subgroup (odds ratio [OR], 1.95; 95% CI, 1.03 to 3.68; P = 0.039; Q = 0.175; Table 4.7). Analysing each co-detected virus separately showed that only infants with HHV-6 co-detection were more likely to require intensive care compared with those without HHV-6 co-detection after adjusting for the above-mentioned variables (OR, 7.60; 95% CI, 1.38 to 62.92; P = 0.030; Q = 0.234). RSV-infected hospitalised infants had a higher total and mean (± SD) viral burden of the other virus(es) than RSV-infected outpatient infants after adjusting for age (total viral burden in a $\log_{10}$ scale: 3.1 ± 1.1 vs. 2.6 ± 0.8, P = 0.01). Other clinical features, RSV subgroup, and viral burden of RSV did not correlate with viral burden of co-detected viruses.

4.3.5 Respiratory bacterial microbiota

The mean (SD) number of bacterial species identified in each RSV-infected infant was 3.2 ± 2.1, with *Moraxella* (67% of the RSV-infected infants), *Streptococcus* (55%), and *Haemophilus* (42%) being the most frequently found genera (Supplementary Fig. B.6). Among each genus, *M. catarrhalis*, *S. pneumoniae*, and *H. influenzae* accounted for more than 50% respectively. Supplementary Table A.6 lists the proportion of each bacterial genus found in the infants. Figure 4.5 shows the composition of all bacteria identified in the samples.

Older infants had more bacterial species identified in the nasopharynx than younger infants (3.7 ± 1.8 vs. 3.0 ± 2.2 for infants over and under 6 months of age respectively; Mann–Whitney U test, P < 0.001). Similarly, infants with any of the three most frequently identified bacterial genera were older than those without [median (IQR), 4.6 (1.9–7.6) months vs. 2.0 (1.1–3.4) months; Mann–Whitney U test, P = 0.002]. Clinical features were compared between infants with and without these commonly identified bacterial genera after adjusting for age, sex,

Table 4.7: Demographic and clinical features of the RSV-infected infants with and without co-detection of any other virus.[a]

	Presence of any other virus	**Absence of any other virus**	**P value**	**Q value**[b]
Age—total no.	105	298		
Median (IQR)—mo	4.3 (2.0–8.0)	3.7 (1.7–7.0)	0.048	0.175
Female sex[c]	42/105 (40)	131/298 (44)	0.555	0.718
Comorbidity[c]	16/105 (15)	21/299 (7)	0.021	0.175
RSV-A[c]	55/104 (53)	145/291 (50)	0.674	0.741
ReSVinet score				
Mean ± SD	7.8 ± 5.0	7.7 ± 4.5	0.771[d]	0.771
Distribution[c]			0.353[d]	0.661
0–7	54/103 (52)	164/289 (57)		
8–13	33/103 (32)	85/289 (29)		
14–20	16/103 (16)	40/289 (14)		
Fever[c]	34/104 (33)	97/288 (34)	0.361[d]	0.661
Hospitalisation[c]	52/96 (54)	158/284 (56)	0.587[d]	0.718
PICU admission[c]	25/96 (26)	51/284 (18)	0.039[d]	0.175
Any respiratory support[c]	46/95 (48)	126/281 (45)	0.539[d]	0.718
Mechanical ventilation[c]	21/95 (22)	44/281 (16)	0.093[d]	0.256

[a] Twenty eight patients with equivocal presence of any other virus were removed from the comparisons. IQR denotes interquartile range; SD, standard deviation; PICU, paediatric intensive care unit.

[b] The Benjamini–Hochberg procedure was used to adjust for multiple comparisons and to determine false discovery rate–corrected Q values.

[c] Data are shown as no./total no. (%). Percentages may not total 100 due to rounding.

[d] Multiple linear regression, ordered logistic regression, and logistic regression were used to adjust for age, sex, comorbidity, and RSV subgroup.

comorbidity, and RSV subgroup. Infants with *Haemophilus* spp. had a more serious RSV infection, demonstrated by multiple outcome measures, than those without (Table 4.8). Infants with *S. pneumoniae* were also more likely to have an RSV infection requiring hospitalisation (OR, 1.98; 95% CI, 1.20 to 3.29; P = 0.008; Q = 0.043) and any respiratory support (OR, 1.69; 95% CI, 1.02 to 2.83; P = 0.042; Q = 0.155) than those without. In contrast, infants with *Moraxella* spp. were less likely to have an RSV infection requiring intensive care (OR, 0.42; 95% CI, 0.23 to 0.79; P = 0.007; Q = 0.038) and mechanical ventilation (OR, 0.503; 95% CI, 0.255 to 0.995; P = 0.047; Q = 0.103) than those without.

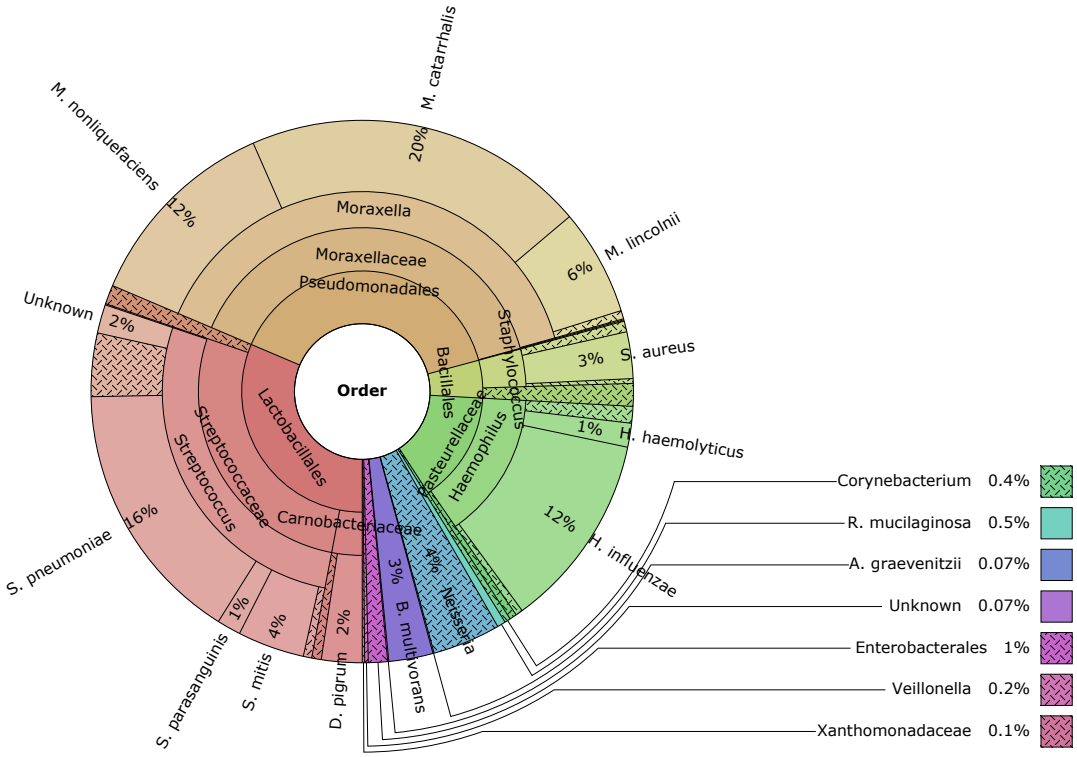


Figure 4.5: Composition of bacteria identified in the nasopharynx of the infants. The percentage of a bacterium was calculated as the number of infants having this bacterium divided by the sum of the number of bacterial species found in each infant. Circles from inside to outside represent the order, family, genus, and species of bacteria with some levels collapsed for better visualisation.

Considering the potential interactions between bacteria (Supplementary Fig. B.6), all the three most frequently identified genera were included in multivariable regression models (or *S. pneumoniae* instead of *Streptococcus* spp. for hospitalisation modelling). Similar significant findings were found except for the associations between fever and *Haemophilus* spp., between hospitalisation and *S. pneumoniae*, and between any respiratory support and *S. pneumoniae*.

There was information available on the use of antibiotics in 357 RSV-infected infants, of which 26% received antibiotics during their RSV episode. A higher proportion of infants with *Haemophilus* spp. received antibiotics than those without (31% vs. 20%; P = 0.024). However, a lower proportion of infants with *S. pneumoniae* received antibiotics than those without (19% vs 31%; P = 0.011). Similarly, a lower proportion of infants with *Moraxella* spp. received antibiotics than those without (20% vs. 43%; P < 0.001).

Table 4.8: Demographic and clinical features of the RSV-infected infants with and without co-detection of any *Haemophilus* sp.[a]

	Presence of any *Haemophilus* sp.	**Absence of any *Haemophilus* sp.**	**Odds ratio (95% CI)**	**P value**	**Q value[b]**
Age—total no.	181	222			
Median (IQR)—mo	5.3 (2.0–8.7)	3.6 (1.7–5.9)		1.2×10^{-4}	3.2×10^{-4}
Female sex[c]	78/181 (43)	100/222 (45)	1.08 (0.71, 1.64)	0.771	0.771
Comorbidity[c]	18/181 (10)	16/223 (7)	1.43 (0.66, 3.09)	0.414	0.455
RSV-A[c]	100/179 (56)	101/216 (47)	0.69 (0.46, 1.05)	0.089	0.109
ReSVinet score					
Mean ± SD	8.4 ± 4.5	6.7 ± 4.5		2.7×10^{-8}[d]	3.0×10^{-7}
Distribution[c]				5.3×10^{-6}[d]	1.9×10^{-5}
0–7	89/178 (50)	139/212 (66)			
8–13	61/178 (34)	48/212 (23)			
14–20	28/178 (16)	25/212 (12)			
Fever[c]	78/178 (44)	53/213 (25)	1.79 (1.13, 2.85)	0.013[d]	0.018
Hospitalisation[c]	104/167 (62)	98/211 (46)	3.71 (2.18, 6.49)	2.4×10^{-6}[d]	1.3×10^{-5}
PICU admission[c]	38/167 (23)	33/211 (16)	2.76 (1.51, 5.14)	0.001[d]	0.002
Any respiratory support[c]	81/164 (49)	82/209 (39)	2.77 (1.64, 4.77)	1.8×10^{-4}[d]	3.3×10^{-4}
Mechanical ventilation[c]	35/164 (21)	25/209 (12)	3.76 (1.92, 7.61)	1.5×10^{-4}[d]	3.3×10^{-4}

[a] Twenty-eight patients with equivocal presence of *Haemophilus* sp. were removed from the comparisons. *Haemophilus* spp. included *H. influenzae* (80%), *H. haemolyticus* (10%), *H. parainfluenzae* (7%), *H. parahaemolyticus* (1%), and unclassified species (1%). CI denotes confidence interval; IQR, interquartile range; SD, standard deviation; and PICU, paediatric intensive care unit.

[b] The Benjamini–Hochberg procedure was used to adjust for multiple comparisons and to determine false discovery rate–corrected Q values.

[c] Data are shown in no./total no. (%). Percentages may not total 100 due to rounding.

[d] Multiple linear regression, ordered logistic regression, and multivariable logistic regression were used to adjust for age, sex, comorbidity, and RSV subgroup.

4.4 Discussion

In this study, I found that chronological age and colonisation with *Haemophilus* spp. were significantly associated with severe RSV disease. Older infants were more likely to have febrile RSV infection than younger ones. While both RSV subgroups had similar viral dynamics and caused similar disease severity, high viral burden and delayed viral clearance were associated with fever and respiratory failure, respectively. Multiple viruses were detected in 24% of the infants, and co-detection of RSV and HHV-6 was associated with severe disease requiring intensive care, although with a small number of HHV-6 positives this study was not powered to make a definitive link.

I demonstrated an inverse relationship between the probability of fever and infant age. A recent study has shown diminished nasopharyngeal type I IFN and IFN-γ levels and enhanced expression of *MUC5AC* and *IL17A* in severe RSV disease[105]. IFN has been known to cause fever[278], and *MUC5AC* and *IL17A*, regulating the production of mucin-5AC and interleukin (IL)-17A respectively, have been related to mucus plugging of the airway, leading to respiratory distress and failure[279]. Extrapolating the local findings to systematic responses of the IFN levels may explain why younger infants have more severe disease but less fever than older ones.

In this dataset, viral burden positively correlated with the probability of fever, but not with other clinical outcome measures. I also found delayed RSV clearance in infants requiring mechanical ventilation. Other studies on viral dynamics similarly showed slow viral clearance and high viral load at later time points during acute RSV infection in infants requiring intensive care and/or mechanical ventilation[102,106]. These findings all suggest that prolonged viral replication is associated with the most severe form of RSV disease.

It has been shown that the presence of RSV reduces the chance of concurrent HRV infection[280,281]. In this dataset, although HRV was the most commonly found virus, the rate of HRV detection (15%) was in accordance with that in previous studies demonstrating an interference between RSV and HRV[280]. Other co-detected viruses were found in <10% of the infants. These viruses could represent either a

non-pathogenic 'bystander,' leftover nucleic acid from a previous infection, or a true pathogen contributing to the disease symptoms. It is difficult to differentiate between these possibilities, but viral burden (i.e., read count) may provide a clue. Co-detected viruses with low viral burden may come from prolonged shedding from a previous infection, whereas those with high viral burden are more likely to cause symptoms.

I found that infants having both RSV and any other virus(es), particularly HHV-6, had an increased risk of requiring intensive care in this dataset. Given that only a small number of infants had HHV-6, I was not able to infer whether HHV-6A, HHV-6B, or both led to the increased probability of requiring intensive care. A previous study found that infants infected with RSV only and those co-infected with RSV and other viruses (not including HHV-6) had similar peripheral blood mononuclear cell IFN-γ responses[268]. This is consistent with my finding that viral co-detection (excluding HHV-6) did not significantly affect disease severity.

I found that presence of *Haemophilus* spp. was significantly associated with severe RSV disease, and colonisation with *S. pneumoniae* was associated with hospitalisation and the need for respiratory support during an RSV infection. These results are consistent with previous studies[273,282]. *Haemophilus* spp. may have caused bacterial secondary infection in the infants since a greater proportion had fever and received antibiotics than those without *Haemophilus* spp. In contrast, the evidence for clinically-suspected bacterial secondary infection with *S. pneumoniae* or *Moraxella* spp. was insufficient. Overrepresentation of *H. influenzae* and *Streptococcus* spp. in RSV infection has been associated with enhanced Toll-like receptor (TLR) signalling[282], mucosal chemokine (C-X-C motif) ligand 8 (CXCL-8, also known as IL-8)[283], and IL-17A[284]. IL-8 and IL-17A signalling contributes to macrophage and neutrophil activation and recruitment[282], inducing bronchoalveolar neutrophil infiltration[285].

Despite being the most frequently found bacteria, *Moraxella* spp. were associated with less severe RSV infection that did not require intensive care or mechanical ventilation in this study, suggesting that these organisms might be components of healthy respiratory microflora. Previous studies also showed that *Moraxella* spp.

were more frequently found in outpatients than hospitalised children[282], but were significantly associated with fever in infants with RSV lower respiratory infections[273]. Further studies are required to clarify the association between *Moraxella* spp. and RSV disease severity.

It is worth noting that studies on the association between RSV disease severity and viral load[102–106], viral co-infection[268,269,281,286], or the respiratory microbiota[282,287] have sometimes yielded conflicting results. This may indicate the variability in viral and bacterial determinants for RSV infection in different host populations and clinical settings. The inconsistency is likely due to differences in study populations, definitions of disease severity, and/or genotypes of the infecting viruses. This study was robust insofar as it included infants with a wide range of disease severity, from self-limited upper respiratory infection to severe lower respiratory tract infection with respiratory failure. Both infants who were previously healthy and those who had comorbidities were also included. I used several clinical outcome measures to account for the variability in disease presentation. Furthermore, the RSV isolates found in this study represented the currently dominant strains across the globe—genotypes ON1 and BA[288,289].

This study represents one of the most comprehensive datasets evaluating the associations between RSV disease severity, host and viral factors, and the nasopharyngeal microbiome. The findings deepen our understanding of the risk factors for severe RSV disease and identify the target populations for therapeutic and preventative measures, particularly antivirals, vaccines, and monoclonal antibodies in late-stage clinical trials. Using antivirals to speed up viral clearance may reduce the probability of developing respiratory failure. Decreasing the colonisation with *Haemophilus* spp. and *S. pneumoniae*, potentially through vaccination, may prevent severe RSV disease. This study also sheds light on the complexity and dynamics of the microbiota during RSV infection. Further studies focusing on the interplay between viruses and bacteria and the underlying transcriptomic and immunological profiles can deepen our understanding of the pathophysiology of RSV disease.

5

Within-host diversity of respiratory syncytial virus

Contents

5.1 Introduction 87
5.2 Methods 88
5.2.1 Sample source and sequencing 88
5.2.2 Genome reconstruction 88
5.2.3 Within-host virus diversity analysis 89
5.2.4 Phylogeny reconstruction 90
5.2.5 Statistical analysis 90
5.3 Results 91
5.3.1 Sample population 91
5.3.2 Minor allele frequencies and minor variants 92
5.3.3 Potential antigenic variants 95
5.3.4 Pairwise nucleotide diversity 98
5.3.5 Genetic distance 100
5.3.6 Temporal change of intrahost virus population 101
5.4 Discussion 102

This chapter is based on a paper submitted to Nature Communications: Lin GL, et al. Distinct patterns of within-host virus populations between two subgroups of human respiratory syncytial virus. This paper has been accepted for publication and is currently under typesetting and production by the journal. The work represents my own contribution to the publication and not that of co-authors.

5.1 Introduction

Two surface glycoproteins of RSV, F and G, are the main targets for antibody-mediated neutralisation. The crystal structure of the F protein has been largely determined, and the F protein has been the focus of the development of vaccines and monoclonal antibodies (mAbs)[290]. Antibodies exclusively targeting prefusion-specific antigenic sites (e.g., sites ∅ and V) are more potent than those targeting sites that can be found in both conformations (e.g., sites I, II, IV)[28]. Nirsevimab (MEDI8897), a recombinant human mAb currently in phase 3 clinical trials, exclusively targets antigenic site ∅[291], and suptavumab (REGN2222), another prefusion-specific monoclonal antibody, binds antigenic site V[205]. Palivizumab and its affinity-enhanced variant, motavizumab[203], target antigenic site II, and antibody 101F binds antigenic site IV[292]. Mutations in the antigenic sites that confer resistance to mAbs have been identified. For example, mutants with N262S/Y, N268I, K272E/N/M/T/Q, or S275F/L in the F protein are less susceptible to palivizumab[258,293,294], and nirsevimab has reduced neutralising activity against mutants with N67I/N208Y, N208S/D, K68N/N201S, or K68N/N208S in the F protein[291].

In contrast, the crystal structure of the G protein is less well defined. The conserved central domain of the G protein, containing antigenic sites $\gamma 1$ and $\gamma 2$, has been shown to be a target for neutralising antibodies[21] and is another focus of vaccine development[24,295]. Multiple antigenic sites have also been found in the mucin-like regions flanking the conserved central domain, although they have yet to be precisely located[26]. The mucin-like region II (i.e., 2nd hypervariable region) has been shown to have hypermutation at the population level, and has thus been used widely in phylogenetic analyses[60].

Molecular epidemiology and evolutionary dynamics of RSV have been extensively studied at the consensus level; however, little is known about virus populations in each infected individual (i.e., within-host or intrahost virus diversity). Using high-throughput whole-genome sequencing, it is now possible to sequence viruses in sufficient depth to obtain a complete picture of within-host populations. A previous study showed that within-host RSV diversity increased in an immunocompromised

infant with persistent RSV infection following a haematopoietic stem cell transplant, and palivizumab escape mutants emerged after multiple administrations of this drug[296]. Another study demonstrated that RSV-A exhibited greater within-host virus diversity in experimentally infected adults than in naturally infected infants[297]. However, these results were limited to RSV-A infection and did not look at natural infections in adult populations. Analysing within-host virus genetic diversity in infections that represent general seasonal epidemics can aid understanding of the patterns of virus evolution and its driving forces, informing the development of preventative and treatment measures.

In this chapter, I intend to characterise within-host virus populations for the two subgroups of RSV using deep sequencing of samples collected from participants in the RESCEU studies. I explore within-host diversity, potential antigenic variants, and temporal changes of intrahost virus populations.

5.2 Methods

5.2.1 Sample source and sequencing

All nasopharyngeal swabs collected from infants and older adults enrolled in the RESCEU studies in the United Kingdom, Spain, and the Netherlands during 2017–2020 were included in this study. The clinical studies and sample collection have been described in detail in section 2.1. Nucleic acid isolation, library construction, and sequencing were performed as described in sections 2.2 and 2.3.

5.2.2 Genome reconstruction

The RSV genome was reconstructed using *shiver* as described in section 2.4. A higher minimum base quality of 35 and mapping quality of 30 were required for a base or an alignment to be counted as mapped in this study, as variant calling is extremely sensitive to systematic errors[298]. Without deduplication, per-position mapped read counts, generated by *shiver*, were used for downstream within-host virus diversity analysis.

5.2.3 Within-host virus diversity analysis

Only samples generating more than 10,000 unique (i.e., deduplicated) RSV reads and containing a single subgroup of RSV were included in within-host virus genetic diversity analysis. Allele frequencies were calculated at each genomic position, excluding those supported by fewer than 200 reads. The choice of this cut-off was based on a predefined criterion that 90% of the included samples had at least 80% of the genome fulfilling this cut-off (Supplementary Fig. B.7). Cumulative minor allele frequency (MAF) was defined as 1 minus major allele frequency, and polymorphic sites were those with a cumulative MAF of ≥3%. Mean cumulative MAF per sample was calculated as the sum of cumulative MAF at each genomic position divided by the total number of positions. Minor variants, or intrahost single nucleotide variants, were defined as variants with an allele frequency of ≥3% and <50%.

Intrahost virus diversity was estimated as pairwise nucleotide diversity (π)[299]. The proportion of pairwise nucleotide differences (D) at each genomic position was calculated as

$$D_i = \frac{A_i \times C_i + A_i \times G_i + A_i \times T_i + C_i \times G_i + C_i \times T_i + G_i \times T_i}{(N_i^2 - N_i)/2}$$

where A_i, C_i, G_i, and T_i represent the copy number of allele A, C, G, and T, respectively, and N_i is the total count of the four alleles (i.e., depth of coverage) at a given locus i, so $N_i = A_i + C_i + G_i + T_i$. Loci with a total count of less than 200 were excluded. Pairwise nucleotide diversity across a genome (π) was then calculated as

$$\pi = \sum_{i=1}^{L} \frac{D_i}{L}$$

where L is the number of genomic positions with a read depth of at least 200×.

Manhattan (L1-norm) distance was used to compare within-host diversity levels between a sample pair, calculated as

$$d_i(p, q) = \sum_{k=1}^{4} |p_k - q_k|$$

$$M = \sum_{i=1}^{N} d_i \times \frac{S}{N}$$

where d_i is the distance between two samples at a given locus i with vectors p and q containing relative frequencies of four possible alleles (i.e., A, C, G, and T), M is the Manhattan distance between coding sequences of two samples, N is the number of coding sequence positions where both samples have the same consensus base and a read depth of at least 200×, and S is the total length of the coding sequence. To remove potential background noise in Manhattan distance calculations, allele frequencies of <3% were changed to 0, and those of >97% were changed to 100%.

Nucleotide positions were numbered from the first base of the coding sequence of each gene according to the NCBI reference sequences with the accession numbers of NC_038235 and NC_001781 for RSV-A and RSV-B, respectively. Amino acid positions were numbered from the first methionine of each protein according to the same NCBI reference sequences.

5.2.4 Phylogeny reconstruction

Maximum likelihood phylogenies of consensus coding sequences, supported by at least two unique (deduplicated) RSV reads, were estimated using RAxML (v8.2.12)[242] with the general time-reversible nucleotide substitution model and gamma-distributed rate heterogeneity. Bootstrapping with 1,000 replicates was used to assess the robustness of tree topologies. Pairwise patristic distances were calculated from the maximum-likelihood trees using the `cophenetic` function of the R package ape (v5.4-1)[246]. Phylogenetic trees were visualised using the R package ggtree (v2.2.4)[247].

5.2.5 Statistical analysis

Continuous variables were summarised using mean, median, maximum, and minimum. All comparisons of continuous variables between groups were conducted by two-tailed Mann–Whitney U tests (two groups) or Kruskal–Wallis tests (three groups). Post hoc application of the Benjamini–Hochberg procedure was used to control false discovery rates for multiple testing. Chi-square tests with Yates' continuity correction were used for contingency analysis; Fisher's exact tests were

performed when the expected value of a cell was less than 5. Logistic regression was employed to model a binary dependent variable while adjusting for a covariate. Two-tailed Pearson correlation analysis was used to evaluate the relationship between two variables. Temporal changes of a variable were determined by ordinary least-squares linear regression. Two approaches were applied to account for batch effects on the comparisons of diversity metrics: (i) including batch as a regression covariate (e.g., regression of pairwise nucleotide diversity on sampling country, sampling season, RSV subgroup, RSV read count, participant age group, disease severity, and 'batch' as in Table 5.3); and (ii) standardising the values within each batch to z-scores, that is, to a mean of zero and a standard deviation of 1 (e.g., Mann–Whitney U test on z-score standardised pairwise nucleotide diversity as in Fig. 5.3, panels A and B). Missing data were imputed using the `aregImpute` function, implemented in the R package Hmisc (v4.5-0)[300]. All statistical analyses were performed using R (v4.0.2)[245]. P values or adjusted P values of less than 0.05 were considered to indicate statistical significance.

5.3 Results

5.3.1 Sample population

A total of 861 nasopharyngeal swabs collected from 462 RSV-infected patients in the United Kingdom, Spain, and the Netherlands during 2017–2020 were sequenced. Of these, 327 samples had sufficient viral load to generate more than 10,000 unique (deduplicated) RSV reads. After removing five samples containing both RSV-A and RSV-B, 322 samples were included in the within-host virus diversity analysis. These samples were sequenced in four batches, with 11, 113, 41, and 157 of the included samples from each batch respectively (Supplementary Table A.7). The 322 samples were collected from 267 different participants, among which 34 participants had multiple samples (mean 2.6, range 2–5) collected on different days (ranging from 1 to 8 days apart).

5.3.2 Minor allele frequencies and minor variants

Genomic positions with a read depth of less than 200 were excluded from the analysis. Nearly 90% of the samples had >80% of the genome passing this threshold. Three samples had a significantly high mean cumulative MAF per sample: 0.52% (from an RSV-A-infected infant; batch 4), 0.19% (from an RSV-B-infected adult; batch 2), and 0.17% (from an RSV-B-infected infant; batch 4). These samples presumably represented a real or artefactual mixture of genetically distinct strains of the same RSV subgroup and were thus excluded from the following analysis. The sources and sequencing yields of the remaining 319 samples (collected from 264 participants) are shown in Table 5.1.

Table 5.1: Characteristics of RSV samples by subgroup.

	RSV-A (N = 175)	**RSV-B (N = 144)**	**P value**[a]
Host number			0.12[b]
Infants	141	115	
Older adults	1	7	
Host age, median (range)			
Infants (month)	4.5 (0.5–11.6)	4.3 (0.2–11.7)	0.72
Older adults (year)	69	75 (72–78)	0.19
Sample source			0.45
United Kingdom	74	64	
Netherlands	58	53	
Spain	43	27	
Sampling season			<0.001
2017–18	14	33	
2018–19	65	63	
2019–20	96	48	
Days between symptom onset and sampling, median (range)	4 (1–11)	4 (1–9)	0.11
Number of unique RSV read pairs ($\log_{10}$), median (range)	4.6 (4.0–5.8)	4.7 (4.0–5.9)	0.22
Batch 1	4.9 (4.1–5.5)	5.3 (4.4–5.6)	0.50
Batch 2	4.6 (4.0–5.6)	4.6 (4.0–5.9)	0.98
Batch 3	4.4 (4.0–4.8)	4.5 (4.0–5.5)	0.18
Batch 4	4.7 (4.0–5.8)	4.9 (4.0–5.6)	0.12
Minimum genome coverage	99.9%	100%	0.37
Average depth of coverage, median (range)	3,372 (696–7,897)	3,650 (525–7,930)	0.41
Batch 1	2,940 (696–6,823)	4,975 (1,295–7,601)	0.63
Batch 2	3,561 (1,092–7,452)	3,469 (1,091–7,930)	0.68
Batch 3	2,045 (803–3,224)	2,258 (525–7,157)	0.37
Batch 4	3,736 (847–7,897)	4,505 (719–7,798)	0.23

[a] Unless otherwise specified, chi-square tests with Yates' continuity correction or Fisher's exact tests were used for contingency analysis, and Mann–Whitney U tests were used to compare numeric variables between subgroups. One RSV-B-infected infant had missing information on age. Six RSV-A-infected infants and five RSV-B-infected infants had missing information on date of symptom onset.

[b] Logistic regression was used to adjust for sampling season. Samples were collected from older adults only in 2017–18 and 2018–19 RSV seasons, when RSV-B was the predominant circulating subgroup.

The median of the mean cumulative MAF per sample was 0.039% (range 0.025%–0.068%) for the 319 samples. The distributions of the mean cumulative MAF per sample were significantly different between samples from different sequencing batches (Fig. 5.1, panel A), likely due to the differences in the ratio of duplicate reads to total RSV reads (percent duplication rate) between batches (Supplementary Table A.7). After adjusting for the observed batch effects (e.g., Fig. 5.1, panel B), RSV-B samples had a higher mean cumulative MAF per sample than RSV-A samples (median of the original data: 0.042% vs. 0.037%; multiple linear regression with batch and the number of unique RSV reads as covariates, P = 0.016; Mann–Whitney U test on the standardised data, P = 0.016).

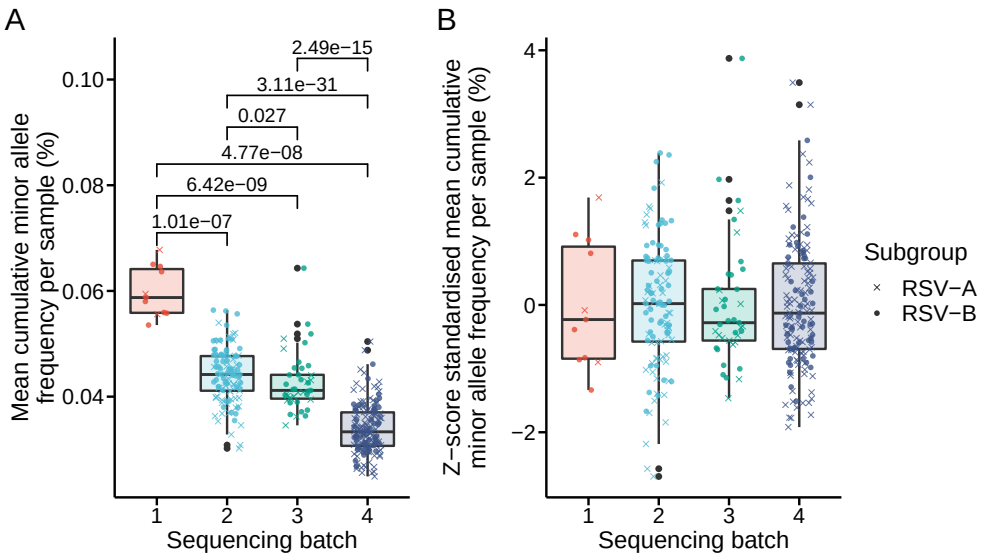


Figure 5.1: Adjusting for batch effects on cumulative minor allele frequency. **A.** Before adjustment. **B.** After z-score standardisation of the data. Each dot represents an individual sample. There were 11, 112, 41, and 155 samples from each batch respectively. Mann–Whitney U tests with the Benjamini–Hochberg procedure were used to evaluate the significance of the batch effects. Adjusted P values of less than 0.05 are shown above the box plots. The centre line of each box denotes the median; box limits, the first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers.

On average, each sample had 3.7 minor variants (range 0–30; defined as variants with a frequency of ≥3%). 18.8% of the samples (60/319) did not have any minor variants. An inverse correlation was noted between the number of unique RSV reads

and the number of minor variants ($r = -0.41$, $P = 4.2 \times 10^{-14}$; Supplementary Fig. B.8), consistent with a greater variance of MAF when the sampling fraction was small (i.e., few unique reads were sequenced)[301]. Variation rarely occurred at the same genomic position in different samples. Among all minor variants found in this study, only 5.9% (57/972) were shared by multiple samples (excluding 17 minor variants only shared by sequential samples from the same participants), usually no more than five samples. However, there was one minor variant shared by 59% (85/144) of the RSV-B samples, with a frequency between 3% and 11%. This minor variant had a G to A substitution at position 3403 of the L gene, causing an amino acid alteration from glutamic acid to lysine at position 1135 (E1135K) of the RNA-dependent RNA polymerase.

5.3.3 Potential antigenic variants

The sequences encoding the antigenic sites of the F protein were highly conserved at the consensus level in this study. However, two RSV-B isolates from two infant participants, both of whom had only one sample collected, had an A to T substitution at nucleotide position 204 of the F gene. This substitution results in an amino acid alteration from lysine to asparagine (K68N), which in a previous study was associated with a 4-fold reduction in susceptibility to nirsevimab neutralisation *in vitro*[291]. No minor variant was found at this position in these two samples.

The frequencies and distribution of all minor variants across the coding sequence of the F gene are shown in Fig. 5.2, panel A. There were one, eight, two, and three minor variants identified in the antigenic sites ∅, II, IV, and V of the F protein, respectively (Table 5.2). 0%, 6.0% (6/100), and 1.6% (2/124) of the participants had potential antigenic variants (i.e., minor variants encoding a nonsynonymous substitution in the antigenic sites) in the 2017–18, 2018–19, and 2019–20 RSV seasons, respectively. One of these minor variants had two co-existing nucleotide substitutions with a frequency of ≥3% in a single codon, encoding an amino acid substitution from isoleucine to threonine at position 261 of the F protein (I261T).

Other minor variants identified in the antigenic sites were from different samples. To date, none of these variants have been reported to confer resistance to mAbs.

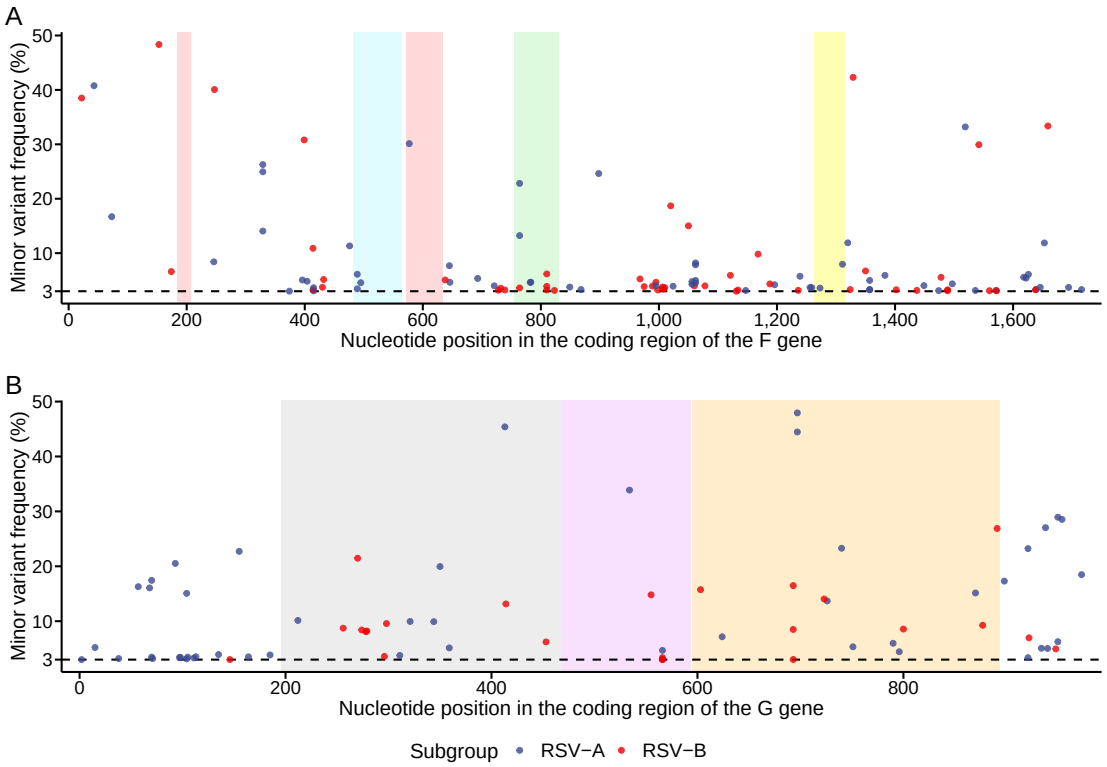


Figure 5.2: Minor variants in the coding regions of the F and G genes among 175 RSV-A and 144 RSV-B samples. **A.** F gene. Shaded regions represent known antigenic sites (neutralising epitopes in particular): red, prefusion-specific antigenic site ∅ (target for nirsevimab); green, site II (target for palivizumab and motavizumab); yellow, site IV (target for 101F); and blue, prefusion-specific site V (target for suptavumab). **B.** G gene. The purple region represents the conserved central domain (target for 3D3 and 2D10), flanked by highly variable mucin-like regions I (grey) and II (orange). Nirsevimab, palivizumab, motavizumab, 101F, suptavumab, 3D3, and 2D10 are RSV-specific mAbs. Each dot denotes a minor variant, coloured by subgroup. Black dashed lines represent minor allele frequency of 3%, used to define a minor variant. Positions are numbered from the first base of the coding sequence of each gene according to the NCBI reference sequence (accession number NC_038235).

Table 5.2: Characteristics of minor variants within the antigenic sites of the fusion protein.

Nucleotide position[a]	Codon change	Amino acid change[b]	Antigenic site	Subgroup/Country/Season/ Minor allele frequency (%)[c]
489	GAA:GAt	E163D	V	A/GB/2018–19/3.4 A/GB/2018–19/6.1
495	AAC:AAt	N165	V	A/GB/2018–19/4.6
577	CCA:tCA[d]	P193S	∅	A/GB/2018–19/30.1
764	AGT:AaT	S255N	II	A/ES/2019–20/13.2[e] A/ES/2019–20/22.8[e] B/GB/2018–19/3.6
782	ATC:Act	I261T	II	A/ES/2018–19/4.6[f]
783	ATC:Act	I261T	II	A/ES/2018–19/4.7[f]
810	CAG:CAa	Q270	II	B/NL/2017–18/3.2 B/NL/2017–18/6.2[g] B/NL/2018–19/3.9
823	TCA:cCA	S275P	II	B/GB/2018–19/3.1
1273	TCA:cCA	S425P	IV	A/GB/2019–20/3.6
1311	AAC:AAt	N437	IV	A/NL/2019–20/8.0

[a] Positions are numbered from the first base of the coding sequence of the F gene according to the NCBI reference sequence (accession number NC_038235).

[b] Positions are numbered from the first methionine of the fusion protein according to the NCBI reference sequence (accession number NC_038235).

[c] GB denotes the United Kingdom; ES, Spain; and NL, the Netherlands.

[d] 55.7% (98/176) of the RSV-A samples and all RSV-B samples had a consensus base of T at this position.

[e] These two variants were found in samples collected from the same participant on day 2 (13.2%) and day 5 (22.8%) of hospitalisation, respectively. Samples collected from this participant on other days (days 1, 3, and 4) did not have variants with a frequency of ≥3% at this position.

[f] These two were co-occurring mutations, identified in the same minor variant.

[g] Except for this variant, found in an adult sample, other minor variants were identified in infant samples.

I also looked at the frequencies and distribution of minor variants in the coding region of the G gene (Fig. 5.2, panel B). The median frequency of minor variants was significantly higher in the G gene than in the F gene, either at potential antigenic sites (median: 9.3% vs. 4.6%; Mann–Whitney U test, P = 0.022) or across the whole coding sequences (median: 8.3% vs. 4.4%; P = 0.004), consistent with previous studies identifying the G gene as the most variable gene in the virus genome[9]. The median minor variant frequency in the mucin-like region II of the G gene (13.7%) was greater than that in the mucin-like region I (9.2%), which was greater than that in the conserved central domain (4.0%). However, these differences were not statistically significant (Kruskal–Wallis test, P = 0.20).

5.3.4 Pairwise nucleotide diversity

Within-host virus genetic diversity was estimated as pairwise nucleotide diversity (see section 5.2.3 Within-host virus diversity analysis). Pairwise nucleotide diversity did not correlate with the number of unique RSV reads after adjusting for the batch effects (Table 5.3 and Supplementary Fig. B.9, panel A), but was highly consistent with the mean cumulative MAF per sample (r = 0.997, $P < 2.2 \times 10^{-16}$; Supplementary Fig. B.9, panel B). The median pairwise nucleotide diversity of the whole dataset was 0.0007 (range 0.0005–0.0014). Gene-wise comparisons showed that the L gene had significantly higher pairwise nucleotide diversity than the NS2, P, SH, and G genes, but the other genes did not have a significant difference in pairwise nucleotide diversity between each other (Supplementary Fig. B.10). These differences were by definition due to the mean proportion of pairwise nucleotide differences at each genomic position within the L gene instead of the length of the L gene.

RSV-B had greater pairwise nucleotide diversity than RSV-A (multiple linear regression, P = 0.044; Fig. 5.3, panel A; Table 5.3), and older adults had a more diverse intrahost RSV-B population than infants (multiple linear regression, P = 0.0006; Fig. 5.3, panel B; Table 5.3). The subgroup difference was still significant if excluding adult samples (Mann–Whitney U tests on standardised data, P = 0.039).

The number of RSV reads and the duration of infection were similar between both RSV subgroups and between both age groups. Samples collected from different countries or seasons or patients with different severity of RSV infections did not have significant differences in pairwise nucleotide diversity (Table 5.3).

Table 5.3: A multiple linear regression model to evaluate the association between pairwise nucleotide diversity and other variables.[a]

Variable	Coefficient	Standard error	T value	P value
Intercept	1.2×10^{-3}	8.1×10^{-5}	15.20	$< 2 \times 10^{-16}$
Country				
Spain (reference)				
Netherlands	-1.3×10^{-5}	1.7×10^{-5}	−0.77	0.44
United Kingdom	8.1×10^{-6}	1.4×10^{-5}	0.57	0.57
Season				
2017–18 (reference)				
2018–19	1.2×10^{-5}	2.1×10^{-5}	0.60	0.55
2019–20	3.0×10^{-5}	3.6×10^{-5}	0.83	0.41
Subgroup				
RSV-A (reference)				
RSV-B	2.3×10^{-5}	1.1×10^{-5}	2.02	0.044
Batch				
First (reference)				
Second	-3.3×10^{-4}	3.6×10^{-5}	−9.11	$< 2 \times 10^{-16}$
Third	-3.3×10^{-4}	3.9×10^{-5}	−8.58	4.8×10^{-16}
Fourth	-5.3×10^{-4}	4.5×10^{-5}	−11.62	$< 2 \times 10^{-16}$
RSV read ($\log_{10}$)	1.1×10^{-5}	1.3×10^{-5}	0.83	0.41
Age group				
Older adult (reference)				
Infant	-1.2×10^{-4}	3.5×10^{-5}	−3.45	0.0006
Severity[b]				
Mild (reference)				
Moderate	1.1×10^{-5}	1.2×10^{-5}	0.94	0.35
Severe	1.7×10^{-5}	1.6×10^{-5}	1.03	0.30

[a] $F(12, 306) = 46.5$, $P < 2.2 \times 10^{-16}$, adjusted $R^2 = 0.63$.

[b] Thirteen patients had missing information on disease severity. These missing data were imputed using the `aregImpute` function, implemented in the R package Hmisc.

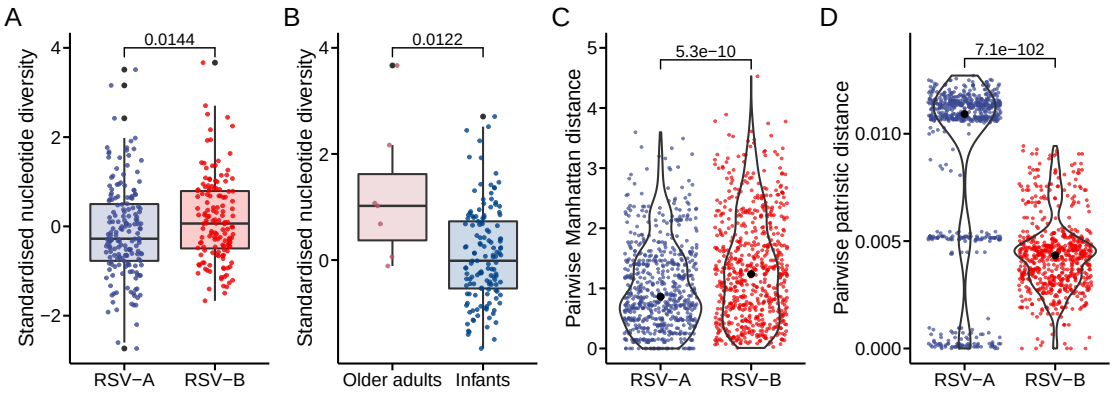


Figure 5.3: Standardised pairwise nucleotide diversity and pairwise genetic distances. **A.** Comparison of z-score standardised pairwise nucleotide diversity between 175 RSV-A and 144 RSV-B samples. **B.** Comparison of z-score standardised pairwise nucleotide diversity of RSV-B between 7 adult samples and 137 infant samples. RSV-A isolates were excluded from this comparison because only one adult had RSV-A infection. **C.** Comparison of pairwise Manhattan distances. **D.** Comparison of pairwise patristic distances. Only pairwise distances between samples from the second sequencing batch, the same country, the same season, and different participants were included in **C** and **D** (650 RSV-A pairs and 656 RSV-B pairs). Each dot represents an individual sample in **A** and **B**, and a sample pair in **C** and **D**. Mann–Whitney U tests were used to evaluate the significance of the differences. P values are shown above the plots. For **A** and **B**, the centre line of each box denotes the median; box limits, the first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers. For **C** and **D**, the violin plots summarise the distribution of the data, and the black dots denote the median value of each group.

5.3.5 Genetic distance

Within-host diversity levels between samples were compared using pairwise Manhattan distances[302] at consensus-identical positions, where allele frequencies below the 3% threshold were converted to 0. In contrast, consensus variations between samples were compared using pairwise patristic distances, which are phylogenetic distances on RSV phylogenies (Supplementary Fig. B.11). To eliminate the batch effects, I only included pairwise distances between samples in the second batch (n = 112; excluding one outlier). To reduce potential bias from geographical and temporal differences, only pairwise distances between samples from the same country and the same season were calculated.

Serial sample pairs had within-host diversity levels comparable to those of samples from different participants (range: 0–3.34 vs. 0–5.03), despite having identical or nearly identical consensus sequences, as indicated by their small patristic distances (range 2.0×10^{-6}–7.5×10^{-5}). Excluding the serial sample pairs, RSV-B sample pairs had significantly greater within-host diversity levels than RSV-A pairs (median: 1.24 vs. 0.86), whereas the comparison of consensus sequences showed the opposite effect (Fig. 5.3, panels C and D). Pairwise patristic distances between RSV-A samples formed three clusters, corresponding to the three main clades of the phylogenetic tree (Supplementary Fig. B.11, panel A). When using all allele frequencies, including those below 3% MAF, to calculate Manhattan distances, RSV-B samples still had significantly greater pairwise Manhattan distances than RSV-A samples (median: 20.46 vs. 18.19, Supplementary Fig. B.12).

5.3.6 Temporal change of intrahost virus population

Putting all samples together, standardised pairwise nucleotide diversity did not have a significant temporal change within 7 days of symptom onset ($R^2 = 0.008$; $P = 0.122$). For the 34 participants with multiple samples collected daily during hospitalisation, pairwise nucleotide diversity was also evaluated in each set of serially collected samples, excluding those sequenced in different batches (Fig. 5.4). No significant trend was noted either in each participant or when combining all samples and adjusting for the batch effects, except for the samples from GB-058, where pairwise nucleotide diversity increased by 0.000063 daily (95% confidence interval, 0.000046 to 0.000080; $P = 0.004$). This patient was a 19-day-old preterm neonate (gestational age of 33 weeks 6 days) with severe RSV infection requiring intensive care and mechanical ventilation.

The changes in minor variants and variant frequencies in the serial samples were also evaluated at polymorphic sites where minor alleles were identified at more than three time points (Fig. 5.5). 79% of these minor variants had a nonsynonymous substitution. Only one minor variant with a G to A substitution at position 3403 of the L gene from participant NL-091, which was shared by 71 participants

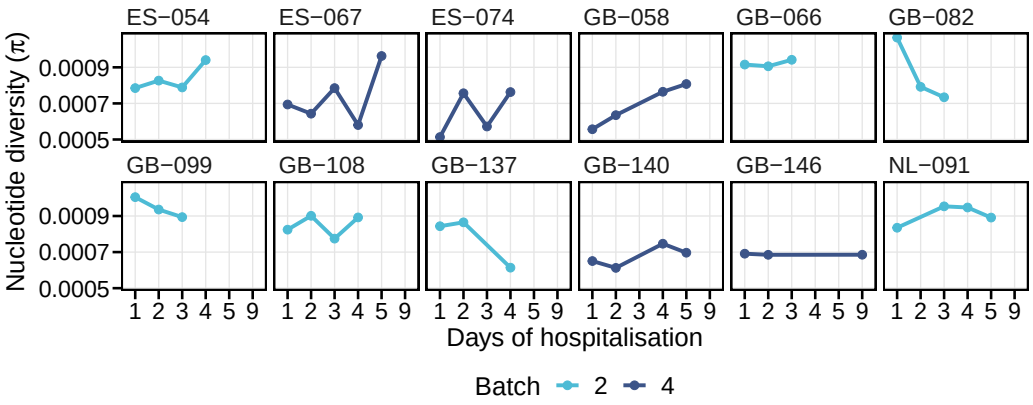


Figure 5.4: Temporal change of pairwise nucleotide diversity. Pairwise nucleotide diversity of serial samples collected at more than two time points and sequenced in the same batch are shown here. Three participants whose samples were sequenced in different batches and 19 participants who had only two samples collected are not shown. Each panel is labelled with the participant ID.

(85 samples), remained above the 3% threshold throughout the sampling period. This patient was a 42-day-old previously healthy infant with severe RSV infection requiring intensive care and mechanical ventilation. All other variants (including the aforementioned variant in other participants) were only detected either early, late, or intermittently, during the course of sample collection.

5.4 Discussion

In this study, nasopharyngeal samples collected in three clinical studies during 2017–2020 were sequenced to profile within-host RSV populations. I demonstrated that RSV-B had greater within-host diversity than RSV-A, whereas RSV-A had greater consensus diversity than RSV-B. Two RSV-B isolates' consensus sequences had a mutation in the F protein (K68N), previously associated with reduced susceptibility to nirsevimab neutralisation. Several other minor variants were also identified in the antigenic sites of the F protein. None of these variants have been reported before except for S255N[303], whose susceptibility to monoclonal antibodies has not been examined. Stochastic (random) patterns were found in the temporal changes of within-host virus diversity and minor variants.

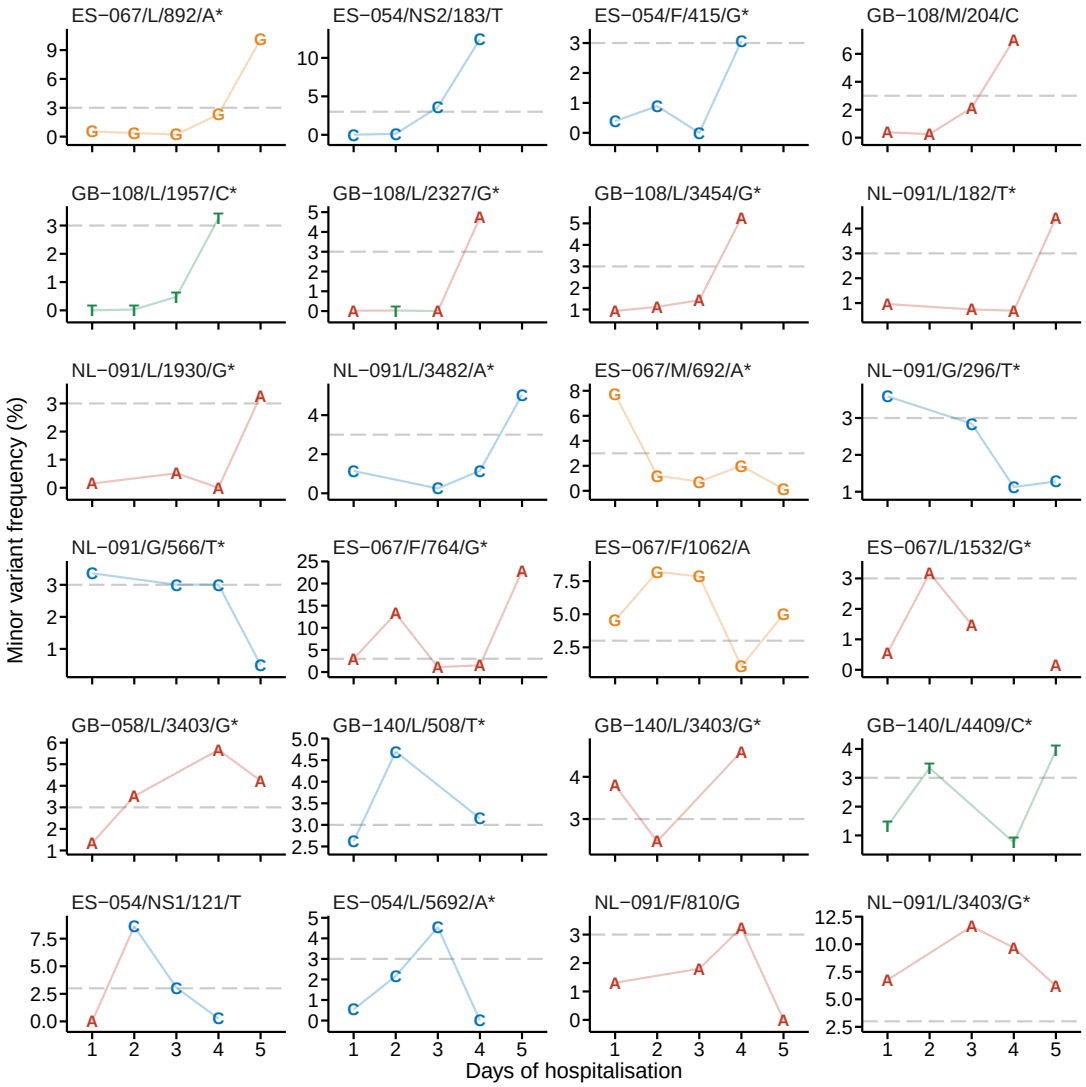


Figure 5.5: Temporal change in minor alleles. Minor alleles and allele frequencies are shown at polymorphic sites within the coding sequence of the serial samples, where minor alleles were detected at ≥3 time points. The grey dashed lines represent the 3% threshold, which defines a minor variant. Panels are labelled with the participant ID, followed by the gene name, the nucleotide position, and the consensus base. Asterisks denote nonsynonymous substitutions. Letters in the plots denote minor allele bases. Panels are ordered by the trend of the change: increased, decreased, and fluctuated. Positions are numbered from the first base of the coding sequence of each gene according to the NCBI reference sequences with the accession numbers of NC_038235 and NC_001781 for RSV-A and RSV-B, respectively.

Low input genetic material (i.e., viral load) has been shown to reduce the sensitivity and specificity of variant calling[298]. In this study, I applied the quantitative methodology of targeted metagenomics to library construction and used the number

of unique RSV reads as a proxy for viral load[1]. The inclusion criterion of more than 10,000 unique RSV reads corresponded with a viral load of approximately 2.4×10^6 copies/mL and above (based on the PCR assay performed by GlaxoSmithKline; Fig. 4.1, panel B), sufficient input levels for accurate minority variant calling[304]. Given the large number of samples in this study, batching was required for sequencing, resulting in variable percent duplication rates and hence some batch effects on diversity metrics. I adopted two approaches to account for the batch effects on the comparisons of mean cumulative MAF per sample and pairwise nucleotide diversity: (i) including batch as a regression covariate and (ii) standardising the values within each batch to z-scores (see section 5.2.5 for details). Both methods showed the same significant findings, making cross-batch comparisons robust. To avoid any residual bias, for pairwise comparisons of genetic distances I used only samples from the same batch (batch 2), which had very high percent duplication rates and similar read numbers for RSV-A and RSV-B (Supplementary Table A.7 and Table 5.1), consistent with capture saturation, and from which I could be confident of recovering the full range of intrahost diversity.

The extent of intrahost virus diversity depends not only on the rate of virus evolution (partly associated with the ability of proofreading for viral replication errors) but also on the duration of infection. RNA viruses generally have a higher mutation rate than DNA viruses[80], and are usually not able to correct the errors of viral replication, which DNA viruses can[305]. In this study, RSV had greater pairwise nucleotide diversity than has been reported for influenza virus, another RNA virus causing acute respiratory infection (range 0.0005–0.0014 vs. 0–0.0002[306]). RSV intrahost diversity appears to be comparable with, or slightly higher than, that of the DNA viruses in the family *Herpesviridae*, which cause chronic infections[307], but up to one to two orders of magnitude lower than that of persistent RNA viruses (e.g., hepatitis C virus, human immunodeficiency virus) and persistent DNA viruses (e.g., hepatitis B virus), which generally have pairwise nucleotide diversity above 0.005[307].

Neutralisation escape mutants have been isolated in 0.7% of immunoprophylaxis-naïve RSV-infected subjects[294], 5–9% of RSV-breakthrough patients receiving

palivizumab[258,308], and 8% of RSV-breakthrough cases receiving nirsevimab[33]. In this study, isolates collected from 0.8% (2/264) of the immunoprophylaxis-naïve participants were found to contain a nirsevimab resistance-associated substitution at the consensus level. I also identified an RSV-B minor variant with an amino acid change from serine to proline at position 275 (S275P) of the F protein. Other amino acid substitutions at this position have demonstrated resistance to palivizumab (S275F/L)[258]. Whether the mutation S275P also alters the neutralising activity of palivizumab requires further investigation; however, all three mutations at this position replaced a polar amino acid with a nonpolar one, which may result in significant conformational or functional changes. It is important to identify neutralisation escape mutants in immunoprophylaxis-naïve children in the era before RSV monoclonal antibodies become extensively used. It indicates the circulation of escape mutants in the community even though they generally have a selective disadvantage in the absence of mAbs[294].

The findings that RSV-B had greater pairwise nucleotide diversity and pairwise Manhattan distances than RSV-A both indicate that, at least in this dataset, RSV-B had a more diverse intrahost virus population than RSV-A. These results do not correlate with the duration between symptom onset and sample collection (Table 5.1), but are consistent with previous studies on global RSV strains, which found that RSV-B has a higher genome-wide evolutionary rate than RSV-A (7.47–7.76 $\times 10^{-4}$ substitutions/site/year vs. 5.68–6.47 $\times 10^{-4}$ substitutions/site/year)[82,86]. This difference extends below the 3% threshold for minority variant calling (Fig. B.12). On the basis of this finding, a plausible hypothesis is that RSV-B is subject to greater immune pressure (e.g., by innate immunity, neutralising antibodies, or T cell-mediated cytotoxicity) than RSV-A. This hypothesis is in line with previous studies showing that intrahost RSV diversity increased in response to an established immunity[296], and that RSV-B has more amino acid alterations[309], predicted O-glycosylation site changes[309], and indel mutations[86] in the G gene than RSV-A, suggesting a stronger selective pressure acting on RSV-B than RSV-A.

RSV-B exhibited higher within-host diversity in older adults than in infants in response to different immune pressures between the two age groups. Of note, this dataset included only eight adults, and this comparison was limited to 7 adults and 137 infants who had RSV-B infection. Further studies enrolling more adults would be of value to delineate the difference in within-host diversity between different age groups. Furthermore, the temporal changes of pairwise nucleotide diversity and minor variants were stochastic within each infected individual, suggesting the driving force of evolutionary dynamics in global RSV populations is more likely from the selective pressure imposed at the population level than within an individual host. Only samples that yielded sufficient RSV reads were included in this study, so these temporal trends were confined to samples collected within a short time frame (mostly 5 days). Nonetheless, a study on seasonal influenza virus also found limited evidence of positive selection at the within-host evolutionary scale[302].

The greater within-host viral diversity observed in RSV-B warrants separate testing and close monitoring for the anti-RSV-B efficacy of vaccines and mAbs that are being developed. This is because the development of several RSV vaccines in preclinical or clinical trials is based on the nucleotide sequences or structure of RSV-A strains[310–312]. Some studies have also shown that RSV-B had more fixed mutations in the antigenic sites of the F protein at the consensus level[313], resulting in more variable *in vitro* and clinical susceptibility to mAbs than RSV-A. For example, in a phase 2b trial of nirsevimab, the drug had reduced neutralising activity against two RSV-B isolates collected from its recipients; one had a mutation of N208S and the other had multiple mutations of I64T, K68E, I206M, and Q209R in the F protein[33]. A phase 3 trial of another investigational RSV mAb, suptavumab, failed to meet its primary end point because all RSV-B strains identified in the trial carried two amino acid changes in the F protein (L172Q and S173L), conferring resistance to the drug[205]. All RSV-B samples in this study also harboured these two amino acid substitutions, except for one that encoded isoleucine instead of leucine at position 173 (a nonpolar-to-nonpolar substitution).

I excluded genomic positions where consensus bases were different from the calculation of Manhattan distance, to ensure that between-host genetic distance would be driven by differences in minor alleles rather than differences at the consensus level[302]. I found that, outside the consensus-different positions, serial samples from the same individual did not have a shorter pairwise Manhattan distance than that of a randomly taken between-host pair from the same country and season. This methodology change makes the results robust to inter-host variation, in contrast to previous studies on influenza virus and RSV, where distance metrics were largely driven by consensus differences[314,315].

This study suggests that RSV-B has a more diverse within-host population than RSV-A, likely driven by selection pressure at the host-population level. This difference between the two subgroups warrants close monitoring for vaccine efficacy and emergence of neutralisation escape variants.

6

The global epidemic history of genotype ON1

Contents

6.1 Introduction 108
6.2 Methods 110
6.2.1 Sequence compilation 110
6.2.2 Maximum-likelihood phylogeny 113
6.2.3 Molecular clock analysis 114
6.2.4 Phylogeographic analysis 114
6.3 Results 115
6.3.1 Sequence compilation 115
6.3.2 Phylogenetic reconstruction of RSV-A 115
6.3.3 Phylogeographic analysis of ON1 117
6.4 Discussion 119

6.1 Introduction

RSV was first identified in an outbreak of respiratory illness in chimpanzees and referred to as the chimpanzee coryza agent in 1956[316]. It was then isolated from infants with acute respiratory disease and designated as respiratory syncytial virus in 1957[317]. Since its discovery, this virus has spread worldwide. Both subgroups, A and B, co-circulate in annual epidemics with an alternating pattern of subgroup

dominance. Multiple genotypes within each RSV subgroup also co-circulate in each epidemic with new ones emerging and old ones disappearing constantly[60,71,72].

Genotypes ON1 (RSV-A) and BA (RSV-B) have gained continued global predominance in recent years[289,318]. Both genotypes have a distinctive large sequence duplication in the second hypervariable region of the G gene; specifically 72 nt in genotype ON1 and 60 nt in genotype BA. Genotype BA was first identified in Buenos Aires, Argentina in June 1999 (GenBank accession no. AY333362)[72], and has since diversified into multiple lineages (e.g., BA1–BA10)[62]. Genotype ON1 was first isolated in Ontario, Canada in December 2010 (accession no. JN257693)[60]. Subsequent studies recovered ON1 sequences from samples collected before its first identification (2010); the oldest known ON1 sequence was recovered in Central Mexico in December 2009 (accession no. MG062683)[319].

It has been shown that genotype ON1 carries strain-specific codon changes, some of which may alter its susceptibility to neutralising antibodies[83,318]. These amino acid changes could potentially affect the effectiveness of palivizumab and/or other RSV monoclonal antibodies and vaccines under development. For example, suptavumab (an RSV monoclonal antibody) failed to meet the primary end point in a phase 3 clinical trial due to two amino acid alternations in all circulating RSV-B strains[205]. Yet there is no clear evidence that genotype ON1 has greater pathogenicity than the preceding ones (e.g., GA2, NA1)[288,318,320]. Studies that analysed global G gene sequences (either full-length or partial) reported various estimates for the time of the most recent common ancestor (TMRCA) of genotype ON1, ranging from 2005 to late 2009[83,86,321,322]. More recent studies using global whole-genome sequences estimated the evolutionary rate of RSV-A at 5.68–6.47$\times 10^{-4}$ substitutions/site/year[82,85,86] and placed the emergence of genotype ON1 in March 2005[85]. However, a systematic study based on whole-genome sequences from across the globe to assess the emergence of genotype ON1 is currently lacking. More importantly, the early transmission and dissemination of genotype ON1 have yet to be formally evaluated.

In this chapter, I undertake a phylogeographic approach to analyse a comprehensive compilation of global RSV-A whole-genome sequences. I seek to uncover the

spatiotemporal origin of genotype ON1 and its epidemic history on a global scale. Elucidating how a new genotype disseminates and how long it takes to establish global epidemics is crucial to our understanding of the transmission dynamics of RSV.

6.2 Methods

6.2.1 Sequence compilation

A total of 2,918 RSV sequences were downloaded from GenBank[232] as of 7th November 2020, with the search terms of 'human orthopneumovirus' or 'respiratory syncytial virus,' and the results limited to sequences with a length between 12500 and 16000. The search syntax was as follows:

```
("Human orthopneumovirus"[Organism] OR "Human orthopneumovirus"[All Fields]) OR ("Respiratory syncytial virus"[Organism] OR "Respiratory syncytial virus"[All Fields]) AND ("12500"[SLEN] : "16000"[SLEN])
```

Multiple sequence alignment was performed using MAFFT (v7.471)[233]. After alignment, intergenic regions were trimmed off, and the 11 open reading frames (ORFs) were concatenated. Sequences that met the following criteria were excluded: (a) sequences that had duplicate records (only one of the duplicates was retained); (b) sequences that were not of RSV; (c) recombinant, attenuated, or mutated sequences; (d) sequences that had unknown (N) bases >10% of its length; (e) sequences that did not have a collection date; (f) unverified sequences; (g) incomplete sequences; (h) sequences that were identical to others collected from the same region and the same RSV season (only the oldest one was retained); (i) sequences that were outliers from the regression of root-to-tip distance on sampling time; (j) sequences that had the 72-nt duplication but were located in a non-ON1 clade; (k) sequences that did not have the 72-nt duplication but were located in the ON1 clade; and (l) RSV-B sequences. Incomplete sequences were defined as those that had <90% of the full-length coding sequence, that is <12,337 nt for ON1 sequences and <12,272 nt for non-ON1 sequences.

Overall, 1,747 sequences were removed, including 10 sequences that had the 72-nt duplication but were located in a non-ON1 clade, and 16 sequences that did not

have this duplication but were located in an ON1 clade. A total of 1,171 complete RSV-A sequences were in the final GenBank collection. The GenBank accession numbers of the 1,171 RSV-A sequences are provided in Supplementary Table A.8.

This GenBank collection was further supplemented with RSV-A concatenated ORF sequences generated from the RESCEU studies from the United Kingdom, Spain, and the Netherlands during the 2017–2020 RSV seasons. The clinical studies[4,219,220] and RSV genome reconstruction[1] are described in detail in sections 2.1 and 2.4. For the RSV-A sequences generated from the RESCEU studies, only those with a genome coverage $\geq$70% were considered to be included (N = 322). Only one sequence was selected from each participant (N = 209). For participants with multiple samples collected, the earliest one with the highest coverage of the RSV genome was selected. Nine identical sequences, one outlier sequence from the regression of root-to-tip distance on sampling time, and eight sequences shorter than 12,337 nt were removed from the RESCEU collection, leaving a final set of 191 sequences.

The final collection consisted of 1,362 complete RSV-A coding sequences (1,171 from GenBank and 191 from RESCEU), with an alignment length of 13,707 nt. A flow diagram for the sequence compilation is shown in Supplementary Fig. B.13. The latest non-ON1 sequence in this collection was sampled in the United States on 7th March 2016. The latest ON1 sequence was collected in the United Kingdom on 13th February 2020. These sequences were classified according to their sampling location into eight groups. The eight geographical regions, modified from the WHO regions (`https://www.who.int/about/who-we-are/regional-offices`), were as follows: African region (AFR), East Asian region (EAR), Eastern Mediterranean region (EMR), European region (EUR), Latin American and the Caribbean region (LAC), North American region (NAR), Oceania region (OCE), and South-East Asian region (SEA). The geographical groups of these sequences are shown in Fig. 6.1.

To avoid potential sampling bias arising from a large number of sequences from the North American, European, and African regions, taxa from each region were downsampled to 61 sequences (the number of sequences available for the

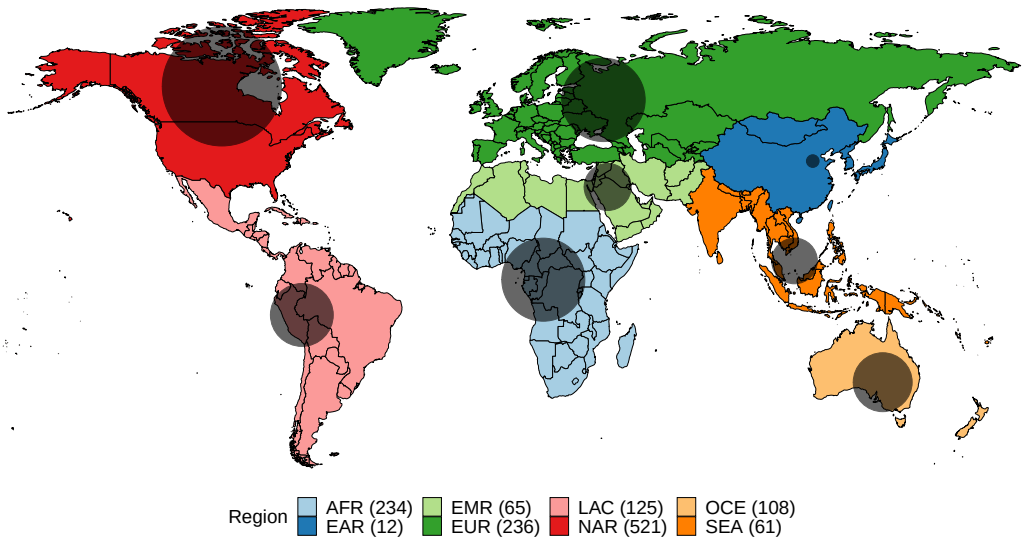


Figure 6.1: Geographical distribution of the 1,362 RSV-A sequences analysed in this study. The area of each black circle is proportional to the number of sequences in each region (shown in brackets in the legend). AFR denotes the African region; EAR, the East Asian region; EMR, the Eastern Mediterranean region; EUR, the European region; LAC, the Latin American and the Caribbean region; NAR, the North American region; OCE, the Oceania region; and SEA, the South-East Asian region.

Southeast Asian region) while balancing the numbers of ON1 and non-ON1 sequences. Specifically, 31 ON1 and 30 non-ON1 sequences were selected from each region. In regions where sequences from one group (e.g., ON1) were fewer than the required number, I supplemented them with sequences from the other group (e.g., non-ON1). In regions where total sequences were fewer than 61, all the sequences were selected without adding sequences from other regions. The downsampled dataset had 439 sequences, including 171 ON1 and 268 non-ON1 sequences. See Supplementary Fig. B.14 panels D, E, F for the number of sequences in each region.

Another downsampling was then performed to obtain 219 ON1 and 220 non-ON1 sequences (total 439 sequences) while balancing the number of sequences from each region. In order to make sure that an equal number of sequences was obtained from each region, it was necessary to divide the total number of required sequences by the number of regions. This would generate an average of 27.5 sequences per region. In cases where a region had fewer than 27 sequences, this was compensated for by taking a greater number of sequences from other regions. Specifically, for ON1

sequences, after selecting all 47 sequences from underrepresented regions (regions with fewer than 27 sequences, i.e., EAR, EMR, LAC, SEA), 43 sequences from each of the remaining four regions were selected. For non-ON1 sequences, after selecting all six sequences from the underrepresented EAR region, 30 or 31 sequences from each of the remaining seven regions were selected.

To increase temporal resolution of the datasets, one of the two oldest RSV-A sequences, sampled from the North American region in 1977, was guaranteed to be included in both downsampled datasets. Each dataset was further supplemented with the six oldest incomplete ON1 sequences (874–1100 nt long) sampled before the oldest available complete ON1 sequence collected in 2011. One was collected in Mexico in December 2009, and the other five were sampled in Panama and Canada in late 2010. The two downsampling processes were repeated twice, generating four datasets with 445 sequences each. Datasets generated from the first downsampling process were referred to as Datasets 1 and 2, and datasets generated from the second downsampling process were referred to as Datasets 3 and 4. The following results were inferred from Dataset 1 unless otherwise specified.

6.2.2 Maximum-likelihood phylogeny

jModeltest 2 (2.1.10 v20160303)[323] was used to evaluate likelihoods of 11 nucleotide substitution models with either equal or unequal base frequencies and either with or without gamma-distributed rate heterogeneity among sites (Γ) with four rate categories. The general time reversible (GTR) nucleotide substitution model with Γ was identified as the best-fitting nucleotide substitution model. Maximum-likelihood phylogenies of the complete coding sequences were reconstructed using PhyML (v3.3.20200621)[324] with this best model and maximum-likelihood estimation of equilibrium nucleotide frequencies and substitution model parameters. PhyML was run with five random starting trees and 100 bootstrap replicates, and final maximum-likelihood trees were estimated with nearest neighbour interchange (NNI) and subtree pruning and regrafting (SPR) heuristic tree search algorithms. Tree topology, branch length, and substitution rate parameters were optimised to maximise the likelihood.

6.2.3 Molecular clock analysis

Clock signal was evaluated by regressing root-to-tip genetic distances obtained from the maximum-likelihood phylogeny against sampling dates with the heuristic residual mean squared regression method[325] available in TempEst (v1.5.3)[326]. Bayesian evolutionary analyses were performed using BEAST (v1.10.4)[327] with the Hasegawa-Kishino-Yano (HKY) model, gamma-distributed rate heterogeneity among sites, and empirical base frequencies as previously described elsewhere[328]. Marginal likelihood estimation using path sampling (PS) and stepping-stone sampling (SS) was performed to identify the best-fitting combination of molecular clock and coalescent models[329,330]. Tested molecular clock models included a strict molecular clock and an uncorrelated lognormal (UCLN) relaxed molecular clock. Tested coalescent priors included a constant population size, an exponential growth, and a Bayesian piecewise-constant skyline tree priors[331]. The most appropriate combination was a UCLN relaxed molecular clock and a Bayesian skyline tree prior (Supplementary Table A.9). The Markov Chain Monte Carlo (MCMC) algorithm was run for 400,000,000 steps with parameters logged every 40,000 steps. BEAST analyses were run in duplicate for each downsampled dataset. Convergence and mixing of the MCMC chains were checked using Tracer (v1.7.1)[332]. LogCombiner (v1.10.4) was used to combine tree samples from the duplicate runs after removing 40,000,000 burn-in for each analysis. The maximum clade credibility tree was then summarised from the combined tree samples using TreeAnnotator (v1.10.4) with node heights being the posterior median node heights. ggtree (v2.4.2)[247] was used for tree visualisation and annotation.

6.2.4 Phylogeographic analysis

For discrete trait analysis, LogCombiner was used to obtain 1,500 trees from two duplicate MCMC runs after removing the first 280,000,000 MCMC steps for each run and resampling at a lower frequency of 160,000 as previously described elsewhere[328]. Two discrete traits were considered: the sampling region and the duplication status in the second hypervariable region of the G gene. Ancestral state reconstruction was performed using BEAST with an asymmetric substitution model and a Bayesian

stochastic search variable selection (BSSVS) approach[333]. The MCMC algorithm was run for 15,000,000 steps in duplicate with parameters logged every 15,000 steps. Tree samples from the duplicate runs were combined after removing 1,500,000 burn-in for each analysis, and maximum clade credibility trees were summarised and visualised in the same way as described above.

6.3 Results

6.3.1 Sequence compilation

A total of 1,171 complete RSV-A sequences downloaded from GenBank were supplemented with 191 complete RSV-A sequences generated from the RESCEU studies, making a final collection of 1,362 sequences. All the RSV-A sequences from the RESCEU studies were genotype ON1, accounting for 30% (191/643) of the ON1 sequences in this final collection. There are more ON1 sequences than non-ON1 sequences from 2012 onwards. The geographical distribution and temporal pattern of the 1,362 RSV-A sequences are shown in Fig. 6.1 and Supplementary Fig. B.14 (panels A, B, and C), respectively. To avoid potential sampling bias arising from a large number of sequences from particular regions, I downsampled taxa from each region while balancing the numbers of ON1 and non-ON1 sequences and the number of sequences from each region (see section 6.2.1 for details). The downsampled dataset consisted of 445 sequences (Supplementary Fig. B.14 panel D, E, and F).

6.3.2 Phylogenetic reconstruction of RSV-A

Maximum likelihood reconstruction using the best-fitting nucleotide substitution model showed that genotype ON1 formed a strongly supported monophyletic clade (bootstrap branch support value of 99) (Fig. 6.2). Root-to-tip genetic distances correlated highly with the sampling times ($R^2 = 0.933$), indicating strong temporal signal of this dataset (inset in Fig. 6.2; Fig. B.15). Bayesian phylogenetic inference using the best-fitting combination of molecular clock and coalescent models estimated the median evolutionary rate of the RSV-A dataset at 6.66 $\times 10^{-4}$ substitutions/site/year (95% highest posterior density [HPD] interval, 6.20 $\times 10^{-4}$

to 7.16 $\times 10^{-4}$ substitutions/site/year). The TMRCA of the RSV-A dataset was estimated to be September 1968 (95% HPD interval, May 1965 to August 1971) (node A, Fig. 6.3). The emergence of genotype ON1 was dated to February 2009 (95% HPD interval, October 2008 to May 2009) (node B, Fig. 6.3; Supplementary Table A.10). Different combinations of molecular clock models and coalescent priors showed similar estimates for the evolutionary rate of RSV-A and the TMRCA of RSV-A and genotype ON1 (Fig. 6.4), indicating that these estimates are robust to the chosen models.

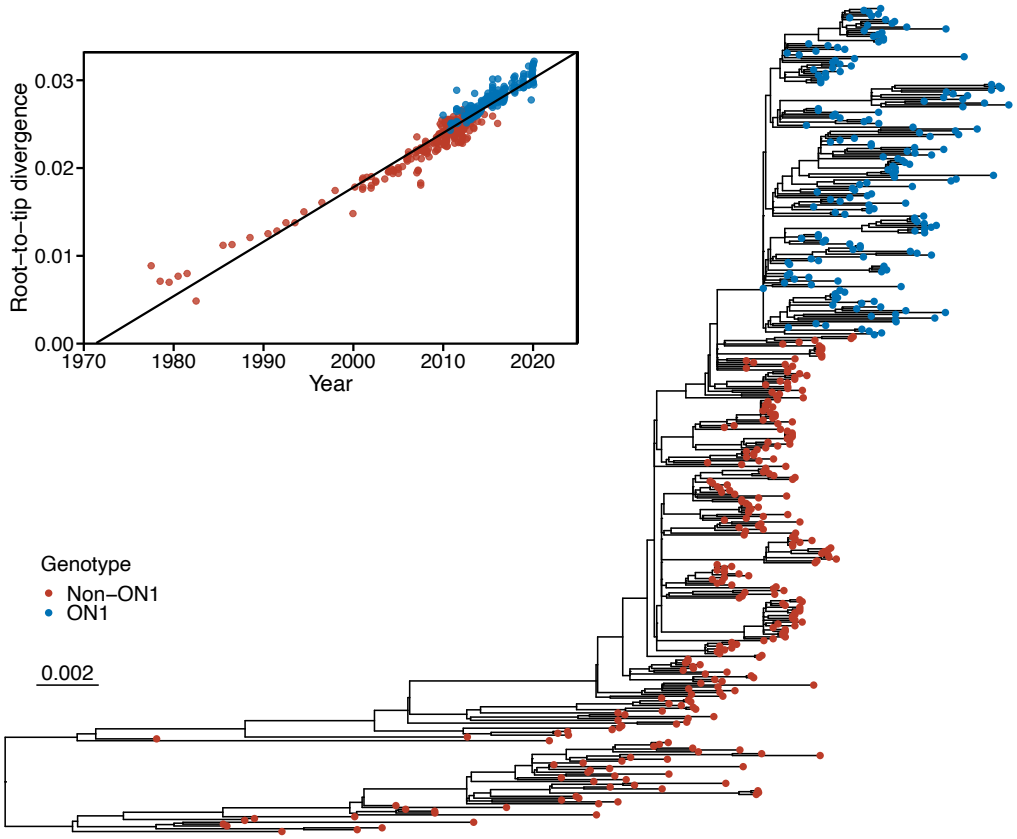


Figure 6.2: Maximum-likelihood phylogeny of Dataset 1 (N = 445). The heuristic residual mean squared regression method implemented in TempEst was used to find the best-fitting root that maximises the correlation coefficient between root-to-tip distance and sampling time. The scale bar represents the number of nucleotide substitutions per site. Non-ON1 and ON1 sequences were coloured according to the legend. The inset shows the regression of root-to-tip distance on sampling time with the correlation coefficient (R) and coefficient of determination (R^2) being 0.966 and 0.933 respectively. The time interval between the first and last samples being taken is 44 years (1977–2020). The x-intercept of the regression line is 1971.28, an estimate for the time of the most recent common ancestor.

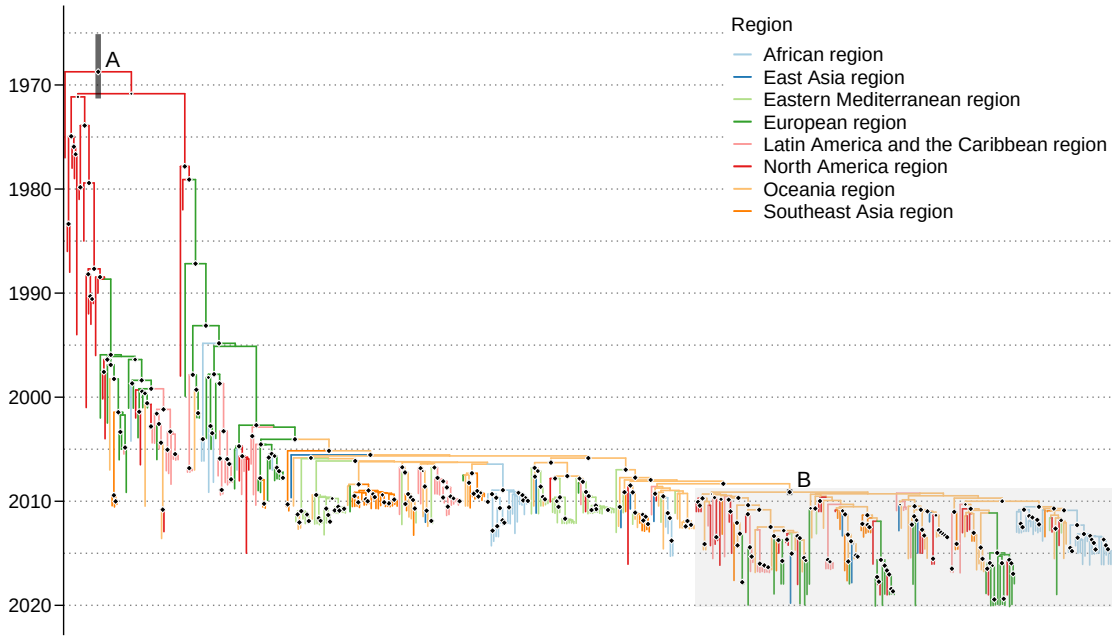


Figure 6.3: Maximum clade credibility tree of 445 RSV-A sequences (Dataset 1), among which 177 were classified as genotype ON1. The ON1 sequences form a monophyletic clade (shaded grey area). Node A is the root of the RSV-A phylogeny and node B is the divergent point of genotype ON1. Branch colours represent the sampling locations or the most probable ancestral locations. Black dots show well-supported nodes with a clade posterior probability over 70% and are sized in proportion to clade posterior probabilities. The clade posterior probabilities of both nodes A and B are 1.00. The estimated ancestral locations for nodes A and B are the North American and Oceania regions with a location posterior probability of 0.97 and 1.00, respectively. The estimated time of the most recent common ancestor (TMRCA) of node A is 1968.73 (95% highest posterior density [HPD] interval, 1965.12 to 1971.30); and 2009.12 (95% HPD interval, 2008.83 to 2009.42) for node B. The grey bars at nodes A and B show the 95% HPD interval of the TMRCA.

6.3.3 Phylogeographic analysis of ON1

All the phylogeographic analyses performed on the four datasets with 445 genome sequences placed the emergence of genotype ON1 in the Oceania region (Supplementary Table A.10). After its emergence, my analysis suggests that genotype ON1 rapidly spread to the Americas (Fig. 6.5, panel A) and was first detected in Mexico on 27th December 2009. Notably, genotype ON1 took less than 2 years to reach every region across the globe with the European and Eastern Mediterranean

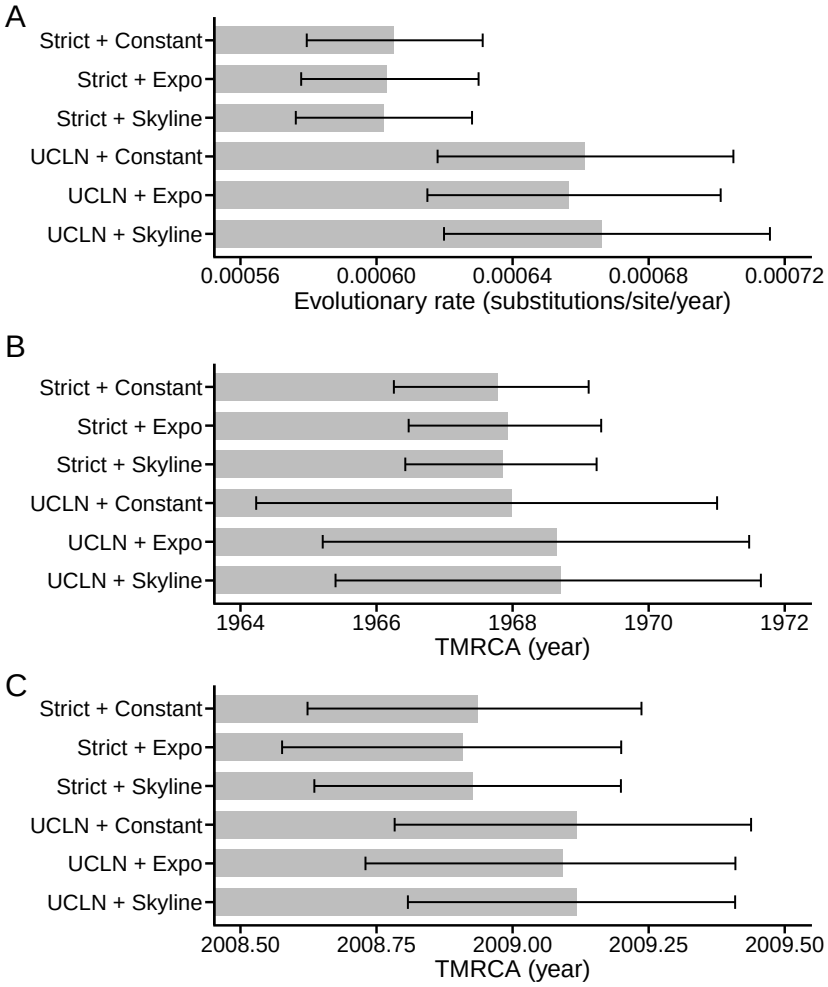


Figure 6.4: Estimated evolutionary rate and time of the most recent common ancestor (TMRCA) using different combinations of molecular clock models and coalescent priors (Dataset 1). **A.** Evolutionary rate of the RSV-A dataset (N = 445). **B.** TMRCA of the RSV-A dataset (N = 445). **C.** TMRCA of genotype ON1 (N = 177). Bar plots show the median value, and error bars show the 95% highest posterior density interval. Tested clock models included strict and uncorrelated lognormal (UCLN) relaxed molecular clocks; tested coalescent priors included a constant population size, an exponential growth (Expo), and a Bayesian piecewise-constant skyline tree priors.

regions being the last to which it spread. Genotype ON1 took over non-ON1 genotypes from mid-2011 onwards and has become the only RSV-A genotype since mid-2016 (Fig. 6.5, panel B).

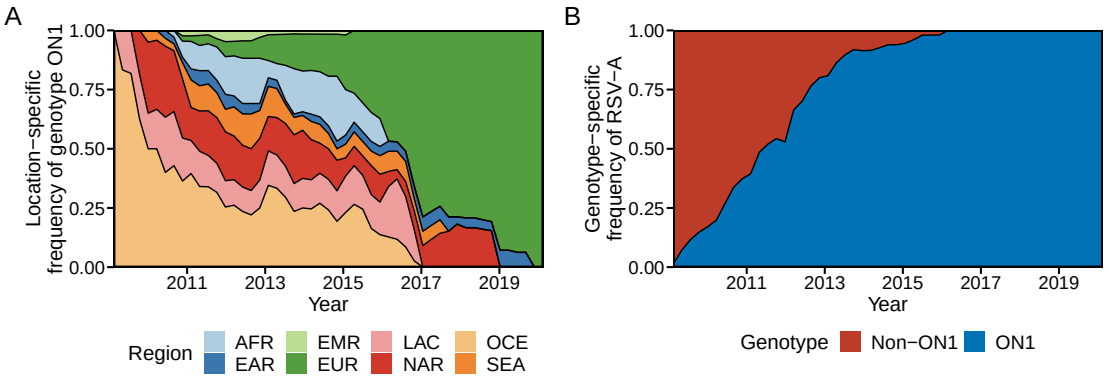


Figure 6.5: Frequency plots since the emergence of genotype ON1 (node B, Fig. 6.3). **A.** Location-specific frequency of genotype ON1, indicating its global spread. For the region abbreviation see Fig. 6.1 legend. **B.** Genotype-specific frequency of RSV-A.

6.4 Discussion

I analysed 445 complete RSV-A sequences collected from across the globe over a period of 44 years, including 177 ON1 sequences with the earliest collected in Mexico in December 2009 and the latest collected in the United Kingdom in February 2020. Genotype ON1 forms a well-supported monophyletic clade. Phylogeographic analysis place the origin of genotype ON1 in the Oceania region (posterior probability = 1.0) in February 2009 (95% HPD interval, October 2008 to May 2009). After its emergence, genotype ON1 first spread to the Americas within 1 year, and soon to every region of the world within 2 years. It established its dominance in mid-2011, and non-ON1 sequences have not been detected since mid-2016.

My analysis suggested that all global ON1 sequences fall in a single monophyletic clade. Of note, 26 sequences that either had the 72-nt duplication but were located in a non-ON1 clade or did not have the 72-nt duplication but were located in the ON1 clade were excluded. These sequences may represent a mixed virus population, an assembly error[334], or incomplete sequencing coverage of the duplication region. A phylogenetic analysis using sequences outside the second hypervariable region of the G gene similarly grouped ON1 sequences (and only ON1 sequences) into a monophyletic clade (data not shown), suggesting the 72-nt duplication is not the only genetic signature that defines genotype ON1, consistent with previous

findings[85]. This further justifies the exclusion of the 'misplaced' sequences from my analysis. However, there have been propositions that the emergence of genotype ON1 was a result of multiple duplication events[86,319]. I could not completely exclude the possibility that these misplaced sequences may have arisen from independent duplication events, though they did not seem to establish persistent transmission that warrants being included in this analysis.

My estimate for the evolutionary rate of RSV-A is comparable to that of previous studies[82,85,86]. However, my estimates for the TMRCA of RSV-A and genotype ON1 are significantly later than that of previous studies. Specifically, previous estimates for the TMRCA of RSV-A were 1951 (95% HPD interval, 1937 to 1964) by Schobel[86] et al. and July 1937 (95% HPD interval, July 1917 to April 1954) by Trovão[87] et al. Otieno et al. estimated the TMRCA of genotype ON1 to be March 2005 (95% HPD interval, December 2003 to June 2006)[85]. In contrast, my estimates for the TMRCA of RSV-A and genotype ON1 were September 1968 (95% HPD interval, May 1965 to August 1971) and February 2009 (95% HPD interval, October 2008 to May 2009), respectively. The differences are likely due to different datasets analysed and, to a lesser extent, different models used in the Bayesian analysis. Temporal signal is a factor of the sampling time frame and the size of alignment. My dataset consisted of 445 RSV-A sequences collected across 44 years (1977–2020), including 177 ON1 sequences collected across 12 years (2009–2020). Schobel et al. analysed 447 RSV-A sequences collected across 33 years (1981–2013), Trovão et al. analysed 669 RSV-A sequences collected across 39 years (1977–2015), and Otieno et al. analysed 344 ON1 sequences collected across 6 years (2011–2016). Temporal signals were similar in my dataset and Trovão's dataset (R^2 of the root-to-tip regression analysis, 0.93 vs. 0.94). The other studies did not specify the temporal signal of their datasets. I also showed that different combinations of molecular clocks and coalescent priors yielded similar estimates in my dataset (Fig. 6.4).

To eliminate the effect of overrepresenting regions on the phylogeographic analysis, I downsampled the RSV-A sequences with two different strategies: (1) prioritising balancing the number of sequences from each region and (2) prioritising

balancing the numbers of ON1 and non-ON1 sequences. The phylogeographic inferences on all datasets led to the same origin estimate with posterior probability of 1.0. However, even with downsampling, the estimated dates and locations are still limited to sequences that survived to be sampled and sequenced, which is inevitable for studies on pathogen evolution using genetic sequences. Furthermore, although the analysed datasets consist of sequences with the most balanced geographical distribution and the longest time interval to date, sequences from certain regions (East Asian region in particular) are still lacking.

Epidemiological studies have shown that RSV transmission is low between October and May in Oceania, and it starts increasing in June and peaks in July and August[335,336]. Combining with these epidemiological data, my analysis suggests that genotype ON1 emerged during an interepidemic trough when the RSV activity was relatively low in Oceania (between October 2008 and May 2009), supported by the results from Datasets 1 to 3. Another possibility is that genotype ON1 emerged from tropical Australia during its peak RSV epidemic in February and March, coincident with rainfall[337]. This 72-nt duplication might occur in an immunocompromised patient with prolonged RSV infection[338] or a zoonotic reservoir[339]. After its emergence, genotype ON1 was brought to the Americas presumably during the winter months in Australia, July and August, when the RSV transmission was high. After arriving in the Americas, genotype ON1 was first detected in Mexico at the end of 2009, and then in Panama and Canada a year later. This rapid spread of genotype ON1 was likely due to continuous global population movements and spread in susceptible populations, as have been seen in the 2009 H1N1[340] and COVID-19[341,342] pandemics. This is also supported by studies reporting multiple importations of genotype ON1 into a geographically defined region[83,87,343] and identical sequences found in distant parts of the world[83] during an epidemic.

The discrete trait analysis showed that genotype ON1 took over other genotypes in mid-2011 (2–3 years after its emergence) on a global scale. Regional and countrywide studies reported the dominance of genotype ON1 at various different times (e.g., 2011/12 season in Kenya[318], 2012/13 season in Germany[288] and South

Korea[344]). This regional variation is presumably due to the spatial dynamics of viral transmission and sampling bias in each region. The fact that genotype ON1 gained dominance may suggest its selective advantage in transmission over the preceding genotypes. Different hypotheses have been proposed to explain this advantage. One functional study showed that binding of RSV genotype BA to host cells is augmented *in vitro* by the presence of the 60-nt duplication in the G gene[345]. This enhanced binding ability may also apply to genotype ON1's G protein, which has the 72-nt duplication. Another possibility is that the longer length of the G protein in genotype ON1 than in other RSV-A genotypes provides more opportunities for amino acid alterations, thus causing antigenic variability and immune evasion[318]. Although this is beyond the scope of my study, future work should investigate whether an enhanced biological fitness of genotype ON1 is linked with its dominance in different regions.

In conclusion, I reconstruct the evolutionary history of genotype ON1. Genotype ON1 was detected less than a year after its emergence, then rapidly spread to every region of the world within 2 years of its emergence and became the dominant strain within 3 years. This study sheds new light on the transmission dynamics of RSV. The rapid spread of a new RSV strain and establishment of its predominance warrants continued genetic surveillance to promptly detect emerging variants that may gain an evolutionary advantage, increase disease severity, or confer resistance to future RSV vaccines or therapeutics.

7

Summary of the insights and contributions of this thesis

Contents

7.1 Summary of the novel and important findings resulting from this thesis 124
7.1.1 Transcriptome profiling of respiratory syncytial virus . . 124
7.1.2 Molecular epidemiology of respiratory syncytial virus in Europe 125
7.1.3 Factors associated with severe RSV disease 126
7.1.4 Patterns of within-host virus diversity 127
7.1.5 Emergence and global spread of genotype ON1 127
7.2 Limitations 127
7.3 Implications and future directions 130
7.4 Closing remarks 133

Since its discovery in 1956, RSV has become the leading cause of childhood respiratory illness and causes substantial morbidity and mortality every year. The global impact of RSV disease in adult populations has also become increasingly recognised. Health organisations and stakeholders have prioritised RSV research and the development of RSV vaccines and monoclonal antibodies (mAbs). WHO has launched a 3-year extension phase of the global RSV surveillance programme since 2019[260]. One of the propositions is to enhance virologic monitoring, to differentiate RSV types, and to identify RSV genetic groups. Many vaccine and mAb candidates

have produced promising results and entered into late stage clinical trials. For example, it was recently announced that a phase 3 trial of prefusion-specific mAb nirsevimab (MEDI8897) has met its primary end point of reducing medically attended RSV-associated lower respiratory tract infection in healthy late preterm and term infants[207]. However, gaps still exist in our understanding of molecular characteristics of RSV[257]. It is crucial to fill these gaps before new therapeutic and preventative measures become available. I addressed these gaps in this work.

In this thesis, I sought to identify viral and microbial factors associated with severe RSV disease, to explore RSV variants and diversity at the individual and population levels, and to reconstruct the global epidemic history of the currently dominant genotype ON1. Nasopharyngeal swabs collected from RSV-infected infants <1 year of age and older adults ≥60 years of age were directly used for targeted metagenomic sequencing without prior culture or purification of viral particles. Sequencing directly on the clinical samples yielded more accurate results, preventing additional mutations introduced during viral isolation[346]. I analysed the sequencing dataset to explore different aspects of the molecular characteristics of RSV, including differential gene expression, viral and microbial determinants of severe disease, potential antigenic variants, and transmission dynamics among others. These samples represented wild type RSV during active clinical infection instead of laboratory strains used in *in vitro* studies. Therefore, the work presented in this thesis reflected the natural genetic characteristics and pathogenetic potential of RSV and broadened our knowledge of different aspects of the genomic features of RSV.

7.1 Summary of the novel and important findings resulting from this thesis

7.1.1 Transcriptome profiling of respiratory syncytial virus

This work started with validating a revised methodology of targeted metagenomic sequencing on RSV-positive nasopharyngeal swabs collected in the RESCEU studies. The RSV genome was successfully sequenced and reconstructed from most of the samples. All the samples with incomplete genome assembly had low viral

load. A strong correlation existed between viral load and the number of uniquely mapped (i.e., deduplicated) RSV reads, indicating the ability of this sequencing method to quantify RSV molecules.

Applying this quantification feature of the sequencing method to RSV transcriptomics showed that the relative transcript abundance of RSV genes generally followed an expected transcriptional gradient derived using *in vitro* transcription systems. However, rather than uniformly decreasing transcript levels, the NS2, P, and G genes had a higher level of transcripts than the genes preceding them, suggesting a more complex post-transcriptional control of transcript abundance. A formal analysis of differential gene expression uncovered significant differences in transcript levels of the P gene and ORF M2-1 (but not of ORF M2-2) between RSV-A and RSV-B. Host disease severity was not associated with differential gene expression. The phosphoprotein (P gene product) is a cofactor of the viral RNA-dependent RNA polymerase, and the M2-1 protein is an anti-terminator, promoting transcription of full-length mRNA molecules and read-through transcripts. In contrast, the M2-2 protein shifts RNA synthesis from transcription to replication.

The observed differences in RSV-A and RSV-B transcription are suggestive of greater differences in the biology of the two subgroups than has been generally appreciated, with potential consequences for the design of vaccines and novel therapeutic agents. I further explored the two subgroups with respect to genetic diversity at the population level and within individual infections, and the association with co-infecting organisms.

7.1.2 Molecular epidemiology of respiratory syncytial virus in Europe

More than 800 samples collected from three European countries (United Kingdom, Spain, and Netherlands) during 2017–2020 were sequenced using the validated targeted metagenomic sequencing method, and RSV genomes were reconstructed. RSV-B dominated the first two seasons, whereas RSV-A became the predominant subgroup in the 2019–2020 season. All the RSV-A strains were genotype ON1,

and all RSV-B strains were genotype BA, indicating the continuing dominance of both genotypes. Genotype ON1 had greater population diversity than genotype BA, whereas genotype BA showed stronger temporal phylogenetic clustering than genotype ON1.

7.1.3 Factors associated with severe RSV disease

The associations were explored between RSV disease severity and host, viral, and microbial factors in the infant participants. Young age was significantly associated with severe RSV disease, demonstrated by multiple clinical outcome variables except for fever. Older infants and infants with high viral load had a higher probability of fever than younger ones and those with low viral load, respectively. RSV burden decreased by 53% daily following the first sample, but the viral clearance was delayed in infants requiring mechanical ventilation than in those who did not (daily decrease, 49% vs. 60%). RSV-A and RSV-B caused diseases of similar severity and exhibited similar viral dynamics.

Concurrent detection of other viruses and bacteria were more common in older infants than in younger ones. At least one other virus was recovered in 24% of the RSV-infected infants, with human rhinovirus being the most frequently co-detected virus, found in 15% of the RSV-infected infants. Infants with co-detection of another virus (particularly human herpesvirus 6) were more likely to require intensive care than those without. *Moraxella*, *Streptococcus*, and *Haemophilus* were the most commonly identified bacterial genera in the RSV-infected infants. After adjusting for confounding factors and the presence/absence of other bacteria, detection of *Haemophilus* spp. strongly correlated with severe RSV infection. Some of the cases may represent bacterial secondary infections. Colonisation with *S. pneumoniae* was also associated with more severe RSV disease requiring hospitalisation, while colonisation with *Moraxella* spp. was associated with less severe disease which did not cause respiratory failure.

7.1.4 Patterns of within-host virus diversity

Deep sequencing analysis of the virus population within each infected individual uncovered distinct patterns of within-host virus diversity between RSV-A and RSV-B. RSV-B exhibited greater intrahost diversity than RSV-A, suggesting that RSV-B may be subject to stronger selective pressure than RSV-A. Consistent with this finding, two RSV-B consensus strains and one minor variant were found to carry amino acid alterations (K68N in the consensus strains and S275P in the minor variant) likely conferring resistance to mAbs, but no RSV-A strains carried known mutations conferring resistance to mAbs. No specific temporal change in within-host virus diversity or minor variants was found, indicating stochastic processes of virus evolution at the individual level.

7.1.5 Emergence and global spread of genotype ON1

The RSV-A strains generated from the RESCEU studies were added to a global collection of RSV-A isolates for a formal phylogeographical analysis. The RESCEU strains accounted for 30% of the ON1 sequences in the final collection, substantially increasing the number of ON1 sequences, especially those in the European region. The analysis suggested that all the ON1 strains fall in a single monophyletic clade. The most recent common ancestor of the RSV-A strains dates back to 1968. The emergence of genotype ON1 was estimated to have occurred between October 2008 and June 2009 in the Oceania region. After its emergence, genotype ON1 rapidly spread to the Americas and was first detected in Mexico in late 2009. Genotype ON1 took about 2 years to reach every region of the world and became the dominant genotype from mid-2011 onwards on a global scale. There has been no non-ON1 sequence isolated in any region of the world since April 2016.

7.2 Limitations

The work presented in this thesis had some limitations. The pilot study validating the methodology of targeted metagenomic sequencing revealed significantly differential

expression of the P gene and ORF M2-1 between RSV-A and RSV-B. These findings were based on only 13 samples with complete RSV genome assembly. Power to determine whether these findings were generalisable to the overall RSV population was insufficient. However, the differentially expressed P gene and ORF M2-1 were identified on the basis of a false discovery rate (FDR) threshold of 5%, giving sufficient power to discover the genuinely differentially expressed genes among this dataset. In addition, the probe set used for target enrichment has been validated and benchmarked[224], and the efficiency of capture has been shown to remain stable at up to 20% sequence divergence between the probes and the target species[225]. A standardised approach to exploring differential gene expression was also applied. Altogether, the results were robust to technical and analytical bias. A confirmation study including more samples would give greater power and provide more details about the RSV transcriptional profiles.

Although a significant correlation was found between viral load and the number of uniquely mapped RSV reads, the PCR assay used to quantify the viral load was likely to have different efficacies or sensitivities between RSV-A and RSV-B, resulting in a generally lower viral load for RSV-A than RSV-B by ~2 $\log_{10}$. This difference complicated the implementation of using read count as a proxy for viral load. However, a small number of samples have been quantified using another RT-PCR assay performed by GlaxoSmithKline. No significant separation of the regression lines was found between RSV-A and RSV-B (Fig. 4.1).

The cohort of older adults in the RESCEU study consisted of a very small number of RSV-positive samples, resulting in insufficient power to evaluate factors associated with severe RSV disease and to examine the nasopharyngeal microbiome in adults. The analysis of within-host virus diversity showed that older adults had a more diverse within-host RSV-B population than infants. It may be due to different immune pressures imposed on RSV between the two age groups. However, this finding was based on only seven samples from the adults. Future studies enrolling more adult patients are required to describe the molecular characteristics of RSV in the adult population.

The targeted metagenomic sequencing selectively enriched 30 bacterial species of high pathogenic potential. Although important bacterial pathogens could be recovered in my study, this sequencing method, unlike 16S rRNA-based methods, was not designed to completely characterise the nasopharyngeal microbiome. Thus, it precludes calculation of relative abundance and diversity metrics of the microbial communities. Furthermore, this sequencing dataset consisted of no sample collected from infants without any respiratory symptoms (healthy controls) and only few samples collected from infants with acute respiratory symptoms but without RSV infection. This resulted in insufficient power to compare the respiratory microbiota and virome between infants with and without RSV infection.

Nucleic acid was extracted at different concentrations and samples were sequenced on different platforms, both of which contributed to batch effects on the analysis of within-host virus diversity. To adjust for the batch effects, two different approaches were applied: (1) linear regression modelling with batch as a covariate, and (2) standardisation of data within each batch using z-score. Both approaches showed the same significant results, indicating the results were robust to the batch effects. The presence of significant batch effects also emphasises the importance of using the same methodology and platform for sequencing whenever possible, and the necessity of adjusting for any confounding factors when analysing sequencing data. These elements should be taken into account when planning future sequencing studies.

The epidemic history of genotype ON1 was reconstructed from a global collection of RSV-A sequences, supplemented by 191 ON1 sequences generated by the sequencing effort of the RESCEU studies. Although this dataset represented one of the most comprehensive datasets to date, there were only few sequences available from certain regions of the world, especially the east and southeast Asian and the eastern Mediterranean regions. The tropics have been shown to be a persistent reservoir for other respiratory viruses including influenza A virus, being the source of annual influenza epidemics in temperate regions[347]. Efforts should be made to increase the number of sequences from the tropics and other regions where

sequence data are lacking, which will improve the accuracy of phylogeographical reconstruction of the RSV epidemic history.

7.3 Implications and future directions

The differential gene expression analysis conducted in the pilot study revealed differential expression of the P gene and the ORF M2-1 between RSV-A and RSV-B and nonconcordant expression of the ORFs M2-1 and M2-2 within individual samples. Based on the functions and expression patterns of these genes and ORFs, plausible hypotheses were that the two ORFs of the M2 gene may represent separate transcriptional units, and the two subgroups of RSV may have different regulatory mechanisms of RNA transcription. In the light of the small sample size in this analysis, a further transcriptomic study is now being planned to analyse the whole RSV sequencing dataset from the RESCEU studies. It would provide greater power to explore RSV gene expression. Preliminary results of the further transcriptomics work are shown in Appendix C.

Manipulation of the M2 gene has been used to design RSV vaccines. Viral-vectored vaccine candidates encoding the F, N, and M2-1 proteins, using modified vaccinia virus Ankara (MVA-RSV)[216], replication-defective chimpanzee adenovirus (PanAd3-RSV)[216], or replication-incompetent chimpanzee adenovirus 155 (ChAd155-RSV)[217], have been shown to be safe, immunogenic, and able to induce RSV-specific humoral and cellular responses in adults. Intriguingly, the differential expression analysis suggested that the ORF M2-1 may be implicated in type-specific differences, which provides an angle to follow up in evaluating the effectiveness of a vaccine that targets, among other things, M2-1.

This targeted metagenomic sequencing method produces quantitative estimates of viral load directly from the sequencing output. The quantitative and high-resolution nature makes it a useful tool in clinical diagnostics and public health surveillance. A single assay can provide viral load estimates and whole-genome sequences from clinical samples. However, the commercial PCR assay that I used in this work showed significant differences in viral load between RSV-A and RSV-B.

GlaxoSmithKline is now carrying out their PCR assay on more samples as planned by the RESCEU studies. These results will be used to re-examine the correlation between read count and viral load.

I uncovered host, viral, and microbial factors significantly associated with severe RSV disease in infants. These results would aid identification of infant populations that are at risk for severe RSV disease where the potential therapeutic and preventative measures could be directed. For example, delayed viral clearance was associated with the need for mechanical ventilation. Early administration of an antiviral may hasten viral clearance and presumably reduce the incidence of respiratory failure. Furthermore, infants colonised or co-infected with *Haemophilus* spp. or *S. pneumoniae* were more likely to have severe RSV infection. Decreasing the colonisation with these bacteria through vaccination, early identification of bacterial secondary infection using multiplex PCR assays (e.g., Filmarray Pneumonia/Pneumonia Plus Panel[348]), or prompt antimicrobial treatment may prevent serious RSV infection and associated complications in infants. I am currently planning to submit the data presented in chapter 4 to peer-reviewed journals, either as two separate papers or as one microbiome/virome paper with the other data provided in supplementary information.

In addition to the differentially expressed RSV gene and ORF, distinct patterns of within-host virus populations were also found between RSV-A and RSV-B. RSV-B exhibited greater intrahost diversity than RSV-A and had a couple of variants that may confer resistance to mAbs. This finding is significant because many RSV vaccines in clinical trials are based on the nucleotide sequences or structure of RSV-A prototypes[310,311], including DS-Cav1[312], a promising subunit vaccine candidate with stably exposed prefusion-specific antigenic site ∅. In addition, a phase 3 clinical trial of prefusion-specific mAb suptavumab failed to meet its primary end point because all RSV-B isolates carried two amino acid changes in the F protein (L172Q and S173L) that conferred resistance to this drug. Therefore, it is important to separately examine the anti-RSV-B efficacy of developing vaccines

and mAbs and to closely monitor the circulating RSV strains to promptly identify potential vaccine and mAb escape variants.

The rapid spread and global establishment of genotype ON1 suggests its fitness advantage over the preceding genotypes. Functional studies examining its replicative and transmission fitness and its susceptibility to human sera would greatly enhance our knowledge of RSV infection dynamics. Another plausible explanation is the interaction with waning host immunity. This could be tested in future work by exploring the genetic history of the G and F genes, both of which are the main targets of human immune responses, in ON1 and in older RSV strains from previous epidemic seasons. Similar genetic features between ON1 and a previously circulating RSV strain would be consistent with waning population immunity to particular viral epitopes.

The RESCEU studies also collected blood samples from RSV-infected participants for immunological and transcriptomic studies. Future work is being planned to investigate the association between RSV strains and host immune responses, through gene expression analysis and antibody profiling. Higher concentrations of prefusion F binding antibodies have been shown to be associated with less severe RSV disease[349]. RSV infection has been linked with overexpression of neutrophils, inflammation, and interferon genes and suppression of T and B cell genes[350]. Whether a certain strain or genotype induces higher concentrations of prefusion F binding antibodies or greater immune dysregulation would provide insight into strain-specific immune protection against severe RSV infection or reinfection and immune selective pressure on RSV evolution.

RSV infection in early infancy has been shown to be associated with recurrent wheeze and asthma in childhood and early adulthood[122–124]. Annual questionnaires regarding the respiratory sequelae of RSV infection are currently being gathered from RSV-infected infant participants enrolled in the RESCEU studies up to 3 years of age. Combing with this information, another potential utilisation of this comprehensive sequencing dataset is to explore the links between viral molecular features and respiratory sequelae. Once the follow-up period ends, one could analyse

the association between the incidence of respiratory sequelae and viral genotypes, variants, and dynamics. This would provide valuable insight into the pathogenesis and prediction of respiratory sequelae following an RSV infection.

7.4 Closing remarks

Several vaccine and mAb candidates have had promising results from clinical trials. It is expected that we will have the first licensed vaccine or more efficacious mAb in the near future. The distinct features of gene expression and within-host virus populations between RSV-A and RSV-B potentially have an impact on the efficacy of these preventative measures. Rapid transmission of an emerging RSV strain or genotype warrants continuing molecular surveillance of the circulating strains. This would be even more important once a vaccine or mAb is rolled out and adds extra selective pressure on RSV. The findings on the changing gene expression levels and diversity among two RSV genotypes provide new insight into the phylogenomic and functional variations of clinical RSV infection and pave the way for a better understanding of infection with RSV and other RNA viruses. The work presented in this thesis makes a substantial contribution towards our knowledge of the molecular characteristics and transmission dynamics of RSV in the pre-RSV vaccination era.

Appendices

A

Supplementary tables

Contents

A.1 Supplementary table for RSV genome and transcriptome sequencing (chapter 3) 136
A.2 Supplementary tables for factors associated with RSV disease severity (chapter 4) 137
A.3 Supplementary table for RSV within-host diversity (chapter 5) 142
A.4 Supplementary tables for global epidemic history of genotype ON1 (chapter 6) 143

A.1 Supplementary table for RSV genome and transcriptome sequencing (chapter 3)

Table A.1: Accession numbers of sequences included in phylogenetic analysis.[a]

RSV-A		**RSV-B**	
JX015481	KJ643581	JX576729	KU316116
JX015482	KM042390	JX576730	KU839638
JX015484	KP218910	JX576731	KU950611
JX015496	KU316171	JX576732	KX655674
JX015499	KX655618	JX576733	KX765889
KF530261	KX655670	JX576734	KY249663
KF826830	KX655672	JX576736	KY249665
KF826831	KX765934	KF530262	KY249666
KF826832	KY654511	KF530265	KY249675
KF826854	MH181909	KF530266	KY249676
KF826856	MH181981	KF826818	KY249677
KJ627316	MH181992	KF826853	LC384999
KJ627343	MH447952	KF826860	MG431252
KJ627358	MH760596	KJ627251	MH594379
KJ643503	MH760608	KP317929	MH760695

[a] Adapted from Lin GL, et al. J Infect Dis 2020;222:S666–S671[1], with permission (licence number 5066600903338).

A.2 Supplementary tables for factors associated with RSV disease severity (chapter 4)

Table A.2: Criteria for the presence status of each target virus.

Virus	Criteria[a]
Human adenovirus	Presence: $\geq$ 50 unique reads & $>$ 30% coverage Equivocal: $<$ 50 unique reads or 10–30% coverage Absence: $<$ 50 unique reads & $<$ 10% coverage
Human coronavirus	Presence: $\geq$ 50 unique reads Equivocal: $<$ 50 unique reads & $\geq$ 10% coverage Absence: $<$ 50 unique reads & $<$ 10% coverage
Human cytomegalovirus	Presence: $>$ 30 unique reads or $\geq$ 30% coverage Equivocal: 10–30 unique reads & $<$ 30% coverage Absence: $<$ 10 unique reads
Enterovirus	Presence: $>$ 20 unique reads or $\geq$ 40% coverage Equivocal: 10–20 unique reads & $<$ 40% coverage Absence: $<$ 10 unique reads
Human herpesvirus 6	Presence: $\geq$ 50 unique reads Equivocal: $<$ 50 unique reads
Influenza C virus	Presence: 1033 unique reads (only one sample)
Human parainfluenza virus	Presence: $>$ 20 unique reads Equivocal: 9 unique reads
Human metapneumovirus	Equivocal: $\geq$1 reads classified as human metapneumovirus by Kraken Absence: no reads classified as human metapneumovirus by Kraken
Human parechovirus	Presence: $\geq$ 25 unique reads Equivocal: $<$ 25 unique reads but $>$ 10 pre-deduplicated reads Absence: $\leq$ 10 pre-deduplicated reads
Human bocavirus	Presence: $>$ 50 unique reads Equivocal: 10–50 unique reads Absence: $<$ 10 unique reads

[a] Coverage represents the percentage of genomic regions covered by a minimum of 2 reads (without deduplication).

Table A.3: Number of RSV-infected infants from each country and each season.

	Birth cohort (N = 146)	**Infant case–control (N = 294)**
2017–18	13	57
United Kingdom	0	6
Spain	10	0
Netherlands	3	51
2018–19	41	122
United Kingdom	12	75
Spain	10	10
Netherlands	19	37
2019–20	92	115
United Kingdom	18	58
Spain	32	15
Netherlands	42	42

Table A.4: Demographic characteristics of the RSV-infected infants, stratified by age group.[a]

	<3 mo (N = 170)	**3 to <6 mo (N = 117)**	**≥6 mo (N = 144)**	**P Value**
Gestational age				
Median (IQR)—wk	39.3 (38.0–40.1)	40.0 (39.0–40.7)	39.7 (38.8–40.7)	0.0003[b]
Distribution[c]				0.24
<32 wk	4/168 (2)	1/115 (1)	3/143 (2)	
≥32 to <37 wk	11/168 (7)	2/115 (2)	4/143 (3)	
≥37 wk	153/168 (91)	112/115 (97)	136/143 (95)	
Female sex—no. (%)	65 (38)	51 (44)	74 (51)	0.06
Weight—total no.[d]	142	64	75	
Mean ± SD—kg	4.6 ± 1.1	6.8 ± 0.9	8.4 ± 1.4	< 0.0001
Comorbidity—no. (%)	19 (11)	5 (4)	14 (10)	0.11

[a] IQR denotes interquartile range; SD, standard deviation.

[b] Pairwise comparisons using Mann–Whitney U tests with the Benjamini–Hochberg procedure showed the youngest group (<3 mo) had a significantly smaller gestational age than the older groups (3 to <6 mo and ≥6 mo) (P = 0.0006 and P = 0.0030 respectively), but no significant difference in gestational age between the older groups (P = 0.55).

[c] Data are shown as no./total no. (%).

[d] Only infants in the case–control study had information about current body weight.

Table A.5: Incidences of the two RSV subgroups in each country and season.

	RSV-A (N = 216)	**RSV-B (N = 207)**	**Mixed (N = 9)**	**Unknown (N = 1)**
2017–18	17	52	1	0
United Kingdom	4	2	0	0
Spain	0	10	0	0
Netherlands	13	40	1	0
2018–19	68	89	4	0
United Kingdom	40	45	2	0
Spain	13	7	0	0
Netherlands	15	37	2	0
2019–20	131	66	4	1
United Kingdom	44	30	0	0
Spain	29	18	0	0
Netherlands	58	18	4	1

Table A.6: The presence and absence of each identified bacterial genus in the RSV-infected infants (N = 433).[a]

	Presence	**Absence**	**Equivocal**
Moraxella	292 (67)	98 (23)	43 (10)
Streptococcus	237 (55)	76 (28)	120 (18)
Haemophilus	181 (42)	224 (52)	28 (6)
Staphylococcus	47 (11)	370 (85)	16 (4)
Neisseria	44 (10)	361 (83)	28 (6)
Burkholderia	37 (9)	362 (84)	34 (8)
Dolosigranulum[b]	34 (8)	382 (88)	17 (4)
Gemella[b]	19 (4)	400 (92)	14 (3)
Enterococcus	15 (3)	406 (94)	12 (3)
Granulicatella[b]	8 (2)	417 (96)	8 (2)
Rothia[b]	7 (2)	422 (97)	4 (1)
Escherichia	5 (1)	427 (99)	1 (<1)
Aggregatibacter[b]	3 (1)	428 (99)	2 (<1)
Corynebacterium[b]	3 (1)	419 (97)	11 (3)
Pasteurella[b]	3 (1)	425 (98)	5 (1)

[a] Data are shown as number (percentage). Percentages may not total 100 due to rounding.

[b] These organisms were not included in the *Castanet* enrichment panel, so their read numbers were expected to be lower.

A.3 Supplementary table for RSV within-host diversity (chapter 5)

Table A.7: Characteristics of RSV samples by sequencing batch.[a]

	Batch 1 (N = 17)	**Batch 2 (N = 375)**	**Batch 3 (N = 94)**	**Batch 4 (N = 375)**
Included samples for this study	11	113	41	157
RSV-A samples	3	53	16	104
RSV-B samples	8	60	25	53
From infants	11	106	39	157
From older adults	0	7	2	0
Sequencing platform	MiSeq	NovaSeq	MiSeq	NovaSeq
Read length	265	151	300	151
Raw read pairs ($\log_{10}$)[b]	6.11 (5.51–6.73)	6.09 (5.44–7.36)	5.39 (4.96–6.43)	6.14 (5.15–7.47)
Unique RSV read pairs ($\log_{10}$)[b]	5.13 (4.13–5.65)	4.57 (4.02–5.91)	4.40 (4.01–5.52)	4.73 (4.01–5.76)
Percent duplication, median (range)	78 (50–91)	91 (80–98)	73 (37–88)	91 (72–98)
Minimum genome coverage (%)	100	100	100	99.9
Depth of coverage[b]	3,763 (696–7,601)	3,561 (1,091–7,930)	2,209 (525–7,157)	3,994 (719–7,897)

[a] Sequencing statistics were based on the included samples only.
[b] Data are shown as median (range).

A.4 Supplementary tables for global epidemic history of genotype ON1 (chapter 6)

Table A.8: GenBank accession numbers of the 1,171 complete RSV-A sequences.

GenBank accession number				
FJ948820	GU591759	JF920046	JF920047	JF920048
JF920049	JF920050	JF920051	JF920052	JF920053
JF920054	JF920055	JF920056	JF920057	JF920058
JF920062	JF920063	JF920065	JF920066	JF920067
JQ901447	JQ901448	JQ901449	JQ901450	JQ901451
JQ901452	JQ901453	JQ901454	JQ901455	JQ901456
JQ901457	JQ901458	JX015479	JX015480	JX015481
JX015482	JX015483	JX015484	JX015485	JX015486
JX015487	JX015488	JX015489	JX015490	JX015491
JX015492	JX015493	JX015494	JX015495	JX015496
JX015497	JX015498	JX015499	JX069798	JX069799
JX069800	JX069801	JX069802	JX069803	JX627336
KC731482	KC731483	KC978856	KF530258	KF530260
KF530261	KF530268	KF530269	KF826816	KF826817
KF826821	KF826823	KF826824	KF826826	KF826827
KF826828	KF826830	KF826831	KF826832	KF826833
KF826836	KF826837	KF826838	KF826840	KF826841
KF826846	KF826847	KF826848	KF826849	KF826850
KF826852	KF826854	KF826855	KF826856	KF973319
KF973331	KF973334	KF973335	KF973337	KF973338
KF973339	KF973340	KJ627245	KJ627246	KJ627248
KJ627250	KJ627252	KJ627253	KJ627255	KJ627257
KJ627258	KJ627259	KJ627260	KJ627263	KJ627264
KJ627266	KJ627267	KJ627268	KJ627269	KJ627270
KJ627271	KJ627272	KJ627273	KJ627274	KJ627275
KJ627276	KJ627279	KJ627281	KJ627282	KJ627283
KJ627284	KJ627286	KJ627287	KJ627288	KJ627289
KJ627290	KJ627291	KJ627292	KJ627293	KJ627294
KJ627295	KJ627296	KJ627297	KJ627298	KJ627300
KJ627303	KJ627304	KJ627305	KJ627306	KJ627307
KJ627309	KJ627311	KJ627312	KJ627313	KJ627314
KJ627315	KJ627316	KJ627318	KJ627319	KJ627320
KJ627321	KJ627322	KJ627324	KJ627325	KJ627327
KJ627328	KJ627329	KJ627333	KJ627335	KJ627336
KJ627337	KJ627338	KJ627339	KJ627343	KJ627344
KJ627345	KJ627349	KJ627350	KJ627351	KJ627352
KJ627353	KJ627354	KJ627355	KJ627357	KJ627358

Continued on next page

Table A.8 – continued from previous page

GenBank accession number				
KJ627360	KJ627361	KJ627362	KJ627366	KJ627369
KJ627370	KJ627371	KJ627372	KJ627374	KJ627647
KJ627649	KJ627650	KJ627652	KJ627653	KJ627654
KJ627655	KJ627656	KJ627657	KJ627658	KJ627660
KJ627661	KJ627662	KJ627663	KJ627666	KJ627668
KJ627669	KJ627670	KJ627671	KJ627672	KJ627673
KJ627676	KJ627677	KJ627678	KJ627680	KJ627681
KJ627682	KJ627685	KJ627686	KJ627688	KJ627689
KJ627690	KJ627691	KJ627692	KJ627693	KJ627694
KJ627695	KJ627697	KJ627698	KJ627699	KJ627700
KJ627701	KJ627702	KJ627708	KJ627710	KJ627711
KJ627712	KJ627713	KJ627714	KJ627716	KJ627717
KJ627719	KJ627720	KJ627721	KJ627723	KJ627726
KJ627727	KJ627729	KJ627730	KJ627731	KJ627732
KJ627733	KJ627738	KJ627739	KJ641590	KJ643462
KJ643464	KJ643467	KJ643468	KJ643472	KJ643476
KJ643479	KJ643482	KJ643485	KJ643487	KJ643489
KJ643490	KJ643492	KJ643493	KJ643496	KJ643498
KJ643499	KJ643500	KJ643502	KJ643504	KJ643505
KJ643509	KJ643513	KJ643515	KJ643516	KJ643518
KJ643522	KJ643523	KJ643527	KJ643529	KJ643532
KJ643533	KJ643536	KJ643538	KJ643539	KJ643540
KJ643541	KJ643542	KJ643544	KJ643547	KJ643548
KJ643551	KJ643557	KJ643558	KJ643559	KJ643560
KJ643561	KJ643562	KJ643563	KJ643565	KJ643573
KJ643575	KJ643576	KJ643577	KJ643578	KJ643579
KJ643580	KJ643582	KJ643584	KJ643585	KJ643586
KJ643587	KJ643588	KJ672424	KJ672426	KJ672427
KJ672428	KJ672429	KJ672431	KJ672432	KJ672433
KJ672434	KJ672435	KJ672436	KJ672437	KJ672439
KJ672440	KJ672441	KJ672442	KJ672443	KJ672444
KJ672446	KJ672447	KJ672448	KJ672449	KJ672450
KJ672451	KJ672452	KJ672453	KJ672455	KJ672457
KJ672458	KJ672459	KJ672462	KJ672463	KJ672464
KJ672465	KJ672467	KJ672468	KJ672469	KJ672471
KJ672472	KJ672474	KJ672475	KJ672477	KJ672478
KJ672479	KJ672480	KJ672482	KJ672483	KJ672484
KJ723462	KJ723464	KJ723465	KJ723468	KJ723473
KJ723483	KJ723486	KJ723487	KJ723488	KJ723490
KJ723492	KJ939935	KJ939936	KJ939937	KJ939938
KJ939939	KJ939940	KJ939941	KJ939942	KJ939943
KJ939944	KJ939945	KJ939946	KJ939947	KJ939948

Continued on next page

Table A.8 – continued from previous page

GenBank accession number				
KJ939949	KJ939950	KJ939951	KJ939952	KJ939953
KJ939954	KJ939955	KJ939956	KJ939957	KJ939958
KJ939959	KJ939960	KJ939961	KJ939962	KJ939963
KJ939965	KJ939966	KJ939967	KJ939968	KJ939969
KJ939970	KJ939971	KM042381	KM042382	KM042383
KM042384	KM042385	KM042386	KM042388	KM042389
KM042390	KM042391	KM042392	KM360090	KM517572
KM578843	KP258696	KP258697	KP258699	KP258700
KP258701	KP258703	KP258704	KP258707	KP258709
KP258710	KP258711	KP258715	KP258722	KP258723
KP258725	KP258726	KP258727	KP258728	KP258732
KP258733	KP258734	KP258737	KP258740	KP258743
KP317920	KP317921	KP317926	KP317933	KP317936
KP317942	KP317944	KP317950	KP317951	KP317953
KP317956	KP663728	KP856967	KP856968	KP856969
KT285064	KU316090	KU316091	KU316092	KU316093
KU316096	KU316098	KU316104	KU316110	KU316118
KU316121	KU316125	KU316126	KU316131	KU316133
KU316135	KU316137	KU316138	KU316139	KU316141
KU316142	KU316143	KU316145	KU316148	KU316149
KU316150	KU316157	KU316161	KU316166	KU316167
KU316170	KU316176	KU316178	KU316180	KU839624
KU839626	KU839630	KU839631	KU950456	KU950459
KU950460	KU950464	KU950468	KU950469	KU950470
KU950472	KU950473	KU950474	KU950475	KU950479
KU950480	KU950483	KU950485	KU950486	KU950487
KU950491	KU950492	KU950493	KU950498	KU950499
KU950501	KU950502	KU950505	KU950506	KU950507
KU950509	KU950511	KU950513	KU950518	KU950520
KU950521	KU950523	KU950524	KU950527	KU950528
KU950529	KU950531	KU950536	KU950537	KU950538
KU950540	KU950541	KU950544	KU950545	KU950546
KU950549	KU950550	KU950554	KU950556	KU950557
KU950560	KU950561	KU950564	KU950566	KU950567
KU950570	KU950572	KU950573	KU950579	KU950581
KU950583	KU950585	KU950590	KU950592	KU950595
KU950596	KU950598	KU950608	KU950609	KU950610
KU950612	KU950615	KU950616	KU950617	KU950620
KU950623	KU950624	KU950628	KU950631	KU950632
KU950634	KU950636	KU950638	KU950641	KU950642
KU950643	KU950644	KU950645	KU950646	KU950649
KU950651	KU950652	KU950653	KU950654	KU950655

Continued on next page

Table A.8 – continued from previous page

GenBank accession number				
KU950661	KU950664	KU950666	KU950667	KU950670
KU950671	KU950674	KU950677	KU950680	KU950692
KU950694	KU950696	KU950698	KX510136	KX510137
KX510138	KX510139	KX510140	KX510141	KX510142
KX510143	KX510144	KX510146	KX510148	KX510154
KX510155	KX510157	KX510159	KX510161	KX510169
KX510173	KX510174	KX510175	KX510176	KX510178
KX510180	KX510184	KX510193	KX510194	KX510197
KX510198	KX510201	KX510205	KX510206	KX510208
KX510209	KX510218	KX510219	KX510223	KX510227
KX510228	KX510231	KX510235	KX510240	KX510244
KX510246	KX510247	KX510253	KX510257	KX510260
KX510261	KX510264	KX655615	KX655616	KX655617
KX655618	KX655619	KX655620	KX655621	KX655622
KX655623	KX655626	KX655627	KX655628	KX655630
KX655631	KX655632	KX655634	KX655635	KX655636
KX655637	KX655638	KX655639	KX655640	KX655643
KX655644	KX655647	KX655650	KX655651	KX655652
KX655656	KX655657	KX655658	KX655659	KX655662
KX655663	KX655665	KX655666	KX655667	KX655668
KX655670	KX655673	KX655676	KX655677	KX655678
KX655679	KX655680	KX655682	KX655687	KX655688
KX655689	KX655691	KX655692	KX655695	KX655696
KX655697	KX655699	KX765886	KX765887	KX765891
KX765896	KX765898	KX765901	KX765902	KX765904
KX765910	KX765913	KX765914	KX765915	KX765916
KX765917	KX765918	KX765919	KX765920	KX765925
KX765926	KX765928	KX765932	KX765933	KX765934
KX765937	KX765939	KX765940	KX765941	KX765942
KX765944	KX765946	KX765948	KX765951	KX765953
KX765954	KX765955	KX765956	KX765958	KX765969
KX765970	KX765971	KX765972	KX765974	KX765977
KX894797	KX894805	KY460517	KY654506	KY654507
KY654508	KY654509	KY654510	KY654511	KY654512
KY654513	KY654514	KY654515	KY654516	KY654517
KY654518	KY782635	KY883566	KY883572	KY967364
KY982516	KY982517	LC474556	LC474557	LC474558
LC530050	MF001038	MF001039	MF001041	MF001043
MF001047	MF001049	MF001050	MF001051	MF001052
MF001053	MF001054	MF001056	MF614946	MF614947
MG027862	MG431254	MG431255	MG642024	MG642026
MG642028	MG642030	MG642031	MG642032	MG642033

Continued on next page

Table A.8 – continued from previous page

GenBank accession number				
MG642034	MG642035	MG642040	MG642048	MG642050
MG642052	MG642055	MG642056	MG642058	MG642061
MG642063	MG642070	MG642071	MG642074	MG642077
MG642080	MG642081	MG642083	MG773267	MG773269
MG773271	MG793382	MG813977	MG813978	MG813980
MG813982	MG813983	MG813984	MG813986	MG813987
MG813988	MG813990	MG839543	MG839546	MH181878
MH181879	MH181880	MH181881	MH181882	MH181883
MH181885	MH181886	MH181888	MH181889	MH181890
MH181891	MH181893	MH181894	MH181895	MH181898
MH181899	MH181902	MH181903	MH181904	MH181905
MH181906	MH181907	MH181908	MH181909	MH181910
MH181911	MH181912	MH181913	MH181914	MH181915
MH181916	MH181917	MH181918	MH181919	MH181920
MH181921	MH181922	MH181923	MH181924	MH181925
MH181926	MH181927	MH181928	MH181929	MH181930
MH181931	MH181932	MH181933	MH181934	MH181935
MH181936	MH181937	MH181938	MH181939	MH181940
MH181941	MH181942	MH181943	MH181944	MH181945
MH181946	MH181947	MH181948	MH181949	MH181950
MH181951	MH181952	MH181953	MH181954	MH181955
MH181956	MH181957	MH181958	MH181959	MH181960
MH181961	MH181962	MH181963	MH181964	MH181965
MH181966	MH181967	MH181968	MH181969	MH181970
MH181971	MH181972	MH181973	MH181974	MH181975
MH181976	MH181977	MH181978	MH181979	MH181980
MH181981	MH181982	MH181983	MH181984	MH181985
MH181986	MH181987	MH181988	MH181989	MH181990
MH181991	MH181992	MH181993	MH181994	MH181995
MH181996	MH181997	MH181998	MH181999	MH182000
MH182001	MH182002	MH182003	MH182004	MH182005
MH182006	MH182007	MH182009	MH182010	MH182011
MH182012	MH182013	MH182014	MH182015	MH182016
MH182017	MH182018	MH182019	MH182020	MH182021
MH182023	MH182024	MH182025	MH182026	MH182027
MH182028	MH182029	MH182030	MH182031	MH182032
MH182033	MH182035	MH182036	MH182037	MH182038
MH182039	MH182040	MH182041	MH182042	MH182043
MH182044	MH182045	MH182046	MH182047	MH182048
MH182049	MH182050	MH182051	MH182052	MH182054
MH182055	MH182056	MH182057	MH182058	MH182059
MH182060	MH182061	MH290724	MH447951	MH447952

Continued on next page

Table A.8 – continued from previous page

GenBank accession number				
MH447953	MH447954	MH447955	MH447956	MH447957
MH447958	MH447959	MH447960	MH760588	MH760589
MH760590	MH760591	MH760592	MH760593	MH760594
MH760595	MH760596	MH760597	MH760598	MH760599
MH760600	MH760601	MH760602	MH760603	MH760604
MH760605	MH760606	MH760607	MH760608	MH760609
MH760610	MH760611	MH760612	MH760613	MH760614
MH760615	MH760616	MH760617	MH760618	MH760620
MH760621	MH760622	MH760623	MH760624	MH760625
MH760626	MH760627	MH760628	MH760629	MH760630
MH760631	MH760632	MH760633	MH760634	MH760635
MH760636	MH760637	MH760638	MH760639	MH760640
MH760641	MH760642	MH760643	MH760644	MH760645
MH760646	MH760647	MH760648	MH760649	MH760650
MH760651	MK109763	MK109772	MK109773	MK109774
MK109775	MK109776	MK109777	MK109785	MK109787
MK167036	MK749867	MK749884	MK749886	MK749890
MK749893	MK749895	MK749909	MK749911	MK749912
MK749913	MK749917	MN071211	MN071212	MN071213
MN071214	MN071215	MN071216	MN071217	MN071218
MN071219	MN071220	MN071221	MN071222	MN071223
MN078106	MN078107	MN078108	MN078110	MN078111
MN078113	MN078114	MN078115	MN078116	MN078117
MN078118	MN078119	MN078120	MN078121	MN306017
MN306021	MN306029	MN306030	MN306031	MN306045
MN306048	MN531557	MN630088	MN630089	MN630092
MN630093	MN630095	MN630096	MN630097	MN630098
MN630100	MN630101	MN630102	MN630103	MN630104
MN630105	MN630106	MN630107	MT422269	MT422270
MT422271				

Table A.9: Model selection using marginal likelihood estimation.[a]

Ranking	Clock model	Coalescent prior	PS	SS
1	UCLN	Bayesian skyline	−98885.18	−98941.36
2	Strict clock	Bayesian skyline	−98977.07	−99031.98
3	UCLN	Exponential growth	−99074.69	−99128.25
4	UCLN	Constant size	−99110.04	−99159.15
5	Strict clock	Exponential growth	−99219.48	−99264.95
6	Strict clock	Constant size	−99271.31	−99330.03

[a] PS denotes path sampling; SS, stepping-stone sampling; UCLN, uncorrelated lognormal relaxed clock.

Table A.10: Time and location of the emergence of genotype ON1.

	Dataset 1	**Dataset 2**	**Dataset 3**	**Dataset 4**
Divergence time				
Median	2009.12	2009.18	2009.07	2008.72
95% HPD interval	2008.83 2009.42	2008.93 2009.41	2008.75 2009.36	2008.19 2009.17
Clade posterior probability	1	1	1	1
Location (posterior probability)	Oceania (1)	Oceania (1)	Oceania (1)	Oceania (1)

B
Supplementary figures

Contents

B.1 Supplementary figures for RSV genome and transcriptome sequencing (chapter 3) 152
B.2 Supplementary figures for factors associated with RSV disease severity (chapter 4) 158
B.3 Supplementary figures for RSV within-host diversity (chapter 5) 160
B.4 Supplementary figures for global epidemic history of genotype ON1 (chapter 6) 166

B.1 Supplementary figures for RSV genome and transcriptome sequencing (chapter 3)

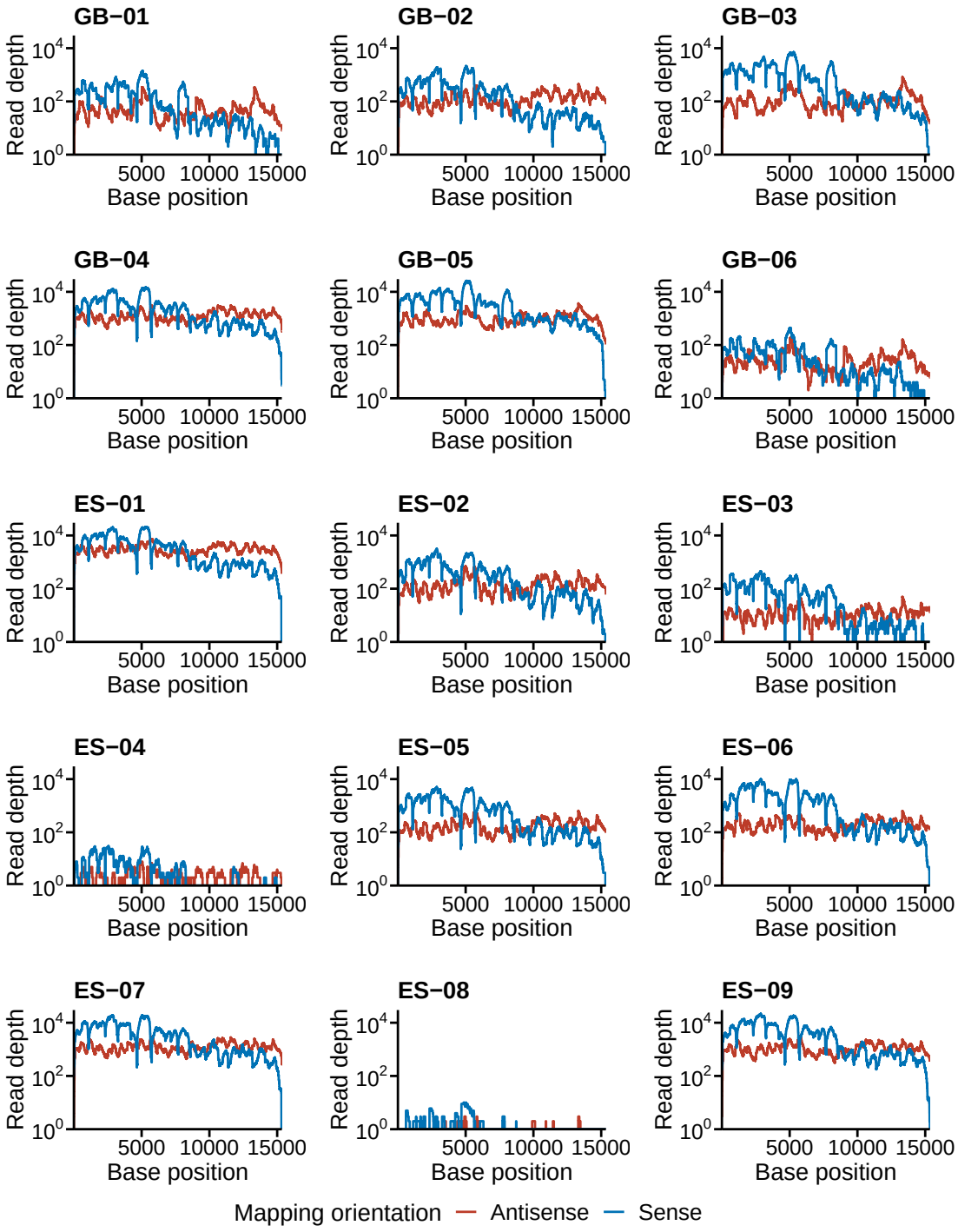


Figure B.1: Coverages of genomic reads (red) and positive-strand reads (blue) for all samples. Among them, ES-04, ES-08, ES-10, and ES-11 failed to generate a complete RSV genome. (continued on next page)

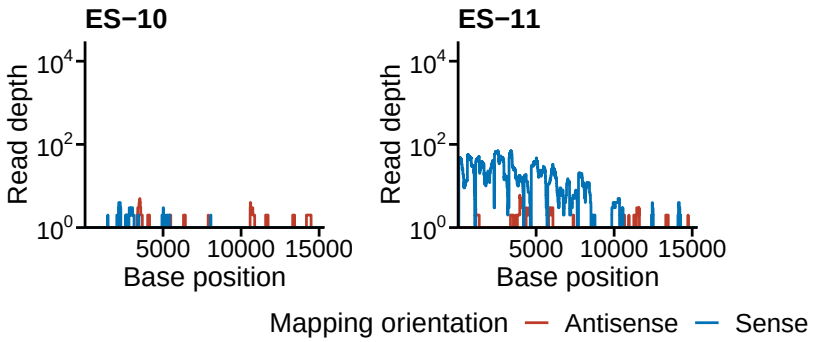


Figure B.1: Coverages of genomic reads and positive-strand reads. (continued)

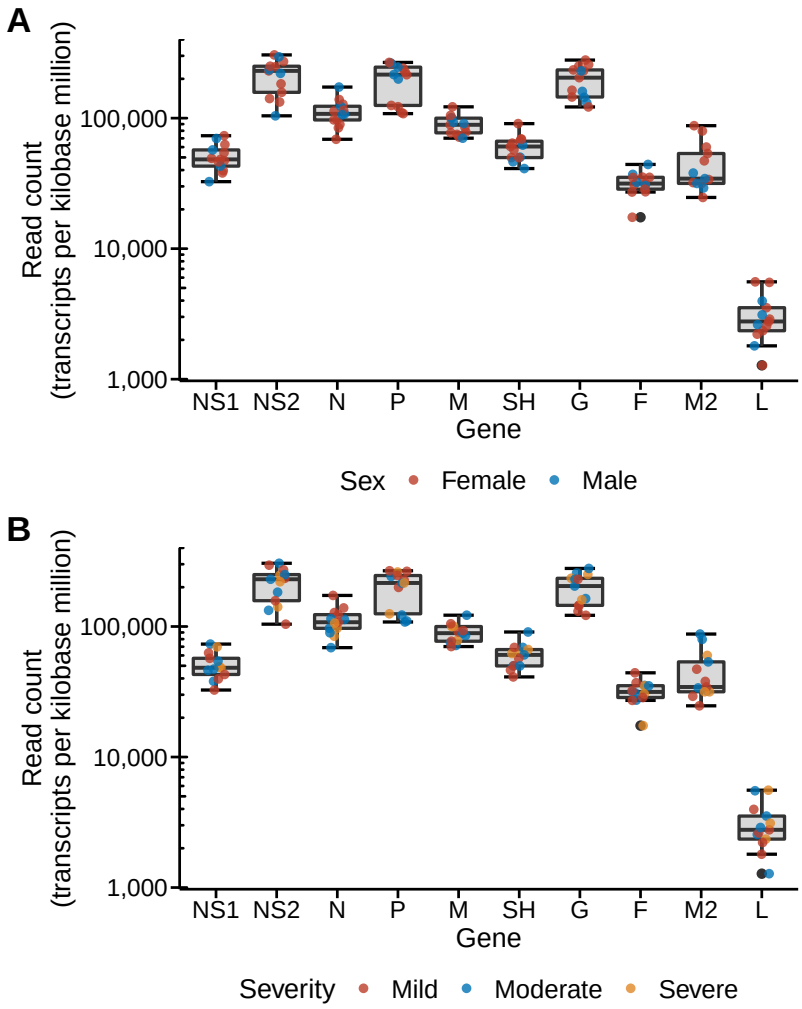


Figure B.2: Relative transcript abundance of RSV genes with different grouping of samples. Total transcript counts for each gene in the 13 samples with complete RSV genome assembly were computed using `htseq-count` and transformed to transcripts per kilobase million RSV-mapped reads (TPM). Each point represents a single sample, coloured by host sex (**A**) or disease severity (**B**). Severity was defined by the ReSVinet scale with a total score of 0–7 representing mild; 8–13 representing moderate; and 14–20 representing severe. The ReSVinet score is determined on the basis of seven clinical parameters, including feeding intolerance, medical intervention, respiratory difficulty, respiratory frequency, apnoea, general condition, and fever. Each parameter ranges from 0 to 2 or 3 points, with the sum of 20 points. RSV genes are displayed from left to right in 3' to 5' genomic order. The centre line of a box denotes the median of the dataset; box limits, first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers.

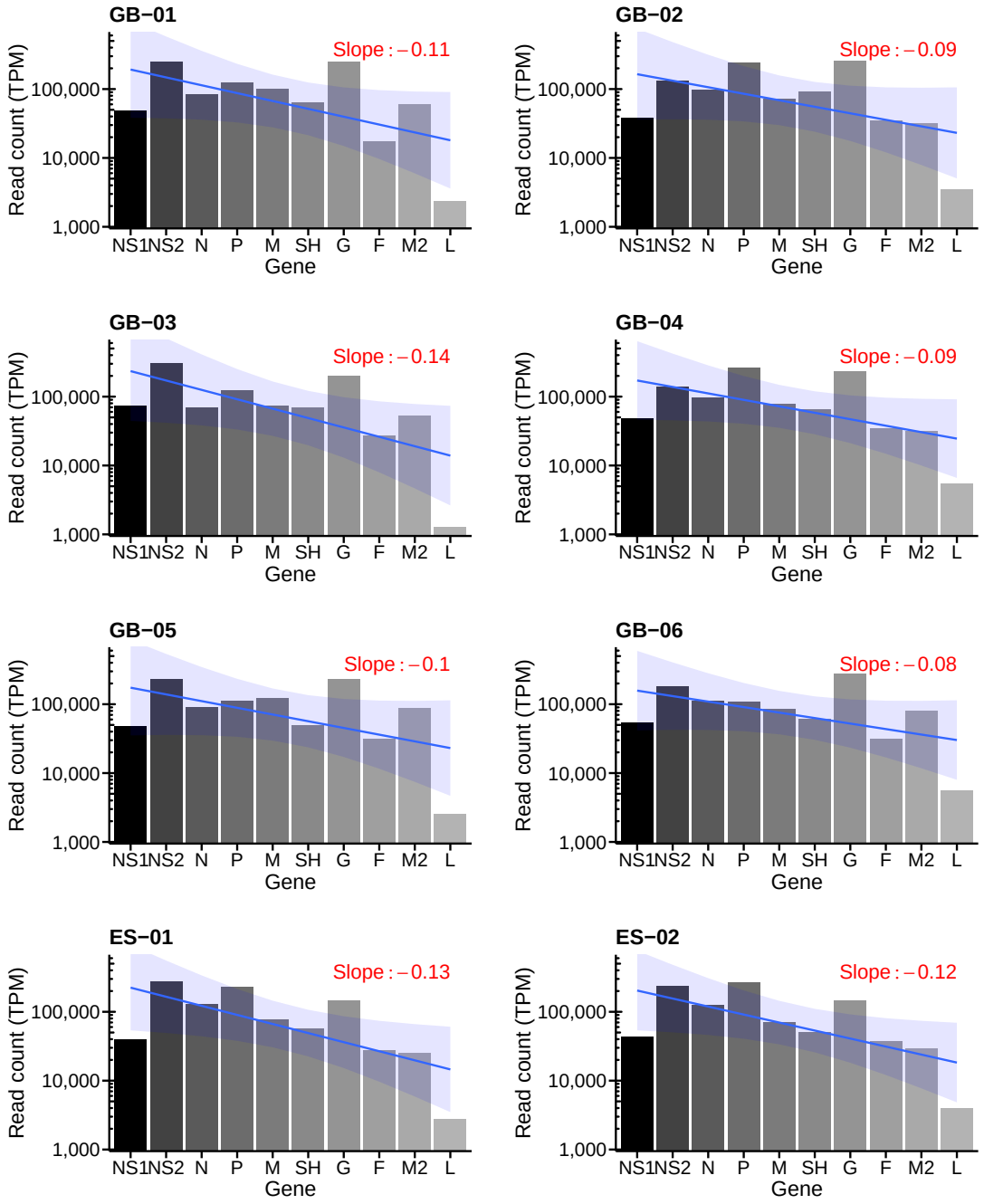


Figure B.3: Relative transcript abundance for each sample. Bar plots show the numbers of transcripts per kilobase million mapped RSV reads (TPM) for the 13 samples with complete RSV genome assembly. The last bar plot shows the expected transcript levels derived using an *in vitro* transcription system[15], with the highest one scaled to the maximum TPM in our dataset. RSV genes are displayed from left to right in 3' to 5' genomic order. The regression lines represent the transcriptional gradients for each sample; the shaded area around a regression line denotes the 95% confidence interval. (continued on next page)

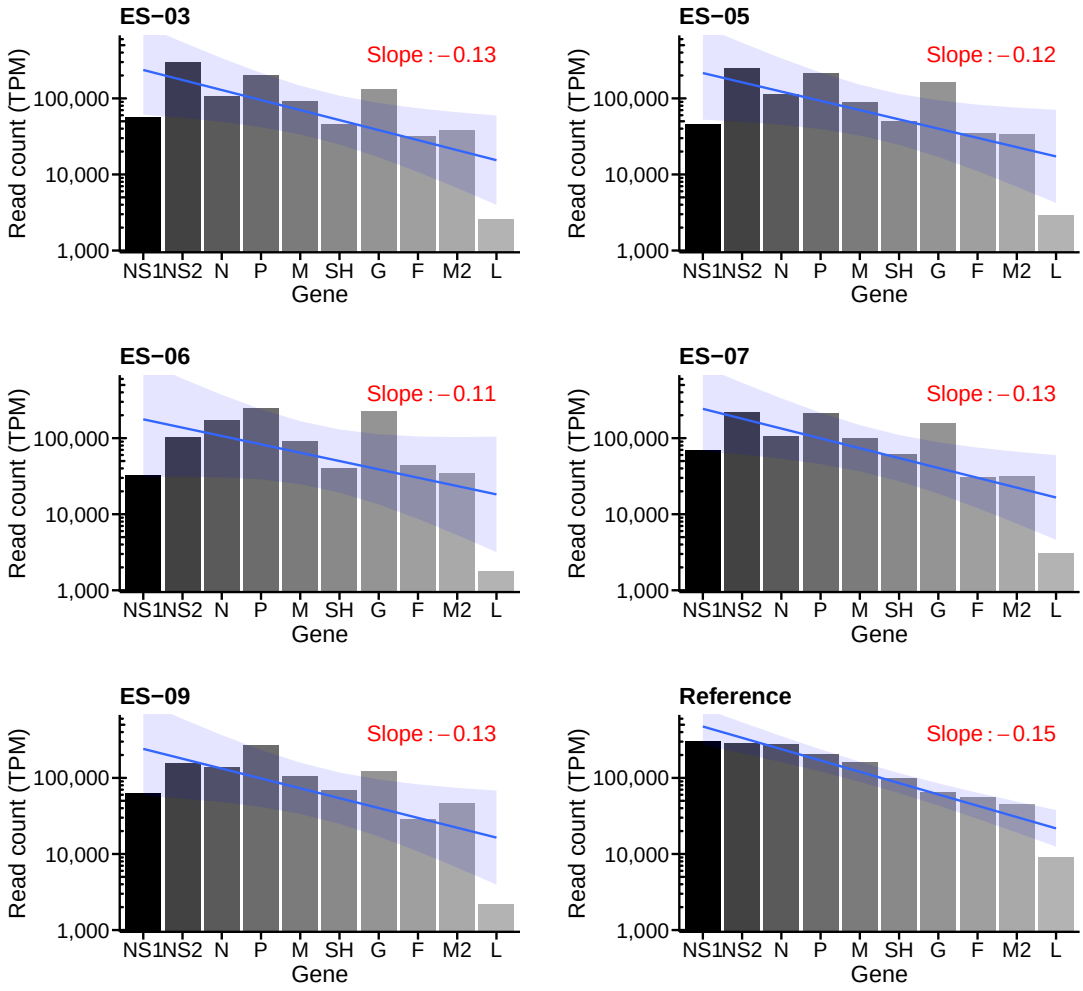


Figure B.3: Relative transcript abundance for each sample. (continued)

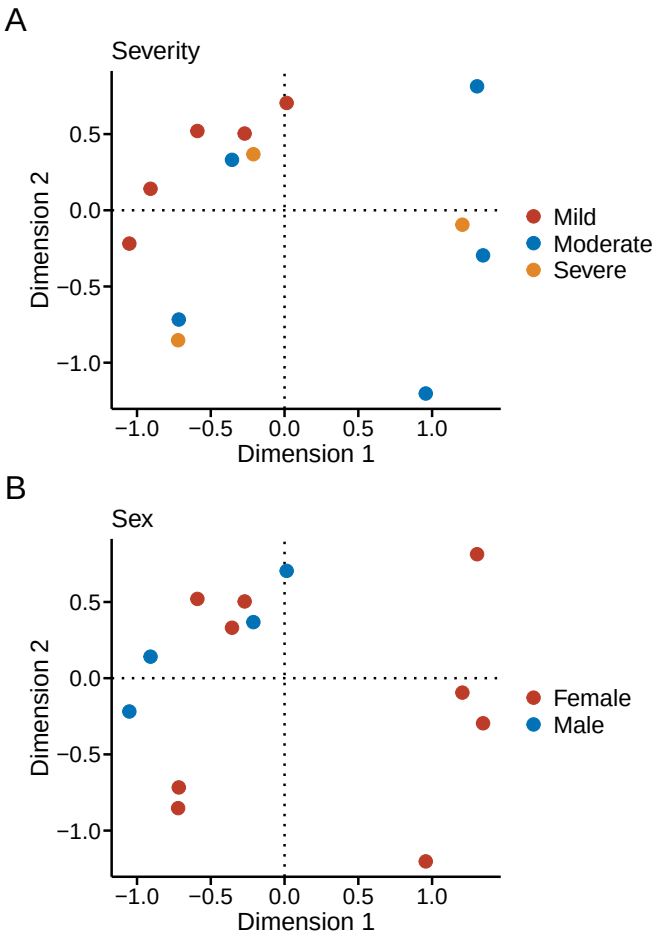


Figure B.4: Gene expression profiles between samples with complete transcriptome assembly. Multidimensional scaling plots demonstrate the relationship between samples on a 2-dimensional scale, where more similar samples tend to cluster together. Each point represents a single sample coloured by disease severity (**A**) and host sex (**B**). Severity was defined by the ReSVinet scale with a total score of 0–7 representing mild; 8–13 representing moderate; and 14–20 representing severe. The ReSVinet scale is determined on the basis of seven clinical parameters, including feeding intolerance, medical intervention, respiratory difficulty, respiratory frequency, apnoea, general condition, and fever.

B.2 Supplementary figures for factors associated with RSV disease severity (chapter 4)

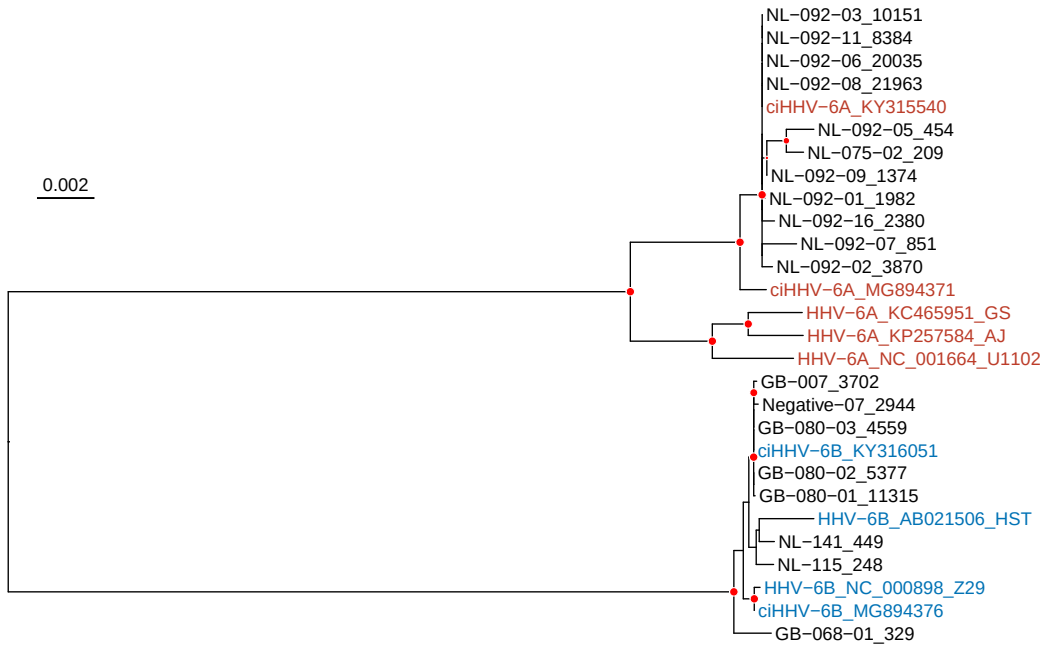


Figure B.5: Phylogenetic reconstruction of human herpesvirus 6 (HHV-6) in samples from this study, contextualised with global HHV-6 sequences. RAxML was used to reconstruct the maximum-likelihood phylogeny with the general time reversible nucleotide substitution model and gamma-distributed rate heterogeneity among sites. Coloured taxa represent reference genomes, labelled as HHV-6 type (ci denotes chromosomally integrated), followed by GenBank accession number, and strain if applicable. Taxa generated from this study were labelled as anonymised patient ID (2 letters followed by 3 digits), order of serial samples if applicable, and total number of HHV-6 reads. Red dots show well-supported nodes with a bootstrap value over 70% and are sized in proportion to bootstrap values. The scale bars represent the number of nucleotide substitutions per site.

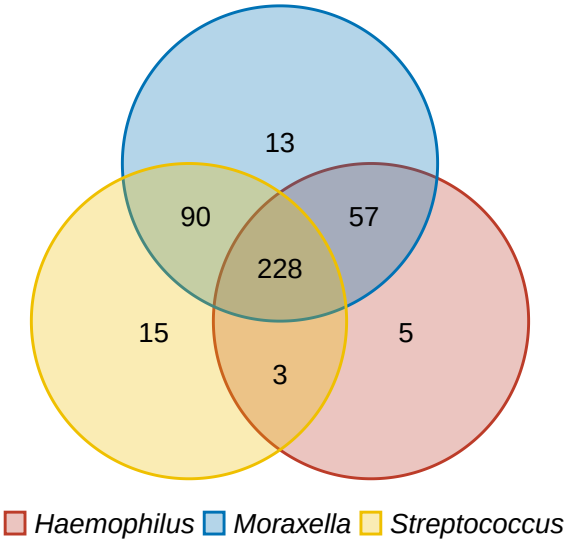


Figure B.6: Venn diagram showing the relationship between the three main bacterial genera found in infants. The number in each area represents the number of infants. 22 infants did not have any of the three bacterial genera.

B.3 Supplementary figures for RSV within-host diversity (chapter 5)

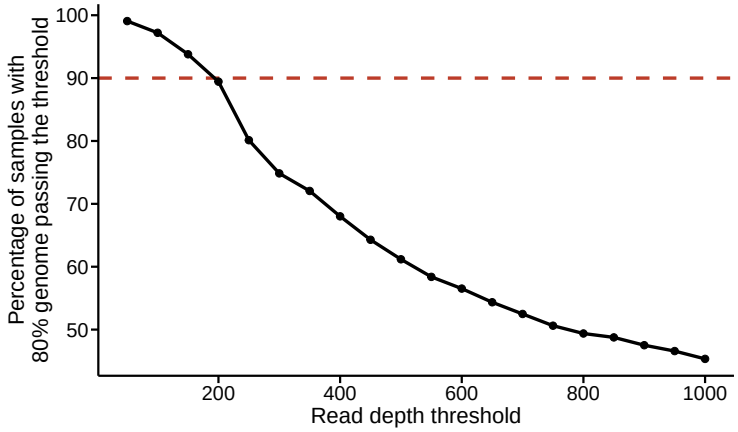


Figure B.7: Percentage of the samples with ≥80% of the genomic positions passing each read depth threshold (from 50 to 1,000). Dashed line shows the predefined criterion of 90% of the samples.

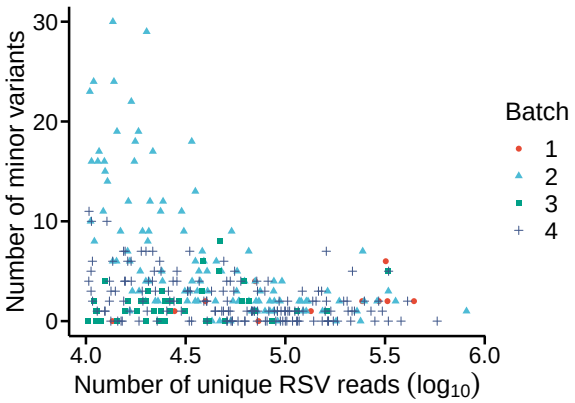


Figure B.8: Correlation between the number of minor variants and the number of unique RSV reads. There were 25 low-burden samples (i.e., those with $\leq 4.5\log_{10}$ uniquely mapped reads) with >10 minor variants in batch 2. Overall, samples in batch 2 had a higher mean unadjusted minor allele frequency (MAF) per sample than batches 3 and 4, as shown in Fig. 5.1, panel A. Therefore, the higher mean MAF per sample in batch 2, together with a greater variance of MAF in low-burden samples, caused a greater number of minor variants in some low-burden samples in batch 2. The samples with the greatest numbers of minor variants in this batch represented both subgroups, with 11 RSV-A and 14 RSV-B, contributing similarly to diversity in both subgroups.

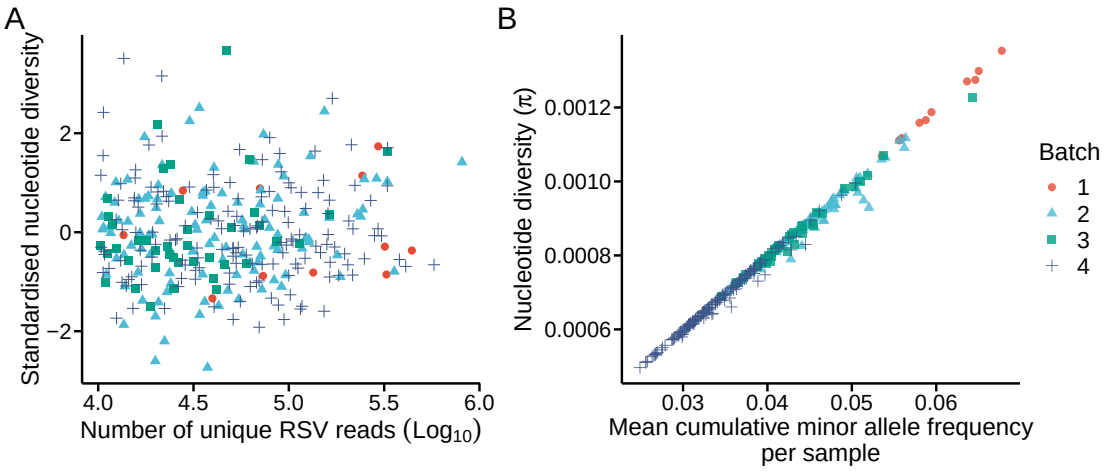


Figure B.9: Correlation between pairwise nucleotide diversity and other variables. **A.** Correlation between z-score standardised pairwise nucleotide diversity and the number of unique RSV reads (Pearson correlation analysis, $r = 0.053$, $P = 0.34$, two-tailed). **B.** Correlation between unadjusted pairwise nucleotide diversity and the mean cumulative minor allele frequency per sample (Pearson correlation analysis, $r = 0.997$, $P < 2.2 \times 10^{-16}$, two-tailed). Each dot represents an individual sample, annotated by sequencing batch. There were a total of 319 samples, with 11, 112, 41, and 155 from each batch respectively.

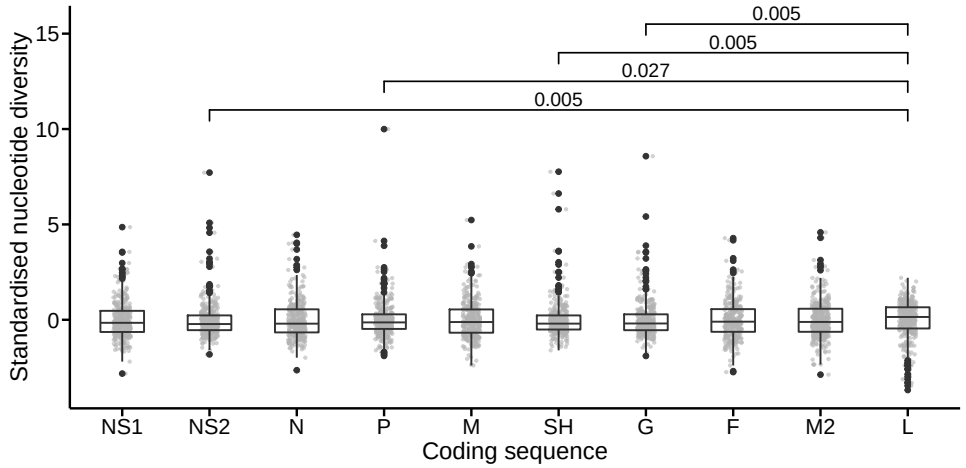


Figure B.10: Gene-wise comparisons of z-score standardised pairwise nucleotide diversity. Each dot represents an individual sample. There were a total of 319 samples. Pairwise two-tailed Mann–Whitney U tests with the Benjamini–Hochberg procedure were used to evaluate the differences between each gene. Adjusted P values of less than 0.05 are shown above the box plots. The centre line of each box denotes the median; box limits, the first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers.

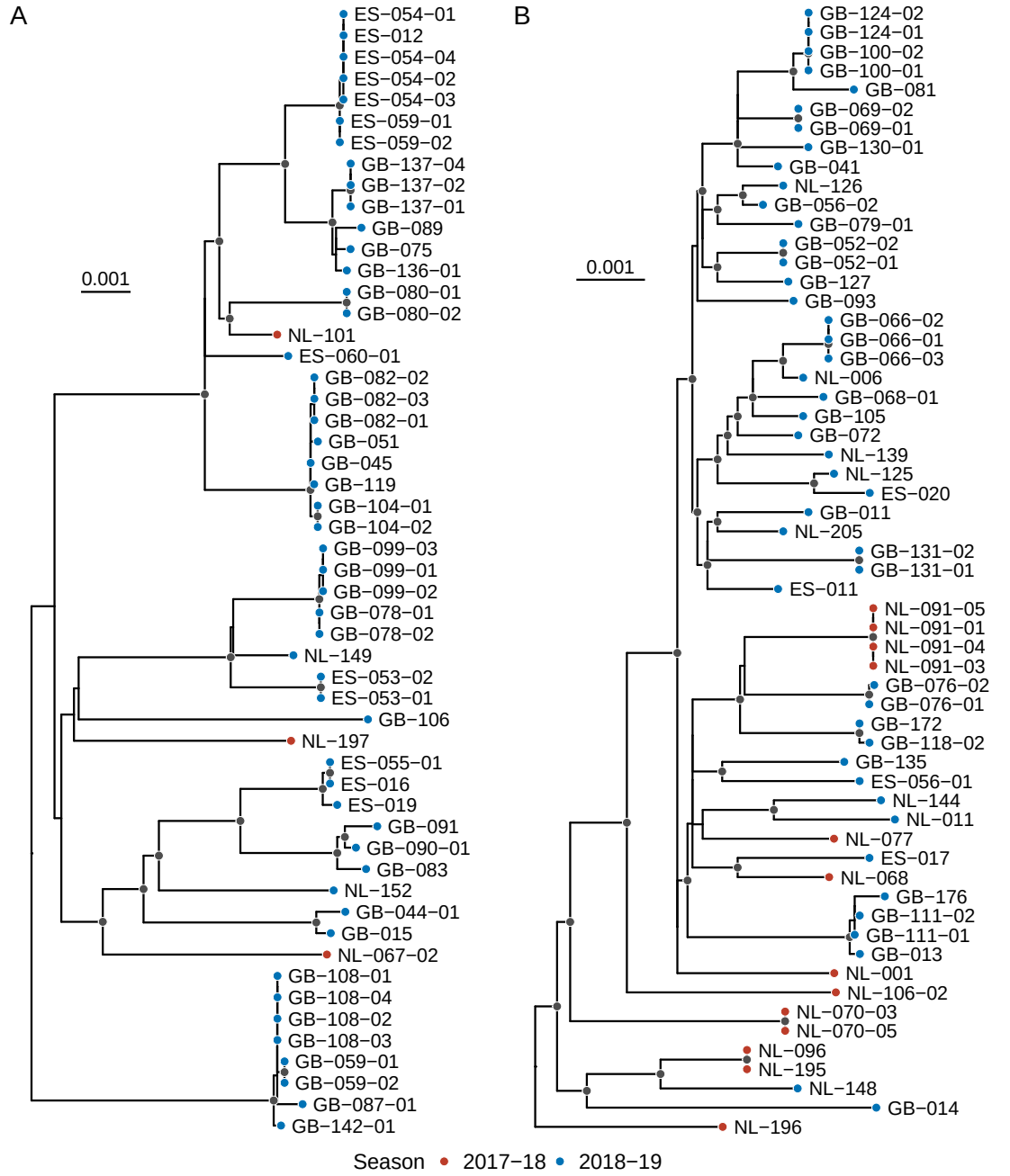


Figure B.11: Maximum-likelihood phylogenies of consensus coding sequences of RSV-A (**A**) and RSV-B (**B**). These phylogenies only included samples from the second sequencing batch (53 RSV-A and 59 RSV-B strains). Taxa are labelled with country (ES, Spain; GB, United Kingdom; and NL, Netherlands), participant ID, and days of hospitalisation if multiple samples were collected. Scale bars show nucleotide substitutions per site. Black dots are well-supported nodes with a bootstrap value of >70%. The phylogenies were rooted to the NCBI reference strains with the accession numbers of NC_038235 and NC_001781 for RSV-A and RSV-B, respectively. The reference strains are not shown here.

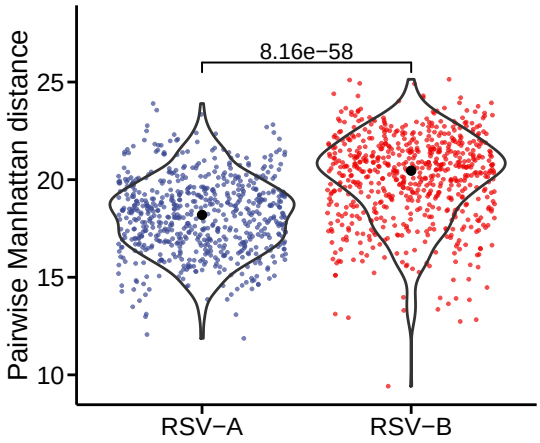


Figure B.12: Manhattan distances between pairs of samples within the RSV-A and RSV-B subgroups. Original allele frequencies were used to calculate Manhattan distances. Only pairwise distances between samples from the second sequencing batch, the same country, the same season, and different participants were included. Each coloured dot is a sample pair (650 RSV-A pairs and 656 RSV-B pairs in total). The violin plots summarise the distribution of the data, and the black dots denote the median value of each group. A two-tailed Mann–Whitney U test was used to evaluate the statistical significance of the difference between the subgroups. P value is shown above the violin plots.

B.4 Supplementary figures for global epidemic history of genotype ON1 (chapter 6)

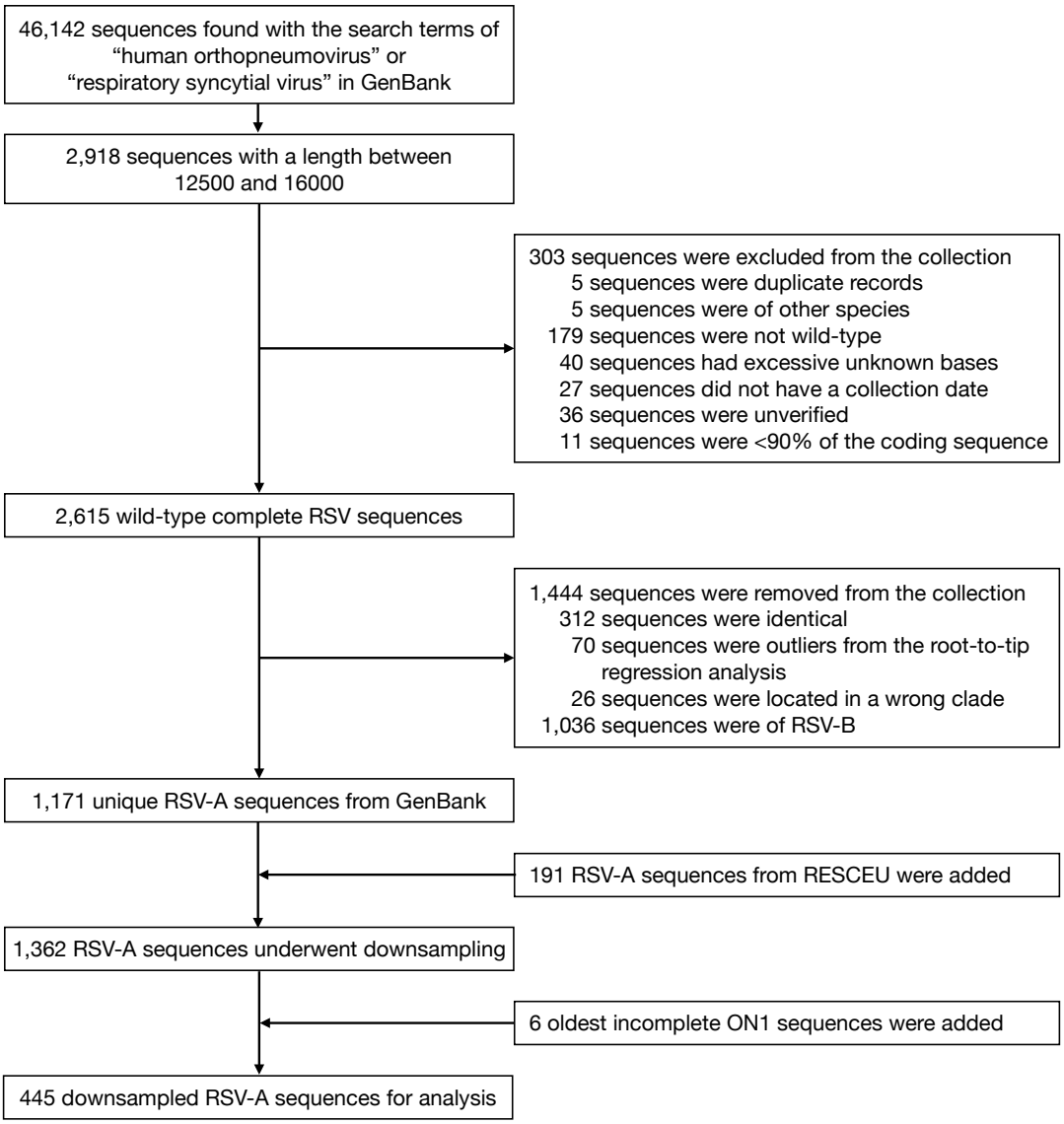


Figure B.13: Flowchart of the sequences included for phylogeographic analysis.

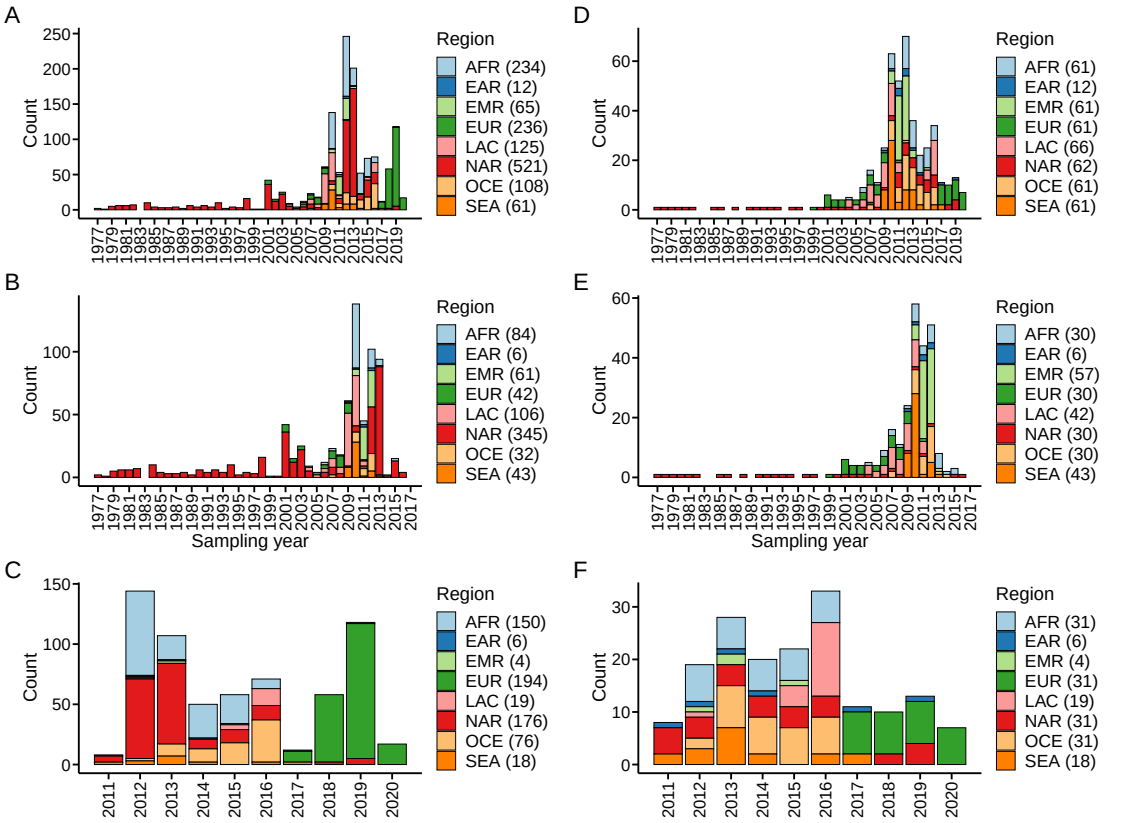


Figure B.14: Geographical distribution and temporal pattern of the sequence collections. **A.** The complete RSV-A collection (n = 1,362) included 719 non-ON1 sequences (**B**) and 643 ON1 sequences (**C**). **D.** Dataset 1 (n = 445) included 268 non-ON1 sequences (**E**), 171 ON1 sequences (**F**), and six additional incomplete ON1 sequences collected in either 2009 or 2010 (not shown in the bar plot). Numbers in brackets in the legends represent the number of sequences for each region. AFR denotes the African region; EAR, the East Asian region; EMR, the Eastern Mediterranean region; EUR, the European region; LAC, the Latin American and the Caribbean region; NAR, the North American region; OCE, the Oceania region; and SEA, the South-East Asian region.

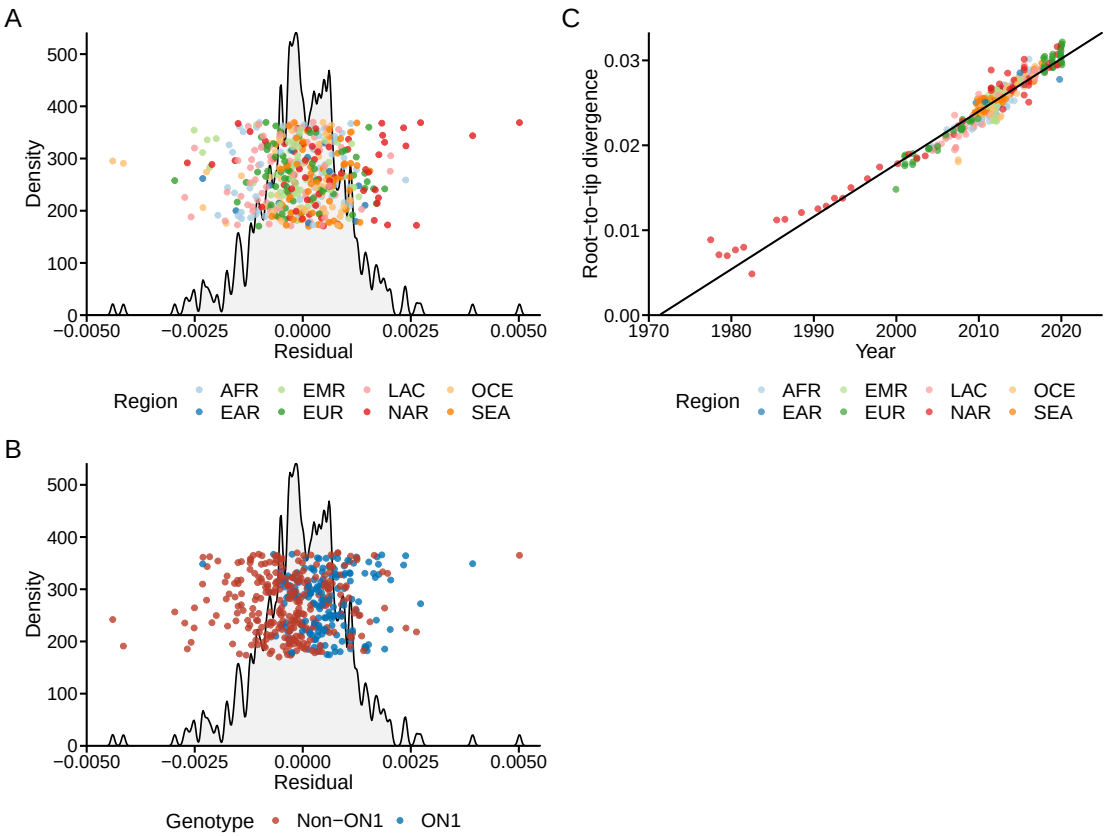


Figure B.15: Regression of root-to-tip divergence on sampling time. Distribution and density of the residuals (**A**, **B**) and the regression line (**C**) for the same regression model as shown in the inset of Fig. 6.2. Each dot is coloured by sampling region (**A**, **C**) or genotype (**B**). AFR denotes the African region; EAR, the East Asian region; EMR, the Eastern Mediterranean region; EUR, the European region; LAC, the Latin American and the Caribbean region; NAR, the North American region; OCE, the Oceania region; and SEA, the South-East Asian region.

C

Supplementary methods and results

Contents

C.1 Supplementary methods for RSV transcriptomics . . . 169
C.2 Preliminary results for RSV transcriptomics 170

C.1 Supplementary methods for RSV transcriptomics

All RSV-positive samples generating enough reads to reconstruct 100% of the genome were included in this preliminary transcriptomic study. As described in section 3.2.2, after RSV genome reconstruction, the pre-deduplicated BAM file from each included sample was divided using samtools into two files: one containing read pairs mapped in the antisense orientation (genomic reads), and the other containing read pairs mapped in the sense orientation (positive-strand reads).

`featureCounts`, a function of the R package Rsubread (v1.22.2)[351], was used to calculate the number of read pairs mapped to each RSV gene in the pre-deduplicated BAM files containing only positive-strand reads. When a read pair was mapped to multiple genes, the gene with the largest number of overlapping bases was the one assigned to the read pair (option: `largestOverlap = TRUE`).

The count data generated from `featureCounts` were normalised by trimmed mean of M values (TMM)[253] and transformed to the $\log_2$ scale as described in section 3.2.4. Multidimensional scaling (MDS) analysis was carried out on the TMM-normalised and $\log_2$-transformed count data to inspect the sample clustering. Differential gene expression analysis between different samples was then performed following the RNAseq123 workflow with the R packages—limma (v3.46.0)[250], Glimma (v2.0.0)[251], and edgeR (v3.32.1)[252]. Fitted linear models were constructed to compare gene expression levels between two RSV subgroups (RSV-A and RSV-B), three severity groups (mild, moderate, and severe), and two host sex groups (male and female). Finally, empirical Bayes moderated t-statistics were applied to compute the significance of differential gene expression with a minimum $\log_2$ fold change of $\log_2 1.2$. P values were corrected using the Benjamini–Hochberg procedure[254]. An adjusted P value of less than 0.05 was considered to indicate statistical significance.

C.2 Preliminary results for RSV transcriptomics

A total of 424 samples generated enough reads to reconstruct the complete RSV genome. Excluding 18 samples with mixed RSV subgroups or missing information on host age or disease severity, 406 samples were included for this preliminary transcriptomic analysis. These samples contained 220 RSV-A and 186 RSV-B. There were 173 (43%) samples collected from female participants. Samples collected from participants with mild, moderate, and severe disease accounted for 49%, 32%, and 20%, respectively.

MDS analysis showed that RSV subgroup, but not host sex or disease severity, was significantly associated with gene expression patterns among these samples (Fig. C.1), consistent with the results in the pilot study (chapter 3). The formal differential gene expression analysis replicated the findings in the pilot study: RSV-A had higher expression of open reading frame (ORF) M2-1 but lower expression of the P gene than RSV-B. The greater power of this preliminary study than that of the pilot study also facilitated the identification of eight other genes/ORFs that were significantly differentially expressed between RSV-A and RSV-B (Fig. C.2).

Only the F gene was not significantly differentially expressed between the two subgroups. Detailed fold changes, 95% confidence intervals, and adjusted P values for all genes/ORFs are shown in Table C.1. No significantly differentially expressed gene or ORF was found between samples grouped by host sex or disease severity.

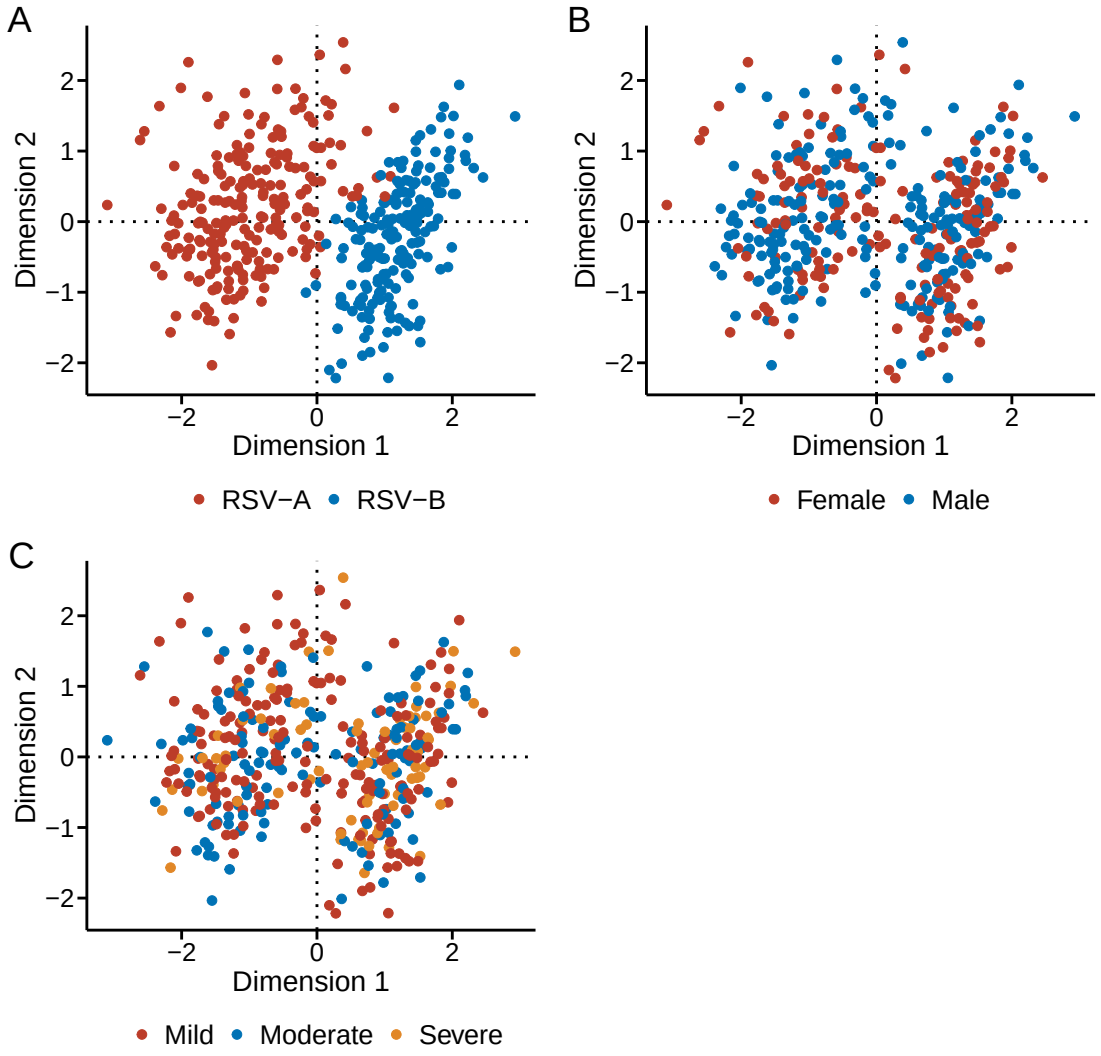


Figure C.1: Multidimensional scaling plots demonstrate the relationship between the 406 samples on a 2-dimensional scale, where more similar samples tend to cluster together. Each point represents a single sample coloured by RSV subgroup (**A**), host sex (**B**), or disease severity (**C**). Severity was defined by the ReSVinet scale with a total score of 0–7 representing mild; 8–13 representing moderate; and 14–20 representing severe. The ReSVinet scale was determined on the basis of seven clinical parameters: feeding intolerance, medical intervention, respiratory difficulty, respiratory frequency, apnoea, general condition, and fever (Table 2.4).

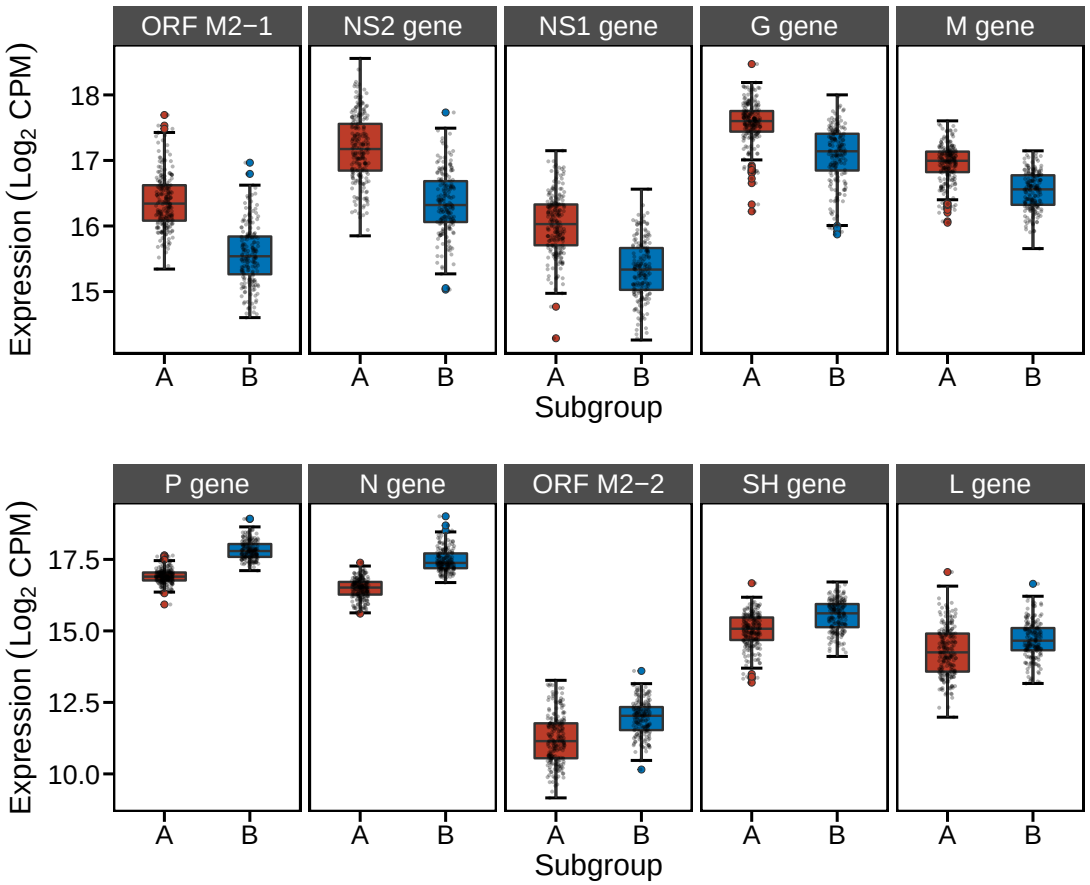


Figure C.2: Differential expression of eight RSV genes and two open reading frames (ORFs). RSV-A (N = 220) had higher expression of ORF M2-1, the NS2, NS1, G, and M genes, but lower expression of the P, N, SH, L genes and ORF M2-2 than RSV-B (N = 186). The fold changes, 95% confidence intervals, and adjusted P values are shown in Table C.1. Each grey dot represents a single sample. Centre line of each box denotes the median; box limits, first and third quartiles; whiskers, the highest and lowest values within 1.5 times the interquartile range from the box limits; and outlying points, outliers. CPM denotes counts per million.

Table C.1: Differential expression of RSV genes and open reading frames (N = 406).

	Fold change[a]	**95% CI**[b]	**Average expression**[c]	**Adjusted P value**
M2-1	1.69	1.60, 1.79	16.00	7.6×10^{-28}
NS2	1.71	1.61, 1.83	16.80	9.9×10^{-24}
NS1	1.60	1.50, 1.70	15.71	4.5×10^{-18}
G	1.39	1.32, 1.47	17.35	1.2×10^{-8}
M	1.29	1.24, 1.34	16.76	2.9×10^{-4}
P	−1.91	−1.84, −1.99	17.33	2.4×10^{-73}
N	−2.06	−1.96, −2.17	16.93	9.6×10^{-67}
M2-2	−1.90	−1.72, −2.11	11.56	3.7×10^{-17}
SH	−1.43	−1.33, −1.54	15.26	3.4×10^{-6}
L	−1.32	−1.20, −1.47	14.47	0.033
F	1.19	1.13, 1.25	16.17	0.599

[a] Fold change shows the ratio of the amount of RSV-A to that of RSV-B.
[b] CI denotes confidence interval.
[c] Average expression shows the $\log_2$ mean counts per million of all samples.

References

1. Lin GL, Golubchik T, Drysdale S, et al. Simultaneous viral whole-genome sequencing and differential expression profiling in respiratory syncytial virus infection of infants. J Infect Dis 2020;222:S666–S671.
2. Lin GL, McGinley JP, Drysdale SB, and Pollard AJ. Epidemiology and immune pathogenesis of viral sepsis. Front Immunol 2018;9:2147.
3. Oner D, Drysdale SB, McPherson C, et al. Biomarkers for disease severity in children infected with respiratory syncytial virus: A systematic literature review. J Infect Dis 2020;222:S648–S657.
4. Jefferies K, Drysdale SB, Robinson H, et al. Presumed risk factors and biomarkers for severe respiratory syncytial virus disease and related sequelae: protocol for an observational multicenter, case-control study from the respiratory syncytial virus consortium in Europe (RESCEU). J Infect Dis 2020;222:S658–S665.
5. Rima B, Collins P, Easton A, et al. ICTV virus taxonomy profile: Pneumoviridae. J Gen Virol 2017;98:2912–3.
6. Kuo L, Grosfeld H, Cristina J, Hill MG, and Collins PL. Effects of mutations in the gene-start and gene-end sequence motifs on transcription of monocistronic and dicistronic minigenomes of respiratory syncytial virus. J Virol 1996;70:6892–901.
7. Collins PL, Olmsted RA, Spriggs MK, Johnson PR, and Buckler-White AJ. Gene overlap and site-specific attenuation of transcription of the viral polymerase L gene of human respiratory syncytial virus. Proc Natl Acad Sci U S A 1987;84:5134–8.
8. Collins PL, Hill MG, and Johnson PR. The two open reading frames of the 22K mRNA of human respiratory syncytial virus: sequence comparison of antigenic subgroups A and B and expression in vitro. J Gen Virol 1990;71:3015–20.
9. Battles MB and McLellan JS. Respiratory syncytial virus entry and how to block it. Nat Rev Microbiol 2019;17:233–45.
10. Morin B, Kranzusch PJ, Rahmeh AA, and Whelan SP. The polymerase of negative-stranded RNA viruses. Curr Opin Virol 2013;3:103–10.
11. Bermingham A and Collins PL. The M2-2 protein of human respiratory syncytial virus is a regulatory factor involved in the balance between RNA replication and transcription. Proc Natl Acad Sci U S A 1999;96:11259–64.
12. Collins PL, Hill MG, Camargo E, Grosfeld H, Chanock RM, and Murphy BR. Production of infectious human respiratory syncytial virus from cloned cDNA confirms an essential role for the transcription elongation factor from the 5' proximal open reading frame of the M2 mRNA in gene expression and provides a capability for vaccine development. Proc Natl Acad Sci U S A 1995;92:11563–7.

13. Collins PL, Hill MG, Cristina J, and Grosfeld H. Transcription elongation factor of respiratory syncytial virus, a nonsegmented negative-strand RNA virus. Proc Natl Acad Sci U S A 1996;93:81–5.

14. Collins PL, Fearns R, and Graham BS. Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease. Curr Top Microbiol Immunol 2013;372:3–38.

15. Barik S. Transcription of human respiratory syncytial virus genome RNA in vitro: requirement of cellular factor(s). J Virol 1992;66:6813–8.

16. Hastie ML, Headlam MJ, Patel NB, et al. The human respiratory syncytial virus nonstructural protein 1 regulates type I and type II interferon pathways. Mol Cell Proteomics 2012;11:108–27.

17. Levine S, Klaiber-Franco R, and Paradiso PR. Demonstration that glycoprotein G is the attachment protein of respiratory syncytial virus. J Gen Virol 1987;68:2521–4.

18. Graham BS, Modjarrad K, and McLellan JS. Novel antigens for RSV vaccines. Curr Opin Immunol 2015;35:30–8.

19. Cane PA. Molecular epidemiology of respiratory syncytial virus. Rev Med Virol 2001;11:103–16.

20. Wertz GW, Collins PL, Huang Y, Gruber C, Levine S, and Ball LA. Nucleotide sequence of the G protein gene of human respiratory syncytial virus reveals an unusual type of viral membrane protein. Proc Natl Acad Sci U S A 1985;82:4075–9.

21. Fedechkin SO, George NL, Wolff JT, Kauvar LM, and DuBois RM. Structures of respiratory syncytial virus G antigen bound to broadly neutralizing antibodies. Sci Immunol 2018;3:eaar3534.

22. Lee YN, Suk Hwang H, Kim MC, et al. Recombinant influenza virus carrying the conserved domain of respiratory syncytial virus (RSV) G protein confers protection against RSV without inflammatory disease. Virology 2015;476:217–25.

23. Jorquera PA, Choi Y, Oakley KE, et al. Nanoparticle vaccines encompassing the respiratory syncytial virus (RSV) G protein CX3C chemokine motif induce robust immunity protecting from challenge and disease. PLoS One 2013;8:e74905.

24. Power UF, Nguyen TN, Rietveld E, et al. Safety and immunogenicity of a novel recombinant subunit respiratory syncytial virus vaccine (BBG2Na) in healthy young adults. J Infect Dis 2001;184:1456–60.

25. Fuentes S, Coyle EM, Beeler J, Golding H, and Khurana S. Antigenic fingerprinting following primary RSV infection in young children identifies novel antigenic sites and reveals unlinked evolution of human antibody repertoires to fusion and attachment glycoproteins. PLoS Pathog 2016;12:e1005554.

26. Lee J, Klenow L, Coyle EM, Golding H, and Khurana S. Protective antigenic sites in respiratory syncytial virus G attachment protein outside the central conserved and cysteine noose domains. PLoS Pathog 2018;14:e1007262.

27. McLellan JS, Ray WC, and Peeples ME. Structure and function of respiratory syncytial virus surface glycoproteins. Curr Top Microbiol Immunol 2013;372:83–104.

28. McLellan JS. Neutralizing epitopes on the respiratory syncytial virus fusion glycoprotein. Curr Opin Virol 2015;11:70–5.

29. Lopez JA, Bustos R, Orvell C, et al. Antigenic structure of human respiratory syncytial virus fusion glycoprotein. J Virol 1998;72:6922–8.

30. Magro M, Mas V, Chappell K, et al. Neutralizing antibodies against the preactive form of respiratory syncytial virus fusion protein offer unique possibilities for clinical intervention. Proc Natl Acad Sci U S A 2012;109:3089–94.

31. Ruckwardt TJ, Morabito KM, Phung E, et al. Safety, tolerability, and immunogenicity of the respiratory syncytial virus prefusion F subunit vaccine DS-Cav1: a phase 1, randomised, open-label, dose-escalation clinical trial. Lancet Respir Med 2021.

32. McLellan JS, Chen M, Joyce MG, et al. Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. Science 2013;342:592–8.

33. Griffin MP, Yuan Y, Takas T, et al. Single-dose nirsevimab for prevention of RSV in preterm infants. N Engl J Med 2020;383:415–25.

34. Graham BS. Vaccine development for respiratory syncytial virus. Curr Opin Virol 2017;23:107–12.

35. Glezen WP, Taber LH, Frank AL, and Kasel JA. Risk of primary infection and reinfection with respiratory syncytial virus. Am J Dis Child 1986;140:543–6.

36. Shi T, McAllister DA, O'Brien KL, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. Lancet 2017;390:946–58.

37. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380:2095–128.

38. Varga SM and Braciale TJ. The adaptive immune response to respiratory syncytial virus. Curr Top Microbiol Immunol 2013;372:155–71.

39. Munywoki PK, Koech DC, Agoti CN, et al. The source of respiratory syncytial virus infection in infants: a household cohort study in rural Kenya. J Infect Dis 2014;209:1685–92.

40. Falsey AR, Hennessey PA, Formica MA, Cox C, and Walsh EE. Respiratory syncytial virus infection in elderly and high-risk adults. N Engl J Med 2005;352:1749–59.

41. Shi T, Denouel A, Tietjen AK, et al. Global disease burden estimates of respiratory syncytial virus-associated acute respiratory infection in older adults in 2015: a systematic review and meta-analysis. J Infect Dis 2020;222:S577–s583.

42. Jackson ML, Scott E, Kuypers J, Nalla AK, Roychoudury P, and Chu HY. Epidemiology of respiratory syncytial virus across five influenza seasons among adults and children one year of age and older—Washington State, 2011/2012–2015/2016. J Infect Dis 2021;223:147–56.

43. Li Y, Reeves RM, Wang X, et al. Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: a systematic analysis. Lancet Glob Health 2019;7:e1031–e1045.

44. Bloom-Feshbach K, Alonso WJ, Charu V, et al. Latitudinal variations in seasonal activity of influenza and respiratory syncytial virus (RSV): a global comparative review. PLoS One 2013;8:e54445.

45. Zamberi S, Zulkifli I, and Ilina I. Respiratory viruses detected in hospitalised paediatric patients with respiratory infections. Med J Malaysia 2003;58:681–7.

46. Lee WK and Young BW. Infectious diseases in children admitted from a residential child care centre. Hong Kong Med J 2006;12:119–24.

47. Waris M. Pattern of respiratory syncytial virus epidemics in Finland: two-year cycles with alternating prevalence of groups A and B. J Infect Dis 1991;163:464–9.

48. Eriksson M, Bennet R, Rotzén-Ostlund M, Sydow M von, and Wirgart BZ. Population-based rates of severe respiratory syncytial virus infection in children with and without risk factors, and outcome in a tertiary care setting. Acta Paediatr 2002;91:593–8.

49. Tsai HP, Kuo PH, Liu CC, and Wang JR. Respiratory viral infections among pediatric inpatients and outpatients in Taiwan from 1997 to 1999. J Clin Microbiol 2001;39:111–8.

50. Tang JW and Loh TP. Correlations between climate factors and incidence—a contributor to RSV seasonality. Rev Med Virol 2014;24:15–34.

51. Public Health England. Weekly national flu reports. 2021. URL: `https://www.gov.uk/government/collections/weekly-national-flu-reports` (visited on Oct. 8, 2020).

52. Viguria N, Martinez-Baz I, Moreno-Galarraga L, Sierrasesumaga L, Salcedo B, and Castilla J. Respiratory syncytial virus hospitalization in children in northern Spain. PLoS One 2018;13:e0206474.

53. Vos LM, Teirlinck AC, Lozano JE, et al. Use of the moving epidemic method (MEM) to assess national surveillance data for respiratory syncytial virus (RSV) in the Netherlands, 2005 to 2017. Euro Surveill 2019;24:1800469.

54. Kuitunen I and Renko M. Lessons to learn from the current pandemic for future non-pharmaceutical interventions against the respiratory syncytial virus—nationwide register-study in Finland. Infect Dis (Lond) 2021;53:476–8.

55. Baker RE, Park SW, Yang W, Vecchi GA, Metcalf CJE, and Grenfell BT. The impact of COVID-19 nonpharmaceutical interventions on the future dynamics of endemic infections. Proc Natl Acad Sci U S A 2020;117:30547–53.

56. Peret TC, Hall CB, Schnabel KC, Golub JA, and Anderson LJ. Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. J Gen Virol 1998;79:2221–9.

57. Venter M, Madhi SA, Tiemessen CT, and Schoub BD. Genetic diversity and molecular epidemiology of respiratory syncytial virus over four consecutive seasons in South Africa: identification of new subgroup A and B genotypes. J Gen Virol 2001;82:2117–24.

58. Peret TC, Hall CB, Hammond GW, et al. Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America. J Infect Dis 2000;181:1891–6.

59. Shobugawa Y, Saito R, Sano Y, et al. Emerging genotypes of human respiratory syncytial virus subgroup A among patients in Japan. J Clin Microbiol 2009;47:2475–82.

60. Eshaghi A, Duvvuri VR, Lai R, et al. Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. PLoS One 2012;7:e32807.

61. Trento A, Viegas M, Galiano M, et al. Natural history of human respiratory syncytial virus inferred from phylogenetic analysis of the attachment (G) glycoprotein with a 60-nucleotide duplication. J Virol 2006;80:975–84.

62. Dapat IC, Shobugawa Y, Sano Y, et al. New genotypes within respiratory syncytial virus group B genotype BA in Niigata, Japan. J Clin Microbiol 2010;48:3423–7.

63. Munoz-Escalante JC, Comas-Garcia A, Bernal-Silva S, Robles-Espinoza CD, Gomez-Leal G, and Noyola DE. Respiratory syncytial virus A genotype classification based on systematic intergenotypic and intragenotypic sequence analysis. Sci Rep 2019;9:20097.

64. Ramaekers K, Rector A, Cuypers L, Lemey P, Keyaerts E, and Van Ranst M. Towards a unified classification for human respiratory syncytial virus genotypes. Virus Evol 2020;6:veaa052.

65. Agoti CN, Otieno JR, Munywoki PK, et al. Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. J Virol 2015;89:3444–54.

66. Bose ME, He J, Shrivastava S, et al. Sequencing and analysis of globally obtained human respiratory syncytial virus A and B genomes. PLoS One 2015;10:e0120098.

67. Arbiza J, Delfraro A, and Frabasile S. Molecular epidemiology of human respiratory syncytial virus in Uruguay: 1985–2001—a review. Mem Inst Oswaldo Cruz 2005;100:221–30.

68. Zlateva KT, Vijgen L, Dekeersmaeker N, Naranjo C, and Van Ranst M. Subgroup prevalence and genotype circulation patterns of human respiratory syncytial virus in Belgium during ten successive epidemic seasons. J Clin Microbiol 2007;45:3022–30.

69. Choi EH and Lee HJ. Genetic diversity and molecular epidemiology of the G protein of subgroups A and B of respiratory syncytial viruses isolated over 9 consecutive epidemics in Korea. J Infect Dis 2000;181:1547–56.

70. White LJ, Waris M, Cane PA, Nokes DJ, and Medley GF. The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. Epidemiol Infect 2005;133:279–89.

71. Cane PA and Pringle CR. Evolution of subgroup A respiratory syncytial virus: evidence for progressive accumulation of amino acid changes in the attachment protein. J Virol 1995;69:2918–25.

72. Trento A, Galiano M, Videla C, et al. Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. J Gen Virol 2003;84:3115–20.

73. Moore ML, Stokes KL, and Hartert TV. The impact of viral genotype on pathogenesis and disease severity: respiratory syncytial virus and human rhinoviruses. Curr Opin Immunol 2013;25:761–8.

74. Tan L, Lemey P, Houspie L, et al. Genetic variability among complete human respiratory syncytial virus subgroup A genomes: bridging molecular evolutionary dynamics and epidemiology. PLoS One 2012;7:e51439.

75. Cane PA, Matthews DA, and Pringle CR. Analysis of relatedness of subgroup A respiratory syncytial viruses isolated worldwide. Virus Res 1992;25:15–22.

76. Garcia O, Martin M, Dopazo J, et al. Evolutionary pattern of human respiratory syncytial virus (subgroup A): cocirculating lineages and correlation of genetic and antigenic changes in the G glycoprotein. J Virol 1994;68:5448–59.

77. Zou L, Yi L, Wu J, et al. Evolution and transmission of respiratory syncytial group A (RSV-A) viruses in Guangdong, China 2008–2015. Front Microbiol 2016;7:1263.

78. Cui G, Zhu R, Deng J, et al. Rapid replacement of prevailing genotype of human respiratory syncytial virus by genotype ON1 in Beijing, 2012–2014. Infect Genet Evol 2015;33:163–8.

79. Song J, Wang H, Shi J, et al. Emergence of BA9 genotype of human respiratory syncytial virus subgroup B in China from 2006 to 2014. Sci Rep 2017;7:16765.

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

81. Rebuffo-Scheer C, Bose M, He J, et al. Whole genome sequencing and evolutionary analysis of human respiratory syncytial virus A and B from Milwaukee, WI 1998-2010. PLoS One 2011;6:e25468.

82. Tan L, Coenjaerts FE, Houspie L, et al. The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. J Virol 2013;87:8213–26.

83. Agoti CN, Otieno JR, Gitahi CW, Cane PA, and Nokes DJ. Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya. Emerg Infect Dis 2014;20:950–9.

84. Rudd PM, Elliott T, Cresswell P, Wilson IA, and Dwek RA. Glycosylation and the immune system. Science 2001;291:2370–6.

85. Otieno JR, Kamau EM, Oketch JW, et al. Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. Virus Evol 2018;4:vey027.

86. Schobel SA, Stucker KM, Moore ML, et al. Respiratory syncytial virus whole-genome sequencing identifies convergent evolution of sequence duplication in the C-terminus of the G gene. Sci Rep 2016;6:26311.

87. Trovao NS, Khuri-Bulos N, Tan Y, et al. Molecular characterization of respiratory syncytial viruses circulating in a paediatric cohort in Amman, Jordan. Microb Genom 2019.

88. Borchers AT, Chang C, Gershwin ME, and Gershwin LJ. Respiratory syncytial virus—a comprehensive review. Clin Rev Allergy Immunol 2013;45:331–79.

89. Chartrand C, Tremblay N, Renaud C, and Papenburg J. Diagnostic accuracy of rapid antigen detection tests for respiratory syncytial virus infection: Systematic review and meta-analysis. J Clin Microbiol 2015;53:3738–49.

90. Zuurbier RP, Bont LJ, Langedijk AC, et al. Low sensitivity of BinaxNOW RSV in infants. J Infect Dis 2020;222:S640–S647.

91. Casiano-Colon AE, Hulbert BB, Mayer TK, Walsh EE, and Falsey AR. Lack of sensitivity of rapid antigen tests for the diagnosis of respiratory syncytial virus infection in adults. J Clin Virol 2003;28:169–74.

92. Deming DJ, Patel N, McCarthy MP, Mishra L, Shapiro AM, and Suzich JA. Potential for palivizumab interference with commercially available antibody-antigen based respiratory syncytial virus diagnostic assays. Pediatr Infect Dis J 2013;32:1144–6.

93. Paton AW, Paton JC, Lawrence AJ, Goldwater PN, and Harris RJ. Rapid detection of respiratory syncytial virus in nasopharyngeal aspirates by reverse transcription and polymerase chain reaction amplification. J Clin Microbiol 1992;30:901–4.

94. Poritz MA, Blaschke AJ, Byington CL, et al. FilmArray, an automated nested multiplex PCR system for multi-pathogen detection: development and application to respiratory tract infection. PLoS One 2011;6:e26047.

95. Hall CB, Weinberg GA, Iwane MK, et al. The burden of respiratory syncytial virus infection in young children. N Engl J Med 2009;360:588–98.

96. Alvarez AE, Marson FA, Bertuzzo CS, Arns CW, and Ribeiro JD. Epidemiological and genetic characteristics associated with the severity of acute viral bronchiolitis by respiratory syncytial virus. J Pediatr (Rio J) 2013;89:531–43.

97. Walsh EE, McConnochie KM, Long CE, and Hall CB. Severity of respiratory syncytial virus infection is related to virus strain. J Infect Dis 1997;175:814–20.

98. Jafri HS, Wu X, Makari D, and Henrickson KJ. Distribution of respiratory syncytial virus subtypes A and B among infants presenting to the emergency department with lower respiratory tract infection or apnea. Pediatr Infect Dis J 2013;32:335–40.

99. Hornsleth A, Klug B, Nir M, et al. Severity of respiratory syncytial virus disease related to type and genotype of virus and to cytokine values in nasopharyngeal secretions. Pediatr Infect Dis J 1998;17:1114–21.

100. Esposito S, Piralla A, Zampiero A, et al. Characteristics and their clinical relevance of respiratory syncytial virus types and genotypes circulating in northern Italy in five consecutive winter seasons. PLoS One 2015;10:e0129369.

101. Martinello RA, Chen MD, Weibel C, and Kahn JS. Correlation between respiratory syncytial virus genotype and severity of illness. J Infect Dis 2002;186:839–42.

102. El Saleeby CM, Bush AJ, Harrison LM, Aitken JA, and Devincenzo JP. Respiratory syncytial virus load, viral dynamics, and disease severity in previously healthy naturally infected children. J Infect Dis 2011;204:996–1002.

103. Houben ML, Coenjaerts FE, Rossen JW, et al. Disease severity and viral load are correlated in infants with primary respiratory syncytial virus infection in the community. J Med Virol 2010;82:1266–71.

104. Hasegawa K, Jartti T, Mansbach JM, et al. Respiratory syncytial virus genomic load and disease severity among children hospitalized with bronchiolitis: multicenter cohort studies in the United States and Finland. J Infect Dis 2015;211:1550–9.

105. Thwaites RS, Coates M, Ito K, et al. Reduced nasal viral load and IFN responses in infants with respiratory syncytial virus bronchiolitis and respiratory failure. Am J Respir Crit Care Med 2018;198:1074–84.

106. Garcia-Maurino C, Moore-Clingenpeel M, Thomas J, et al. Viral load dynamics and clinical disease severity in infants with respiratory syncytial virus infection. J Infect Dis 2019;219:1207–15.

107. Tahamtan A, Askari FS, Bont L, and Salimi V. Disease severity in respiratory syncytial virus infection: Role of host genetic variation. Rev Med Virol 2019;29:e2026.

108. Hull J, Rowlands K, Lockhart E, Moore C, Sharland M, and Kwiatkowski D. Variants of the chemokine receptor CCR5 are associated with severe bronchiolitis caused by respiratory syncytial virus. J Infect Dis 2003;188:904–7.

109. Amanatidou V, Sourvinos G, Apostolakis S, Tsilimigaki A, and Spandidos DA. T280M variation of the CX3C receptor gene is associated with increased risk for severe respiratory syncytial virus bronchiolitis. Pediatr Infect Dis J 2006;25:410–4.

110. Puthothu B, Krueger M, Forster J, and Heinzmann A. Association between severe respiratory syncytial virus infection and IL13/IL4 haplotypes. J Infect Dis 2006;193:438–41.

111. Ventre K and Randolph AG. Ribavirin for respiratory syncytial virus infection of the lower respiratory tract in infants and young children. Cochrane Database Syst Rev 2007:Cd000181.

112. Sinclair SM, Jones JK, Miller RK, Greene MF, Kwo PY, and Maddrey WC. The Ribavirin Pregnancy Registry: An interim analysis of potential teratogenicity at the mid-point of enrollment. Drug Saf 2017;40:1205–18.

113. Ralston SL, Lieberthal AS, Meissner HC, et al. Clinical practice guideline: the diagnosis, management, and prevention of bronchiolitis. Pediatrics 2014;134:e1474–502.

114. National Institute for Health and Care Excellence. Bronchiolitis in children: diagnosis and management. 2015. URL: `https://www.nice.org.uk/guidance/ng9` (visited on May 22, 2015).

115. Hartling L, Bialy LM, Vandermeer B, et al. Epinephrine for bronchiolitis. Cochrane Database Syst Rev 2011:Cd003123.

116. Patel H, Gouin S, and Platt RW. Randomized, double-blind, placebo-controlled trial of oral albuterol in infants with mild-to-moderate acute viral bronchiolitis. J Pediatr 2003;142:509–14.

117. King VJ, Viswanathan M, Bordley WC, et al. Pharmacologic treatment of bronchiolitis in infants and children: a systematic review. Arch Pediatr Adolesc Med 2004;158:127–37.

118. Cade A, Brownlee KG, Conway SP, et al. Randomised placebo controlled trial of nebulised corticosteroids in acute respiratory syncytial viral bronchiolitis. Arch Dis Child 2000;82:126–30.

119. Zhang L, Mendoza-Sassi RA, Wainwright C, and Klassen TP. Nebulised hypertonic saline solution for acute bronchiolitis in infants. Cochrane Database Syst Rev 2017;12:Cd006458.

120. Liu F, Ouyang J, Sharma AN, et al. Leukotriene inhibitors for bronchiolitis in infants and young children. Cochrane Database Syst Rev 2015:Cd010636.

121. Xing Y and Proesmans M. New therapies for acute RSV infections: where are we? Eur J Pediatr 2019;178:131–8.

122. Stein RT, Sherrill D, Morgan WJ, et al. Respiratory syncytial virus in early life and risk of wheeze and allergy by age 13 years. Lancet 1999;354:541–5.

123. Sigurs N, Aljassim F, Kjellman B, et al. Asthma and allergy patterns over 18 years after severe RSV bronchiolitis in the first year of life. Thorax 2010;65:1045–52.

124. Carbonell-Estrany X, Perez-Yarza EG, Garcia LS, Guzman Cabanas JM, Boria EV, and Atienza BB. Long-term burden and respiratory effects of respiratory syncytial virus hospitalization in preterm infants—the SPRING study. PLoS One 2015;10:e0125422.

125. Blanken MO, Rovers MM, Molenaar JM, et al. Respiratory syncytial virus and recurrent wheeze in healthy preterm infants. N Engl J Med 2013;368:1791–9.

126. Scheltema NM, Nibbelke EE, Pouw J, et al. Respiratory syncytial virus prevention and asthma in healthy preterm infants: a randomised controlled trial. Lancet Respir Med 2018;6:257–64.

127. Mochizuki H, Kusuda S, Okada K, Yoshihara S, Furuya H, and Simoes EAF. Palivizumab prophylaxis in preterm infants and subsequent recurrent wheezing. Six-year follow-up study. Am J Respir Crit Care Med 2017;196:29–38.

128. O'Brien KL, Chandran A, Weatherholtz R, et al. Efficacy of motavizumab for the prevention of respiratory syncytial virus disease in healthy Native American infants: a phase 3 randomised double-blind placebo-controlled trial. Lancet Infect Dis 2015;15:1398–408.

129. Russell CD, Unger SA, Walton M, and Schwarze J. The human immune response to respiratory syncytial virus infection. Clin Microbiol Rev 2017;30:481–502.

130. Welliver TP, Garofalo RP, Hosakote Y, et al. Severe human lower respiratory tract illness caused by respiratory syncytial virus and influenza virus is characterized by the absence of pulmonary cytotoxic lymphocyte responses. J Infect Dis 2007;195:1126–36.

131. Heidema J, Lukens MV, Maren WW van, et al. CD8+ T cell responses in bronchoalveolar lavage fluid and peripheral blood mononuclear cells of infants with severe primary respiratory syncytial virus infections. J Immunol 2007;179:8410–7.

132. Wang SZ, Smith PK, Lovejoy M, Bowden JJ, Alpers JH, and Forsyth KD. The apoptosis of neutrophils is accelerated in respiratory syncytial virus (RSV)-induced bronchiolitis. Clin Exp Immunol 1998;114:49–54.

133. Abu-Harb M, Bell F, Finn A, et al. IL-8 and neutrophil elastase levels in the respiratory tract of infants with RSV bronchiolitis. Eur Respir J 1999;14:139–43.

134. Lukens MV, Pol AC van de, Coenjaerts FE, et al. A systemic neutrophil response precedes robust CD8(+) T-cell activation during natural respiratory syncytial virus infection in infants. J Virol 2010;84:2374–83.

135. Johnson JE, Gonzales RA, Olson SJ, Wright PF, and Graham BS. The histopathology of fatal untreated human respiratory syncytial virus infection. Mod Pathol 2007;20:108–19.

136. Cortjens B, Boer OJ de, Jong R de, et al. Neutrophil extracellular traps cause airway obstruction during respiratory syncytial virus disease. J Pathol 2016;238:401–11.

137. Bermejo-Martin JF, Garcia-Arevalo MC, De Lejarazu RO, et al. Predominance of Th2 cytokines, CXC chemokines and innate immunity mediators at the mucosal level during severe respiratory syncytial virus infection in children. Eur Cytokine Netw 2007;18:162–7.

138. Bont L, Heijnen CJ, Kavelaars A, et al. Peripheral blood cytokine responses and disease severity in respiratory syncytial virus bronchiolitis. Eur Respir J 1999;14:144–9.

139. Smyth RL, Mobbs KJ, O'Hea U, Ashby D, and Hart CA. Respiratory syncytial virus bronchiolitis: disease severity, interleukin-8, and virus genotype. Pediatr Pulmonol 2002;33:339–46.

140. Gill MA, Palucka AK, Barton T, et al. Mobilization of plasmacytoid and myeloid dendritic cells to mucosal sites in children with respiratory syncytial virus and other viral respiratory infections. J Infect Dis 2005;191:1105–15.

141. Gill MA, Long K, Kwon T, et al. Differential recruitment of dendritic cells and monocytes to respiratory mucosal sites in children with influenza virus or respiratory syncytial virus infection. J Infect Dis 2008;198:1667–76.

142. Kerrin A, Fitch P, Errington C, et al. Differential lower airway dendritic cell patterns may reveal distinct endotypes of RSV bronchiolitis. Thorax 2017;72:620–7.

143. Weng K, Zhang J, Mei X, et al. Lower number of plasmacytoid dendritic cells in peripheral blood of children with bronchiolitis following respiratory syncytial virus infection. Influenza Other Respir Viruses 2014;8:469–73.

144. Midulla F, Villani A, Panuska JR, et al. Respiratory syncytial virus lung infection in infants: immunoregulatory role of infected alveolar macrophages. J Infect Dis 1993;168:1515–9.

145. Bendelja K, Vojvoda V, Aberle N, et al. Decreased Toll-like receptor 8 expression and lower TNF-α synthesis in infants with acute RSV infection. Respir Res 2010;11:143.

146. Gagro A, Tominac M, Krsulovic-Hresic V, et al. Increased Toll-like receptor 4 expression in infants with respiratory syncytial virus bronchiolitis. Clin Exp Immunol 2004;135:267–72.

147. Smyth RL, Fletcher JN, Thomas HM, and Hart CA. Immunological responses to respiratory syncytial virus infection in infancy. Arch Dis Child 1997;76:210–4.

148. Kawasaki Y, Hosoya M, Kanno H, and Suzuki H. Serum regulated upon activation, normal T cell expressed and presumably secreted concentrations and eosinophils in respiratory syncytial virus infection. Pediatr Int 2006;48:257–60.

149. Lindemans CA, Kimpen JL, Luijk B, et al. Systemic eosinophil response induced by respiratory syncytial virus. Clin Exp Immunol 2006;144:409–17.

150. Bermejo-Martin JF, Garcia-Arevalo MC, Alonso A, et al. Persistence of proinflammatory response after severe respiratory syncytial virus disease in children. J Allergy Clin Immunol 2007;119:1547–50.

151. Brand HK, Ferwerda G, Preijers F, et al. CD4+ T-cell counts and interleukin-8 and CCL-5 plasma concentrations discriminate disease severity in children with RSV infection. Pediatr Res 2013;73:187–93.

152. Sheeran P, Jafri H, Carubelli C, et al. Elevated cytokine concentrations in the nasopharyngeal and tracheal secretions of children with respiratory syncytial virus disease. Pediatr Infect Dis J 1999;18:115–22.

153. Larranaga CL, Ampuero SL, Luchsinger VF, et al. Impaired immune response in severe human lower tract respiratory infection by respiratory syncytial virus. Pediatr Infect Dis J 2009;28:867–73.

154. Bem RA, Bos AP, Bots M, et al. Activation of the granzyme pathway in children with severe respiratory syncytial virus infection. Pediatr Res 2008;63:650–5.

155. Sommereyns C, Paul S, Staeheli P, and Michiels T. IFN-lambda (IFN-lambda) is expressed in a tissue-dependent fashion and primarily acts on epithelial cells in vivo. PLoS Pathog 2008;4:e1000017.

156. Selvaggi C, Pierangeli A, Fabiani M, et al. Interferon lambda 1-3 expression in infants hospitalized for RSV or HRV associated bronchiolitis. J Infect 2014;68:467–77.

157. De Weerd W, Twilhaar WN, and Kimpen JL. T cell subset analysis in peripheral blood of children with RSV bronchiolitis. Scand J Infect Dis 1998;30:77–80.

158. Roe MF, Bloxham DM, White DK, Ross-Russell RI, Tasker RT, and O'Donnell DR. Lymphocyte apoptosis in acute respiratory syncytial virus bronchiolitis. Clin Exp Immunol 2004;137:139–45.

159. Jozwik A, Habibi MS, Paras A, et al. RSV-specific airway resident memory CD8+ T cells and differential disease severity after experimental human infection. Nat Commun 2015;6:10224.

160. Aberle JH, Aberle SW, Dworzak MN, et al. Reduced interferon-gamma expression in peripheral blood mononuclear cells of infants with severe respiratory syncytial virus disease. Am J Respir Crit Care Med 1999;160:1263–8.

161. Pinto RA, Arredondo SM, Bono MR, Gaggero AA, and Diaz PV. T helper 1/T helper 2 cytokine imbalance in respiratory syncytial virus infection is associated with increased endogenous plasma cortisol. Pediatrics 2006;117:e878–86.

162. Isaacs D, Bangham CR, and McMichael AJ. Cell-mediated cytotoxic response to respiratory syncytial virus in infants with bronchiolitis. Lancet 1987;2:769–71.

163. Cherukuri A, Patton K, Gasser R. A. J, et al. Adults 65 years old and older have reduced numbers of functional memory T cells to respiratory syncytial virus fusion protein. Clin Vaccine Immunol 2013;20:239–47.

164. Raiden S, Pandolfi J, Payaslian F, et al. Depletion of circulating regulatory T cells during severe respiratory syncytial virus infection in young children. Am J Respir Crit Care Med 2014;189:865–8.

165. Li B, Wu FL, Feng XB, Sun DK, Cui QQ, and Zhao ZX. Changes and the clinical significance of CD4+ CD25+ regulatory T cells and Th17 cells in peripheral blood of infants with respiratory syncytial virus bronchiolitis. Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 2012;28:426–8.

166. Everard ML, Swarbrick A, Wrightham M, et al. Analysis of cells obtained by bronchial lavage of infants with respiratory syncytial virus infection. Arch Dis Child 1994;71:428–32.

167. Yao S, Jiang L, Moser EK, et al. Control of pathogenic effector T-cell activities in situ by PD-L1 expression on respiratory inflammatory dendritic cells during respiratory syncytial virus infection. Mucosal Immunol 2015;8:746–59.

168. Souza AP de, Freitas DN de, Antuntes Fernandes KE, et al. Respiratory syncytial virus induces phosphorylation of mTOR at ser2448 in CD8 T cells from nasal washes of infected infants. Clin Exp Immunol 2016;183:248–57.

169. Reed JL, Welliver TP, Sims GP, et al. Innate immune signals modulate antiviral and polyreactive antibody responses during severe respiratory syncytial virus infection. J Infect Dis 2009;199:1128–38.

170. Raes M, Peeters V, Alliet P, et al. Peripheral blood T and B lymphocyte subpopulations in infants with acute respiratory syncytial virus brochiolitis. Pediatr Allergy Immunol 1997;8:97–102.

171. Shinoff JJ, O'Brien KL, Thumar B, et al. Young infants can develop protective levels of neutralizing antibody after infection with respiratory syncytial virus. J Infect Dis 2008;198:1007–15.

172. Jensen IP, Thisted E, Glikmann G, et al. Secretory IgM and IgA antibodies to respiratory syncytial virus in nasopharyngeal aspirates: a diagnostic supplement to antigen detection. Clin Diagn Virol 1997;8:219–26.

173. Yamazaki H, Tsutsumi H, Matsuda K, Nagai K, Ogra PL, and Chiba S. Respiratory syncytial virus group-specific antibody response in nasopharyngeal secretions from infants and children after primary infection. Clin Diagn Lab Immunol 1994;1:469–72.

174. Walsh EE and Falsey AR. Humoral and mucosal immunity in protection from natural respiratory syncytial virus infection in adults. J Infect Dis 2004;190:373–8.

175. Falsey AR and Walsh EE. Relationship of serum antibody to risk of respiratory syncytial virus infection in elderly adults. J Infect Dis 1998;177:463–6.

176. Habibi MS, Jozwik A, Makris S, et al. Impaired antibody-mediated protection and defective IgA B-cell memory in experimental infection of adults with respiratory syncytial virus. Am J Respir Crit Care Med 2015;191:1040–9.

177. Tsutsumi H, Matsuda K, Yamazaki H, Ogra PL, and Chiba S. Different kinetics of antibody responses between IgA and IgG classes in nasopharyngeal secretion in infants and children during primary respiratory syncytial virus infection. Acta Paediatr Jpn 1995;37:464–8.

178. McGill A, Greensill J, Marsh R, Craft AW, and Toms GL. Detection of human respiratory syncytial virus genotype specific antibody responses in infants. J Med Virol 2004;74:492–8.

179. Sande CJ, Mutunga MN, Medley GF, Cane PA, and Nokes DJ. Group- and genotype-specific neutralizing antibody responses against respiratory syncytial virus in infants and young children with severe pneumonia. J Infect Dis 2013;207:489–92.

180. Freitas GR, Silva DA, Yokosawa J, et al. Antibody response and avidity of respiratory syncytial virus-specific total IgG, IgG1, and IgG3 in young children. J Med Virol 2011;83:1826–33.

181. Vissers M, Ahout IM, Jonge MI de, and Ferwerda G. Mucosal IgG levels correlate better with respiratory syncytial virus load and inflammation than plasma IgG levels. Clin Vaccine Immunol 2015;23:243–5.

182. Yamazaki H, Tsutsumi H, Matsuda K, Nagai K, Ogra PL, and Chiba S. Effect of maternal antibody on IgA antibody response in nasopharyngeal secretion in infants and children during primary respiratory syncytial virus infection. J Gen Virol 1994;75 (Pt 8):2115–9.

183. Kaul TN, Welliver RC, and Ogra PL. Development of antibody-dependent cell-mediated cytotoxicity in the respiratory tract after natural infection with respiratory syncytial virus. Infect Immun 1982;37:492–8.

184. Kaul TN, Welliver RC, Faden HS, and Ogra PL. The development of respiratory syncytial virus-specific immune complexes in nasopharyngeal secretions following natural infection. J Clin Lab Immunol 1984;15:187–90.

185. Russi JC, Delfraro A, Borthagaray MD, Velazquez B, García-Barreno B, and Hortal M. Evaluation of immunoglobulin E-specific antibodies and viral antigens in nasopharyngeal secretions of children with respiratory syncytial virus infections. J Clin Microbiol 1993;31:819–23.

186. Chung HL and Jang YY. High serum IgE level in the children with acute respiratory syncytial virus infection is associated with severe disease. J Allergy Clin Immunol 2016;137:AB110.

187. Ye Q, Shao WX, Shang SQ, Pan YX, Shen HQ, and Chen XJ. Epidemiological characteristics and immune status of children with Respiratory Syncytial Virus. J Med Virol 2015;87:323–9.

188. Fernandez JA, Roine I, Vasquez A, and Caneo M. Soluble interleukin-2 receptor (sCD25) and interleukin-10 plasma concentrations are associated with severity of primary respiratory syncytial virus (RSV) infection. Eur Cytokine Netw 2005;16:81–90.

189. Semple MG, Dankert HM, Ebrahimi B, et al. Severe respiratory syncytial virus bronchiolitis in infants is associated with reduced airway interferon gamma and substance P. PLoS One 2007;2:e1038.

190. Garcia C, Soriano-Fallas A, Lozano J, et al. Decreased innate immune cytokine responses correlate with disease severity in children with respiratory syncytial virus and human rhinovirus bronchiolitis. Pediatr Infect Dis J 2012;31:86–9.

191. Faber TE, Groen H, Welfing M, Jansen KJ, and Bont LJ. Specific increase in local IL-17 production during recovery from primary RSV bronchiolitis. J Med Virol 2012;84:1084–8.

192. Benten IJ van, Drunen CM van, Koopman LP, et al. RSV-induced bronchiolitis but not upper respiratory tract infection is accompanied by an increased nasal IL-18 response. J Med Virol 2003;71:290–7.

193. Vieira RA, Diniz EM, and Ceccon ME. Correlation between inflammatory mediators in the nasopharyngeal secretion and in the serum of children with lower respiratory tract infection caused by respiratory syncytial virus and disease severity. J Bras Pneumol 2010;36:59–66.

194. Bennett BL, Garofalo RP, Cron SG, et al. Immunopathogenesis of respiratory syncytial virus bronchiolitis. J Infect Dis 2007;195:1532–40.

195. DeVincenzo JP, Wilkinson T, Vaishnaw A, et al. Viral load drives disease in humans experimentally infected with respiratory syncytial virus. Am J Respir Crit Care Med 2010;182:1305–14.

196. Chen ZM, Mao JH, Du LZ, and Tang YM. Association of cytokine responses with disease severity in infants with respiratory syncytial virus infection. Acta Paediatr 2002;91:914–22.

197. Hassan MA, Eldin AM, and Ahmed MM. T-helper 2/T-helper 1 imbalance in respiratory syncytial virus bronchiolitis in relation to disease severity and outcome. Egypt J Immunol 2008;15:153–60.

198. Caballero MT, Serra ME, Acosta PL, et al. TLR4 genotype and environmental LPS mediate RSV bronchiolitis through Th2 polarization. J Clin Invest 2015;125:571–82.

199. Legg JP, Hussain IR, Warner JA, Johnston SL, and Warner JO. Type 1 and type 2 cytokine imbalance in acute respiratory syncytial virus bronchiolitis. Am J Respir Crit Care Med 2003;168:633–9.

200. The IMpact-RSV Study Group. Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. The IMpact-RSV Study Group. Pediatrics 1998;102:531–7.

201. Feltes TF, Cabalka AK, Meissner HC, et al. Palivizumab prophylaxis reduces hospitalization due to respiratory syncytial virus in young children with hemodynamically significant congenital heart disease. J Pediatr 2003;143:532–40.

202. American Academy of Pediatrics. Updated guidance for palivizumab prophylaxis among infants and young children at increased risk of hospitalization for respiratory syncytial virus infection. Pediatrics 2014;134:e620–38.

203. Wu H, Pfarr DS, Johnson S, et al. Development of motavizumab, an ultra-potent antibody for the prevention of respiratory syncytial virus infection in the upper and lower respiratory tract. J Mol Biol 2007;368:652–65.

204. Munoz FM and Englund JA. Chapter 10—Respiratory syncytial virus. In: *Maternal Immunization*. Ed. by Leuridan EE, Nunes MC, and Jones CE. Academic Press, 2020:213–34.

205. Simoes EAF, Forleo-Neto E, Geba GP, et al. Suptavumab for the prevention of medically attended respiratory syncytial virus infection in preterm infants. Clin Infect Dis 2020.

206. Zhu Q, McLellan JS, Kallewaard NL, et al. A highly potent extended half-life antibody as a potential RSV vaccine surrogate for all infants. Sci Transl Med 2017;9:eaaj1928.

207. AstraZeneca. Nirsevimab MELODY Phase III trial met primary endpoint of reducing RSV lower respiratory tract infections in healthy infants. 2021. URL: `https://www.astrazeneca.com/media-centre/press-releases/2021/nirsevimab-phase-iii-trial-met-primary-endpoint.html` (visited on Apr. 26, 2021).

208. Kim HW, Canchola JG, Brandt CD, et al. Respiratory syncytial virus disease in infants despite prior administration of antigenic inactivated vaccine. Am J Epidemiol 1969;89:422–34.

209. PATH. RSV vaccine and mAb snapshot. 2021. URL: `https://www.path.org/resources/rsv-vaccine-and-mab-snapshot/` (visited on Apr. 7, 2021).

210. Mazur NI, Higgins D, Nunes MC, et al. The respiratory syncytial virus vaccine landscape: lessons from the graveyard and promising candidates. Lancet Infect Dis 2018;18:e295–e311.

211. Madhi SA, Polack FP, Piedra PA, et al. Respiratory syncytial virus vaccination during pregnancy and effects in infants. N Engl J Med 2020;383:426–39.

212. McFarland EJ, Karron RA, Muresan P, et al. Live-attenuated respiratory syncytial virus vaccine with M2-2 deletion and with small hydrophobic noncoding region is highly immunogenic in children. J Infect Dis 2020;221:2050–9.

213. McFarland EJ, Karron RA, Muresan P, et al. Live respiratory syncytial virus attenuated by M2-2 deletion and stabilized temperature sensitivity mutation 1030s is a promising vaccine candidate in children. J Infect Dis 2020;221:534–43.

214. Karron RA, Luongo C, Mateo JS, Wanionek K, Collins PL, and Buchholz UJ. Safety and immunogenicity of the respiratory syncytial virus vaccine RSV/ΔNS2/Δ1313/I1314L in RSV-seronegative children. J Infect Dis 2020;222:82–91.

215. Green CA, Scarselli E, Sande CJ, et al. Chimpanzee adenovirus- and MVA-vectored respiratory syncytial virus vaccine is safe and immunogenic in adults. Sci Transl Med 2015;7:300ra126.

216. Green CA, Sande CJ, Scarselli E, et al. Novel genetically-modified chimpanzee adenovirus and MVA-vectored respiratory syncytial virus vaccine safely boosts humoral and cellular immunity in healthy older adults. J Infect 2019;78:382–92.

217. Cicconi P, Jones C, Sarkar E, et al. First-in-human randomized study to assess the safety and immunogenicity of an investigational respiratory syncytial virus (RSV) vaccine based on chimpanzee-adenovirus-155 viral vector-expressing RSV fusion, nucleocapsid, and antitermination viral proteins in healthy adults. Clin Infect Dis 2020;70:2073–81.

218. Sadoff J, De Paepe E, Haazen W, et al. Safety and immunogenicity of the Ad26.RSV.preF investigational vaccine coadministered with an influenza vaccine in older adults. J Infect Dis 2021;223:699–708.

219. Wildenbeest JG, Zuurbier RP, Korsten K, et al. Respiratory Syncytial Virus Consortium in Europe (RESCEU) birth cohort study: defining the burden of infant respiratory syncytial virus disease in Europe. J Infect Dis 2020;222:S606–S612.

220. Korsten K, Adriaenssens N, Coenen S, et al. Burden of respiratory syncytial virus infection in community-dwelling older adults in Europe (RESCEU): an international prospective cohort study. Eur Respir J 2021;57:2002688.

221. Justicia-Grande AJ, Pardo-Seco J, Cebey-Lopez M, et al. Development and validation of a new clinical scale for infants with acute respiratory infection: the ReSVinet scale. PLoS One 2016;11:e0157665.

222. Bonsall D, Golubchik T, Cesare M de, et al. A comprehensive genomics solution for HIV surveillance and clinical monitoring in low income settings. J Clin Microbiol 2020;58:e00382–20.

223. Faircloth BC and Glenn TC. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PLoS One 2012;7:e42543.

224. Goh C, Golubchik T, Ansari MA, et al. Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection. bioRxiv 2019:716902.

225. Bonsall D, Ansari MA, Ip C, et al. ve-SEQ: robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. F1000Res 2015;4:1062.

226. Jolley KA, Bliss CM, Bennett JS, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology (Reading) 2012;158:1005–15.

227. Wymant C, Blanquart F, Golubchik T, et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. Virus Evol 2018;4:vey007.

228. Bolger AM, Lohse M, and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.

229. Hunt M, Gall A, Ong SH, et al. IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics 2015;31:2374–6.

230. Nurk S, Meleshko D, Korobeynikov A, and Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res 2017;27:824–34.

231. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.

232. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW. GenBank. Nucleic Acids Res 2016;44:D67–72.

233. Katoh K and Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80.

234. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.

235. Broad Institute. Picard Tools. 2021. URL: `https://broadinstitute.github.io/picard/` (visited on May 1, 2021).

236. Gall A, Ferns B, Morris C, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol 2012;50:3838–44.

237. Prachayangprecha S, Schapendonk CM, Koopmans MP, et al. Exploring the potential of next-generation sequencing in detection of respiratory viruses. J Clin Microbiol 2014;52:3722–30.

238. Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 2010;7:709–15.

239. Zhu YY, Machleder EM, Chenchik A, Li R, and Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques 2001;30:892–7.

240. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.

241. Gotoh O. Optimal alignment between groups of sequences and its application to multiple sequence alignment. Comput Appl Biosci 1993;9:361–70.

242. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.

243. Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences 1986;17:57–86.

244. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 1994;39:306–14.

245. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2018. URL: `https://www.R-project.org/`.

246. Paradis E and Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019;35:526–8.

247. Yu G, Smith DK, Zhu H, Guan Y, and Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 2017;8:28–36.

248. Yu G, Lam TTY, Zhu H, and Guan Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol Biol Evol 2018;35:3041–3.

249. Anders S, Pyl PT, and Huber W. HTSeq: a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–9.

250. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.

251. Su S, Law CW, Ah-Cann C, Asselin-Labat ML, Blewitt ME, and Ritchie ME. Glimma: interactive graphics for gene expression analysis. Bioinformatics 2017;33:2050–2.

252. Robinson MD, McCarthy DJ, and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.

253. Robinson MD and Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25.

254. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 1995;57:289–300.

255. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag, 2016. URL: `http://ggplot2.org`.

256. Law CW, Alhamdoosh M, Su S, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res 2016;5.

257. Kim L, Rha B, Abramson JS, et al. Identifying gaps in respiratory syncytial virus disease epidemiology in the United States prior to the introduction of vaccines. Clin Infect Dis 2017;65:1020–5.

258. Zhu Q, McAuliffe JM, Patel NK, et al. Analysis of respiratory syncytial virus preclinical and clinical variants resistant to neutralization by monoclonal antibodies palivizumab and/or motavizumab. J Infect Dis 2011;203:674–82.

259. Gilbert BE, Patel N, Lu H, et al. Respiratory syncytial virus fusion nanoparticle vaccine immune responses target multiple neutralizing epitopes that contribute to protection against wild-type and palivizumab-resistant mutant virus challenge. Vaccine 2018;36:8069–78.

260. World Health Organization. WHO meeting to launch Phase-2 of the RSV Surveillance Pilot based on the Global Influenza Surveillance and Response System. 2019. URL: `https://www.who.int/influenza/rsv/who_rsv_surveillance_2nd_phase/en/` (visited on Sept. 17, 2019).

261. Aljabr W, Touzelet O, Pollakis G, et al. Investigating the influence of ribavirin on human respiratory syncytial virus RNA synthesis by using a high-resolution transcriptome sequencing approach. J Virol 2016;90:4876–88.

262. Grosfeld H, Hill MG, and Collins PL. RNA replication by respiratory syncytial virus (RSV) is directed by the N, P, and L proteins; transcription also occurs under these conditions but requires RSV superinfection for efficient synthesis of full-length mRNA. J Virol 1995;69:5677–86.

263. Fearns R and Collins PL. Role of the M2-1 transcription antitermination protein of respiratory syncytial virus in sequential transcription. J Virol 1999;73:5852–64.

264. Richard CA, Rincheval V, Lassoued S, et al. RSV hijacks cellular protein phosphatase 1 to regulate M2-1 phosphorylation and viral transcription. PLoS Pathog 2018;14:e1006920.

265. Karron RA, Luongo C, Thumar B, et al. A gene deletion that up-regulates viral gene expression yields an attenuated RSV vaccine with improved antibody responses in children. Sci Transl Med 2015;7:312ra175.

266. Wang Z, Gerstein M, and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.

267. Martin ET, Kuypers J, Wald A, and Englund JA. Multiple versus single virus respiratory infections: viral load and clinical disease severity in hospitalized children. Influenza Other Respir Viruses 2012;6:71–7.

268. Aberle JH, Aberle SW, Pracher E, Hutter HP, Kundi M, and Popow-Kraupp T. Single versus dual respiratory virus infections in hospitalized infants: impact on clinical course of disease and interferon-gamma response. Pediatr Infect Dis J 2005;24:605–10.

269. Yoshida LM, Suzuki M, Nguyen HA, et al. Respiratory syncytial virus: co-infection and paediatric lower respiratory tract infections. Eur Respir J 2013;42:461–9.

270. Mansbach JM, Piedra PA, Teach SJ, et al. Prospective multicenter study of viral etiology and hospital length of stay in children with severe bronchiolitis. Arch Pediatr Adolesc Med 2012;166:700–6.

271. Greer RM, McErlean P, Arden KE, et al. Do rhinoviruses reduce the probability of viral co-detection during acute respiratory tract infections? J Clin Virol 2009;45:10–5.

272. Nickbakhsh S, Mair C, Matthews L, et al. Virus-virus interactions impact the population dynamics of influenza and the common cold. Proc Natl Acad Sci U S A 2019;116:27142–50.

273. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. Cell Host Microbe 2015;17:704–15.

274. Prevaes SM, de Steenhuijsen Piters WA, de Winter-de Groot KM, et al. Concordance between upper and lower airway microbiota in infants with cystic fibrosis. Eur Respir J 2017;49.

275. Kroneman A, Vennema H, Deforche K, et al. An automated genotyping tool for enteroviruses and noroviruses. J Clin Virol 2011;51:121–5.

276. Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 2011;12:385.

277. Aswad A, Aimola G, Wight D, et al. Evolutionary history of endogenous human herpesvirus 6 reflects human migration out of Africa. Mol Biol Evol 2021;38:96–107.

278. Dinarello CA, Cannon JG, and Wolff SM. New concepts on the pathogenesis of fever. Rev Infect Dis 1988;10:168–89.

279. Fujisawa T, Velichko S, Thai P, Hung LY, Huang F, and Wu R. Regulation of airway MUC5AC expression by IL-1beta and IL-17A; the NF-kappaB paradigm. J Immunol 2009;183:6236–43.

280. Achten NB, Wu P, Bont L, et al. Interference between respiratory syncytial virus and human rhinovirus infection in infancy. J Infect Dis 2017;215:1102–6.

281. Karppinen S, Toivonen L, Schuez-Havupalo L, Waris M, and Peltola V. Interference between respiratory syncytial virus and rhinovirus in respiratory tract infections in children. Clin Microbiol Infect 2016;22:208.e1–208.e6.

282. Steenhuijsen Piters WA de, Heinonen S, Hasrat R, et al. Nasopharyngeal microbiota, host transcriptome, and disease severity in children with respiratory syncytial virus infection. Am J Respir Crit Care Med 2016;194:1104–15.

283. Ederveen THA, Ferwerda G, Ahout IM, et al. Haemophilus is overrepresented in the nasopharynx of infants hospitalized with RSV infection and associated with increased viral load and enhanced mucosal CXCL8 responses. Microbiome 2018;6:10.

284. Gulraiz F, Bellinghausen C, Bruggeman CA, and Stassen FR. Haemophilus influenzae increases the susceptibility and inflammatory response of airway epithelial cells to viral infections. Faseb j 2015;29:849–58.

285. Coultas JA, Smyth R, and Openshaw PJ. Respiratory syncytial virus (RSV): a scourge from infancy to old age. Thorax 2019;74:986–93.

286. Luchsinger V, Ampuero S, Palomino MA, et al. Comparison of virological profiles of respiratory syncytial virus and rhinovirus in acute lower tract respiratory infections in very young Chilean infants, according to their clinical outcome. J Clin Virol 2014;61:138–44.

287. Rosas-Salazar C, Shilts MH, Tovchigrechko A, et al. Differences in the nasopharyngeal microbiome during acute respiratory tract infection with human rhinovirus and respiratory syncytial virus in infancy. J Infect Dis 2016;214:1924–8.

288. Streng A, Goettler D, Haerlein M, et al. Spread and clinical severity of respiratory syncytial virus A genotype ON1 in Germany, 2011–2017. BMC Infect Dis 2019;19:613.

289. Trento A, Casas I, Calderon A, et al. Ten years of global evolution of the human respiratory syncytial virus BA genotype with a 60-nucleotide duplication in the G protein gene. J Virol 2010;84:7500–12.

290. Ruckwardt TJ, Morabito KM, and Graham BS. Immunological lessons from respiratory syncytial virus vaccine development. Immunity 2019;51:429–42.

291. Zhu Q, Lu B, McTamney P, et al. Prevalence and significance of substitutions in the fusion protein of respiratory syncytial virus resulting in neutralization escape from antibody MEDI8897. J Infect Dis 2018;218:572–80.

292. Wu SJ, Schmidt A, Beil EJ, et al. Characterization of the epitope for anti-human respiratory syncytial virus F protein monoclonal antibody 101F using synthetic peptides and genetic approaches. J Gen Virol 2007;88:2719–23.

293. Zhao X, Chen FP, Megaw AG, and Sullender WM. Variable resistance to palivizumab in cotton rats by respiratory syncytial virus mutants. J Infect Dis 2004;190:1941–6.

294. Zhu Q, Patel NK, McAuliffe JM, et al. Natural polymorphisms and resistance-associated mutations in the fusion protein of respiratory syncytial virus (RSV): effects on RSV susceptibility to palivizumab. J Infect Dis 2012;205:635–8.

295. Choi Y, Mason CS, Jones LP, Crabtree J, Jorquera PA, and Tripp RA. Antibodies to the central conserved region of respiratory syncytial virus (RSV) G protein block RSV G protein CX3C-CX3CR1 binding and cross-neutralize RSV A and B strains. Viral Immunol 2012;25:193–203.

296. Grad YH, Newman R, Zody M, et al. Within-host whole-genome deep sequencing and diversity analysis of human respiratory syncytial virus infection reveals dynamics of genomic diversity in the absence and presence of immune pressure. J Virol 2014;88:7286–93.

297. Lau JW, Kim YI, Murphy R, et al. Deep sequencing of RSV from an adult challenge study and from naturally infected infants reveals heterogeneous diversification dynamics. Virology 2017;510:289–96.

298. McCrone JT and Lauring AS. Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant calling. J Virol 2016;90:6884–95.

299. Nelson CW and Hughes AL. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. Infect Genet Evol 2015;30:1–7.

300. Harrell Jr FE. Hmisc: Harrell Miscellaneous. R package version 4.5-0. 2021. URL: `https://CRAN.R-project.org/package=Hmisc` (visited on May 12, 2021).

301. Lythgoe KA, Hall M, Ferretti L, et al. SARS-CoV-2 within-host diversity and transmission. Science 2021;372:eabg0821.

302. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, and Lauring AS. Stochastic processes constrain the within and between host evolution of influenza virus. Elife 2018;7:e35962.

303. Tabor DE, Fernandes F, Langedijk AC, et al. Global molecular epidemiology of respiratory syncytial virus from the 2017-2018 INFORM-RSV study. J Clin Microbiol 2020;59:e01828–20.

304. Xue KS, Moncla LH, Bedford T, and Bloom JD. Within-host evolution of human influenza virus. Trends Microbiol 2018;26:781–93.

305. Sanjuan R and Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci 2016;73:4433–48.

306. Valesano AL, Fitzsimmons WJ, McCrone JT, et al. Influenza B viruses exhibit lower within-host diversity than influenza A viruses in human hosts. J Virol 2020;94:e01710–19.

307. Cudini J, Roy S, Houldcroft CJ, et al. Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. Proc Natl Acad Sci U S A 2019;116:5693–8.

308. Papenburg J, Carbonneau J, Hamelin M, et al. Molecular evolution of respiratory syncytial virus fusion gene, Canada, 2006–2010. Emerg Infect Dis 2012;18:120–4.

309. Matheson JW, Rich FJ, Cohet C, et al. Distinct patterns of evolution between respiratory syncytial virus subgroups A and B from New Zealand isolates collected over thirty-seven years. J Med Virol 2006;78:1354–64.

310. Smith G, Raghunandan R, Wu Y, et al. Respiratory syncytial virus fusion glycoprotein expressed in insect cells form protein nanoparticles that induce protective immunity in cotton rats. PLoS One 2012;7:e50852.

311. Pierantoni A, Esposito ML, Ammendola V, et al. Mucosal delivery of a vectored RSV vaccine is safe and elicits protective immunity in rodents and nonhuman primates. Mol Ther Methods Clin Dev 2015;2:15018.

312. Crank MC, Ruckwardt TJ, Chen M, et al. A proof of concept for structure-based vaccine design targeting RSV in humans. Science 2019;365:505–9.

313. Bin L, Liu H, Tabor DE, et al. Emergence of new antigenic epitopes in the glycoproteins of human respiratory syncytial virus collected from a US surveillance study, 2015–17. Sci Rep 2019;9:3898.

314. Poon LL, Song T, Rosenfeld R, et al. Quantifying influenza virus diversity and transmission in humans. Nat Genet 2016;48:195–200.

315. Githinji G, Agoti CN, Kibinge N, et al. Assessing the utility of minority variant composition in elucidating RSV transmission pathways. bioRxiv 2018:411512.

316. Blount R. E. J, Morris JA, and Savage RE. Recovery of cytopathogenic agent from chimpanzees with coryza. Proc Soc Exp Biol Med 1956;92:544–9.

317. Chanock R, Roizman B, and Myers R. Recovery from infants with respiratory illness of a virus related to chimpanzee coryza agent (CCA). I. Isolation, properties and characterization. Am J Hyg 1957;66:281–90.

318. Otieno JR, Kamau EM, Agoti CN, et al. Spread and evolution of respiratory syncytial virus A genotype ON1, coastal Kenya, 2010–2015. Emerg Infect Dis 2017;23:264–71.

319. Comas-Garcia A, Noyola DE, Cadena-Mota S, Rico-Hernandez M, and Bernal-Silva S. Respiratory syncytial virus-a ON1 genotype emergence in Central Mexico in 2009 and evidence of multiple duplication events. J Infect Dis 2018;217:1089–98.

320. Panayiotou C, Richter J, Koliou M, Kalogirou N, Georgiou E, and Christodoulou C. Epidemiology of respiratory syncytial virus in children in Cyprus during three consecutive winter seasons (2010-2013): age distribution, seasonality and association between prevalent genotypes and disease severity. Epidemiol Infect 2014;142:2406–11.

321. Duvvuri VR, Granados A, Rosenfeld P, Bahl J, Eshaghi A, and Gubbay JB. Genetic diversity and evolutionary insights of respiratory syncytial virus A ON1 genotype: global and local transmission dynamics. Sci Rep 2015;5:14268.

322. Hirano E, Kobayashi M, Tsukagoshi H, et al. Molecular evolution of human respiratory syncytial virus attachment glycoprotein (G) gene of new genotype ON1 and ancestor NA1. Infect Genet Evol 2014;28:183–91.

323. Darriba D, Taboada GL, Doallo R, and Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 2012;9:772.

324. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307–21.

325. Drummond A and Rodrigo AG. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol Biol Evol 2000;17:1807–15.

326. Rambaut A, Lam TT, Max Carvalho L, and Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016;2:vew007.

327. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, and Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 2018;4:vey016.

328. Faria NR, Rambaut A, Suchard MA, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 2014;346:56–61.

329. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, and Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol 2012;29:2157–67.

330. Baele G, Li WL, Drummond AJ, Suchard MA, and Lemey P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Mol Biol Evol 2013;30:239–43.

331. Drummond AJ, Rambaut A, Shapiro B, and Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 2005;22:1185–92.

332. Rambaut A, Drummond AJ, Xie D, Baele G, and Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol 2018;67:901–4.

333. Lemey P, Rambaut A, Drummond AJ, and Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol 2009;5:e1000520.

334. Vodovar N, Goic B, Blanc H, and Saleh MC. In silico reconstruction of viral genomes from small RNAs improves virus-derived small interfering RNA profiling. J Virol 2011;85:11016–21.

335. Obando-Pacheco P, Justicia-Grande AJ, Rivero-Calle I, et al. Respiratory syncytial virus seasonality: a global overview. J Infect Dis 2018;217:1356–64.

336. Saravanos GL, Sheel M, Homaira N, et al. Respiratory syncytial virus-associated hospitalisations in Australia, 2006-2015. Med J Aust 2019;210:447–53.

337. Paynter S, Ware RS, Sly PD, Weinstein P, and Williams G. Respiratory syncytial virus seasonality in tropical Australia. Aust N Z J Public Health 2015;39:8–10.

338. Tabatabai J, Thielen A, Lehners N, Daeumer M, and Schnitzler P. Respiratory syncytial virus A in haematological patients with prolonged shedding: Premature stop codons and deletion of the genotype ON1 72-nucleotide-duplication in the attachment G gene. J Clin Virol 2018;98:10–7.

339. Unwin S, Chatterton J, and Chantrey J. Management of severe respiratory tract disease caused by human respiratory syncytial virus and Streptococcus pneumoniae in captive chimpanzees (Pan troglodytes). J Zoo Wildl Med 2013;44:105–15.

340. Khan K, Arino J, Hu W, et al. Spread of a novel influenza A (H1N1) virus via global airline transportation. N Engl J Med 2009;361:212–4.

341. Candido DS, Claro IM, Jesus JG de, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science 2020;369:1255–60.

342. Plessis L du, McCrone JT, Zarebski AE, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science 2021;371:708–12.

343. Di Giallonardo F, Kok J, Fernandez M, et al. Evolution of human respiratory syncytial virus (RSV) over multiple seasons in New South Wales, Australia. Viruses 2018;10:476.

344. Yun KW, Choi EH, and Lee HJ. Molecular epidemiology of respiratory syncytial virus for 28 consecutive seasons (1990–2018) and genetic variability of the duplication region in the G gene of genotypes ON1 and BA in South Korea. Arch Virol 2020;165:1069–77.

345. Hotard AL, Laikhter E, Brooks K, Hartert TV, and Moore ML. Functional analysis of the 60-nucleotide duplication in the respiratory syncytial virus Buenos Aires strain attachment glycoprotein. J Virol 2015;89:8258–66.

346. Kumaria R, Iyer LR, Hibberd ML, Simoes EA, and Sugrue RJ. Whole genome characterization of non-tissue culture adapted HRSV strains in severely infected children. Virol J 2011;8:372.

347. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, and Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. Nature 2008;453:615–9.

348. Murphy CN, Fowler R, Balada-Llasat JM, et al. Multicenter evaluation of the BioFire FilmArray Pneumonia/Pneumonia Plus Panel for detection and quantification of agents of lower respiratory tract infection. J Clin Microbiol 2020;58:e00128–20.

349. Capella C, Chaiwatpongsakorn S, Gorrell E, et al. Prefusion F, postfusion F, G antibodies, and disease severity in infants and young children with acute respiratory syncytial virus infection. J Infect Dis 2017;216:1398–406.

350. Mejias A, Dimo B, Suarez NM, et al. Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. PLoS Med 2013;10:e1001549.

351. Liao Y, Smyth GK, and Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res 2019;47:e47.