RSV sequence variation and within-host minority variant dynamics

A thesis submitted for the degree of Doctor of Philosophy

Inne Nauwelaers

November 2019

Department of Respiratory Medicine

National Hearth and Lung Institute at St. Mary's

Imperial College London, Norfolk Place, London, W2 1PG

Declaration of originality

I hereby declare that the work in this thesis was my own and the thesis itself was written by me. Where others have contributed to this work, it has been stated and they have been acknowledged.

Copyright statement

Acknowledgements

Moving to another country to do a PhD project is quite an adventure and even though I moved all on my own, there are many people who have made this adventure possible for me and they deserve my deepest gratitude. There will definitely be people I have met on my way and are not mentioned here, but I will carry them with me in my fondest memories.

First, I would like to thank Peter Openshaw, Christopher Chiu and Monica Galiano for deciding so promptly that I would be a good fit for this project. After a one hour Skype interview I was called back within 10 minutes to receive the great news that I was moving to the UK for a PhD at Imperial College London. There have been few moments in my life that I was so excited and proud of myself. Thank you for the opportunity to work on this project and thank you to Maria Zambon also. I am especially grateful to Monica for the pep talks and scientific brainstorm sessions in the hallway of Public Health England. You have kept me believing it was possible.

The group I was part of is the most hardworking group I have ever encountered. Thank you, Aggie, for making time when there was none, to take me under your wing at the very beginning of my time at St. Mary's and the patience you showed. Thank you Zoe, Anakin, Mohini, Suzy,Iris and Emma for the support I got from you all, it has been an absolute pleasure to be part of the team and I never felt excluded, even though I was on another campus most of the time, thanks to you all. Steph, you deserve my gratefulness for listening when I was struggling, for cheering me up, but most importantly to show me I did do well, regardless of my thoughts. It means the world to me that you cared so much.

My time in the office was always fun, even when the work did not work or analysis failed or presentations did not go well. Anabel, Kate, Dawid, Freja and Sophie, I loved being part of such a great group of bright scientists (and great pub-goers) and I miss you all so much! I know every single one of you will get to present an awesome thesis, because you are all such capable and bright minds. A great thanks to Sasha, Susana and Seema as well, without whom there would be no lab to work in.

One of the people most important people that have kept me going is Gale. I cannot thank you enough for laughs and giggles, great pub times and even better talks. Gale, you are one in a million. You have saved many by being you and you might even save the UK. Said by the little Belgian from the imaginary country. You are a real-life angel.

I did not only work when I moved to London, I managed to combine my PhD with a social life and it was a rich one with Supernova in the centre of it. Thank you all for so willingly and kindly accepting me as part of the team(s). Korfball has kept me sane during this adventure and Supernova has added more experiences to the adventure. Thank you Jenny and Woody, Eiti, Becki Fish, Joe Carroll, Martin and many, many more people that have enriched my life. It was a privilege to be a part of Supernova and I will always be proud of it.

Obviously, all of this would not be possible without my family's help. Mum and dad, Nele and Arne, thank you for your support, the skype calls, the visits, the messages and the warm weekends at home. You have always given me exactly when I needed when I needed it, even when I did not know what I needed myself. Nele, you are always welcome for advice on your own PhD adventure, I will help in any way I can.

In a similar fashion, thank you Liesbeth, Laurine, Aurélie, Ans and Amy, and also Tim and Carl for visiting and for great Christmas parties when I was home over the holidays. It never felt like I had been away for too long when we were catching up. I want to thank Liesbeth especially, who has been one of my closest friends since we were teenagers. She probably knows me better than I know myself and is the only reason I dared to venture into the great unknown. She always has my back and she could always bring me back to my senses. My dearest friend, I wish you all the love in the world, you are my superhero.

I have spent the last year of my PhD writing at home, which was a small adventure in itself. In the space of a year I somehow managed to move back to Belgium, to care for the best grandma in the world, to find a job, to find a place and move there, to play some more korf, visit my London friends, go on holiday and write a thesis. For this mini-adventure I have people to thank as well. My korfball team at home have adopted me halfway through the season and embraced me as their own immediately, thank you for that.

A special thanks as well to Cleo, who has been finishing up at a similar time as I did and knows the struggles and pains of writing up when there is so much more research to do and work to be done, and knows the feeling of guilt when taking time to write instead of helping anyone else. Thank you for your understanding talks, I would not have kept my calm if it were not for you.

Another person I want to thank, especially for the last couple of months of finishing up, is Mariska, aka 'de Marre'. Thank you for taking the time to struggle together with me. Your time at mine kept me focussed and helped me complete this work. It would not have been finished in the way it is right now if not for you.

And last, but not least, I want to thank José Melero, a pioneer in RSV research and the kindest scientist I have ever met. I regret only having known you for such a short time, but you gave me the feeling that my research was worth researching. You sparked my fire and have given me the feeling of worthiness, and you leaving this world broke me more than I could have imagined it ever would.

Abstract

Respiratory syncytial virus (RSV) infection is a common disease that causes the most severe disease in the extremes of age. Parts of the RSV genome are extremely variable, however, origination of genomic variation of RSV is not studied very well. Epidemiology studies have shown rapidly changing RSV strains and grouped these in genotypes based on a part of the RSV genome that consists of the (partial) G gene.

In this thesis, the genotyping system was inspected and it showed that the part of the G gene previously used for genotyping did not contain enough information to reliably determine which genotype a strain belonged to. Phylogenetic analysis was performed to determine the necessary and sufficient part of the genome to determine the genotype reliably, which was full G. Other proteins were investigated for variability as well and both F and L carried plenty of variation as well.

The amount of variation within a patient has been understudied. Therefore, a new method was optimised to detect the prevalence of minority variations in clinical samples. The prevalence of minority variants was examined in a community cohort and hospital cohort from season 2015-2016 of which all samples were spatiotemporally and age-matched. The detected genotypes were GA2 and ON1. Most clinical samples in this study did carry minority variants, however, there was no difference in the amount of variation between community and hospital samples. The gene that displayed the most variations per nucleotide, and most non-synonymous variations was G.

This research also demonstrates that these variations can be transmitted or develop during acute infection. Consecutive samples from volunteers inoculated with a known RSV strain showed that both synonymous and non-synonymous variations can occur and their frequency can increase, decrease or remain stable over time. The F gene rarely developed non-synonymous variations in this study.

List of figures

List of tables

List of abbreviations

Abbreviation	Explanation
A	Adenine
AA	Amino acid
ATM	Amplicon tagment mix
bp	Base pairs
BWA	Burrows-Wheeler Aligner
C	Cytosine
cDNA	copy DNA
CPE	Cytopathic effect
CRV	Conserved region V
ddNTP	Dideoxynucleoside triphosphate
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
F	Fusion protein
G	Guanine
G	Glycoprotein
GFP	Green Flueorescent Protein
HEp2	Human Epithelial type 2 cells
hpi	Hours post infection
HT1	Hybridisation buffer 1
HVR2	Hyper variable region 2
IFN	Interferon
JC69	Jukes-Cantor model
K2P	Kimura 2-parameter model
Kb	kilobase
L	Polymerase protein
LNA1	Library normalisation additives 1
LNB1	Library normalisation beads 1
LNS1	Library normalisation storage buffer 1
LNW1	Library normalisation wash 1
LRTI	Lower respiratory tract infection
M	Matrix protein
M2-1	Matrix protein 2-1
M2-2	Matrix protein 2-2
M37	Respiratory syncytial virus Memphis-37 strain
MEM	Maximum Exact Matches
ml	Millilitre
ML	Maximum likelihood
MMLV	Moloney Mouse Leukemia Virus
MOI	Multiplictiy of infection
MP	Maximum parsimony
MWW	Mann-Whitney-Wilcoxon test

N	Nucleoprotein
NCBI	National Center for Biotechnology Information
NJ	Neighbour-joining
NNI	Nearest neighbour interchange
NPM	Nextera PCR Master Mix
NS1	Non-structural protein 1
NS2	Non-structural protein 2
nt	Nucleotide
NT	Neutralise tagment buffer
P	Phosphoprotein
PBS	Phosphate Buffered Saline
PCR	Polymerase chain reaction
PFU	Plaque forming unit
PHE	Public Health England
PRR	Pattern recognition receptors
qPCR	Quantitative PCR
Q-RT-PCR	Quantitative real-time PCR
RNA	Ribonucleic acid
rpm	Rotations per minute
RSB	Resuspension buffer
RSV	Respiratory syncytial virus
RT	Reverse transcription
SAM	Sequence Alignment Map
SH	Small hydrophobic protein
SMRT	Single Molecule Real Time
SOP	Standard Operating Protocol
SPR	Subtree pruning and regrafting
TBR	Tree bisection and reconnection
TD	Tagment DNA buffer
T	Thymine
VCF	Variant calling file

Table of contents

1. Introduction

1.1. Respiratory syncytial virus

1.1.1. Clinical RSV disease

1.1.1.1. Disease burden in the UK and worldwide

Respiratory syncytial virus (RSV) is a ubiquitous pathogen that, despite decades of research, continues to represent an enormous burden of disease worldwide. Its widespread circulation means that almost all children are infected with RSV at least once by the age of two (1). While most infections in healthy, older children and young adults are mild, every year over 450,000 GP episodes, 29,000 hospitalisations and 83 deaths are attributed to RSV in those under 18 years of age in the UK alone (2).

Globally, 28% of lower respiratory tract infections (LRTI) in children under the age of 5 are due to RSV, which translates to 33.1 million RSV LRTI worldwide (3, 4). 3.2 million of those infections require hospitalisation, of which nearly half are in infants under the age of 6 months (Figure 1.1). Preterm infants have a 3 times greater risk to end up in hospital compared to full term babies (5). It is estimated that nearly 150,000 deaths occur in children under the age of 5 worldwide each year (4).

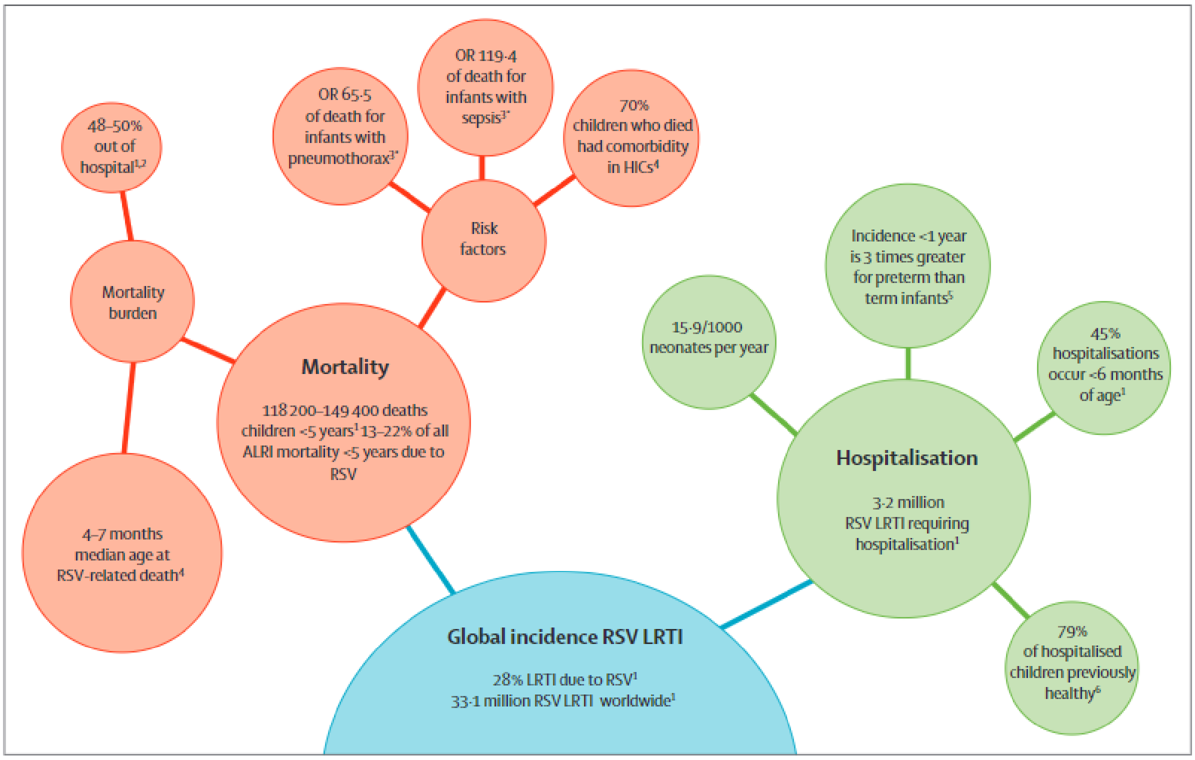

48–50% out of hospital1,2
OR 65·5 of death for infants with pneumothorax3*
OR 119·4 of death for infants with sepsis3*
70% children who died had comorbidity in HICs4
Risk factors
Mortality burden
Mortality
118 200–149 400 deaths children <5 years1 13–22% of all ALRI mortality <5 years due to RSV
4–7 months median age at RSV-related death4
Incidence <1 year is 3 times greater for preterm than term infants5
15·9/1000 neonates per year
45% hospitalisations occur <6 months of age1
Hospitalisation
3·2 million RSV LRTI requiring hospitalisation1
79% of hospitalised children previously healthy6
Global incidence RSV LRTI
28% LRTI due to RSV1
33·1 million RSV LRTI worldwide1

Figure 1.1: The global trends regarding the incidence, hospitalisation and mortality rates due to RSV LRTI in children under the age of 5 shown an enormous burden of disease. Figure taken from Mazur et al. (3).

In comparison, influenza caused 20 million episodes and 1 million hospitalisations per year, while hospitalisations with RSV reached 3.2 million only in children under the age of 5 annually. Deaths were estimated at 28,000-111,500 for influenza (6, 7).

It has become clear that RSV causes severe disease in the elderly as well with approximately 487,000 RSV-related GP episodes, 18,000 hospitalisations and 8,482 deaths occurring in UK adults, with most hospitalisations (79%) and almost all deaths (93%) in those over 65 years (8).

It is generally believed that RSV also greatly contributes to a largely underestimated burden of mortality in this population. Older adults with COPD are a particularly vulnerable population, since RSV infection can cause exacerbations of their disease and an accelerated decline of FEV1 (9). It is even estimated that RSV infection is as common in elderly as non-pandemic influenza A infections (10).

1.1.1.2. Symptoms

Disease in people infected with RSV is highly variable. Some have asymptomatic infections, while others develop lethal pneumonia and infants can develop a condition known as bronchiolitis. Mild disease presents as a cough, rhinorrhoea and mild fever, but can evolve to severe disease with acute lower respiratory tract manifestations (11). Bronchioles may become obstructed by mucus and severe coughing can develop. Ciliated epithelial cells of the airways undergo necrosis and submucosal oedema builds up. Difficulty with breathing is frequently associated with wheezing, hypoxia and cyanosis (12). However, these more severe features only occur in a minority of infected individuals and the differential disease severity seen, even between seemingly similar hosts, is still incompletely understood.

Early events are key in determining the course of disease and these may be influenced by both virus and host responses. For example, early viral proteins interfere with the innate immune system, promoting high viral load, which causes more severe disease (13). Furthermore, studies in mice have suggested that infection at a very young age is a particular risk factor for more severe disease, but that reinfection later in life can also escalate to more severe disease, indicating that both the maturity of the immune system and its interaction with RSV are important determinants (14). Similar trends are noted in humans. Neonates under the age of 6 months seem particularly prone to developing severe disease. This might be partially contributed to infection of a specific subset of neonatal regulatory B cells. Upon RSV infection, these cells start producing IL-10 which inhibits further immune responses (15).

Contrasting to Influenza infection, the same RSV strain can reinfect the same host (16). It was suggested this might be due to several factors, including a dysregulated inflammatory environment (17). However, it is unclear which underlying immune dysfunction is the main cause of the possibility of reinfections.

1.1.2. Immunology

1.1.2.1. Immune responses to RSV infection

Infection with RSV triggers the immune response in several ways and severe RSV disease is thought to be due to an exaggerated inflammatory reaction rather than direct damage by viral cytopathology (18). It has long been thought that interferons (IFNs) are the first line of defense in viral infection and most research has been focused on this aspect of the immune system, but recent studies have found an indication that O-linked glycans might activate anti-viral immune responses dependent on neutrophils first (19).

After RSV entry into ciliated epithelial cells, it is recognised by pattern recognition receptors (PRR) which causes the release of reactive oxygen species which upregulate the STAT pathway. RSV proteins cause NFκB induction, which activates an anti-viral response by transcription of IFNα/β and chemokines via IκB kinase (Figure 1.2). This innate immune response then triggers and coordinates mechanisms for early elimination of the virus and adaptive immunity (13, 20, 21).

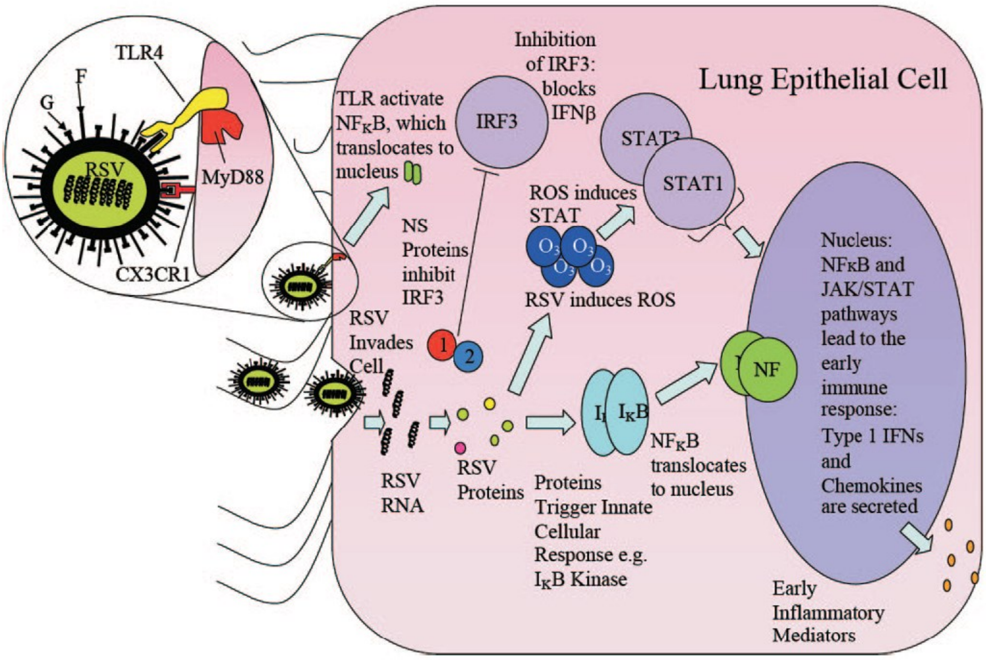

TLR4
F
G
RSV
MyD88
CX3CR1
Inhibition of IRF3: blocks IFNβ
Lung Epithelial Cell
TLR activate NF$_{\kappa}$B, which translocates to nucleus
IRF3
STAT2
STAT1
ROS induces STAT
O_3
RSV induces ROS
NS Proteins inhibit IRF3
RSV Invades Cell
1
2
RSV RNA
RSV Proteins
Proteins Trigger Innate Cellular Response e.g. I$_{\kappa}$B Kinase
I I$_{\kappa}$B
NF$_{\kappa}$B translocates to nucleus
NF
Nucleus: NFκB and JAK/STAT pathways lead to the early immune response: Type 1 IFNs and Chemokines are secreted
Early Inflammatory Mediators

Figure 1.2: Intracellular immune response to RSV infection: RSV entry induces ROS, which induces STAT and causes increased interferon (IFN) and chemokine transcription. NFκB is activated via IκB upregulation and also increases IFN and chemokine transcription. NS1 and NS2 RSV proteins inhibit IRF3 and IFN production. Figure taken from Openshaw *et al.* (13).

However, RSV encodes a number of mechanisms to counteract host immunity. NS1 and NS2 proteins act in a species-specific manner to partially block the early antiviral response by impairing signalling of the type I interferon pathway. Different levels of the cellular IFN pathway are hampered: TRAF3, STAT2 and IKKε are decreased by either or both NS1 and NS2 proteins (22).

While most respiratory viruses primarily induce Th1-biased immunity, impairment of type I interferons by RSV may alter the balance of Th1 and Th2 responses (23). When a Th2 response is dominant over the Th1 response, it causes higher levels of IL-4, IL-5 and IL-13, more inflammation and more mucus production.

Once CD8⁺ T cells arrive on the scene, a decrease in viral load is noted. Nevertheless, memory CD8⁺ T cells in the blood were poorly correlated with protection, possibly due to limited functionality or a short life span of these cells. In the airways, epitope-specific resident memory CD8⁺ T cells were correlated with reduction in disease severity (24).

The activation of cellular immunity is not without problems either. The N protein of RSV has been shown to interfere with antigen presenting cells and formation of the immunological synapse, thus potentially impairing cell-mediated immunity (25). Furthermore, the G protein contains a fractalkine-like motif that may alter immune cell chemotaxis and inflammation. Dysregulation of the immune response by RSV is thought to contribute to immunopathology by other cell types including cytotoxic CD8⁺ T cells that help with viral clearance, but can also enhance disease, as well as other innate cells, including NK cells and γδ T cells (Figure 1.3) (13).

Once B cells are involved, viral clearance should speed up, but it might also be affected when specific neonatal B cells are infected and the immune system is inhibited (15). Furthermore, impaired IgA responses to RSV infection have been discovered, which might explain why recurrent infection are so common (26).

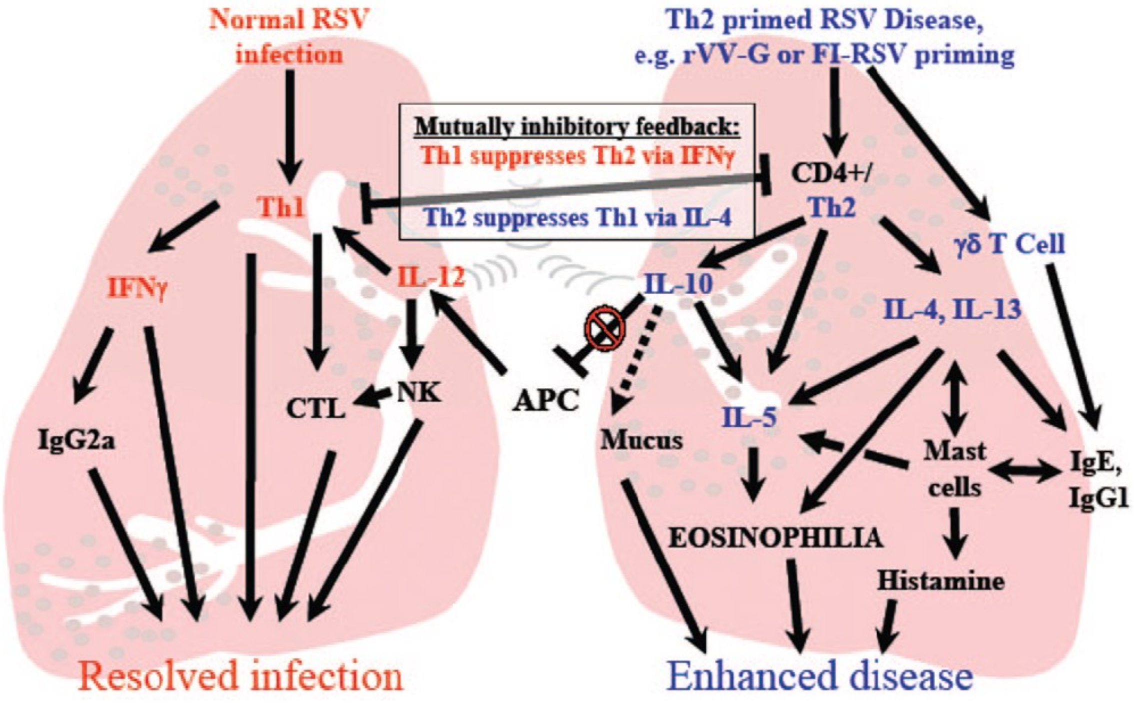


Figure 1.3: Interactions of T cells and cytokines during RSV infection. When the immune system exhibits a Th1 response, RSV disease is mild or asymptomatic. Inhibition of interferons by NS1 and NS2 proteins of RSV cause a shift to Th2 response and enhanced disease. Figure taken from Openshaw *et al.* (13).

1.1.2.2. Risk factors for severe disease and persistent infection

Several risk factors have been identified that are linked to higher risks of getting infection, but also of getting more severe infections. Risk factors for severe RSV disease in children include living with school aged siblings, birth order, exposure to tobacco smoke (either while pregnant or after birth), being younger than 3 months old at the onset of the RSV season, low birth weight and male sex (27-29).

Mice infected very early in life develop more severe disease and develop worse disease when re-infected later in life, suggesting that the age of first onset may have an imprinting effect on host immunity (14). Other factors affecting disease severity include polymorphisms in genes for cytokines and chemokines or their receptors, such as IL-13, RANTES (CCL5), CCR5 and TLR4 (30-32). An extensive genome-wide association study has identified several polymorphisms in genes that might predispose to bronchiolitis (33). It has been shown that polymorphisms in human genes can interfere with the formation of protein complexes in different ways. The associated phenotype can be located at an interaction site affecting protein binding or it can affect monomer stability (34). Polymorphisms in immunity-related genes can affect the immune responses when it encounters infectious agents, like RSV.

Other risk factors in children include differences in the baseline production of cytokines and chemokines like TNFα, interferon-α/β and interferon-γ, Th2 cytokines, IL-6 and IL-8 (30). It is therefore believed that a major contributor to disease is the host's immune response (18, 26). Nevertheless, in experimentally infected adult volunteers, IL-6 and IL-8 levels correlate with viral load, suggesting that interaction with the virus is also critical (35). Furthermore, immunocompromised patients also develop severe disease, which implies host immunity is not the sole determinant of severe disease and that different disease mechanisms may exist in immunosuppressed versus immunocompetent individuals (13).

1.1.3. Virology

1.1.3.1. Taxonomy and structure

According to the most recent release of the International Committee of Virus Taxonomy, the official species name of RSV is *Human orthopneumovirus*. This species is part of the Genus *Orthopneumovirus*, Family of *Pneumoviridae*, Order of *Mononegavirales*, Class of *Monjiviricetes*, Subphylum of *Haploviricotina*, Phylum of *Negamaviricota*, which is part of the Realm of *Riboviria*. This last one is the only realm currently found in the taxonomy of viruses, but the taxonomy is being expanded quickly and yearly updates are published online (36).

RSV was first isolated in 1956 from chimpanzees and was initially called the Chimpanzee Coryza Agent (37). It is a negative-sense, single stranded RNA virus with a non-segmented genome. The length of the genome is about 15,200 nucleotides (nt) long, depending on the subtype (A or B) and genotype. It codes for 11 proteins in the following order: 2 non-structural proteins (NS1 and NS2), nucleoprotein (N), phosphoprotein (P), matrix protein (M), small hydrophobic protein (SH), glycoprotein (G), fusion protein (F), M2 ORF 1 protein (M2-1), M2 ORF 2 protein (M2-2) and large protein or polymerase (L) (Figure 1.4).

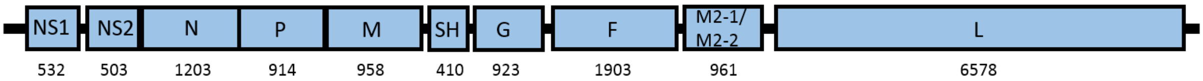


Figure 1.4: Genome of RSV and protein length in nucleotides for the A2 strain with a total genome length of 15,222 nucleotides. NS = non-structural protein; N = nucleoprotein; P = phosphoprotein; M = matrix protein; SH = small hydrophobic protein; G = glycoprotein; F = fusion protein; L = polymerase.

1.1.3.2. Protein structure and function

The structure of NS1 has only recently been discovered and it showed that NS1 contains unique regions that are responsible for the altered host responses in a number of ways(38).

NS1 and NS2 interfere with the host immune response in a number of ways, including by inhibiting the production of and signalling by type I interferons on many levels (21, 22, 39, 40).

NS1 also interferes with glucocorticoid treatment. The activation of the GR complex by glucocorticoid steroids does not lead to increased anti-inflammatory gene expression to reduce the inflammation response caused by RSV infection (41). Furthermore, NS1 modifies miR-24 expression via TGFβ (42) and facilitates replication through miR29a-mediated inhibition of the IFNα receptor (43). In fact, several miRNAs have been shown to be up- or downregulated in mice due to RSV vaccination both at priming and boosting, and later upon challenge as well (44).

NS2 is the cause for detachment of infected cells which reduces viral load, but it also does induce obstruction of bronchioles and contributes to disease severity in this way (45). Moreover, it is responsible for the hijacking of the ubiquitin system to use it to the virus' advantage by labelling STAT2 for degradation (46).

Recombinant viruses lacking NS1 and NS2 showed improved levels of IFNα and IFNβ, but it also showed that NS1 and NS2 function independently of each other, because viruses lacking only one of the proteins still showed reduced IFN levels (20). These two proteins are masters in immune dysregulation upon infection of human cells and more ways of immune suppression are still being discovered today.

There are three proteins expressed on the surface of the RSV virion: G, F and SH. G is responsible for viral attachment and is assisted by F which mediates cell membrane fusion. G is the most variable gene of RSV and codes for a type II glycoprotein. G is anchored into the virion membrane with a hydrophobic domain, although it has been shown that soluble G can be generated (47), which lacks the first part of the protein as it starts translation at the methionine located at position 48 in the middle of the anchor region. The signal peptide, which is located at the N-terminus, and anchor domain are cleaved off after translation (48). The ectodomain of G is heavily glycosylated with both N- and O-linked oligosaccharides. The 32 kDa protein can carry up to 40 O-linked glycans and 5 N-glycans, which means that about 60% of the G protein's molecular mass is carbohydrate (49, 50). These carbohydrates are necessary for its function (51). The middle of the ectodomain contains four cysteine residues that form a loop and is thought to be the binding site of G(52). Most of the variation in G is found in the two protein segments flanking the four cysteine residues, which are the two mucin-like regions that carry the carbohydrate groups.

Both G and F are a major target for antibody-mediated immunity. Human antibodies recognise several epitopes of G. These epitopes can be divided into three types: conserved epitopes that are present in all RSV strains, group-specific epitopes that are only found within the same antigenic group, and strains-specific epitopes that are only found in certain strains of the same antigenic group and are mostly found in the second hypervariable region of G (53). First, only conserved epitopes were discovered. These target the central conserved region of G (54), however, in 1997, Cane *et al.* showed that the immune system also expresses antibodies again several epitopes against the C-terminal end of G, which contains potential N-glycosylation sites (55). These sites can differ between genotypes and glycosylation is dependent on the cell type that is infected (56). This suggests RSV can overcome immune pressure against G by allowing variation at certain positions and the presence of glycosylation sites point to another mechanism of defence against the immune system. There are theories suggesting variation in the G protein is antibody-driven and therefore influences RSV evolution, however, Trento *et al.* could not find any experimental support for this theory (57).

F is a lot more conserved than G, but amino acid substitutions are still observed in all six antigenic sites, the signal peptide, p27, heptad repeat domain 2 and transmembrane domain (58, 59).

For F to be activated, it has to form a trimer, which is only possible after p27 is cleaved out (60). The crystal structure of this trimeric conformation was determined and showed the preservation of certain neutralizing epitopes after fusion (61). The structure of F in the pre-fusion and post-fusion conformation showed to be different and thus has to be taken into account when designing vaccines (62, 63). The function of SH is currently still unknown. Neither SH nor G are strictly necessary for viral replication *in vitro*, although lack of these genes does decrease effectivity (64).

There are four proteins associated with the nucleocapsid and replication: N, P, M2-1 and L. These proteins direct RNA replication and transcription after infection of the host cell. The P and N proteins are necessary (and sufficient) to form inclusion bodies where replication takes place (65). P and L form a complex in those inclusion bodies to replicate the RSV genome. P is also necessary to recruit M2-1 to the inclusion bodies. Then, P hijacks PP1 to dephosphorylate and thus activate M2-1. Disruption of that hijacking manoeuvre impairs viral transcription (66).

In those inclusion bodies, a specific region on L, the conserved region V (CRV), is part of the capping domain of L. Mutations in this domain result in shortened transcripts due to the termination of many transcripts, because they have not been capped. Even more so, those same mutations also cause partial or even completely defect RNA replication. The RNA is not elongated within the promotor region (67). M2-2 is thought to be a regulatory factor that decides when to change from RNA transcription to RNA replication for virion production (68).

The M protein is also located in the inclusion bodies for viral transcription, but only when M2-1 proteins are present (65). It is thought to have a function in organising the envelope (Figure 1.5).

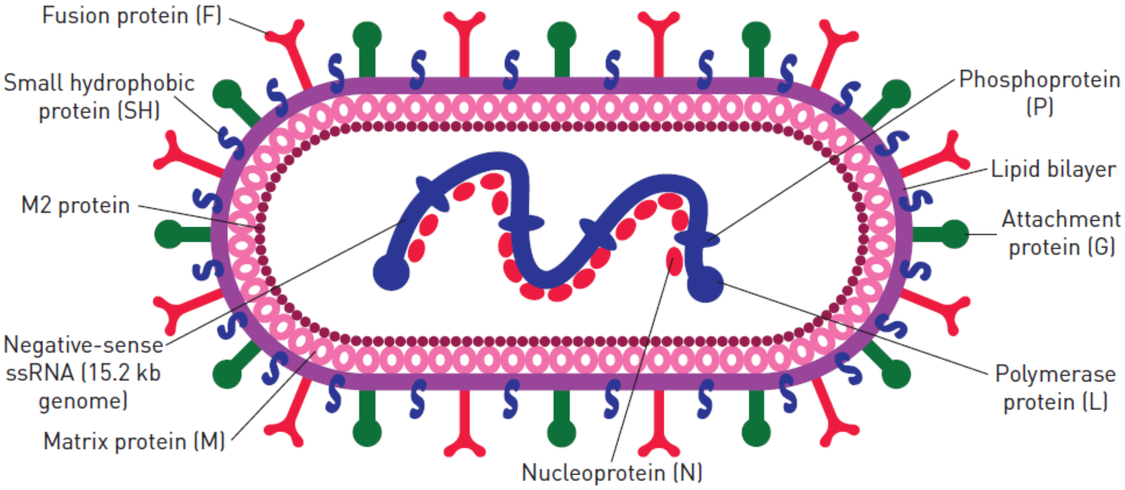


Figure 1.5: RSV virion with all proteins indicated where they are located. G, F and SH are found on the outside of the virion. M and M2 are thought to be important for envelop development. P, N and L are involved with genomic replication. NS1 and NS2 (not shown here) are mainly involved in immune dysregulation. Ss = single-stranded. Figure taken from Jha *et al.* (12).

1.1.3.3. Genotypes

The evolutionary rate of the entire RSV A genome is 6.47 x 10^{-4} substitutions/site/year (69). However, the G gene has higher evolutionary rates, namely 3.58 x 10^{-3} nt substitutions/site/year in the ectodomain (70), which corresponds to the greater variability in this gene.

Variability of RSV has been detected in several ways in the past and was and is mostly focused on G and F. It was first described at an antigenic level by Coates *et al.* in 1966 by performing a plaque reduction neutralization test (71). Almost 20 years later, this variability of RSV was demonstrated again by dividing the virus into two different groups, namely RSV A and RSV B, based on serological differences (72-74). Within a couple of years, it became apparent that serological differences could also be detected with monoclonal antibodies within those two groups of RSV (75-77). The variation that was noticed with monoclonal antibody reactivity was also confirmed by restriction mapping (78, 79), which was more detailed and could show more patterns than monoclonal antibody studies could (80).

Sequencing confirmed these findings and more systems of grouping strains emerged (78, 79, 81).

In 1998, Peret *et al.* started the currently most used genotyping system based on the G gene (82) It used a 270 nucleotide long region in particular, the second highly variable region (HVR2), of the G gene located at the C-terminus (Figure 1.6).

[redacted]

Figure 1.6: Highly variable regions (HVR) of the G gene of RSV are separated by a more conserved region. HVR2 is used for the classification of RSV strains into genotypes. Based on a figure from Trento *et al*. (83)

The first 5 genotypes of RSV A that were identified were named GA1 - GA5 and the first 4 genotypes for RSV B, GB1 - GB4. Since then, more genotypes have been characterised (Table 1.1).

Grouping RSV strains has helped determining which strains are common and showed that the same lineages can be found all over the world, but also that multiple lineages can be detected during the same epidemic and recur during following epidemics (78, 81, 84). The G gene of RSV strains of the same lineage does acquire amino acid changes over time, which indicates genetic evolution that might be related to immune selection (70, 84, 85).

For example, most recently, the BA and ON strains have been notable, with 60-nucleotide and 72-nucleotide duplications in G respectively. Since their emergence, they have rapidly spread all over the world, suggesting some evolutionary advantage although it has not yet been fully unravelled (83, 86, 87).

Venter *et al.* proposed that a genotype should consist of sequences clustering together with bootstrap values of 70-100% for well-supported nodes and a p-distance of less than 0.07 (proportion of different amino acids) to all other sequences in the same cluster (88). Over ten years later, Trento *et al.* used the GA1 genotype as an example to determine the criteria for new genotypes. GA1 includes some of the oldest RSV strains including laboratory strains such as A2 and Long, and within this group the intragenotypic p-distance is never higher than 0.049. They therefore used this as a minimal cut-off to divide new strains into genotypes (57).

However, not everyone has followed this last rule and using this criterion, several recognised genotypes should be considered to be one and the same. For RSV B, there are a number of different genotypes, which are in reality minimally different.

Questions arose as to what should be included in a genotype. For example, the ON1 strain and NA1 strain are very similar apart from the 72 nt duplication. Nevertheless, there are some amino acid substitutions that can be found as differences between these two genotypes. Six of them are in the G protein, one substitution is located in the P protein, one in F and two in L (89). Should these be part of the genotyping system?

It has long been the rule that HVR2 should be considered the minimum necessary region for genotyping, however, recent research showed that 20% of strains should be considered to be exactly the same based on the HVR2 only, even though these are part of 5 different genotypes, namely NA1, ON1, GA5, BA9 and BA10 (90). These are all still prevalent and even common genotypes today, so only the HVR2 region should no longer be considered the minimum necessary region for genotyping strains.

Table 1.1: Genotypes of RSV A and RSV B in the currently used genotyping system. Depending on the criteria used, some genotypes are not considered new genotypes.

SUBGROUP	GENOTYPE	YEAR OF DISCOVERY	PLACE OF DISCOVERY	COMMENTS	REFERENCE
RSV A					
	GA1 – GA7	1998, 2000	Rochester, USA		(82, 91)
	SAA1 – SAA2	2001, 2012	South Africa		(88, 92)
	NA1 – NA4	2009	Niigata, Japan and Beijing, China	Considered part of GA2 depending on criteria	(93, 94)
	ON1 – ON2	2012, 2014	Ontario, Canada	Considered part of GA2 depending on criteria, **contains 72 nt duplication**	(86, 95)
	TN1 – TN2	2016	Tennessee, USA	Part of ON1	(96)
RSV B					
	GB1 – GB4	1998	Rochester, USA		(82)
	BA1 – BA10	2003, 2006, 2010	Buenos Aires, Argentina and Niigata, Japan	Come all together in one genotype BA, **contains 60 nt duplication**	(83, 87, 97, 98)
	CB1	2013	Beijing, China		(94)
	BA-C	2013	Beijing, China	Considered part of BA	(94)
	SAB1 – SAB4	2001, 2011	South Africa, Cambodia		(88, 99)
	URU1 – URU2	2005	Uruguay		(100)

1.1.3.4. Epidemiology

It was shown in 1991 by Cane *et al.* that genetically different lineages are found during the same epidemic, that these lineages can reappear in later years and that these lineages can be found all over the world (78, 81). Genetic variation accumulates in the viral genome and is thought to be driven by immune selection (85). Different lineages might exist now because of evolutionary survival of certain genotypes, while others disappeared (84).

Numerous papers have described the prevalence of genotypes worldwide over de last decade. Two genotypes are of particular interest as they have developed a duplication in the G gene that seems to benefit the virus (87, 101-103). One of the best tracked genotypes is an RSV A genotype, namely ON1. It was first discovered in 2012 in Ontario, Canada (86) and subsequently found all over the world; in Germany (104), Kenya (105), Japan (106), Thailand (107), and many more. Comas-García *et al.* report that a 72-nucleotide duplication was found in samples from Central Mexico from 2009 and speculate that there might be at least three independent duplication events in the G gene of a GA2 strain (108). However, this was refuted by Furuse *et al.* who think that the polymorphisms in the duplicated region (used as an argument by Comas-Garcías *et al.)* might have developed under positive selection pressure (109).

The number of ON1 strains expanded quickly in the years following its discovery and genetic variation has been detected in the duplicated region as well (110). The rapid replacement of most other GA2 strains indicates there might be a fitness advantage of ON1 compared to GA2. Luckily, there is no clear evidence found of altered pathogenicity of ON1 compared to its ancestor GA2 (111).

Even though ON1 has spread all over the world, it is not introduced numerous times at the same location, but rather is introduced a minimal amount of times and then spreads locally picking up variation over time. This was shown in Kilifi, Kenya where more local strains rather than global strains (112) were detected and once a new strain was introduced in the household, this strain spread and developed variations within the household (113).

A similar effect has been seen with the RSV B genotype BA that has a 60-nucleotide duplication and was discovered in Buenos Aires, Argentina (83). BA strains have replaced most other RSV B strains in the years following its discovery and can now be found on all continents (92, 114, 115). Within the group of BA strains multiple lineages have been described (see above) of which the BA9 genotype is the most common. The estimated time to the most recent common ancestor of the BA group is 1995. For BA9 specifically, this is estimated to be 2000 (116). More detailed studies revealed several sites under positive selection as well as sites under negative selection (116).

The genetic differences in the viral genome are driven by the immune system and have shaped the genotype landscape as it is today. It should be noted that most epidemiology studies are focused on the G gene and to a lesser extend F. However, it has been shown, at least for RSV A strain ON1, that other genes also cluster distinctly compared to GA2. This indicates that other genes can also play a role in RSV adaptation to its host (117). It suggests that all kinds of genetic variability can have an evolutionary and epidemiological impact and perhaps a clinical impact as well.

1.1.3.5. Viral sequence variation and clinical impact

The evolutionary rate of RSV is approximately 6.47 x 10^{-4} substitutions/site/year (69). The variability of RSV and particularly of G is affected by the immune system as it is partially driven by immune pressure (both positive and negative selection occurs), which also affects antigenic variability. This in turn influences epidemiology and evolution of RSV.

So far, most sequencing efforts have gone towards the G gene, which has a higher evolutionary rate than the rest of the viral genome, to establish which subgroup and genotype is most prevalent in an epidemic. These differences might cause variability in disease severity, just like certain polymorphisms in human genes can have an effect.

Several viruses are already known to have altered virulence caused by single amino acid substitutions in different kinds of viral proteins (118). Several pandemic influenza strains have altered virulence due to amino acid substitutions, for example, one of the virulence factors of the notorious 1918 Spanish flu was unravelled and shown to come down to the Glu-Ser-Glu-Val amino acid sequence at the C-terminal end of NS1 protein (119). The 2009 pandemic H1N1 virus on the other hand had a single E47K substitution in the HA protein, which affected its infectivity by making it a more stable virus with increased viral fusion capabilities (120). Even the Ebola virus strain that caused an enormous outbreak in West Africa killing tens of thousands of people between 2014 and 2016 was shown to have increased mortality rates compared to known strains due to an a single A82V substitution in the glycoprotein (121, 122).

There has been some research to investigate the clinical impact of protein deletions, specific sequence variations of RSV in humans (123) and whether certain genotypes cause more severe disease or not. Some studies have suggested that RSV A causes more severe disease (124-128), although not all studies agree (129). Some studies are even more specific and suggest NA1 specifically causes more severe disease (130).

In vitro

Several *in vitro* studies have shown that sequence variation alters the infectivity, replication rate, viral attachment and binding affinity. After growth in Vero cells, Kwilas *et al.* noted a truncation in the G protein. Virus with this truncated G protein was shown to be 600-fold less effective in infecting cultured human airway epithelial cells (131). Conversely, comparison of BA strains with the 60-nucleotide duplication and recombinant viruses without the duplication, showed that the duplicated region increased virus attachment (96, 132).

Another study comparing RSV A strain A2 and clinical strain A2001/2-20 (2-20) suggested that the G protein of RSV 2-20 had a greater binding activity than the G protein from A2 both in the clinical virus 2-20 and a chimeric A2 virus engineered to express the 2-20 G protein instead of its own (133).

Culturing a recombinant RSV strain without the M2-2 gene resulted in four mutations. The K66E mutation in the F gene appeared to cause differences in growth and cytopathic effect in Vero cells (134).

Moore *et al.* also performed several studies *in vitro* showing that viral sequence variation in F influences viral attachment, fusogenicity and mucus production. During a study in 2014, they identified five mutations in the F gene of RSV line19 by comparing the sequence with strains A2 and Long. By engineering chimeric viruses, two of those mutations were shown to increase fusogenicity when both were present and two other mutations increased mucus induction (135).

However, results *in vitro* should always be interpreted with care, especially when working with cell lines compared to primary cells. Besides the lack of an immune system, it was also shown that the expression profiles of cell lines differ greatly and this affects viral replication and even virion assembly. For example, Vero cell lines are based on kidney cells from an African green monkey and showed to have a 100-fold higher activity of the protease cathepsin L (compared to human HeLa cells derived from cervical cancer). Cathepsin L cleaves the G protein of RSV after infection and thus affects infectivity (136).

In vivo

Some of these studies show that these effects are also seen *in vivo*. Whitehead *et al.* created a chimeric virus based on the A2 strain with G and F proteins from the B1 strain. They discovered several attenuating mutations and designed a virus that carried three of those mutations, which was highly attenuated in both upper and lower airways of chimpanzees (137). Five mutations were detected in F and L proteins, which attenuated the virus as well (138). These mutations were used to create a recombinant virus that could be used to infect chimpanzees and establish the mutation effects. The virus showed less efficient replication both in upper and lower respiratory tract and chimpanzees displayed less rhinorrhoea and coughing (139).

Further experiments in chimpanzees showed that deletion of entire proteins like NS1 or M2-2 only decreased viral replication by 10-fold or even just delayed it. However, combining the deletion of these two proteins resulted in 2,200 – 55,000-fold decrease in replication (140). Deletion of NS2 and SH proteins also resulted in less replication in the lower respiratory tract and less rhinorrhoea in chimpanzees (141). These efforts were aimed at creating an attenuated virus that could be used for further vaccine research and development. In the meantime, it showed that individual mutations in certain proteins do affect its replication and immunogenicity features. Current research does no longer involve chimpanzees, but rather mice, rats and ferrets.

Infection in BALB/c mice showed that infection with the line19 strain increased lung IL-13 and mucus expression compared to infection with strain A2 and Long (142). Infection with a recombinant RSV strain A2 carrying the line19 F gene, showed that same increase in lung IL-13 and mucus expression in BALB/c mice (143). When comparing the levels of lung IL-13 and gob-5 along with disease severity between the clinical strain 2-20, strains A2, Long, line19 and another clinical strain A2001/3-12, 2-20 caused more severe disease and higher levels of IL-13 and gob-5 (144).

Further studies indicated an increase in necrotic damage in the airways and higher neutrophil infiltration in BALB/cJ mice when infected with strain 2-20 compared to infection with strain A2 (145). All these studies indicate that sequence variation in the viral genome can have a clinical impact either by directly altering the fitness of the virus or by influencing the immune response.

1.1.4. Treatments and vaccines

To date, there are no licensed vaccines or effective treatments for RSV. Ribavirin has been given sporadically, but due to toxicity, cost and lack of strong evidence for efficacy, it is only used in select patient populations (146). Nevertheless, it has been shown in one study that Ribavirin increases the amount of mutations in the RSV genome due to decreased precision of the L protein, and a decrease in functional virus production (147).

Treatment with glucocorticoid receptors have shown very little relieve from inflammation and the cause of that was studied recently. The study showed that the NS1 protein can interact with IPO13, which is necessary for nuclear translocation of the glucocorticoid receptor (GR). Since NS1 can block the interaction of IPO13 with the GR, its activation does not lead to increased anti-inflammatory gene expression to reduce inflammation caused by RSV infection (41).

It has been suggested that therapies should aim to reduce viral load, for example by inhibiting ARP2 in the airways. It decreases replication of RSV at later stages of the infection due to reduced spread to neighbouring cells as ARP2 is thought to be necessary for filopodia formation which shuttles RSV to neighbouring cells (148). Unfortunately, there are still no RSV specific treatments available to reduce viral load.

The only FDA-approved prophylactic for RSV is the monoclonal antibody palivizumab, which reduces hospitalisation of premature infants, but is only effective when given in monthly injections as a preventative measure (149). Although no vaccines are yet available, recent advances in our understanding of RSV epidemiology and antigenic structures mean that there are now nearly 50 candidates in development (Figure 1.7) (150-152).

Nevertheless, a number of potential problems remain including the phenomenon of vaccine-enhanced disease that was caused by formalin-inactivated RSV trialled in the 1960s, where more severe disease developed in children who received the vaccine than in children who did not (18, 153). RSV vaccine development has also been problematic as natural infection causes only transient and partial immunity to reinfection, meaning that correlates of protection remain obscure and vaccines have to overcome that transient and partial protection to be approved (16). In a study of experimentally infected adult volunteers, Jozwik *et al.* noted reduced memory CD8$^+$ T cell functionality compared to Influenza (24), while the IgA memory response also showed evidence of impairment, both of which might contribute to recurrent infections (26).

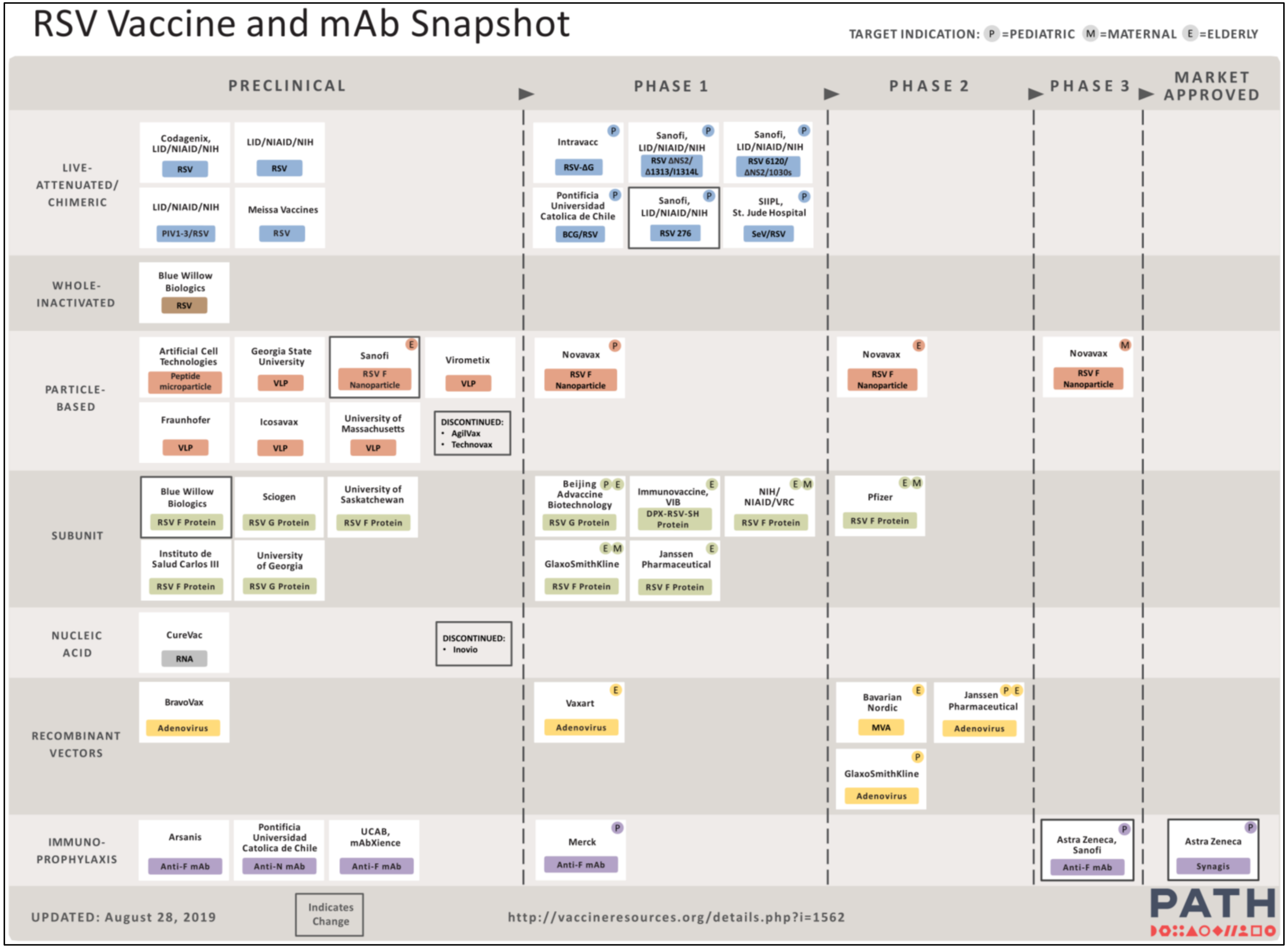
Treatment and vaccines are based on our current knowledge on antigenic epitopes and neutralizing antibodies. It has been shown that D25 antibody neutralises both RSV A and B, while 5C4 preferentially neutralises RSV A (154). RSV antibody repertoires have been examined and showed that certain antibodies bound 100 times more potently to pre-F than palivizumab (155). These types of developments should be taken into account when new vaccines are designed.

The F protein is the most targeted protein for vaccine development, although antigenic epitopes have been detected for all proteins. Nonetheless, the proteins with the most detected epitopes were F (37%) and G (35%), while NS1, SH and P only produced 1% of epitopes each (156). Compared to G, F is much more conserved and therefore a more logical target for vaccine development. Yet, amino acid substitutions have been observed in all antigenic sites of F before (58, 59).

All kinds of vaccines are in trial phases right now (Figure 1.7), like the subunit vaccine from GalaxoSmithKline which is a maternal vaccine in phase II trial (157). Novavax was also trialling a maternal vaccine, which is a nanoparticle-based vaccine, however, it recently failed in their second phase III trial (79). All live-attenuated vaccines are still in phase I. An intranasal, subunit vaccine with replication deficient RSV has been trialled as well and showed persistent antibody responses in phase I trial (158).

None of these have proven to be effective enough yet. Perhaps, defective particles might be of use for vaccine development as it has been demonstrated that defective, non-replicating particles can activate a strong immune response against RSV (159).

Regardless of all the bumps in the road, it is clear that working vaccines are necessary to prevent hospitalisations and deaths caused by this virus. Even when young infants cannot be vaccinated, maternal vaccination could produce antibodies and pass these on to the baby. Spreading in the community could be reduced, which would protect new-borns from getting in contact with the virus at an age where they cannot fight this infection yet. It would also reduce the variability seen in the population as it was shown that usually one variant was introduced into the household and following that, new variations developed which could then be spread into the community (113). This increase in variation in the population makes it more difficult to find a vaccine that can prevent infection and disease and increases the chances of developing more virulent, more infectious and more dangerous strains.



RSV Vaccine and mAb Snapshot

TARGET INDICATION: P =PEDIATRIC M =MATERNAL E =ELDERLY

	PRECLINICAL	PHASE 1	PHASE 2	PHASE 3	MARKET APPROVED
LIVE-ATTENUATED/ CHIMERIC	Codagenix, LID/NIAID/NIH · RSV LID/NIAID/NIH · RSV LID/NIAID/NIH · PIV1-3/RSV Meissa Vaccines · RSV	Intravacc · RSV-ΔG · P Sanofi, LID/NIAID/NIH · RSV ΔNS2/ Δ1313/I1314L · P Sanofi, LID/NIAID/NIH · RSV 6120/ ΔNS2/1030s · P Pontificia Universidad Catolica de Chile · BCG/RSV · P Sanofi, LID/NIAID/NIH · RSV 276 · P SIIPL, St. Jude Hospital · SeV/RSV · P			
WHOLE-INACTIVATED	Blue Willow Biologics · RSV				
PARTICLE-BASED	Artificial Cell Technologies · Peptide microparticle Georgia State University · VLP Sanofi · RSV F Nanoparticle · E Virometix · VLP Fraunhofer · VLP Icosavax · VLP University of Massachusetts · VLP DISCONTINUED: AgilVax, Technovax	Novavax · RSV F Nanoparticle · P	Novavax · RSV F Nanoparticle · E	Novavax · RSV F Nanoparticle · M	
SUBUNIT	Blue Willow Biologics · RSV F Protein Sciogen · RSV G Protein University of Saskatchewan · RSV F Protein Instituto de Salud Carlos III · RSV F Protein University of Georgia · RSV G Protein	Beijing Advaccine Biotechnology · RSV G Protein · P E Immunovaccine, VIB · DPX-RSV-SH Protein · E NIH/ NIAID/VRC · RSV F Protein · E M GlaxoSmithKline · RSV F Protein · E M Janssen Pharmaceutical · RSV F Protein · E	Pfizer · RSV F Protein · E M		
NUCLEIC ACID	CureVac · RNA DISCONTINUED: Inovio				
RECOMBINANT VECTORS	BravoVax · Adenovirus	Vaxart · Adenovirus · E	Bavarian Nordic · MVA · E Janssen Pharmaceutical · Adenovirus · P E GlaxoSmithKline · Adenovirus · P		
IMMUNO-PROPHYLAXIS	Arsanis · Anti-F mAb Pontificia Universidad Catolica de Chile · Anti-N mAb UCAB, mAbXience · Anti-F mAb	Merck · Anti-F mAb · P		Astra Zeneca, Sanofi · Anti-F mAb · P	Astra Zeneca · Synagis · P

UPDATED: August 28, 2019

Indicates Change

http://vaccineresources.org/details.php?i=1562

Figure 1.7: Snapshot of RSV vaccines and monoclonal antibodies which are in clinical trials on August 28th, 2019. Taken from VaccineResources.org (160).

1.2. Sequencing

Variation in RSV strains was detected by studying antigenic variability way before sequencing viral RNA was an option (71). Serology experiments showed a clear distinction between two groups of RSV now known as RSV A and RSV B (72, 73). More details became apparent by using monoclonal antibodies and restriction mapping (76, 78). The importance of those differences was already highlighted in 1991 by comparing antigenically similar and different strains and studying them over time. Similar antigenic strains can cause epidemics in multiple seasons (78), which could be explained by immune driven accumulation of genetic alterations that cause evolutionary survival and adaptation to the human population (84, 85).

However, nowadays, exact changes can be detected by sequencing the virus in more detail. When the G gene varies, this can affect the RSV phenotype and consequently the viral protein function and its recognition by the host immune system.

Sequencing a genome means that the order of nucleotides in the DNA is determined. However, the Watson and Crick model of DNA was only proposed in 1953, which is not that long ago and just three years before RSV was discovered. In 1977, Maxam and Gilbert published their first DNA sequencing method (161), which worked by modifying the DNA chemically and then cleaving the DNA at specific bases. By separating the fragments on a gel, the length of the fragment reveals the base at that position.

1.2.1. Sanger sequencing

1.2.1.1. Technology

In that same year, a more known method was published and refined soon after to become the more favoured method: Sanger sequencing (162). This method used a lot less chemicals and was easier to work with. By adding dideoxynucleoside triphosphates (ddNTP) during DNA elongation by the DNA polymerase, the elongation was stopped when ddNTPs were incorporated and, similarly to Maxam and Gilbert's method, the fragments could then be separated on a gel to find out which base was located at each position. This method was later optimised by using fluorescent dyes attached to each ddNTP. These would be cleaved when incorporated so that each ddNTP lit up in another colour on the gel and only one reaction was necessary to read all nucleotides of the DNA. This is referred to as dye-terminator sequencing and is currently most often used, albeit more automated since further advances in 1991 (163).

The details of the technology are based on the fact that DNA is built by adding one deoxynucleoside triphosphate (dNTP) at a time. When a ddNTP is incorporated, there is no –OH group to which the next dNTP can be attached and therefore the elongation of the DNA strand stops (Figure 1.8). In the earlier methods, the same sample was treated in 4 different reactions with each containing a proportion of ddNTPs of one of the four nucleobases, adenine (A), cytosine (C), guanine (G) or thymine (T). By separating the DNA fragments of different sizes on a gel, it is clear where a ddNTP of A, C, G or T was incorporated and lining up the reactions next to each other showed the DNA sequence (Figure 1.8). This was a very tedious job and only small fragments were sequenced. By attaching fluorescent dyes to the ddNTPs, all ddNTPs could be used in the same reaction and separated on one lane on the gel. This method is still used, but now the light of the fluorophores is measured automatically in the sequencer and a chromatogram shows how strong the emitted light is at each position.

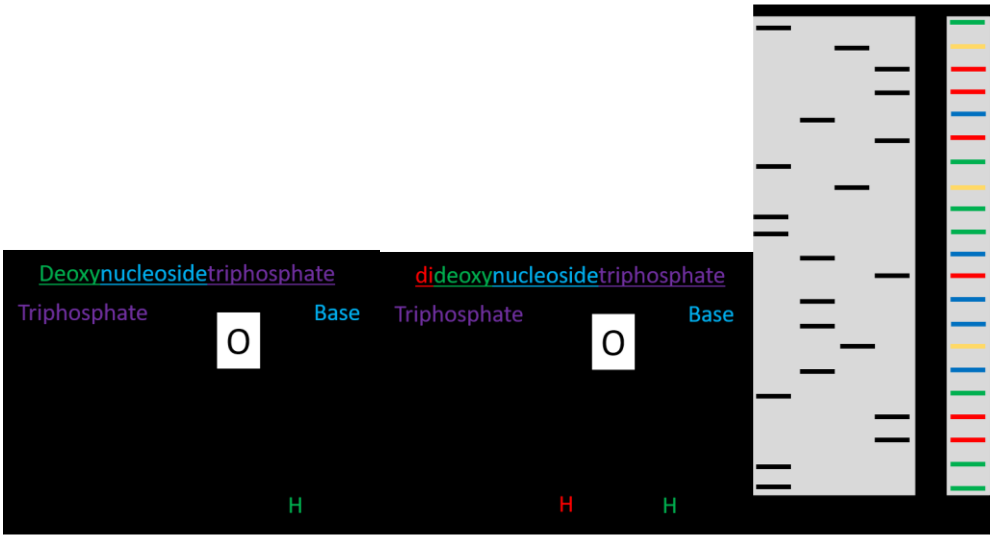


Figure 1.8: In a regular elongation reaction, deoxynucleosidetriphosphates are incorporated. When dideoxynucleosidetriphosphates are incorporated, the elongation terminates as the next triphosphate group cannot bind the -OH group. The fragments can be separated by size on a gel and sequences can be read from those gels.

With current technologies, the sequence can be extracted in a fasta file, which contains sequences of variable sizes. Regularly, these sequences have to be combined into a longer, more complete sequence, which can be done automatically by several programs. Once the complete sequence is assembled, it can be saved in a fasta file and further research can be carried out with this information.

1.2.1.2. History

In 1995, the first whole genome of an organism was sequenced, namely from *Haemophilus influenza* (164). In 1999, for the first time, a whole human chromosome was sequenced, chromosome 22 (165). Two years later, the entire human genome was sequenced (166). This was an amazing accomplishment. Shortly after, the 1000 Human Genomes Project set off and in 2010 they announced the results of their trial phase (167) and two years later they achieved the goal of the project and had sequenced >1000 human genomes (69).

Immediately after, a follow-up project was announced, the 100,000 genomes project. In December 2018, less than 20 years after the first human genome was sequenced, a milestone was reached when the 100,000th human genome was sequenced for this project.

During this time span, new techniques have been developed for sequencing. One of the 'Next-generation sequencing' methods that is commonly used is Illumina sequencing.

1.2.2. Illumina MiSeq sequencing

1.2.2.1. History

Illumina sequencing uses a different method, which was invented by Balasubramanian and Klenerman at Cambridge, who founded the company Solexa in 1998. In 2006, their first sequencer, Genome Analyzer, was launched. Only one year later, Illumina acquired the company and used the technology to produce more powerful sequencers time and again. One of those sequencers is Illumina MiSeq, which is now frequently used worldwide.

1.2.2.2. Technology

Illumina sequencing is based on sequencing short reads (Figure 1.9). The sample DNA is first fragmented into small, 150 bp pieces and each of them is tagged with adapters that contain information on the sample that was sequenced by adding a barcode, but also primer sites for amplification and flow cell oligos that are complementary to the oligos that are attached to the flow cell. Once these adapters are added to each piece of DNA, the library is ready to be loaded onto the flow cell.

Each fragment binds to the flow cell by the complementary oligos that were added when the adapter was ligated to the fragment. Clusters of the same fragment are produced by several rounds of bridge amplification, which allows to form packed bundles of the same fragment and amplify the fluorescent signal. Next, each fragment in each cluster goes through sequencing cycles which produces fluorescent light signals that are detected and linked to which nucleotide is being incorporated. Unlike with Sanger sequencing, these nucleotides are not ddNTPs, so more nucleotides can be incorporated after the fluorescent signal has been released. This happens all over the flow cell until each fragment is sequenced. All that information is exported in a fastq file which can be used for further analysis.

Since each fragment is sequenced multiple times and all the information is kept and exported rather than combined before exporting. This allows each position to be sequenced multiple times and displays all this information. Minority variations can be detected this way, while that information would otherwise be considered background noise and would be lost. Therefore, this technique is also called deep-sequencing.

[redacted]

Figure 1.9: Flow chart of general steps of Illumina next-generation sequencing methods.

1.2.2.3. NGS to study RSV

NGS sequencing has been used to sequence RSV and to learn from the extra information that can be obtained from deep-sequencing. It has to be kept in mind that all the steps of the process towards obtaining a sequence can introduce errors. Everything from extraction, reverse transcription, amplification and sequencing itself might affect the outcome, however, with the right checks in place, deep-sequencing still provides us with a lot of information, like minority variant identification (168). It can also give information about drug resistance mutations (169) or a clear view of strains before vaccines enter the market after which the effect of vaccines on circulating strains can be unravelled (170).

A paper from 2017 described the presence of diversity in the RSV genome on the amino acid level in adults infected with RSV M37 and naturally infected infants (171). They noted one non-synonymous minority variant with a population >5% and they found that variant in 58% of their subjects. They found a lot of diversity in the samples following inoculation which was not contributable to viral load. The proteins containing the most variants were NS2 and M2. The naturally infected infants showed a lot less diversity than the adults and most strains belonged to the GA5 clade.

In 2014, a study was published on the effect of immune presence or absence on RSV genome diversity. A child with severe combined immune deficiency syndrome and a persistent RSV infection was examined and 26 upper airway samples were sequenced using Illumina HiSeq 2000 technology (172). Samples were taken over a period of 78 days, which was both before and after a bone marrow transplant. It showed that diversity of the RSV genome in the investigated samples increased significantly more after engraftment. This suggests that the immune system might drive viral diversity.

1.2.3. Possibilities of other NGS platforms

Sequencing technology is improving at lightning speed. At the moment, there are several third-generation sequencers available already. These technologies eliminate the need to puzzle small fragments of the genome together as they can produce long reads of several thousands of nucleotides in parallel. This solves the issues that arose concerning repetitive regions, however, the technology is not quite at its top yet as these long reads sequencers still produce increasing amounts of random errors with growing read lengths. These techniques have been around and are being optimised since 2008 (173).

PacBio have developed a long-read sequencer based on SMRT technology, which stands for Single Molecule Real Time sequencing, and can detect nucleotide incorporation based on dye colours. In 2019, they released a new SMRT cell, which can detect up to 8 million fragments in parallel and each fragment can be up to 50,000 nucleotides long.

Oxford Nanopore Technologies have produced MinION, which is an extremely small sequencer, which fits in your pocket and can be linked to a laptop via USB 3.0. As long as the laptop is powerful enough and has enough storage capacity, the sequencing can be executed anywhere, no laboratory required. This technology measures the difference in electrical field when DNA passes through a nanoscale pore, which is different for each nucleotide. Other technologies by Oxford Nanopore are GridION and PromethION, which are larger versions of MinION and can carry up to 144,000 nanopores and thus fragments. SmidgION is a mobile phone sequencer, which is even smaller than MinION and is currently under development.

These technologies can be used to sequence more samples cheaper and faster. Long reads require less assembly and less room for errors during analysis. The fact that they still introduce too many errors is not ideal for RNA viruses like RSV where the genomes evolve rapidly. Nevertheless, in 2018, Keller *et al.* managed to sequence the complete coding genome of Influenza A from its RNA (174). This opens the possibility of detecting sequences without having to convert RNA to cDNA first, which is a source of errors also and can cause loss of information we do not know about yet.

1.3. Big data analysis: how to find the relevant information in data?

1.3.1. Fasta/fastq files

Fasta files only contain sequence information of one or more sequences, while fastq file also carry information about the quality of each nucleotide at each position. If a cluster on the flow cell was not clearly readable as either one of the nucleotides, the quality score (based on ASCII coding) will be lower.

Quality scores that are added to the fastq files allow for detailed quality checks of each sequence individually and the whole sample generally. If a sample is not of good quality, there will not be a good quality sequence at the end of the run. But sometimes, certain reads did not amplify well enough to produce a clear signal. These can be excluded from analysis based on their quality scores

Fastq files contain sequences from the sample, but because of its processing to be able to sequence the sample, lots of extra information was added to each sequence. The reads contain adapters with barcodes, flow cell oligos and primer sites, which are not part of the sample. These have to be removed before the sample sequence can be put together.

1.3.2. Assembly

Fasta files from Sanger sequencing can quite easily be assembled by programs that are available, but fastq files from NGS contain a lot of short sequences and other methods are necessary to combine all that information into one long consensus sequence.

It is like making a puzzle, you could do it without seeing the picture up front, but it is easier, faster and less error-prone if you have a reference to compare your own puzzle to. The same is true for read assembly from fastq files. It is easier with a reference strain to locate and order the reads. Nevertheless, it can be done without a reference strain, although it will take longer and might not be complete due to low coverage regions.

Software which can be used for assembly without a reference is the SPAdes program (175). It will compare all the reads to each other and overlap the ones with a similar part that can lead to a longer sequence in the end, called a contig. If all goes well, the longest contig is the full sequence you wanted to determine. The contigs produced by this program can be visualised using Bandage (176). These contigs can be compared to a similar strain from what you expect, without influencing your original sequence.

The easier method would be to already suspect or know what you are looking for and having a similar sequence at hand. This can be used to compare each read to and can be done using the Burrows-Wheeler Aligner (BWA) algorithm Maximum Exact Matches (MEM) software (177). By running this through the samtools software (178), a multitude of information is stored in the resulting Sequence Alignment Map (SAM) file, like read length, position compared to the reference sequence, sequence quality as well as mapping quality and more.

1.3.3. Alignment

Once all the sequences have been put together, it can be useful to compare these to each other and see where differences can be found between sequences. To do this, all sequences have to be compared to one another and sometimes, gaps have to be introduced when a sequence has an insertion or deletion compared to another one. When all sequences have been laid out so that their similarities all line up, an alignment is produced.

There are numerous algorithms that do so, each with a slightly different emphasis on how important similarities are and how big the penalty has to be for a gap or multiple gaps. The most used algorithm today is probably NCBI BLAST+. The basis for this algorithm was founded in 1990 (179) and has been optimised 20 years later (180). It is used to search through the entire online database NCBI where all published sequences have to be submitted.

Other alignment softwares that can be used are MUSCLE, which claims to have high accuracy and throughput characteristics (181), but also reduced time and space complexity (182). Another much used software is MAFFT, which is based on fast Fourier transformations (183). Clustal Omega (184) was also tested, which is the latest successor of the Clustal series. The first one was built in 1988, which was supposed to be a locally run program (185). An updated version, Clustal V, was released in 1992 (186) and another update in 1994, namely Clustal W (187), which was similar to Clustal V, but with some improvements and an interactive user interface. The next version was Clustal X, which made improvements to the user interface (188). The method for alignment was never updated to be able to handle the big data that is available now until the release of Clustal Omega (189).

All these alignment programs have one thing in common; they calculate the best alignment based on their unique variable values and penalty scores. They will get the best possible alignment they can find, but cannot distinguish between two alignments with the same scores, even when it is obvious to human minds. Therefore, it is important to always manually check the entire alignment and not accept it as complete coming straight out of the program.

The quality of the alignment should be checked as well. Technically, an alignment full of ambiguous base pairs is a good alignment, however, it does not give any useful information. Likewise, an alignment filled with gaps can be the best alignment the program can produce, but that does not mean the alignment is fit for further research. Alistat is a program that is part of a project between Commonwealth Scientific and Industrial Research Organisation and Zentrum für Molekulare Biodiversitätsforschung (https://github.com/thomaskf/AliStat). It calculates the completeness of the alignment, whether it is incomplete because of gaps or ambiguous nucleotides, and the program will indicate issues and where they are located.

1.3.4. Phylogenetic analysis

Once sequences are obtained, an alignment is built, and the completeness is checked, phylogenetic analysis can be performed. Phylogenetic analysis shows the relatedness of sequences to each other.

The analysis can be visualised in phylogenetic trees, which are common to determine which genotype a strain belongs to, *e.g.* to which group of known strains an unknown strain is most related to.

1.3.4.1. Fitness check

However, just like with alignments, a program can output anything, but that does not mean the input was fit for this type of analysis. To be able to determine relatedness between sequences, there has to be some variability between the sequences. Sequences that are all identical cannot be distinguished from each other and it is impossible to determine any relatedness. However, too much variation is not useful either. If there are not enough similarities to determine relatedness from, the program will output a phylogenetic tree that does not show any trustworthy information.

To check the fitness of an alignment for phylogenetic analysis, a quartet analysis can be done. This will be much faster compared with building the tree and then checking the bootstrap values (see below). Likelihood mapping will take four random sequences and determine if it can make an informative phylogenetic tree of those four sequences. It will run this analysis as many times as indicated and return the result in a triangular likelihood mapping figure that shows the number of informative and uninformative quartets (Figure 1.10). This method was first described by Strimmer *et al.* in 1997 (190) and is still used today.

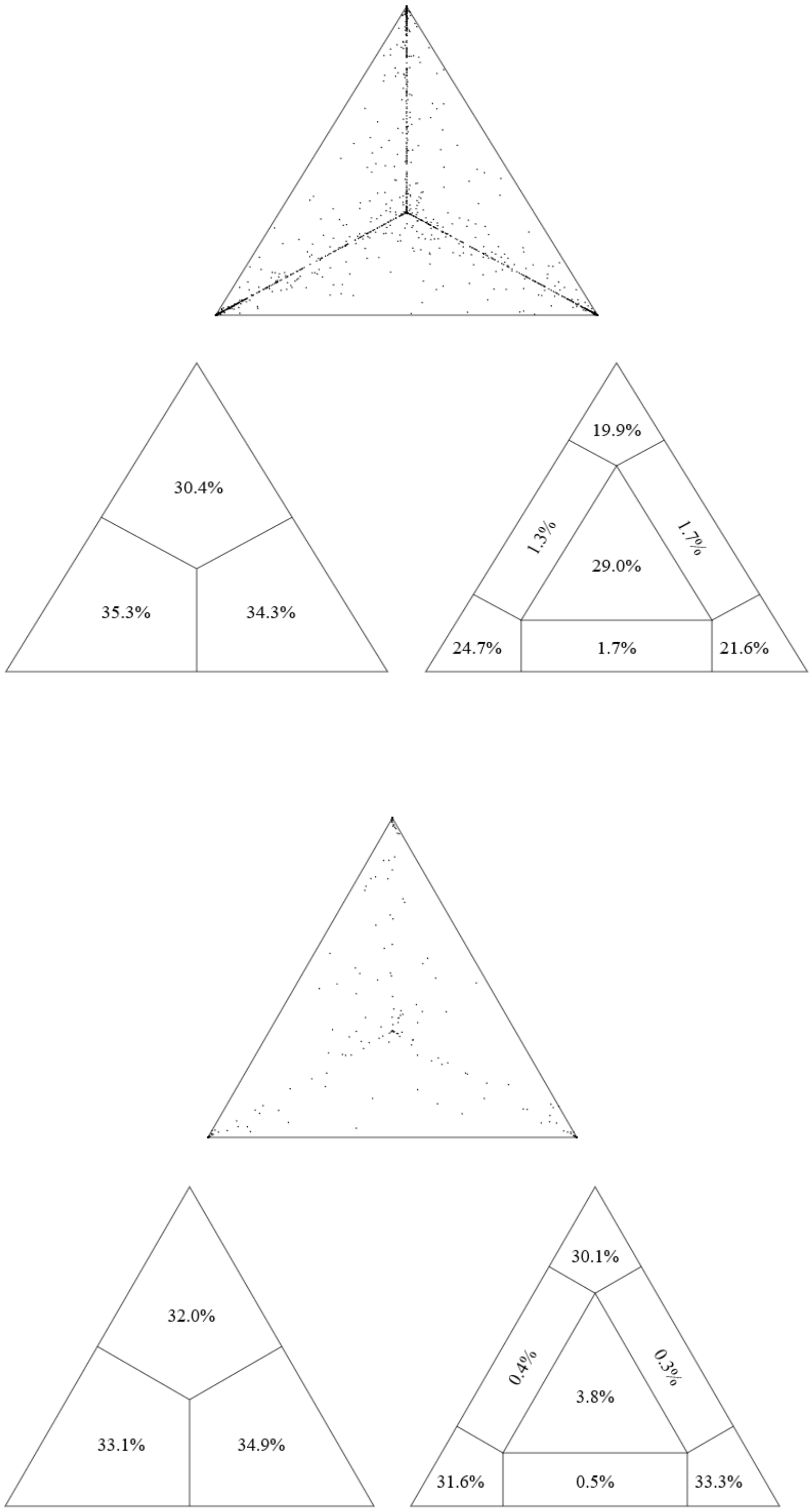


Figure 1.10: A bad (top) and good (bottom) dataset for phylogenetic as test by triangular likelihood mapping. Triangular likelihood mapping tests whether a quartet (four randomly selected sequences from the dataset) produces an informative phylogenetic tree. Each dot in the top triangle is a quartet. The three parts of the left triangle indicate whether any subsets (if they were selected) have a bigger number of quartets. The right triangle shows uninformative quartets in the middle and informative quartets in the corners. Partial informative quartets are shown on the sides of the triangle.

1.3.4.2. Evolutionary models

Next, the model to base the phylogenetic analysis on should be identified. Each dataset contains their own best fitting model and hundreds of models have been described so far. The first model in 1969 assumed that all nucleotides were present in equal frequencies and transitions and transversions happened at equal rates; this was the Jukes-Cantor (JC69) model (Figure 1.11)(191). However, transitions and transversions do not happen in equal rates and the Kimura 2-parameter (K2P) model took that into account in 1980 (192). Various new models have been described since and extra characteristics can be added to a model to accommodate for specific features of a dataset, like the feature that the rate variation among sites is gamma distributed or the proportion of invariable sites, that indicates the extent of unchanging sites in a dataset. Programs like ModelFinder will determine the best fitting model for each dataset (193).

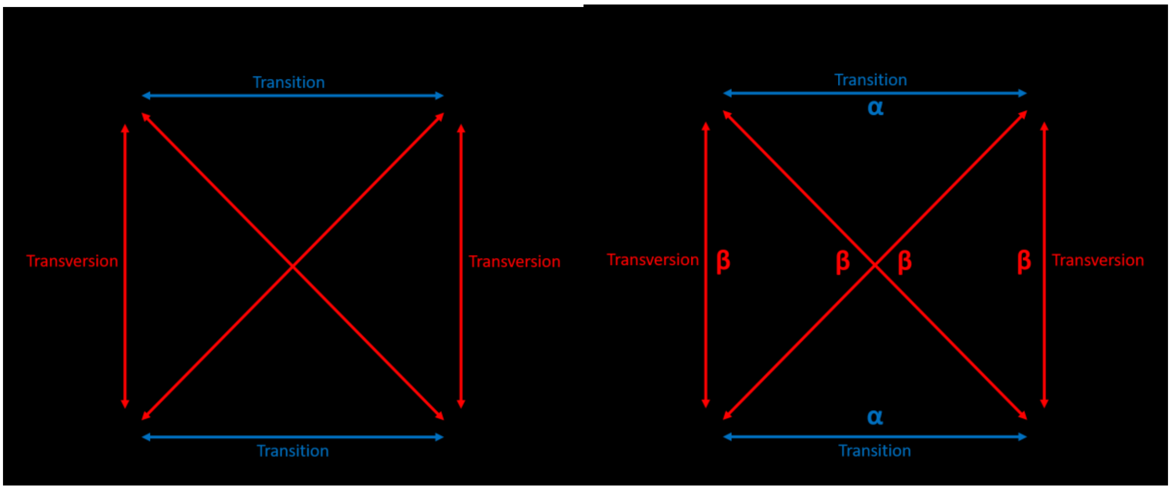


Figure 1.11: The first evolutionary model was JC69 and assumes an equal substitution rate α for all transitions and transversions. The Kimura 2-parameter model assumes equal substation rates within transitions (α) or transversions (β), but not between those.

1.3.4.3. Types of trees

There are different types of phylogenetic trees that can be calculated. They can have different properties, are calculated in different ways or have special features. The start of each tree is to determine whether to calculate a rooted or unrooted tree. A rooted tree assumes a common ancestor to all sequences in the tree that will be calculated, while an unrooted tree makes no assumptions about ancestry. Trees can be bifurcating or binary trees, which means that each node splits in exactly two descendants, while multifurcating trees can have multiple descendants in each node. A tree can be labelled to show which sequence is located where in the tree or unlabelled to focus on the tree topology only.

The types of trees that are most common are Neighbour-Joining (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML) or Bayesian Inference trees. Some of these need model selection first, others do not. NJ is based on a distance-matrix method (as is UPGMA) and is the simplest and fastest to calculate. It is often used as a guide tree, but does not use an evolutionary model in its calculation.

MP trees assume some form of evolution in the dataset, namely the one with the fewest changes, hence the shortest possible tree that fits the data will be outputted as the final MP tree. However, for a dataset containing sequences with high evolutionary rates, like sequences from RNA viruses, this is not a good method.

ML trees do base their calculations on the selected evolutionary model as one of the parameters and return more correct evolutionary trees. However, the calculation of these trees is also more power-intense and takes longer.

Bayesian trees use even more power and take even longer as they not only produce a maximum likelihood tree, but also improve it by calculating the posterior probability of each branch based on extra information.

The branching of these trees are typically tested during tree building by Nearest Neighbour Interchange (NNI), which means that each set of four subtrees is exchanged and tested for improved phylogenetic trees (194, 195). This is the slowest, but most optimising way of searching for the best possible tree if done meticulously (Figure 1.12). Another way is the Subtree Pruning and Regrafting (SPR) method (196). It removes a subtree and creates a new node elsewhere in the tree. A third method, known as the Tree Bisection and Reconnection (TBR) method, removes a subtree and reconnects it elsewhere (197). However, instead of keeping the root node from the subtree, it connects a branch from the subtree to a branch of the original tree. This creates a new node and new tree topology of the subtree. This method is even more power intensive than NNI, but could also produce a more correct tree if all subtree options are tested. More methods for optimising branching have been described, but are less common.

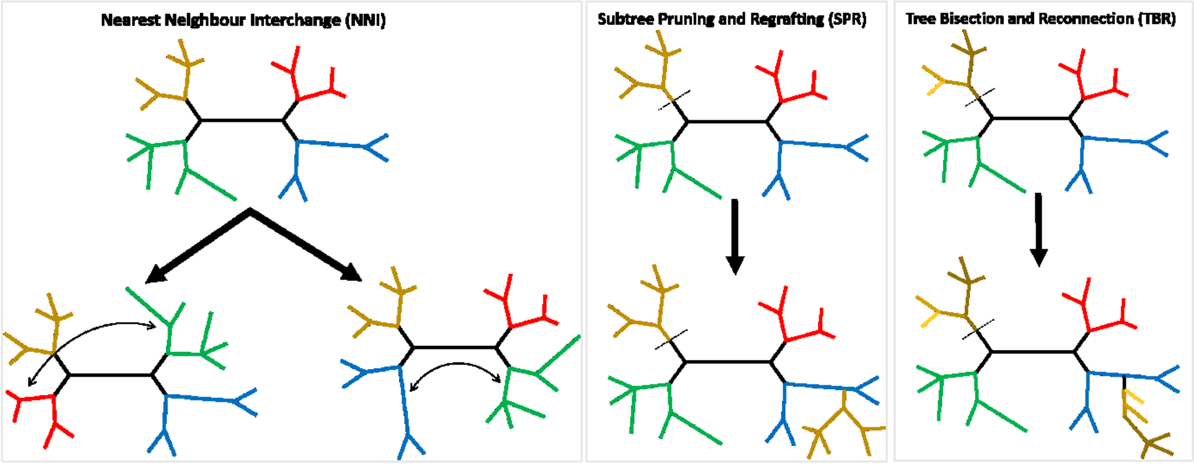

Nearest Neighbour Interchange (NNI)
Subtree Pruning and Regrafting (SPR)
Tree Bisection and Reconnection (TBR)

Figure 1.12: The three most common types of branching optimisation are Nearest Neighbour Interchange, Subtree Pruning and Regrafting, and Tree Bisection and Reconnection.

All of these methods are heuristic and to be able to identify whether a tree is useful or not, several extra calculations can be done, like showing the number of bootstrap values calculated by ultrafast bootstrap approximation (198) at each branch or approximate likelihood ratio test values to check how often this conformation was the best fitting conformation (199).

1.3.5. Variant calling

Fastq files can also be used for in-depth variant analysis, which is usually not done with fasta files as they only show consensus sequences. Once the fastq files have been compared to a reference sequence using samtools, it can also be used to determine which positions are different from the reference and which proportion of the reads carry that variation. This is done by bcftools, which is part of the samtools software (200). It produces a Variant Calling File (VCF) that contains information about all the nucleotides that are different from the reference sequence. It will contain information on the position of the variation, which nucleotide(s) was detected, which quality this variant has and filters can be applied to leave out variants that are too close to an indel and therefore still untrustworthy. There is even detailed information available on the absolute number of reads that carry the reference versus alternative nucleotide. This allows for in-depth investigation of the variations in a sample and allows for further research in other domains, like protein conformation analysis.

1.4. Experimental rationale

RSV is known to be a variable RNA virus, however, part of its genome is more variable, while other regions are more preserved. It has been shown in numerous epidemiological studies that RSV strains change year after year in the population. The variation in this virus' genome is used to divide it into genotypes.

However, it is not studied very well when these changes originate as RSV is a virus that causes short, acute infections. Human RSV is only known to infect humans and causes acute infections after which the virus is cleared. When, how and how often those variations occur is under studied and studies on the effect of RSV genome variations on disease severity have not been agreeing with each other.

1.4.1. Hypotheses

1. Genotypes can be confirmed from known sequences in online databases and determined from unknown RSV strains from clinical samples.
2. Clinical samples contain minority variants that can be detected by deep-sequencing using NGS methods.
3. The difference in disease severity between patients is caused by a difference in the amount or type of minority variants present during infection.
4. During acute infection, the appearance and disappearance of minority variants can be detected.

1.4.2. Aims

a) Investigate the current genotyping system with known sequences and test it vigorously with known and new strains.
b) Set up new NGS methods that allow for minority variant detection.
c) Detect presence of minority variants and compare their prevalence and characteristics between community and hospital patients.
d) Obtain deep-sequencing data from consecutive time points during acute infection of healthy, adult volunteers.

2. Materials and Methods

2.1. Reagents, materials and equipment

2.1.1. Buffers and reagents

Table 2.1: Reagents and buffers, their usage and their supplier.

Reagent	**Usage or recipe**	**Supplier and product code**
Absolute methanol	Ready to use	Sigma 322415
Agarose MP	Dilute 1g in 100 ml TBE	Sigma Aldrich 11388991001
Agencourt AMPure XP	Ready to use	Beckman Coulter A63881
Ampicillin (200mg, stock at 10mg/mL)	Reconstitute according to instructions	Invitrogen 11593-027
Biotinylated polyclonal goat anti-RSV IgG	Ready to use	Biotin AbD Serotec 7950-0104
BlueJuice™ Gel Loading Buffer (10X)	Dilute 1:50 with 1X TBE buffer	Thermo Fisher Scientific 10816015
DNA ladder, 1 kb	Mix with BlueJuice as needed	Invitrogen 15615016
dNTPs mix	Ready to use, 10 mM	Thermo Fisher Sientific 10297018
Dulbecco's Modified Eagle Medium (DMEM) – high glucose	Ready to use, supplement as required	Sigma D6429
Dulbecco's Phosphate Buffered Saline (PBS)	Ready to use, supplement as required	Sigma D8537
ExtrAvidin-Peroxidase	Ready to use	Sigma E2886
FastRuler High Range DNA Ladder	Ready to use	Thermo Fisher Scientific SM1123
Heat-Inactivated Fetal Calf Serum (HI-FCS), 10%	Heat-inactivated at 56^{o}C for 30 minutes, filtered, aliquoted and stored at -80^{o}C	Source Bioscience
Hydrogen peroxide (H2O2), 2%	Ready to use	Sigma H1009
L-Glutamine, 100X	Dilute to 1X in cell culture media	Gibco 25030-024

Lysis buffer	Guanidine thiocyanate + Triton X100	Roche lysis buffer (Roche MagNa Pure Total Nucleic Acid Isolation Kit cat# 03246779001
Massruler Loading Dye, 6X	Ready to use	Thermo Fisher Scientific R0621
MMLV Reverse Transcriptase enzyme	Ready to use	Thermo Fisher Scientific 28025013
Molecular Biological Grade water	Ready to use	Thermo Fisher Scientific CMC-750-030M
Nuclease-free water	Ready to use	Promega & Severn Biotech
NucliSENS easyMAG magnetic silica	Ready to use with NUCLISENS® easyMAG®	Biomérieux 280133
NucliSENS Extraction buffer 1	Ready to use with NUCLISENS® easyMAG®	Biomérieux 280130
NucliSENS Extraction buffer 2	Ready to use with NUCLISENS® easyMAG®	Biomérieux 280131
NucliSENS Extraction buffer 3	Ready to use with NUCLISENS® easyMAG®	Biomérieux 280132
NucliSENS Lysis buffer	Ready to use with NUCLISENS® easyMAG®	Biomérieux 280134
Penicillin + Streptomycin, 100X	Dilute to 1X in cell culture media	Sigma P4333
PhiX Control v3, 10nM	Prepare according to instructions	Illumina FC-110-3001
rRNAsin	Ready to use	Promega N2111
S-(5'-Adenosyl)-L-methionine (SAM) elution buffer	PBS + 1% BSA + 0.5 Triton X + 0.05% sodium azide	Sigma A9384
SIGMAFASTTM 3-3’-Diaminobenzidine (DAB) tablets	Dissolved in nanopure water according to instructions	Sigma D4418-5SET
Sodium acetate, 3M	Ready to use	Sigma S2889
Sodium hydroxide solution, 10M	Diluted to 0.1M NaOH with Molecular Biological Grade water	Sigma 72068

Sodium Hydroxide, 10M	Ready to use	Sigma-Aldrich 72068
TaqMan Gene Expression Master Mix	Ready to use	Applied Biosystems 4369016
Trypan Blue solution, 0.4%	Ready to use	ThermoFisher Scientific T10282
Tween® 20 Bioxtra, viscous liquid	Ready to use	Sigma-Aldrich P7949
UltraPure™ TBE Buffer, 10X	Dilute 1:10 in nanopure water	Thermo Fisher Scientific 15581044
Virus Transport Medium (VTM)	Ready to use	Public Health England In-house

2.1.2. Commercially available kits

Table 2.2: Commercially available kits used during this project.

Kit Name	**Contents**	**Supplier and product code**
Agencourt CleanSEQ Kit	CleanSEQ Reagent	Beckman Coulter A29161
BigDye® Terminator v3.1 Cycle Sequencing Kit	BigDye® Terminator v3.1 Ready Reaction Mix, M13 primer, pGEM Control DNA 5X Sequencing Buffer	Thermo Fisher Scientific 4337457
GX/GXII		
High-capacity cDNA Reverse Transcription Kit with RNase Inhibitor	10X RT Buffer, 10X RT random primers, 25X dNTP Mix (100 mM), MultiScribeTM Reverse Transcriptase (50 U/µl), RNase Inhibitor	Applied Biosystems 4374966
HT DNA High Sensitivity LabChip® Kit LabChip	DNA Dye Concentrate, Chip Storage buffer, DNA HiSens Gel Matrix, 10X DNA HiSens Ladder, DNA HiSens Marker, Spin Filters	PerkinElmer CLS760672
Illumina® Nextera® XT DNA Sample Preparation Kit	ATM, TD, NPM, RSB, LNA1, HT1, LNW1, LNB1, NT, LNS1 *	Illumina FC-131-1096
Illumina® Nextera® XT Index Kit	96 unique Indexes	Illumina FC-131-1002

KAPA SYBR® FAST Universal qPCR Kit	Universal qPCR Master Mix, 50X ROX High, 50 ROX Low, DNA Standards 1-6, 10X Primer Mix **, 2X KAPA SYBR® FAST qPCR Master Mix	KAPA Biosystems, KK4824
MiSeq Reagent Kit v2	Reagent cartridge, Hybridization buffer (HT1), incorporation buffer and flow cell	Illumina MS-102-2002
PfuUltra II Fusion HS DNA Polymerase Kit	PfuUltra II Fusion HS DNA Polymerase, 10X PfuUltra II Fusion HS DNA Polymerase buffer, provides final 1X Mg2+ concentration of 2mM	Agilent Technologies 600674
Platinum™ Quantitative PCR SuperMix-UDG	40 U/ml UDG, 60 U/ml Platinum® Taq DNA Polymerase, 40 mM Tris-HCl, 100 mM KCl, 6 mM MgCl2, 400 µM of dATP, dCTP, dGTP and 800 µM of dUTP	Invitrogen 11730-017
QIAamp Viral RNA mini Kit	Buffers AVL, AW1, AW2, AE, carrier RNA, QIAamp Mini Spin Columns, Collection tubes	Qiagen 52904
QIAquick PCR purification Kit	250 QIAquick Spin Columns, 2 ml Collection Tubes, PB Buffer, PE Buffer, EB Buffer, pH indicator I	Qiagen 28106
Quant-iT™ DNA Assay Kit, High Sensitivity	Concentrated assay reagent, dilution buffer, DNA standards	Thermo Fisher Scientific Q33120
Qubit™ dsDNA HS Assay Kit	dsDNA HS buffer, dsDNA HS reaction dye, pre-diluted DNA standards	Thermo Fisher Scientific Q32854
SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity DNA	SuperScript® III RT/Platinum® Taq High Fidelity Enzyme Mix, 2X Reaction Mix (containing 0.4 mM of each dNTP, 2.4 mM MgSO4), 5 mM Magnesium Sulfate	Thermo Fisher Scientific 12574030

SuperScriptTM III One-Step RT-PCR System with PlatinumTM Taq High Fidelity DNA Polymerase Kit	SuperScriptTM III One-Step RT/PlatinumTM Taq High Fidelity Enzyme Mix, 2X Reaction Mix containing 0.4mM each dNTP and 2.4 mM MgSO4, 5mM Magnesium Sulfate	Invitrogen 12574-035
SuperScript™ VILO™ cDNA Synthesis Kit	5X VILO Reaction Mix with random primers, dNTPs and MgCl2, 10X SuperScript III Reverse Transcriptase, RNaseOUT Recombinant Ribonuclease Inhibitor and proprietary helper protein	Thermo Fisher Scientific 11754050

2.1.3. Laboratory materials

Table 2.3: : Laboratory materials used in this project.

Material	**Supplier and product code**
Agencourt SPRIPlate 96R Ring Magnet Plate	Beckman Coulter A29164
BioClean Pipette tips Green-Pak™	Rainin 17002410, 17002413, 17002414, 17002429, 17002431
Corning™ Thermowell™ 96-Well Polypropylene PCR Microplates	Corning™ 6551
E-Gel™ 48 Agarose Gels, 1%	Thermo Fisher Scientific G800801
Falcon™ 50mL Conical Centrifuge Tubes	Thermo Fisher Scientific 10788561
Greiner CELLSTAR® 96 well plates	Sigma M0812
Lens Cleaning Tissues	Thermo Fisher Scientific CMC-750-030M
MicroAmp Fast 96-well reaction plates	Applied Biosystems 4346907
Qubit™ Assay Tubes	Thermo Fisher Scientific Q32856
RNase-free microfuge tubes	Thermo Fisher Scientific AM12400
RotorGene Strip Tubes and Caps 0.1 ml	Qiagen 981103
Tissue culture flasks, 175cm^2 (T175)	Nunc 178883

V-bottom thermal cycler-compatible polypropylene plates	Beckman Coulter 609801

2.1.4. Equipment

Table 2.4: Equipment used during this project.

Equipment name	**Supplier**
3730xl DNA Analyzer	Thermo Fisher Scientific
ABI ViiA7	Thermo Fisher Scientific
Biomek NXP Automated Workstation	Beckman Coulter
Countess Automated cell counter	Invitrogen
GloMax® Discover System	Promega
Incubator	Binder
Inverted light microscope (Leica DME)	Leica Microsystems
LabChip GX Touch HT Nucleic Acid Analyzer	Perkin Elmer
Laboratory Centrifuge	Thermo Fisher Scientific
Laboratory Centrifuge 5810 R	Eppendorf
Molecular Imager® Gel DocTM XR System	Bio-Rad
Mother E-BaseTM device	Invitrogen
NanoDrop™ 1000 Spectrophotometer	Thermo Fisher Scientific
NucliSENS® easyMag®	Biomérieux
Qubit® 3.0 Fluorometer	Thermo Fisher Scientific
Real-time PCR cycler ABI 7500 Fast	Applied Biosystems
RotorDisc 100 with QIAgility set-up	Qiagen
Sonicator	Ultrawave Ltd
TETRAD2 Thermocycler	MJ Research
TruSeq Index Plate Fixture Kit	Illumina FC-130-1005
Veriti™ 96-well Thermocycler	Thermo Fisher Scientific

2.1.5. Software

Table 2.5: Software used in this project.

Software	**Supplier**	**Reference**

7500 Software v2.3 for qPCR analysis	Applied Biosciences	
AliStat v1.7	CSIRO	https://github.com/thomaskf/AliStat/blob/ master/AliStat_manual.pdf
Bandage v0.8.1	Holt Lab – Microbial Genomics	doi.org/10.1093/bioinformatics/btv383
Bcftools v1.8	Broad Institute	10.1093/bioinformatics/btr509
BWA v0.7.12	Broad Institute	10.1093/bioinformatics/btp324
IGV v2.4.10	Broad Institute	10.1038/nbt.1754

IQ-Tree v1.6.6	http://www.iqtree.org/	https://doi.org/10.1093/molbev/msu300;
		https://doi.org/10.1038/nmeth.4285;
		https://doi.org/10.1093/molbev/msx281

MEGA7	MEGA	10.1093/molbev/msw054
Picard Tools v2.18.2	Broad Institute	https://broadinstitute.github.io/picard/
Quantity One Software for Gel Doc XR	Bio-Rad	www.bio-rad.com/en-uk/product/quantity-one-1-d-analysis-software
R v3.4.3 & RStudio v1.0.153	R Foundation for Statistical Computing	https://www.r-project.org/
Samtools v1.8	Broad Institute	10.1093/bioinformatics/btp352
SeqMan Pro (VERSION)	DNASTAR®	www.dnastar.com/t-seqmanpro.aspx
SPAdes v3.5.0	Center for Algorithmic Biotechnology	10.1089/cmb.2012.0021

2.2. Laboratory techniques and assays

Laboratory assay protocols were protocols tested and optimised at the laboratory of Respiratory Infections at the National Heart and Lung Institute of Imperial College London or at the Respiratory Virus Unit of at the National Infection Services of Public Health England. These in-house protocols were used unless stated otherwise.

2.2.1. Viruses

2.2.1.1. *Inoculum for human challenge study*

RSV A Memphis-37 (M37) was used by Max Habibi, a previous PhD student, to inoculate healthy, adult volunteers with RSV. He purchased this strain from Meridian Lifesciences and produced a laboratory virus stock (26). This stock was used for the optimisation of cell culture experiments. hVivo (www.hvivo.com) provided us with RNA extract of the original stock of RSV A M37-b, GMP RSV P88023 (LOT 070504). This was used for deep sequencing of the complete RSV M37 genome and served as the baseline control for in-host evolution analysis.

2.2.1.2. *Laboratory strains*

Respiratory syncytial virus (RSV) strains used for testing and optimisation of sequencing protocols were first grown *in vitro* in HEp-2 cell culture to amplify available stocks in the lab. Four laboratory strains were selected, namely RSV A2, RSV A Long, RSV B 9320 and RSV B N2. These were frozen at -80°C immediately after harvesting to create a viral stock.

2.2.1.3. *Clinical samples*

Clinical samples were obtained from three different populations. The first cohort contained samples from hospitalised patients. Nasopharyngeal aspirates and nasal swabs were collected from patients in Imperial College Healthcare NHS Trust hospitals in London between November 2015 and February 2016 and stored at -80°C at the biobank at Charing Cross hospital until processing.

The second cohort contained samples from community patients with mild symptoms. Nasal and throat swabs were collected by Public Health England in collaboration with the Royal College of General Practitioners on a yearly basis. These surveillance samples from 2009-2018 were diagnostically tested for RSV and stored at -80°C in separate aliquots of 1.5 ml. Samples with high enough viral loads were selected for further processing.

The third cohort contained samples from healthy volunteers that were infected with RSV Memphis-37 and daily nasal lavage samples were taken for 10 consecutive days with follow-up samples taken at day 14 and day 28 post-infection (26). RSV RNA was quantified and samples with high enough viral loads were selected for deep sequencing.

2.2.2. Cell and viral culture

RSV was grown in HEp-2 cell culture. Cells were incubated at 37°C and 5% CO_2 in complete media, which consisted of Dulbecco's Modified Eagle Medium (DMEM) supplemented with 1% Penicillin and Streptomycin, and 1% L-glutamine. 10% heat-inactivated foetal calf serum (HI-FBS) was added as well. Cells were split when they were 90% confluent. Splitting cells was done by removing media, washing cells with 10 ml phosphate buffered saline (PBS) and adding trypsin after PBS was discarded. Trypsin was left to incubate for 5 minutes, then 10 ml of complete media was added to inactivate the trypsin. All cells were collected and spun down for 5 minutes at 1500 rpm. The pellet was resuspended in complete media and 1 million cells were plated in 24 ml of media in a T175 flask.

To grow RSV A M37, 10 million HEp-2 cells were seeded in a 175 cm^2 flask in 25 ml of complete media and left overnight to grow to 80% confluence. The cells were washed with serum free media and infected with 4 ml of virus with MOI = 0.1. This meant that 1 viral particle for every 10 cells was added. The flask was rotated every 15 minutes. After 2 hours, the serum free media was topped up with 26 ml of complete media. The media was removed after 24 hours, centrifuged for 5 minutes at 1500 rpm and the pellet was resuspended in 25 ml of serum free media, then incubated for another 24 hours at 37^{o}C. To harvest virus, the cells were scraped off the flask when the cytopathic effect was estimated to be 50%, sonicated for 20 seconds and spun down for 5 minutes at 1500 rpm. The supernatant was aliquoted into 1 ml cryovials and snap frozen in liquid nitrogen. All vials were stored in liquid nitrogen tanks at -196^{o}C.

2.2.3. Plaque assay

Plaque assays were performed to assess the infectivity of the virus. A monolayer was grown overnight at 37^{o}C in a flat-bottomed 96-well plate by adding 20,000 HEp-2 cells, 10% HI-FCS and complete media to each well. The virus was diluted 12 times in two-fold dilutions starting at 1 in 10. When the cells were 90% confluent, they were washed with serum free media and 50 µl of the viral dilution series was added to the plate in triplicates. The virus was left for 1.5 - 2 hours at 37^{o}C, then 150 µl of complete media was added to each well and the plates were incubated at 37^{o}C and 5% CO_2 overnight. Cells were washed with 1X PBS and fixed with absolute methanol and 2% H_2O_2 for 20 minutes. Next, the cells were washed twice with PBS complemented with 1% BSA and then washed again with a PBS/BSA/Azide solution. The plates were refrigerated at 4^{o}C overnight. Biotinylated anti-RSV goat antibody was added to each well and incubated for 1 hour at room temperature. Then, the plate was washed twice with PBS/BSA and a 1:500 dilution of extravidin peroxidase was added to the plate for 30 minutes. The plate was washed twice with PBS/BSA and 3’3-diaminobenzidine (DAB) was added for 20 minutes to 24 hours at room temperature in the dark. The plate was monitored for development of background staining and the reaction was stopped by washing with PBS when staining was successful, but before background staining set in. The cells were kept in PBS to count. All plaques were manually counted through the microscope and plaque forming units (pfu) were calculated by Equation 2.1.

$$pfu = \ mean\, number\ of\ plaques * fold\ dilution * \frac{20}{ml} of\ virus\ inoculum$$

Equation 2.1: Plaque forming units were calculated based on dilution and amount of inoculum used.

2.2.4. Extraction

2.2.4.1. Manual extraction of nucleic acids with QIAamp Viral RNA Mini Kit

Viral RNA was extracted from cell culture samples using the QIAamp Viral RNA Mini Kit. The kit protocol was adapted as follows. The AVL buffer was prepared by mixing 1:1 AVE buffer and lyophilized carrier RNA. 1000 µl prepared AVL buffer was added to each microcentrifuge tube followed by 250 µl of sample. The tubes were vortexed and incubated at room temperature for 10 minutes, then spun down. Next, 1000 µl of 96% ethanol was added to the sample, which was then vortexed and briefly spun.

After preparation of the sample, 600 µl was added to the spin column and centrifuged for 1 minute at 6000 g. Flow through was discarded and another 600 µl of prepared sample was added to the spin column. The process was repeated until the entire sample was loaded onto the spin column. The spin column was washed with 600 µl AW1 buffer and centrifuged for 1 minute at 6000 g. Flow through was discarded and the spin column was washed with AW2 buffer by centrifuging for 3 minutes at 21,100 g. Flow through was discarded and the spin column was spun dry for 1 minute at 21,100 g. For elution, 60 µl of nuclease-free water was added to the spin column and incubated for 1 minute, then spun 1 minute at 6000 g.

Extracted nucleic acids were put on ice straight away. To estimate nucleotide quantity, 1.5 µl of the sample was measured with the NanoDrop 1000 Spectrophotometer for indicative purposes.

2.2.4.2. Automated complete nucleic acid extraction with NucliSENS® easyMag®

Sequence optimisation and clinical samples were first aliquoted into 150 µl aliquots to which 50 µl of Magnapure lysis buffer was added. Extractions with a NucliSENS® easyMag® instrument was performed according to protocol with the following settings:

- Matrix = other
- Extraction volume = 0.200 ml
- Elution volume = 100 µl
- Protocol = Generic 2.0.1
- Type = Primary
- Priority = Normal

- On board lysis

This method used the BOOM technology, which is based on solid phase extraction and used silica beads to bind nucleic acid in a four-phase process. The sample was first lysed and homogenized by NucliSENS Lysis Buffer, secondly NucliSENS Extraction buffer 1, chaotropic substance, and silica beads were added to bind nucleic acids. Thirdly, the silica beads were washed with NucliSENS Extraction buffer 2 to remove contaminants. Finally, the bonded nucleic acids were separated from the silica beads and eluted into NucliSENS Extraction buffer 3.

2.2.5. Reverse transcriptase reaction

2.2.5.1. Reverse transcriptase reaction with High-Capacity cDNA Reverse Transcription Kit with RNase Inhibitor

Once cell culture samples were extracted, RNA was converted using the High-Capacity cDNA Reverse Transcription Kit with RNase Inhibitor, which was adapted as follows. For each reaction, 7 µl of 2X Reverse Transcription (RT) Mix was mixed with 13 µl of sample RNA. The 2X Reverse Transcription Mix contained 2µl of 10X RT Buffer, 0.8 µl of 25X dNTP Mix at 100 nM concentration, 2 µl of 10X RT random primers, 1µl of MultiScribe^{TM} Reverse Transcriptase, 1 µl of RNase Inhibitor and 0.2 µl of nuclease-free water for each reaction. Sample RNA was mixed with the RT mix by pipetting up and down and spun down for 2 minutes at 2000 rpm. The program used on the thermal cycler was 10 minutes at 25^{o}C followed by 120 minutes at 37^{o}C and 6 minutes at 85^{o}C. When finished, the samples were kept at 4^{o}C.

2.2.5.2. Reverse transcriptase with Moloney Murine Leukemia Virus (MMLV)

To determine the viral load of clinical samples, extracted RNA was converted to cDNA using Moloney Murine Leukemia Virus (MMLV) reverse transcriptase. For each reaction, a reverse transcription mix was made by mixing 4 µl of 10X PCR buffer, 6 µl of $MgCl_2$, 6µl of dNTPs (at 10mM concentration), 0.4 µl of random primers, 0.4 µl of RNAsin and 1 µl of MMLV. 22.2 µl of sample RNA was added and the mix was incubated at room temperature for 10 minutes and then heated in the thermal cycler for 45 minutes at 37^{o}C. The reverse transcriptase was heat inactivated at 95^{o}C for 5 minutes. Samples were kept at 4^{o}C.

2.2.5.3. Reverse transcriptase with VILO enzyme

After easyMag® extraction of clinical samples and confirmation of RSV RNA presence, reverse transcriptase was performed using VILO enzyme, which is a genetically engineered version of MMLV Reverse Transcriptase with reduced RNase H activity. This enzyme had improved thermostability for more efficient and precise cDNA synthesis. For each reaction, 12 µl of 5X VILO reaction mix, 27 µl of nuclease-free water and 6 µl of Superscript III Enzyme mix was added to 15 µl of extracted sample

RNA. The thermal cycler ran for 10 minutes at 25ºC, 60 minutes at 42ºC and 5 minutes at 85ºC. cDNA was kept at 4ºC.

2.2.6. Quantitative real-time polymerase chain reaction (Q-RT-PCR)

2.2.6.1. N gene-based RSV A Q-RT-PCR

After reverse transcription with High-Capacity cDNA Reverse Transcription Kit with RNase Inhibitor, quantitative real-time PCR (Q-RT-PCR) was performed to determine the viral load of laboratory grown virus samples based on presence of the nucleoprotein gene (N). For each sample, 10 µl of qPCR master mix was prepared with 900 nM of forward primer for N, 900 nM of reverse primer for N and 200 nM of N probe. Then 2.5 µl of cDNA was added to the master mix. All samples were tested in duplicate. The RSV A forward primer sequence was 5'-CATCCAGCAAATACACCATCCA-3' (Invitrogen), the RSV A reverse primer sequence was 5'-TTCTGCACATCATAATTAGGAGTATCAA-3' (Invitrogen) and the RSV Pan A N gene probe was 5'-FAM-CGGAGCACAGGAGAT-TAMRA-3' (Eurofins MWG/Operon).

The standard curve of N was prepared as follows. 10 µl of N gene plasmid stock at a concentration of 1^{10} was diluted 1:10 in nuclease-free water and 9 further 10-fold dilutions were prepared with nuclease-free water. The last 7 out of 10 dilutions were used as standard curve and measured in duplicate. 12.5 µl of standards and samples were added to the plate. Positive control was cDNA extracted from M37 lab grown strain and negative control was qPCR master mix with nuclease-free water. The plate was spun for 2 minutes at 2000 rpm. The thermal cycler ran for 2 minutes at 50ºC, 10 minutes at 95ºC, then 40 cycles of 15 seconds at 95ºC and 1 minute at 60ºC. To end the run, the thermal cycler kept samples at 4ºC.

Applied Biosciences 7500 Software v2.3 was used to calculate the linear regression line through the standards. Quantity of samples was determined based on this standard regression.

2.2.6.2. N gene-based diagnostic RSV/hMPV RT-PCR with internal control (SBCMV)

Detection of RSV in clinical samples was performed with a laboratory developed multiplex assay which tested RSV A, RSV B, hMPV A and hMPV B presence. Soil-Borne Cereal Mosaic Virus (SBCMV) was used as internal control. Each reaction contained 5 µl of cDNA, 12.5 µl of 2X Real-Time Platinum qPCR SuperMix-UDG, 1.5 µl of 50 mM $MgCl_2$ (3 mM final conc), 0.1 µl of RSV A forward primer (400 nM final conc), 0.1 µl of RSV A reverse primer (400 nM final conc), 0.0625 µl of RSV A probe (25 nM final conc), 0.05 µl of RSV B forward primer (200 nM final conc), 0.5 µl of RSV B reverse primer (200 nM final conc), 0.375 µl of RSV B probe (150 nM final conc), 0.075 µl of hMPV A forward primer (300 nM final conc), 0.075 µl of hMPV A reverse primer (300nM final conc), 0.125 µl of hMPV A probe (50 nM final conc), 0.0125 µl of hMPV B forward primer (50 nM final conc), 0.1 µl of hMPV B reverse primer (400 nM final conc), 0.0625 of hMPV B probe (25 nM final conc), 0.0125 µl of SBCMV forward primer (50 nM final conc), 0.0125 µl of SBCMV reverse primer (50 nM final conc), 0.0313 µl of SBCMV probe (12.5 nM final conc) and 4.7563 µl of deionized water. All primer and probe sequences are specified in Table 2.6. Each sample was tested in duplicate on a RotorGene thermal cycler using the following program: 2 minutes at 50 ^{0}C, 2 minutes at 95^{0}C and 40 cycles of 5 seconds at 94^{0}C and 20 seconds at 60^{0}C.

RSV A was detected in the red channel (Quasar 670), RSV B in the yellow channel (JOE), hMPV in the green channel (FAM), hMPV in the crimson channel (Quasar 705) and the internal control SBCMV was detected in the orange channel (ROX). The detection thresholds were set at 0.07236 for RSV A, 0.05759 for RSV B, 0.05665 for hMPV A, 0.09565 for MPV B and 0.15795 for the internal control. The control RSV A primers were based on RSV A Long strain, control RSV B primers were based on RSV B 9320 strain.

Table 2.6: List of primers and probes used in diagnostic multiplex assay for RSV and hMPV.

	Sequence	**Final concentration**
RSV A forward primer	5’ CAG AGG TGG CAG TAG AGT TGA 3’	400 nM
RSV A reverse primer	5’ CCT GCA CCA TAG GCA TTC AT 3’	400 nM
RSV A probe	5’Quasar670 – AGG GAT TTT TGC AGG ATT GTT T - BHQ2 3’	25 nM
RSV B forward primer	5’ GCA TTG CAC AAT CAT CCA CA 3’	200 nM
RSV B reverse primer	5’ CAG CTT CTC CTC CCA ACT TC 3’	200 nM
RSV B probe	5’JOE – TCC AAG CAG AAA TGG AGC AAG TTG - BHQ1 3’	150 nM
hMPV A forward primer	5’ CAG CAC CAG ACA CAC CTA TAA T 3’	300 nM
hMPV A reverse primer	5’ TGC ATC ACT TAG TAC ACG GTT AG 3’	300 nM
hMPV A probe	5’FAM – AGT GGG ATT AGA GAC CAC AGT CAG AAG A - BHQ1 3’	50 nM
hMPV B forward primer	5’ GGG TGT CAT TGC CAG ATC AT 3’	50 nM
hMPV B reverse primer	5’ CCA GAT TCA GGA CCC ATT TCT C 3’	400 nM
hMPV B probe	5’ Quasar705 – AGG GCA TGT ATC TGT GCA AGC TGA - BHQ2 3’	25 nM
SBCMV forward primer	5’ CAC TCA GGA CGG TGA CGA GAT 3’	50 nM
SBCMV reverse primer	5’ GTG ATA CTG TGA GTC TGG TGA TGA TTT 3’	50 nM
SBCMV probe	5' ROX - TTT TGT GAC CTT GGA GGT GAG GCA GTT ATG - BHQ2 3'	12.5 nM

2.2.7. Amplification PCR for WGS

2.2.7.1. Amplification PCR in 16 fragments

Once RNA was converted to cDNA with VILO enzyme, the viral RNA was amplified using 16 primer pairs to cover the entire RSV genome (Figure 2.1). Primers were designed and validated at PHE and bought from Thermo Fisher Scientific (Table 2.7). The first fragment of the genome was covered by two primer pairs because of the variation in the 5' UTR of the genome. Each reaction contained 37.75 µl of water, 5 µl of 10X Pfu buffer, 1.25 µl of 10 mM dNTPs mix, 1 µl of Pfu Ultra II fusion HS DNA polymerase and 2 µl of primer mix (0.6 µM final conc) and 3 µl of cDNA. The tubes were spun down briefly and transferred to the thermal cycler. Samples were heated for 1 minute to 95^{o}C, then 40 cycles of 20 seconds at 95^{o}C, 20 seconds at 55^{o}C and 45 seconds at 72^{o}C followed. Final extension was 5 minutes at 72^{o}C.

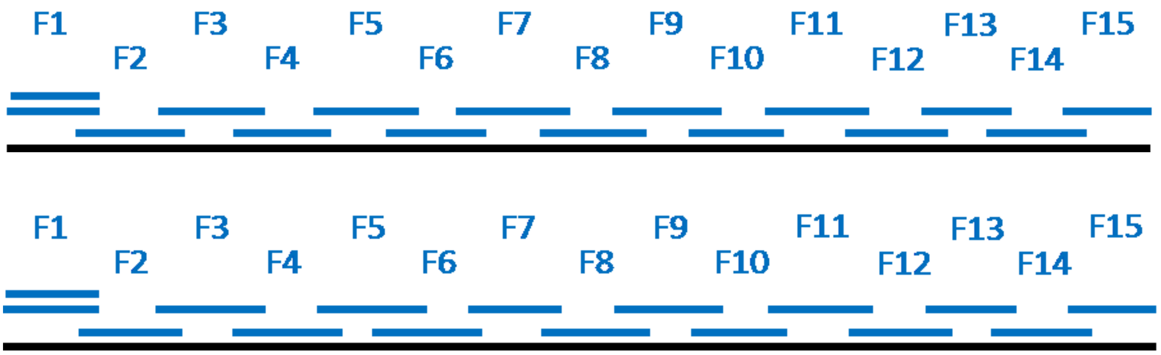


Figure 2.1: Genome coverage with 16 primers pairs for amplification of RSV A (top) and RSV B (bottom). Black = RSV genome, blue is amplified fragment.

Table 2.7: Primers used for amplifying RSV A with PCR. W= A or T, R = A or G, M is C or A, Y = T or C. All primers were ordered from Metabion. Italics underlined = M13.

Binding region	Fragment	Primer name	Primer sequence	Nucleotide position	Fragment size
Leader, NS1, NS2, N	1	AF1F	ACGCGAAAAAATGCGTACAAC	1	1215
		AF1R	TGRATRGTRTATTTRCTRGA	1216	
		AF1Falt	*GTAAAACGACGGCCAG*GGGGCAAATAAGAATTTGAT	45	1171
		AF1Ralt	*CAGGAAACAGCTATGAC*TGRATRGTRTATTTRCTRGA	1216	
NS2, N, P	2	AF2F	*GTAAAACGACGGCCAG*ATCATGATGGGTTCTTAGAATGC	920	1446
		AF2R	*CAGGAAACAGCTATGAC*TTCAGGAGCAAACTTTTCCAT	2366	
N, P, M	3	AF3F	*GTAAAACGACGGCCAG*AAAAATTGGGTGGWGAAGCA	2014	1417
		AF3R	*CAGGAAACAGCTATGAC*CCCTTGGGTGTGGATATTTG	3431	
P, M, SH	4	AF4F	*GTAAAACGACGGCCAG*AACCTRTTGGAAGGGAATGA	3019	1301
		AF4R	*CAGGAAACAGCTATGAC*AGGCCAGAATTTGCTTGAGA	4320	
M, SH, G	5	AF5F	*GTAAAACGACGGCCAG*ACMAACMCTCTGTGGTTCAA	4076	1389
		AF5R	*CAGGAAACAGCTATGAC*TTGTGSGTTCTGGATTTCCTG	5465	

G, F	6	AF6F	*GTAAAACGACGGCCAG*RAGTCAACCCYRCAATCCAC	5054	1328
		AF6R	*CAGGAAACAGCTATGAC*GCATTAACACTAAATTCCCTGGT	6382	
F	7	AF7F	*GTAAAACGACGGCCAG*AGCCAGAAGAGAACTACCAAG	5978	1520
		AF7R	*CAGGAAACAGCTATGAC*CTTAAAAAGTGTAAGTGAGATGGTT	7498	
F, M2, L	8	AF8F	*GTAAAACGACGGCCAG*YCCATTAGTRTTCCCYTCTG	7097	1421
		AF8R	*CAGGAAACAGCTATGAC*TCCAYYAATAATGGGATCCATT	8518	
M2, L	9	AF9F	*GTAAAACGACGGCCAG*TRCCDGCAGAYGTRYTGAA	8060	1435
		AF9R	*CAGGAAACAGCTATGAC*TTTATTATGTAGAASCCYTCATTRTG	9495	
L	10	AF10F	*GTAAAACGACGGCCAG*CAATGCAACATCCTCCATCA	9084	1268
		AF10R	*CAGGAAACAGCTATGAC*TGCCTGTCAATGATACCACAT	10352	
L	11	AF11F	*GTAAAACGACGGCCAG*AGTGGATCTTGAAATGATCATAAAT	10087	1380
		AF11R	*CAGGAAACAGCTATGAC*GGGATCACCACCACCAAATA	11467	
L	12	AF12F	*GTAAAACGACGGCCAG*AGTGGGACCGTGGATAAACA	11164	1360
		AF12R	*CAGGAAACAGCTATGAC*TGACTGTAAGGCGATGCAAA	12524	
L	13	AF13F	*GTAAAACGACGGCCAG*TGGACATCAAATATACWACAAGCA	12180	1200
		AF13R	*CAGGAAACAGCTATGAC*TTAACAACCCAAGGGCAAAC	13380	
L	14	AF14F	*GTAAAACGACGGCCAG*AAAAAGATTGGGGAGAGGGATA	13041	1334
		AF14R	*CAGGAAACAGCTATGAC*TGCAYTTTCTTACATGCTTGC	14375	
L, end	15	AF15F	*GTAAAACGACGGCCAG*GGTGAAGGAGCAGGGAATTT	14054	1168
		AF15R	*CAGGAAACAGCTATGAC*ACGAGAAAAAAAGTGTCAAAAACT	15222	

Table 2.8: Primers used for amplifying RSV B with PCR. W= A or T, R = A or G, M is C or A, Y = T or C. All primers were ordered from Metabion. Italics underlined = M13.

Binding region	**Fragment**	**Primer name**	**Sequence**	**Nucleotide position**	**Fragment size**
Leader, NS1, NS2, N	1	BF1F	*GTAAAACGACGGCCAG*ACGCGAAAAAATGCGTACTAC	1	1255
		BF1Falt	*GTAAAACGACGGCCAG*ACACTCGAAAAAAATGGGGC	30	1226
		BF1rev	*CAGGAAACAGCTATGAC*ACATCATAATTGGGAGTGTCAA	1256	
NS2, N, P	2	BF2F	*GTAAAACGACGGCCAG*AACCCGTAACTTCCAACAAA	992	1374
		BF2R	*CAGGAAACAGCTATGAC*TTCAGGTGCAAACTTCTCCAT	2366	
N, P, M	3	BF3F	*GTAAAACGACGGCCAG*GAAGTTGGGAGGAGAAGCT	2013	1447
		BF3R	*CAGGAAACAGCTATGAC*CCTTTGGGCGTAGAGATCTG	3460	
	4	BF4F	*GTAAAACGACGGCCAG*GACTTGTTGGAAGACAACGA	3018	1330

P, M, SH		BF4R	*CAGGAAACAGCTATGAC*GGGCCAAAATTTGCTTGTGA	4348	
M, SH, G	5	BF5F	*GTAAAACGACGGCCAG*TCCCACTCAAAATCCAAAATCAC	4137	1286
		BF5R	*CAGGAAACAGCTATGAC*GGTTGATGGTGGTTTCTTTT	5423	
G, F	6	BF6F	*GTAAAACGACGGCCAG*CATCTCTGCCAATCACAAAG	4873	1512
		BF6R	*CAGGAAACAGCTATGAC*GCATTGACACTAAATTCTCTGGT	6385	
F	7	BF7F	*GTAAAACGACGGCCAG*GTAGGATCTGCAATAGCAAG	6136	1226
		BF7R	*CAGGAAACAGCTATGAC*GTTGATTTGGGATTGGTGGTCA	7362	
F, M2, L	8	BF8F	*GTAAAACGACGGCCAG*CCCTCTAGTGTTTCCTTCTG	7100	1428
		BF8R	*CAGGAAACAGCTATGAC*TCCATTAATAATGGGATCCATTT	8528	
M2, L	9	BF9F	*GTAAAACGACGGCCAG*TACCAGCAGACGTGCTGAAG	8071	1434
		BF9R	*CAGGAAACAGCTATGAC*TTTATTATGTAAAAGCCTTCATTATG	9505	
L	10	BF10F	*GTAAAACGACGGCCAG*CAATGCAACATCCTCCATCA	9094	1268
		BF10R	*CAGGAAACAGCTATGAC*TTCTTTACCAGTTAGTGATAC	10362	
L	11	BF11F	*GTAAAACGACGGCCAG*AGTGGATCTTGAAATGATAATAAAT	10097	1380
		BF11R	*CAGGAAACAGCTATGAC*AGGATCACCACCACCAAACA	11477	
L	12	BF12F	*GTAAAACGACGGCCAG*GAGTAGGTCCATGGATAAATAC	11173	1361
		BF12R	*CAGGAAACAGCTATGAC*TGACTGTTAACCGGTGTAAATA	12534	
L	13	BF13F	*GTAAAACGACGGCCAG*TGGACATTAAATATACAACTAGCA	12190	1200
		BF13R	*CAGGAAACAGCTATGAC*TTAACAACCCAAGGGCATAC	13390	
L	14	BF14F	*GTAAAACGACGGCCAG*AAAAAGATTGGGGAGAGGGGTA	13051	1334
		BF14R	*CAGGAAACAGCTATGAC*TGCACTTTCTTACATGCTTACTC	14385	
L, end	15	BF15F	*GTAAAACGACGGCCAG*GGTGAAGGAGCTGGTAACTT	14064	1161
		BF15R	*CAGGAAACAGCTATGAC*ACGAGAAAAAAAGTGTCAAAAACT	15225	

2.2.7.2. RNA conversion and cDNA amplification in one-step PCR reaction

In this reaction, the RNA was converted and subsequently amplified in one step. 1 µl of the enzyme mix (containing both SuperScript III Reverse transcriptase and Platinum Taq HiFi) was added to 25 µl of 2X reaction mix, 4 µl of primer mix and 15 µl of RNase-free water. 5 µl of RNA was added to the master mix and then transferred to the thermal cycler. The samples were heated to 42^{o}C for 60 minutes, then 94^{o}C for 2 minutes for the reverse transcription (RT). Immediately after, the samples entered 5 cycles of 94^{o}C for 30 seconds, then 44^{o}C for 30 seconds and then 68^{o}C for 3 minutes. Next, the samples entered 31 cycles of 94^{o}C for 30 seconds, 57^{o}C for 30 seconds and 68^{o}C for 3 minutes.

Lastly, the samples were kept at 68^{o}C for 5 minutes. Afterwards, the samples were kept at 4^{o}C until further processing.

This method used RSV sequence-specific primers rather than random primers for RT and were also used for amplification of 6 fragments each covered by 4 or 6 possible primer pairs (Figure 2.2). The primers used were published by Agoti *et al.* in 2015 (112). Primer mixes for each fragment contained 10 µM of each forward and reverse primer (Table 2.9 and 2.10).

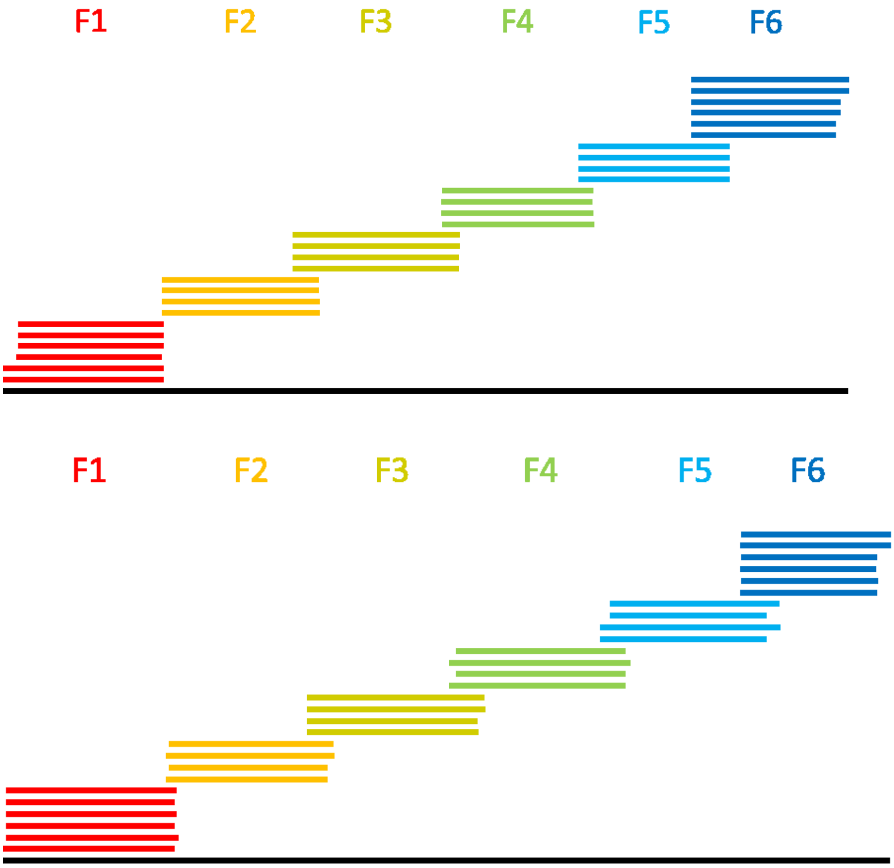


Figure 2.2: Coverage of RSV A (top) and RSV B (bottom) genome in 6 amplicons. Each fragment is covered by multiple slightly different amplicons depending on the primers used. The primers all have different hybridisation positions, although all closely together for each amplicon.

Table 2.9: Primers used for RSV A genome amplification in 6 fragments with size depending on primer pairing (112).

Fragment	Primer name	Sequence (5' to 3')	Nucleotide position	Fragment sizes
F1	AStartF	ACGCGAAAAAATGCGTACAAC	1	2877, 2878, 2612, 2609,
	A52F	TGTGCATGTTATTACAAGTAGTGATATTTG	266	

	A50F	GCATGTTATTACAAGTAGTGATATTTGCC	269	2609, 2610
	A175R	TTCTCTTAAACCAACCATGGCATCT	2878	
	A39R	CTTCTCTTAAACCAACCATGGCATC	2879	
F2	A117F	ATAAGAGATGCCATGGTTGGTTTAAGA	2849	2825, 2823
	A86F	AAGAGATGCCATGGTTGGTTTAAGA	2851	
	A1644R	CAACTCCATTGTTATTTGCCCC	5674	
	A1688R	CAACTCCATTGTTATTTGCCCCA	5674	
F3	A1820F	GCAGCATATGCAGCAACAATC	5207	2989, 2988, 2996, 2995
	A1914F	CAGCATATGCAGCAACAATCCAA	5208	
	A341R	TTGTCAGGTAGTATCATTATTTTTGGCATG	8196	
	A312R	AGGATATTTGTCAGGTAGTATCATTATTTTTGG	8203	
F4	A704F	ATGTGTTGCCATGAGCAAACTC	7893	2727, 2729, 2720, 2722
	A731F	GCCATGAGCAAACTCCTCACT	7900	
	A497R	GCTTGATTGAATTTGCTGAGATCTGT	10620	
	A539R	ATGCTTGATTGAATTTGCTGAGATCTG	10622	
F5	A374F	AAGAGAACTCAGTGTAGGTAGAATGTTT	10360	2710, 2715, 2707, 2712
	A350F	AGAACTCAGTGTAGGTAGAATGTTTG	10363	
	A364R	TTATATATCCCTCTCCCCAATCTTTTTCAAA	13070	
	A385R	ATCAGTTATATATCCCTCTCCCCAATCTT	13075	
F6	A1220F	GATTGGGTGTATGCATCTATAGATAACAAG	12386	2597, 2596, 2677, 2676, 2837, 2836
	A1232F	ATTGGGTGTATGCATCTATAGATAACAAG	12387	
	A4066R	GTTGTATAACAAACTACCTGTGATTTTAATCAG	14983	
	A5632R	TAACTATAATTGAATACAGTGTTAGTGTGTAGC	15063	
	AEndR	ACGAGAAAAAAAGTGTCAAAAACTAATA	15223	

Table 2.10: Primers used for RSV B genome amplification in 6 fragments with size depending on primer pairing.

Fragment	Primer name	Sequence (5' to 3')	Nucleotide position	Fragment sizes
F1	BStartF	ACGCGAAAAAATGCGTACTACA	1	2936, 3962, 2893, 2919 2892, 2918
	B3F	TGGGGCAAATAAGAATTTGATAAGTGC	44	
	B1021F	GGGGCAAATAAGAATTTGATAAGTGCTATT	45	
	B50R	AGTCTTGCCATAGCCTCTAACCT	2937	
	B95R	CCATTTTTTCGCTTTCCTCATTCCTA	2963	

F2	B33F	ATATTAGGAATGCTCCATACATTAGTAGTTG	2777	2771, 2707, 2885, 2821
	B71F	TAAGAGATGCTATGGTTGGTCTAAGAGA	2841	
	B7442R	GATGTGGAGGGCTCGGATG	5548	
	B7423R	CCATGGTTATTTGCCCCAGATTTAAT	5662	
F3	B7884F	AGTATATGTGGCAACAATCAACTCTG	5202	2928, 2924, 3045,3041
	B7996F	TATGTGGCAACAATCAACTCTGC	5206	
	B3652R	GCTTATGGTTATGCTTTTGTGGATATCTAAT	8130	
	B3660R	GCAATCATGCTTTCACTTGAGATCAA	8247	
F4	B3762F	AGAGGTCATTGCTTGAATGGTAGAA	7642	3031, 2911, 3104, 2984
	B3712F	AAGAGCATAGACACTTTGTCTGAAATAAG	7762	
	B47R	CCATGCAGTTCATCTAATACATCACTG	10673	
	B168R	TGCATGTCTATATGTACATATTATTGTGACAAG	10746	
F5	B32F	AAGAAGAGTACTAGAGTATTACTTGAGAGATAA	10236	2852, 2676, 3089, 2913
	B52F	AAATCCAAATCTTAGCAGAGAAAATGATAG	10412	
	B27R	TTAATGAACATATGATCAGTTATATACCCCTCT	13088	
	B60R	AACTTAAAACTGTGACAGCCTTTTATTCT	13325	
F6	B651F	ATCGACATTGTGTTTCAAAATTGCATAAG	12640	2337, 2338, 2324, 2325, 2576, 2563
	B165F	TTCAAAATTGCATAAGTTTTGGTCTTAGC	12653	
	B1199R	ATAGTACACTACCTGTTATTTTAATCAGCTTCT	14977	
	B989R	TATAGTACACTACCTGTTATTTTAATCAGCTTC	14978	
	BEndR	ACGAGAAAAAAAGTGTCAAAAACTAATGT	15216	

2.2.8. Agarose gels and E-gels

2.2.8.1. 1% agarose gels

The presence of each fragment was checked on 1% agarose gels. To prepare the 1% agarose gels, 1g of agarose was mixed with 100 ml of 1X TBE (10X TBE diluted 1:10 with pure water) and heated to just under boiling temperature until all crystals were dissolved. The liquid was poured into a gel rack and left to cool and set. Once cooled, the gel was put in the running unit and submerged in 1X TBE. 10 µl of each fragment was mixed with 2 µl of 6X Massruler Loading Dye and loaded onto the gel. 2 µl of 1 Kb ladder was mixed with 2 µl of 6X Massruler Loading Dye and added to either side of the sample lanes. The gel was run at 150 V and 100 mA for 60 minutes. After the run, the gel was stained with Ethidium Bromide for 15 minutes to let the UV detectable dye intercalate with the cDNA fragments.

Lastly, bands were visualised using the Molecular Imager® Gel Doc^{TM} XR System and pictures taken with Quantity One Software for Gel Doc XR.

2.2.8.2. Pre-made 1% agarose E-gels

5 µl of each fragment was mixed with 10 µl of loading dye. The loading dye consisted of 10X BlueJuice™ diluted 1:50 in 1X TBE. The sample mix was loaded onto the pre-made 1% agarose E-gel together with 1 µl of 1 Kb ladder in 14 µl of diluted Blue Juice in the designated marker lanes on either side of the gel. The E-gel was run for 21 minutes on the Mother E-Base**™** device on the PG program for E-gels. For visualisation, the gel was taken to the Molecular Imager® Gel Doc^{TM} XR System straight away because of the Ethidium Bromide which is present in the E-gel.

2.2.9. Clean-up of PCR products

2.2.9.1. AmPure XP

For Sanger sequencing, samples were cleaned with AMPure XP before they were quantified. 90 µl of Agencourt AMPure was added to 50 µl of sample and mixed thoroughly. Then, the beads were separated from solution on the Agencourt SPRIPlate 96R. The supernatant was discarded and the beads were washed twice with 200 µl of 70% ethanol which was removed after 30 seconds of incubation. The samples were then air-dried for 10 minutes after which 40 µl of RNase-free, deionized water was added to detach the sample amplicons from the beads.

2.2.9.2. QIAquick PCR purification Kit

Samples for Illumina sequencing were cleaned up using the QIAquick PCR purification Kit with the following adjustments. 600 µl of pH indicator I was added to 150 ml of PB buffer and 96-100% ethanol was added to PE buffer to prepare all buffers. Sample was diluted 1 in 5 with prepared PB buffer. When samples showed an orange colour (meaning they were too basic), 3 µl of 3 M sodium acetate at pH = 5 was added and the sample mixed to decrease the pH to optimal levels. The sample was added to a QIAquick spin column and spun at 13000 rpm for 1 minute. The flow through was discarded and 750 µl of PE buffer was added to the column and incubated for 1 minute. The tube was spun for 1 minute at 13000 rpm and the flow through was discarded. The column was then spun dry for another minute at 13000 rpm. The amplicons were eluted by adding 65 µl of RNase-free water and incubating for 2 minutes before spinning the column for 1 minute at 13000 rpm. Cleaned PCR products were kept at 4^{o}C.

2.2.10. Amplicon quantification

For sequencing with Illumina MiSeq, 6 ng/µl of amplicons was needed. To quantify the amplicons of cleaned PCR products, the Qubit™ dsDNA HS Assay Kit was used according to protocol. Master mix was prepared by adding 199 µl of dsDNA HS buffer to 1 µl of dsDNA HS reaction dye for each sample and for standards. 10 µl of each pre-diluted standard was mixed with 190 µl of Master mix and 1 µl of sample was mixed with 199 µl of Master mix and then measured on a Qubit® 3.0 Fluorometer using dsDNA at High Sensitivity. The target quantity was 6 ng/µl. If the amplicon content was too high, samples were diluted and quantified again.

2.2.11. Sanger sequencing

For Sanger sequencing, 4 or 5 nested primers and forward and reverse M13 primers were used per fragment to cover the whole RSV genome in 15 fragments (see Table 2.11 and 2.12). The nested sequencing PCR was performed by adding 1 µl of nested primer, 6 µl of RNase-free, deionized water and 8 µl of BigDye® Terminator Ready Reaction Mix to 5 µl of quantitated PCR product of the sample. Two pGEM internal controls were added to each plate. Sequencing reactions were run on Veriti™ 96-well Thermocycler. Afterwards, PCR products were cleaned-up using 10 µl of CleanSeq per sample. After adding 62 µl 85% ethanol and throughout mixing, the beads were separated from solution with a magnet. The supernatant was removed and the beads were washed twice with 100 µl of 85% ethanol. The samples were then air-dried to remove all left-over ethanol. The products were eluted by adding 40 µl of RNase-free water and incubating for 2 minutes. The supernatant was transferred to the sequencing plate and loaded onto the 3730xl DNA Analyzer.

2.2.11.1. Nested primers for Sanger sequencing

Table 2.11: Nested primers for Sanger sequencing of RSV A.

	Direction	ID	Sequence 5' --> 3'	Position* (5' end)
M13	Forward		GTAAAACGACGGCCAG	
M13	Reverse		CAGGAAACAGCTATGAC	
F1	Reverse	A F1 STARTrev	TGGCATTGTTGTGAAATTGG	341
	For1	A F1 n1 for85	TCCCTTGGTTAGAGATGG	85
	For2	A F1 n2 for448	TAAGTGATTCAACAATGACC	448
	Rev2	A F1 n2 rev822	ATAGTTGACCAGGAATGTAA	822
	Rev3	A F1 n3 rev1166	ATTCAACTTGACTTTGCTAA	1166
F2	For1	A_F2_I1_For_1360	TAGGAAGAGAAGACACC	1360
	For2	A_F2_I2_For_1596	TCTCCTGATTGTGGRAT	1596
	Rev1	A_F2_I1_Rev_1969	CACTAGCRTGTCCTARC	1969
	Rev2	A_F2_I2_Rev_1629	TCCTAATCACRGCTGTA	1629
F3	For1	A_F3_I1_For_2368	TTCCATGGAGAAGAYGC	2368
	For2	A_F3_I2_For_2564	YAGGGAACAAGCCCAA	2564
	Rev1	A_F3_I1_Rev_2994	GAGAGACACTTCATCTG	2994

	Rev2	A_F3_I2_Rev_2826	GTAGGTCCTGCACTYG	2826
F4	For1	A_F4_I1_For_3362	ATATGGGTGCCCATGTT	3362
	For2	A_F4_I2_For_3751	CYTAAGATCCATYAGYG	3751
	Rev1	A_F4_I1_Rev_4025	CCATGGGTTTGATTGCA	4025
	Rev2	A_F4_I2_Rev_3677	GGGTTGAGTGTYTTCAT	3677
F5	For1	A_F5_I1_For_4349	CTGGCCTTACTTTACAC	4349
	For2	A_F5_I2_For_4701	AAAMCAAGGACCAACGC	4701
	For3	A_F5_I3_For_5026	CACCACCATACTAGCTT	5026
	Rev2	A_F5_I2_Rev_4891	GTTTGCCGAGGCTATG	4891
	Rev3	A_F5_I3_Rev_4601	GGYTTGCATGGTGRGA	4601
F6	For1	AF6F1_5081	AAGACCAAAAACACAACAA	5081
	For11	A F6 n1 for5134	ACAACGCCAAAACAAAC	5137
	Rev1	AF6R1_5481	AAGGTTTCCATTTGACTTG	5481
	Rev2	AF6R2_6003	TAAACCTTGGTAGTTCTCTT	6003
	Rev3	AF6R3_6370	AATTCCCTGGTAATCTCTAG	6370
F7	For1	A F7 n1 for6605	ACTACACACATCCCCTC	6605
	Rev1	A F7 n1 rev6479	TTTGAACATTGTTGGACATT	6479
	Rev2	A F7 n2 rev6967	TAATCGCACCCGTTAGA	6967
	Rev3	A F7 n3 rev7325	TGCTTAGTGTGACTGGT	7325
F8	For1	AF8F1_7112	CTCTGATGAATTTGATGCAT	7112
	For3	AF8F3_7981	AGAGTGTACAATACTGTCAT	7981
	Rev1	AF8R1_7648	AATGACCTCGAATTTCAAAT	7648
	Rev2	AF8R2_8123	ATGGTTATGCTCTTATGGAT	8123
F9	For1	A F9 n1 for8083	ACCATCAAAAACACATTGG	8083
	For2	A F9 n2 for8331	TCCATTGGACCTCTCAA	8331
	Rev2	A F9 n2 rev8696	ACTTAGATATTAAGGACTGTGT	8696
	Rev4	A F9 n4 rev9402	AACATTATTGAATCCGCATC	9402
F10	For1	A F10 n1 for9218	GGATTTCAATTTATTTTGAATCA	9218
	For2	A F10 n2 for9537	CAGAAGAAGATCAATTCAGA	9537
	For3	A F10 n3 for10015	GGAACTTACAGAAAGAGATT	10015
	Rev1	A F10 n1 rev9424	GATAACATTATTGAATCCGC	9424
	Rev2	A F10 n2 rev9897	TATATAAAGGCACCTCTTAAC	9897
F11	For1	A F11 n1 for10453	ACAATTCTTTCCTGAAAGTC	10453

	For2	A F11 n2 for10723	AATATGCACATATAGGCATG	10723
	Rev1	A F11 n1 rev10540	TGATTTGTTACTTATTCCTGC	10540
	Rev2	A F11 n2 rev10802	ATCCACTTTGTTCATCTACA	10802
F12	For1	A_F12_I1_For_11442	CCCATGTTATTTGGTGG	11442
	For2	A_F12_I2_For_11752	TGAGTACAGCTCCAAAC	11752
	Rev1	A_F12_I1_Rev_12164	GGTGATGTAACACCAAC	12164
	Rev2	A_F12_I2_Rev_11745	CBGTAACTGCCAGTCTA	11745
F13	For1	A_F13_I1_For_12531	AGACCATGTGAATTCCC	12531
	For2	A_F13_I2_For_12770	TCCCATATTCACAGGTG	12770
	Rev1	A_F13_I1_Rev_13003	GAATCCAATGTCCAGCY	13003
	Rev2	A_F13_I2_Rev_12675	GGCCAAARCTTATACAG	12675
F14	For1	A_F14_I1_For_13365	TGCCCTTGGGTTGTTAA	13365
	For2	A_F14_I2_For_13618	ATCCTACACCWGAAACY	13618
	For3	A_F14_I3_For_13933	TGCTTCCTTGGCATCAT	13933
	Rev1	A_F14_I1_Rev_13869	GGTTRGATTTGGCTGTA	13869
F15	For1	A_F15_I1_For_14493	AAGTTAAAGGGRTCKGA	14493
	For2	A_F15_I2_For_14759	AGCTGGACGKAATGAAG	14759
	Rev1	A_F15_I1_Rev_14739	CTCCRCTAACAACACTC	14739
	Rev2	A_F15_I2_Rev_14538	CAGGACCTATWGTAAGG	14538

Table 2.12: Nested primers for Sanger sequencing RSV B.

	Direction	ID	Sequence 5' --> 3'	Position* (5' end)
M13	Forward		GTAAAACGACGGCCAG	
M13	Reverse		CAGGAAACAGCTATGAC	
F1	For1	B_F1_F1_17	ACTACAAACTTGCACAYT	17
	For2	B_F1_F2_390	TGCTCTCAATTAAAYGGTCTA	390
	For3	B_F1_F3_911	TTTATCAATCATGGCGGG	911
	Rev2	B_F1_R2_851	GTACTCCCTACTTTGTG	851
	Rev3	B_F1_R3_1257	ACATCATAATTGGGAGTGT	1257
F2	For1	B_F2_F1_993	ACCCGTAAMTTCCAACAA	993
	For2	B_F2_F2_1388	AAGATGCTGGATATCATGTT	1388
	Rev2	B_F2_R2_1965	ACTAGCATGWCCTAGCAT	1965
	Rev3	B_F2_R3_2360	TGCAAACTTCTCCATGTT	2360

F3	For1	B_F3_F1_2026	GAAGCTGGWTTCTACCATAT	2026
	For2	B_F3_F2_2430	GCATCATCCAAAGATCCTAA	2430
	Rev1	B_F3_R1_2464	ATGCTATCTTTCTTCTTAGGA	2464
	Rev2	B_F3_R2_3027	CCAACAAGTCACTCAATTTT	3027
F4	For1	B_F4_F1_3034	ACGATAGYGACAATGATC	3034
	For2	B_F4_F2_3410	AGAACTTGCAAGCATCAA	3410
	Rev2	B_F4_R2_3958	TCTTTTTCTAGGTAGGCWC	3958
	Rev3	B_F4_R3_4346	GCCAAAATTTGCTTGTGA	4346
F5	For1	B_F5_F1_4139	CCACTCAAAATCCAAAATCA	4139
	For2	B_F5_F2_4454	AACAAAACTCTTGAACWAGG	4454
	Rev2	B_F5_R2_5050	ATTDGGTGATATTGTGGCTG	5050
	Rev3	B_F5_R3_5330	CTCTTTTGTTTGTGGTTTTG	5330
F6	For1	B_F6_F1_4875	TCTCTGCCAATCACAAAG	4875
	For2	B_F6_F2_5334	CHAAAACACYAGCCAAAA	5334
	Rev2	B_F6_R2_5943	GTAATTCTGTYACTGCATTCT	5943
	Rev3	B_F6_R3_6369	CTCTGKTGATTTCCAACA	6369
F7	For1	B_F7_F1_6056	GAAGAGGAAACGAAGATTTC	6056
	For2	B_F7_F2_6397	CACCTTTAAGCACTTACATG	6397
	Rev2	B_F7_R2_7007	ATCTACTCCTTTGTTTGACA	7007
	Rev3	B_F7_R3_7358	ACTTAGTTGGTCTTTGCTTA	7358
F8	For1	B_F8_F1_7110	TTTCCTTCTGATGAGTTTGA	7110
	For2	B_F8_F2_7512	CACAACTAAGCTAGATCCTT	7512
	For3	B_F8_F3_8082	GTGCTGAAGAAGACAATAAA	8082
	Rev2	B_F8_R2_8101	TTTATTGTCTTCTTCAGCAC	8101
	Rev3	B_F8_R3_8514	GATCCATTTTGTCCCATAAC	8514
F9	For1	B_F9_F1_8082	GTGCTGAAGAAGACAATAAA	8082
	For2	B_F9_F2_8481	ACTRTCTTAATAAGGTTATGGGA	8481
	Rev1	B_F9_R1_8548	TAGATAYACATTAGCAGAG	8548
	Rev2	B_F9_R2_9076	AAGAGTGTTGTTTTGATAKTGA	9076
	Rev3	B_F9_R3_9498	TGTARAAGCCTTCATTATGA	9498
F10	For1	B_F10_F1_9095	AATGCAACATCCTCCATC	9095
	For2	B_F10_F2_9414	GYGGATTCAATAATGTTGTG	9414
	Rev2	B_F10_R2_10037	TTCTGTGATTTCAAGTAGAGA	10037

	Rev3	B_F10_R3_10348	ACCACRTGRTTAGAGTTG	10348
F11	For1	B_F11_F1_10111	TGATAATAAATGACAAAGCYA	10111
	For2	B_F11_F2_10493	TGGTGATCTAGAGCTTCA	10493
	Rev2	B_F11_R2_11086	GTCTCWGTTCCCTTRAGC	11086
	Rev3	B_F11_R3_11462	AAACAGCATAGGCAARTT	11462
F12	For1	B_F12_F1_11173	GAGTAGGTCCATGGATAAAT	11173
	For2	B_F12_F2_11579	ACAAGATAAGCTCCAGGA	11579
	Rev2	B_F12_R2_12145	AAYGACCAAGATCTTTCTCT	12145
	Rev3	B_F12_R3_12533	GACTGTTAAVCGGTGTAA	12533
F13	For1	B_F13_F1_12207	ACTAGCACTATAGCCAGT	12207
	For2	B_F13_F2_12516	TTACACCGBTTAACAGTC	12516
	Rev2	B_F13_R2_13076	TATYTACCCCTCTCCCCA	13076
	Rev3	B_F13_R3_13381	CAAGGGCATACGGTAAAT	13381
F14	For1	B_F14_F1_13059	TGGGGAGAGGGGTAYATA	13059
	For2	B_F14_F2_13405	CAACACAYATGAAAGCTATA	13405
	Rev2	B_F14_R2_13997	GCATCCTGTGGAACTAAA	13997
	Rev3	B_F14_R3_14385	TGCACTTTCTTACATGCT	14385
F15	For1	B_F15_F1_14066	TGAAGGAGCTGGTAACTT	14066
	For2	B_F15_F2_14368	AGCATGTAAGAAAGTGCA	14368
	For3	B_F15_F3_14894	AGAGTCCACATATCCTTACT	14894
	Rev2	B_F15_R2_14955	TCTTGAGCTCATTGGTTG	14955
	Rev3	B_F15_R3_15212	TGTCAAAAACTAATGTCTCG	15212

2.2.12. Illumina sequencing

2.2.12.1. Sample library preparation with Nextera XT NGS Library Prep

The sample containing 6 ng/µl of cDNA was diluted to 0.2 ng/µl. The first step of library preparation was tagmentation. This was performed by adding 5 µl of Tagment DNA buffer and 2.5 µl of amplicon tagment mix to 2.5 µl of diluted sample cDNA, which was heated to 55^{0}C for 5 minutes and then cooled down to 10^{0}C. 2.5 µl of Neutralise Tagment buffer was added, mixed thoroughly and the plate was spun down at 280g for 1 minute. Then, the samples were incubated at room temperature for 5 minutes.

The second step of library preparation was PCR amplification. For each sample, 7.5 µl of Nextera PCR Master Mix, 2.5 µl of i7 index primers and 2.5 µl of i5 index primers were added and mixed. Samples were transferred to a thermal cycler and run at 72°C for 3 minutes, 95°C for 30 seconds, then 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds, 72°C for 30 seconds, and ended with 72°C for 5 minutes. Afterwards, the products were kept at 10°C.

The third step of library preparation was a clean-up with AMPure XP beads. 15 µl of mixed AMPure XP beads were added to the samples, mixed thoroughly and incubated at room temperature for 5 minutes. Then, the plate was transferred to a magnetic stand for 2 minutes after which the supernatant was removed. The beads were washed twice with 80% fresh ethanol and after 30 seconds of incubation, all the supernatant was removed. Following the second wash, the beads were air dried for 15 minutes. Next, the plate was removed from the magnetic stand and 25 µl of resuspension buffer was added and mixed thoroughly. After 2 minutes of incubation at room temperature, the plate was put back on the magnetic stand for 2 minutes and 25 µl of the supernatant was transferred for the next step.

After library preparation the fourth step was to check the quality of the libraries produced. This was dependent on fragment size and library concentration.

Fragment size was determined using the HT DNA High sensitivity LabChip kit. First, the gel-dye solution was prepared. 1 ml of DNA gel matrix was mixed with 13 µl of HT DNA Dye Concentrate and spun for 10 minutes in a filter tube at 9200g. Then, the sample buffer was prepared. 600 µl of Molecular Biological Grade water was mixed with 150 µl of DNA sample buffer. Next, the DNA ladder was prepared. 108 µl of Resuspension Buffer was mixed with 12 µl of 10X DNA HiSens Ladder. A second tube was prepared with a 0.2X DNA HiSens Ladder by adding 96 µl of Resuspension Buffer to 24 µl of the 1X DNA HiSens Ladder. The chip was loaded by adding 75 µl of gel-dye solution to well numbers 3, 7, 8 and 120 µl in well 10 (Figure 2.3). In well number 4, 75 µl of DNA HiSens Marker is added. Then, the chip was loaded on the LabChip GX instrument. The prepared sample buffer and 0.2X DNA Ladder were added to the LabChip GX as well and the DNA Chip was then primed for 10 minutes.

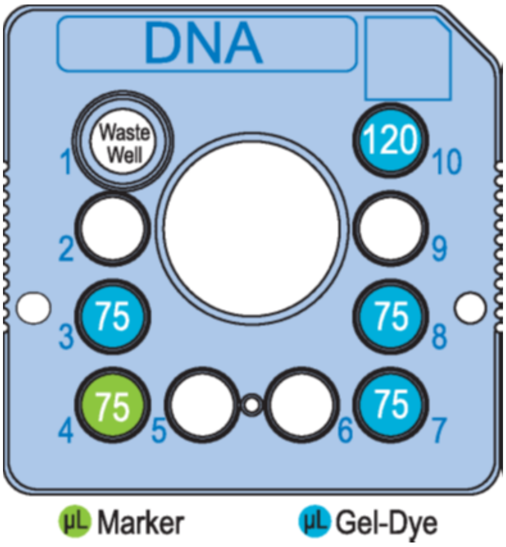


Figure 2.3: LabChip diagram showing how to load the gel-dye solution and marker for fragment size analysis. Figure taken from Perkin Elmer manual "HT DNA High Sensitivity LabChip Kit LabChip GX/GXII User Guide".

To prepare the DNA samples, 6 µl of Molecular Biological Grade water was mixed with 4 µl of DNA sample. After 10 minutes of priming of the DNA Chip, the prepared samples were also loaded onto the LabChip GX and the HT DNA High Sensitivity assay was run. The output file "smearanalysis" was used to calculate the fragment size with the parameters set as follows. Start size = 200, end size = 4000, colour = Red, name = Region[200-4000], property displayed in well table = Size (bp) and apply to wells = all. The resulting files with fragment sizes for each well were exported and used to calculate the average fragment size.

The library concentration was determined with the GloMax® Discover System and Quant-iT™ DNA Assay, High Sensitivity kit. First, 8 standards of the Quant-iT™ DNA Assay Kit were prepared. 500 µl of 4 ng/µl standard was mixed with 500 µl of Molecular Biological Grade water to make a 1 in 2 dilution. 6 more 2-fold dilutions were prepared with the last dilution being 0.003125 ng/µl. 500 µl of Molecular Biological Grade water was used as a negative control. The Quant-iT Working Solution was prepared by mixing 1 µl of Quant-iT reagent with 199 µl of Quant-iT buffer for each sample and each standard. 10 µl of standards and 5 µl of samples were mixed with 190 µl or 195 µl of Quant-iT working solution respectively. The samples were incubated for 2 minutes at room temperature and then loaded onto the GloMax® Discover System. The parameters used were as follows. Protocol = Quant-it HS 2ng/µl 96 well, protocol component = Fluor, optical kit = Blue, reading = 1, loops = 1 and plate type = 96. The output file was a csv file which contained the concentrations of the libraries.

The fragment size corrected concentration was calculated, which was dependent on fragment size and library concentration and calculated using the formula below (Equation 2.2) and if library preparation was successful, the libraries were diluted to 0.5 nM in Molecular Biological Grade water.

$$nM = \frac{Library\ concentration\ (\mu g/\mu l) * 1000000}{Average\ fragment\ size * 660\ g/mol}$$

Equation 2.2: Formula used to calculate the corrected concentration of fragment size based on library concentration and fragment size.

The fifth step was library normalisation which was performed using the Nextera XT DNA Library Prep Kit. 44 µl of Library Normalisation Additives 1 buffer and 8 µl of Library normalisation beads 1 was mixed. 45 µl of the bead working solution was combined with 20 µl of sample library (as prepared in step 3) and mixed thoroughly for 30 minutes. The libraries were put on a magnetic stand and incubated for 2 minutes. All supernatant was removed and the beads were washed twice with Library Normalisation Wash 1 by adding 45 µl of LNW1 and mixing for 5 minutes, then incubating on a magnetic stand for 2 minutes. The supernatant was discarded. Next, 30 µl 0.1M of NaOH was added to the beads and mixed for 5 minutes to denature and elute the libraries. Lastly, the samples were put back on the magnetic stand and incubated for 2 minutes and the supernatant was transferred into 30 µl of Library Normalisation Storage buffer 1.

2.2.12.2. Preparation PhiX control library and pooling of sample libraries

After normalisation of the libraries, a pooled library could be produced. First, a control library with 20 pM of PhiX was prepared to be pooled with the sample libraries. 2 µl of the 10nM PhiX library was mixed with 3 µl of EBT buffer, which consists of 10mM TRIS and 0,1%Tween20, and 5 µl of 0.2M NaOH. This was vortexed and centrifuged at 280g for 1 minute and then incubated at room temperature for 5 minutes to denature the PhiX library. Lastly, 990 µl of hybridisation buffer was added to create a 20 pM denatured PhiX library.

To pool the libraries, 5 µl of each sample library were added together to create the Pooled Amplicon Library (PAL). PAL concentration was determined using KAPA Real-Time PCR. First, the prepared qPCR master mix was made by adding 1 ml of primer mix to 5 ml of KAPA SYBR® FAST qPCR master mix. Then, the samples were prepared by adding 1 µl of PAL to 999 µl of Molecular Biological Grade water. 12 µl of the prepared qPCR master mix was mixed with 3.6 µl of Molecular Biological Grade water, 0.4 µl of 50X ROX Low and 4 µl of prepared sample or Standard 1-6 or MBG water as a negative control. Standards 1-6 had concentrations of 20000 ng/µl, 2000 ng/µl, 0.2 ng/µl, 0.02 ng/µl, 0.002 ng/µl and 0 ng/µl respectively. Samples were then loaded onto the ViiA7 qPCR machine and samples were run at 95^{o}C for 5 minutes and 25 cycles of 95^{o}C for 30 seconds and 60^{o}C for 45 seconds.

PAL was then used to create 600 µl of Diluted Amplicon Library (DAL) with a concentration of 16 pM. The volume of PAL necessary was calculated with the formula below (Equation 2.3):

$$Required\ PAL\ volume = \frac{Final\ concentration\ of\ DAL\ (16\ pM) * Total\ volume\ of\ DAL\ (600\ \mu l)}{Concentration\ of\ PAL}$$

Equation 2.3: Calculation determine the required volume of pooled amplicon library based on concentration and volume of diluted amplicon library and concentration of the pooled amplicon library.

Next, the required volume of denatured PAL was diluted with HT1 and gently mixed to create a DAL. DAL was incubated in an ice water bath for 5 minutes and 19 µl of the denatured PhiX library was added. The finished pooled library was loaded onto Illumina MiSeq for sequencing.

2.2.12.3. Running the Illumina MiSeq

The MiSeq reagents were thawed and the flow cell was thoroughly rinsed with Molecular Biological Grade water to remove all salts and dried carefully with lint-free lens cleaning tissues. Then, the MiSeq cartridge was inverted 10 times to mix all buffers and then gently tapped to remove air bubbles. Using a 1 ml pipette tip, the seal of the loading well was pierced and 600 µl of DAL was added to the loading well. The cartridge was transferred to the Illumina MiSeq Instrument. The sample sheet was created using following parameters:

- Category: Small Genome Sequencing
- Application: Plasmids
- Chemistry: Default
- Reagent Cartridge Barcode: Barcode on MiSeq cartridge
- Sample Prep Kit: Nextera XT
- Index Reads: 2
- Read Type: Paired end
- Read length Read 1 and 2: 151
- Adapter Trimming: Yes
- Index1: i7
- Index2: i5

Then the flow cells was loaded onto the Illumina MiSeq Instrument and so were the PR2 and waste bottles, and the MiSeq cartridge. Lastly, the run was started. The fastq files were produced as output files and used for further data analysis.

2.3. Data analysis

2.3.1. Sanger sequencing analysis

2.3.1.1. Quality check and assembly of data

The types of files returned from Sanger sequencing were ab1 files which contained the electropherograms and consensus DNA sequence of each reaction. All ab1 files of the same sample were opened in SeqMan Pro. The ends of each electropherogram were trimmed at the ends to remove bad quality data. Then, all sequences were assembled to form a contig of the complete RSV genome. The contig was cleaned up by deleting any conflicting sequences where overlapping fragments contradicted each other and then reassembled. The complete contig was blasted to check that the sequence was indeed RSV. When multiple contigs could not be reconciled into one, a reference sequence was used to create a scaffold. Once a complete sequence was established, it was exported as a fasta file.

2.3.1.2. Alignment of sequences

The genomes were imported into MEGA7 and aligned using ClustalW with a Gap Opening Penalty of 15, Gap Extension Penalty of 6.66, DNA Weight Matrix was IUB, Transition Weight was 0.5 ad no Negative Matrix was used. Aligned sequences were exported as fasta files or meg files. These alignments showed regions of variability and what those variations were.

2.3.1.3. Phylogenetic analysis

The aligned sequences were uploaded into MEGA7 and used for indicative phylogenetic analysis. Neighbour-joining trees were created with a bootstrap value of 1000, based on the Kimura 2-parameter model with nucleotide substitutions of both transitions and transversions. A uniform rate is assumed with complete deletion of missing data.

The calculation of Maximum Likelihood trees was based on the IQ-Tree software and trees were annotated using FigTree software.

2.3.2. NGS sequencing analysis

The raw data files produced by Illumina MiSeq were fastq files. These contained sequence and quality values for each base of each read. Several software packages were used to assemble and analyse those reads. To keep the explanation as simple as possible, I have named example files in the codes “foo”.

2.3.2.1. Read assembly

Clinical samples from natural infections contained an unknown strain of RSV. Therefore, assembly of reads was initially performed without using a reference sequence. Later, the best results from *de* novo assembly were used to find the best fitting reference and use that reference as a reference for more complete assembly results (Figure 2.4).

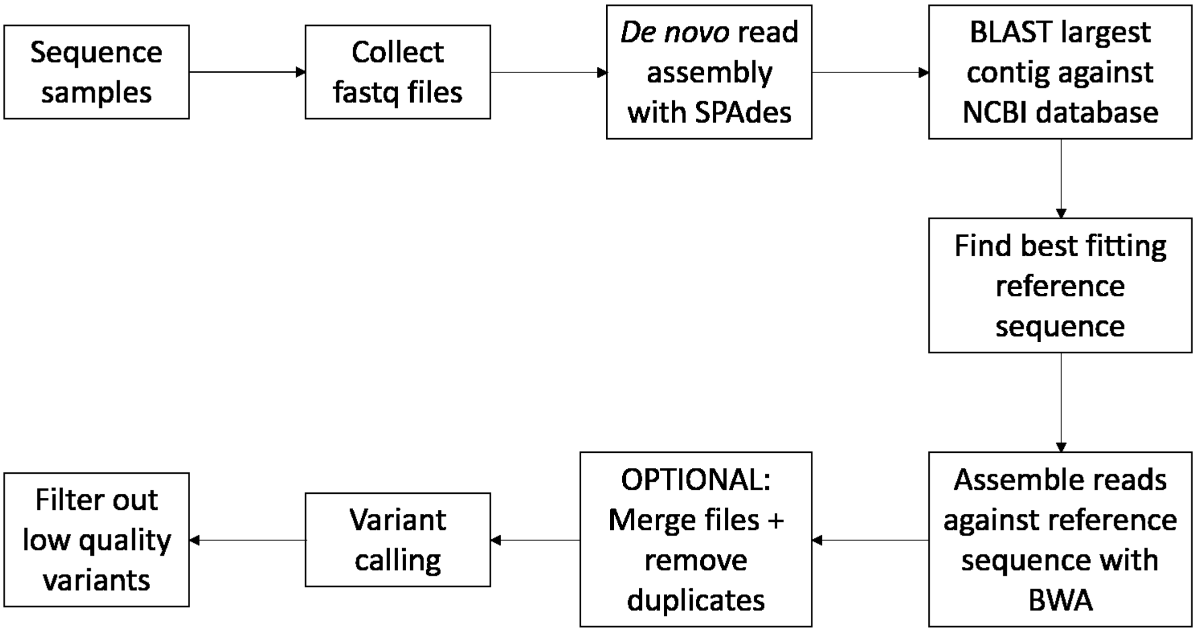

Sequence samples
Collect fastq files
De novo read assembly with SPAdes
BLAST largest contig against NCBI database
Find best fitting reference sequence
Assemble reads against reference sequence with BWA
OPTIONAL: Merge files + remove duplicates
Variant calling
Filter out low quality variants

Figure 2.4: Workflow for read assembly of RSV samples with unknown strains.

To assemble reads *de novo*, SPAdes (version 3.5.0) was used which required an input fastq file containing forward reads and an input fastq files containing reverse reads for each sample.

```
spades.py -k 21,33,55,77,99 -careful -pe1-1 foo1.fq -pe1-2 foo2.fq –
o foo.donovo
```

The resulting output folder contained files with contigs and scaffolds in fasta and fastq formats as well as log files, parameter files and output files with the corrected reads. These contigs were visualised using Bandage v0.8.1 (Figure 2.5 - 2.9) and blasted using web NCBI Blastn to determine what the best reference strain was. The fasta file of the closest strain was used as a reference in the next step.

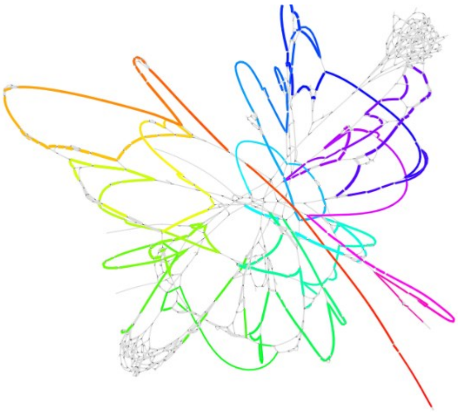


Figure 2.5: Example of RSV genome of volunteer #10 on day 6. 21-mers were used for assembly of contigs. The RSV sequence is coloured according to mapping positions to the reference sequence (red - orange – yellow – green – blue – purple). Using 21-mers resulted in several short contigs.

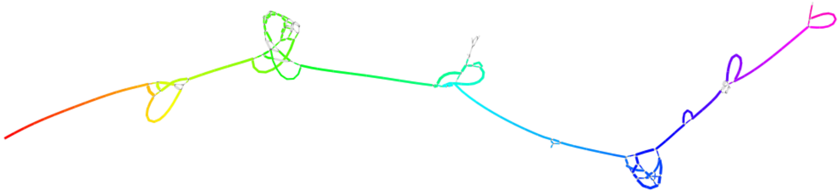


Figure 2.6: Example of RSV genome of volunteer #10 on day 6. 33-mers were used for assembly of contigs. The RSV sequence is coloured according to mapping positions to the reference sequence (red – orange – yellow – green – blue – purple). Using 33-mers resulted in a contig covering the entire RSV genome, although not positions of the genome are clearly mapped. Repeats are shown as single nodes with multiple inputs and outputs.

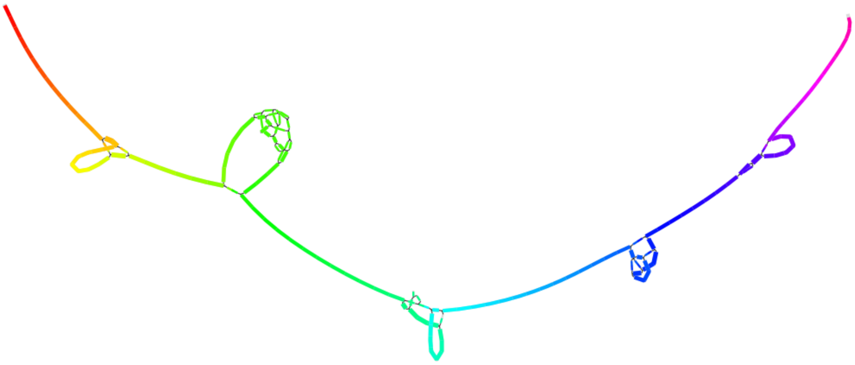


Figure 2.7: Example of RSV genome of volunteer #10 on day 6. 55-mers were used for assembly of contigs. The RSV sequence is coloured according to mapping positions to the reference sequence (red – orange – yellow – green – blue – purple). Using 55-mers resulted in a contig covering the entire RSV genome, has reduced the number of problematic regions and has improved on the difficult regions of this assembly.

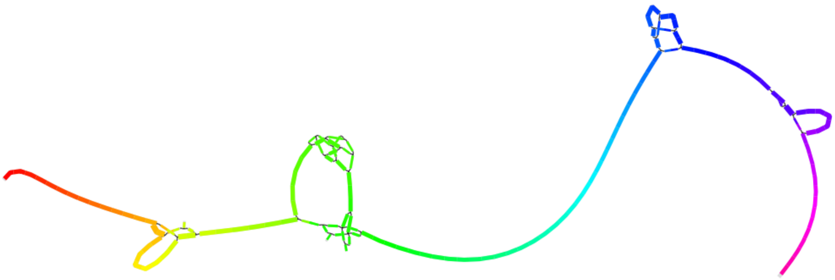


Figure 2.8: Example of RSV genome of volunteer #10 on day 6. 77-mers were used for assembly of contigs. The RSV sequence is coloured according to mapping positions to the reference sequence (red – orange – yellow – green – blue – purple). Using 77-mers resulted in a contig covering the entire RSV genome and shows similar results to an assembly using 55-mers.

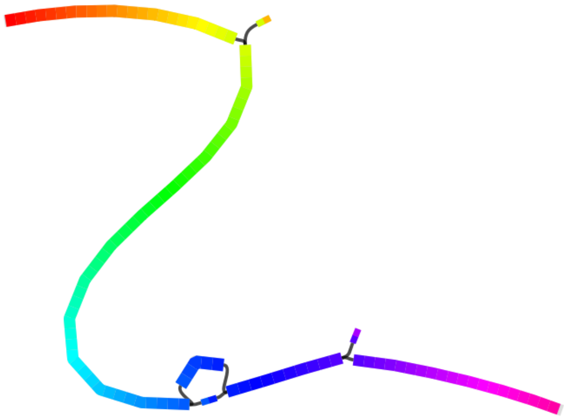


Figure 2.9: Example of RSV genome of volunteer #10 on day 6. 99-mers were used for assembly of contigs. The RSV sequence is coloured according to mapping positions to the reference sequence (red – orange – yellow – green – blue – purple). Using 99-mers resulted in a contig covering the entire RSV genome. It further reduced the number of problematic regions compared to 55- and 77-mers. Repeated regions are visualised as single notes with multiple inputs and outputs.

Burrow-Wheels Alignment (BWA) was used to assemble the reads and map them to the reference. This created more complete genomes than *de novo* assembly and the output was a sam file. Since .sam files were big files and were slow to work with, they were compressed to .bam files. The reads in the bam file were sorted according to co-ordinate position and the file was indexed.

```
bwa mem –M –R ’@RG\tID:foo\tSM:samplefoo’ reference.fa foo1.fq foo2.fq
> foo.sam
```

```
samtools view –O bam foo.sam > foo.bam
```

```
samtools sort foo.bam > foo.sorted.bam
```

```
samtools index foo.sorted.bam
```

When samples were sequenced twice, bam files were merged by the MergeSamFiles function from Picard Tools. Next, the resulting file was indexed with samtools and duplicates were marked and removed with the MarkDuplicates function from Picard Tools.

```
java –Xmx1g –jar picard.jar MergeSamFiles I=foo1.bam I=foo2.bam
O=libraryfoo.bam
```

```
samtools index libraryfoo.bam
```

```
java –Xmx1g –jar picard.jar MarkDuplicates I=libraryfoo1.bam
O=libraryfoo.markdup.bam M=libraryfoo.markdup.metrics.txt
```

To check the quality of the alignment, samtools was used to produce stats on the alignment and then plotted to show these stats visually. These stats were also visualised using FastQC v0.11.7 and rechecked.

```
samtools stats foo.sorted.bam > foo.stats.txt

plot-bamstats –p plot foo.stats.txt
```

Quality checks were performed and sequences were considered of high enough quality to work with if:

- Number of reads mapped in pairs: >80%
- Number of bases mapped: >90%
- Number of reads with adaptor: <0.01%
- Number of duplicate fragments: <5%
- Error rate: <0.02
- Indel ratio: <1
- Indels per cycle: <2000, no peaks
- Fragment size distribution: fragments of 150 bp and larger
- GC versus depth: should be equal overall

Next, the sorted bam files were used to call variants and create a bcf file. This file contained all information about any variations compared to the reference sequence. The binary files were converted to vcf files which could then be read. The variants in these files were filtered based on their proximity to indels and their quality given by the original base calling algorithm of Illumina.

```
samtools mpileup –g –f reference.fa foo.sorted.bam > foo.bcf

bcftools call –c –v foo.bcf > foo.vcf

cat foo.vcf | bcftools filter –m+ -sLowQual –e”%QUAL<=10” –g3 –G10 –
Ov > foo.filtered.vcf
```

Typical VCF calling biases were checked for:

- Strand bias: variation that was only present on strands going one direction and not on the other direction.
- End distance bias: variation that was only present when near the end of the read.
- Consistency across replicates/libraries: variation only found in one library.
- Variant distance bias: Splice-site artefacts.

The read depth and position of variations was visualised with IGV viewer to find regions that were not covered and visualising variations that were recurrent in most reads or between different samples.

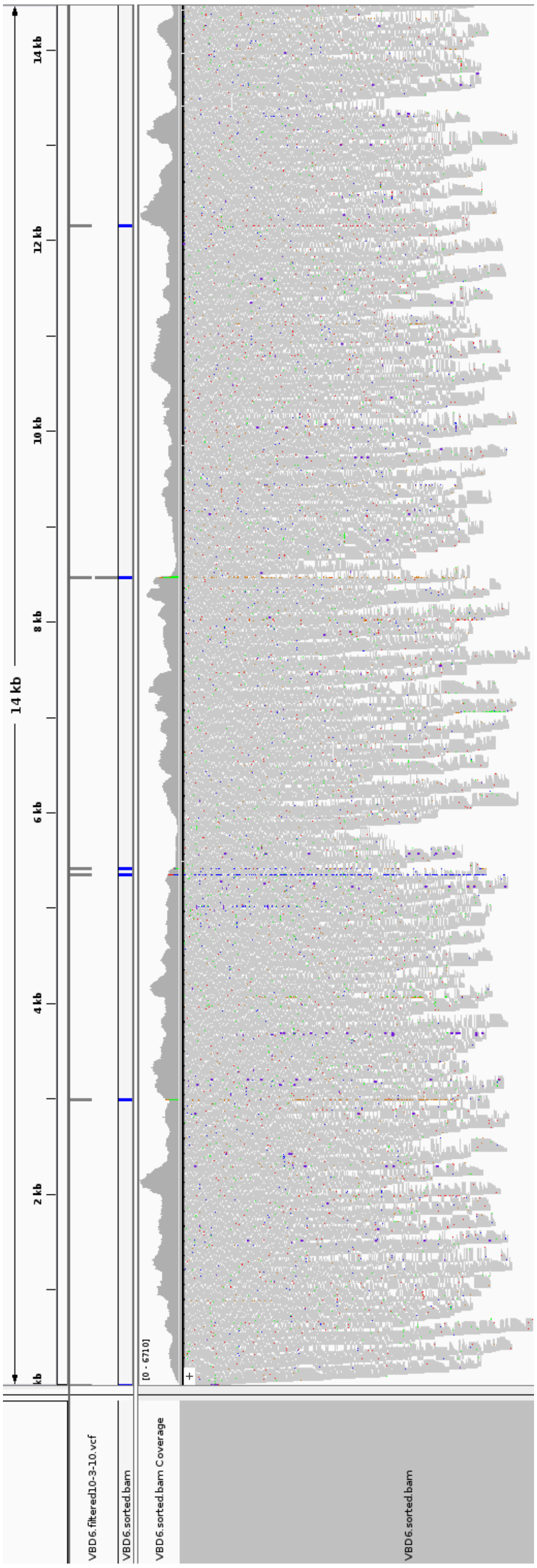


Figure 2.10: Example of IGV viewer of volunteer #10 showing filtered variations in the top panel, read depth graph in the middle panel and all reads in the bottom panel with variations from the reference coloured.

2.3.3. Phylogenetic analysis

2.3.3.1. Build dataset

Phylogenetic analysis was performed on several datasets, each one put together to answer a specific question. Datasets were built by selecting sequences from online databases. The applied filters selected for sequences from the database as follows (where "gene" stood for the gene being investigated and the sequence length ranged from the length of the gene to full genome):

```
(((Respiratory syncytial virus) AND complete "gene") AND xxx:15300[Sequence Length]) AND genomic dna rna[Filter]) AND nuccore pubmed[Filter]
```

This filtering ensured that only genome RNA or DNA was selected and only sequences published on PubMed (and therefore peer-reviewed) were selected. All sequences that passed the filtering process were added to the dataset.

Next, an annotation file was compiled of all sequences in the dataset. This annotation file contained the name, subtype and genotype of virus, country and city of sample collection, year, month and day of isolation, and gender and age of the infected person. Based on this annotation file, the dataset was divided into RSV A and RSV B subtypes. Manual inspection of the data was performed to remove any sequences from non-human hosts, mutant or recombinant viruses and non-RSV sequences. After this process, the final dataset was constructed.

2.3.3.2. Alignment of dataset

The following step was alignment of all sequences in the dataset. Three different alignment algorithms were run to check for inconsistencies between alignments. These alignment algorithms were performed by programs MUSCLE v3.8.31, Clustal Omega v1.2.2 and MAFFT v7.4. MAFFT was run with the E-INS-I option for genes with known variable regions. After running all alignments, manual editing of the alignments in AliView v1.23 and JalView v2.10.4 produced a final alignment which was used in all further steps. Command line arguments were:

```
muscle3.8.31_i86win32.exe -in foo.fasta -out TEMP_OUT_FILE
```

```
mafft.bat --localpair --reorder --maxiterate 1000 -o TEMP_OUT_FILE foo.fasta
```

```
clustalo.exe -i foo.fasta -o TEMP_OUT_FILE --force
```

2.3.3.3. Quality checks of alignment

When the final alignment was obtained, quality checks were performed to make sure this alignment was of good quality and had little ambiguous sites. This quality test was performed using Alistat v1.7.

This program showed completeness scores for the overall alignment (Ca), completeness scores for individual sequences, individual sites and completeness scores for pairs of sequences. A triangle heatmap visualised the completeness scores for pairs of sequences. The command line argument was:

```
Alistat.exe foo.fasta 1 -t 1,2,3 -r row -i 1 -d
```

2.3.3.4. *Suitability of alignment for phylogenetic analysis*

When alignments showed sufficient completeness scores to base analysis on, the suitability of the alignment for phylogenetic analysis was checked for. Sequences that were too similar or too dissimilar could not be ordered in a tree with high enough confidence intervals to conclude any information. Therefore, IQ-Tree v1.6.6 was used to perform likelihood mapping of quartets which was visualised in triangle plots. When the sequences could not be ordered in these mini-trees, the quartets were deemed ‘uninformative’. When enough ‘informative’ quartets were available, chances for trustworthy trees with high confidence were larger. An indicative neighbour-joining tree was constructed in the process as well. The command line argument was:

```
iqtree -s foo.fasta -lmap 1500 -n 0
```

2.3.3.5. *Phylogenetic tree building*

If all quality checks were passed, the best fitting model for the dataset was determined by ModelFinder (193). This process was incorporated in the IQ-Tree command that was used and forced all subsequent analysis steps to use the selected model based on the Bayesian information criterion (BIC).

Next, a maximum likelihood tree was built based on the best fitting model for the dataset using bootstrap values of 1000 and performing approximate likelihood tests (199) 1000 times for internal branches comparing the current configuration with the two alternatives by Nearest-Neighbour-Interchange (NNI) jumps. The outcome of these tests showed if specific branches in the tree were confident or doubtful configurations. The command line argument was:

```
iqtree -s foo.fasta -m MFP -alrt 1000 -bb 1000
```

3. Comprehensive bioinformatics analysis on RSV classification based on genetic variation

3.1. Introduction

RSV is a common virus that causes an acute and rapid infection with cold-like symptoms in otherwise healthy individuals. The genetic sequence of the virus is variable, at least in some parts of the genome. These variants most likely arise during acute infection after which they have to be successfully transmitted to be able to spread in the general population. Currently, RSV genomes are mostly studied on the population level, rather than on an individual level, so development of *de novo* variants is not well-studied.

It is known that certain regions of the RSV genome, like the G gene, are more prone to showing variation than others. In general, conserved genomic regions are linked to necessary features to maintain functionality of a virus, while some variable genes can be linked to immune evasion, like the hemagglutinin gene of Influenza viruses (201). These variable genes are necessary in order to replicate the viral genome before being detected and to avoid immediate activation of the host's immune system and elimination of the virus.

In this chapter, the main aim is to construct a reference dataset of known RSV genotypes to be able to genotype clinical samples. To accomplish this, several objectives were set out.

First, the original sequences for each discovered genotype were compiled into a dataset. These were mostly partial G gene sequences which ranged in length.

Then, these sequences were used to find complete genomes that were the same or most similar to the original sequences in that region. The variability for each protein was investigated and the best protein for genotyping was selected. Based on this dataset, three phylogenetic trees were constructed from either whole genomes, full G genes or the second hypervariable region (HVR2) of G. Then, their tree-branching structure was compared and investigated to study which part of the genome is necessary and sufficient for distinguishing genotypes. A set of reference sequences was selected from this dataset.

Lastly, this newly set up reference set was used to genotype 112 clinical samples collected in England between 2014 and 2018, which were sequenced by Sanger sequencing to obtain their complete genome. The G genes were used to detected the genotype of each sample by phylogenetic analysis.

3.2. History of RSV genotyping and a brief meta-analysis

RSV strains were long categorised by serotyping only and therefore divided into RSV A and RSV B. The first paper studying the difference in antigenic reactivity between different virus strains came out in 1966, 10 years after the virus was first isolated (71). The following years, it became clear there were two major groups in RSV based on serology studies, RSV A and RSV B (72, 73). This grouping of RSV strains is still used today. Experiments based on monoclonal antibody reactivity (76), restriction mapping (78, 80) and RNase A digestion in combination with electrophoresis (202) showed more detailed differences in later years. Grouping of RSV strains has become more detailed since sequencing viral genomes became more feasible and could be used to determine genotypes. Several papers described different systems of grouping strains (78, 79, 81). The system currently used the most is described in the next subchapter.

3.2.1. Current genotypes for RSV A and RSV B

The first paper written about the current genotyping system to group RSV strains based on sequencing was by Peret *et al.* in 1998 (82). Sequencing had becoming more accessible and sequences that were determined showed variability and based on these variations new groups, *i.e.* genotypes, were made. This system is still employed today and new genotypes are still being described. However, the criteria to identify a new genotype are not defined yet. Therefore, everyone can claim they have found a new genotype based on their own criteria. Thus, it is possible not all described genotypes in literature are discussed in the following part of this chapter.

Today, there are about 18 RSV A genotypes and 29 RSV B genotypes (Table 3.1), depending on the definition of what a new genotype entails. The most recently discovered genotype of importance is ON1, which was described by Eshaghi *et al.* in 2012 (86) (Figure 3.1). This RSV A genotype is clearly distinguishable from other genotypes by a 72-nucleotide duplication in the G gene. Since its discovery in 2012, it has been detected all over the world and increased in prevalence over the years.

Genotypes have been established quite arbitrarily and voices are being raised to standardise the criteria for appointment of new genotypes and to reform the current system. Each time, the discoverers have decided on names for new genotypes and there is no clear path that was followed by the scientific community. Some genotypes have only been described in one paper in one specific dataset without any confirmation from any other research group, while others have been found by many research groups. This means some sub-genotypes might have been presented as new genotypes which complicates genotyping attempts by other researchers (Table 3.1). These issues will not be discussed in this thesis, however, it is good to keep in mind that they might change the currently known genotypes in the coming years as workgroups have been established and are working on harmonisation at this very moment.

Table 3.1: Overview of papers announcing newly described genotypes of RSV A and RSV B. The first author, year of publication, described genotype, DOI and reference are listed.

First author	Year	RSV A	RSV B	DOI	Reference
Coates	1966	A	B	10.1093/oxfordjournals.aje.a120586	(71)
Anderson	1985	A	B	10.1093/infdis/151.4.626	(72)
Mufson	1985	A	B	10.1099/0022-1317-66-10-2111	(73)
Cane	1991	A	B	10.1099/0022-1317-72-2-349	(78)
Garcia	1994	A	B		(79)
Cane	1994	A	B		(81)
Peret	1998	GA1 – GA5	GB1 – GB4	10.1099/0022-1317-79-9-2221	(82)
Peret	2000	GA6 – GA7		10.1086/315508	(91)
Venter	2001	SAA1	SAB1 – SAB3	10.1099/0022-1317-82-9-2117	(88)
Trento	2003		BA	10.1099/vir.0.19357-0	(83)
Blanc	2005		URU1 – URU2	10.1007/s00705-004-0412-x	(100)
Trento	2006		BA1 – BA4	10.1128/JVI.80.2.975-984.2006	(97)
Shobugawa	2009	NA1 – NA2		10.1128/JCM.00115-09	(93)
Trento	2010		BA5 – BA6	10.1128/JVI.00345-10	(87)
Dapat	2010		BA7 – BA10	10.1128/JCM.00646-10	(98)
Arnott	2011		SAB4	10.1128/JCM.01131-11	(99)
Eshaghi	2012	ON1		10.1371/journal.pone.0032807	(86)
Baek	2012	CB-A	BA11, CB-B	10.1007/s00705-012-1267-1	(203)
Pretorius	2013	SAA2		10.1093/infdis/jit477	(92)
Cui	2013	NA3 – NA4	BA-C, CB1	10.1371/journal.pone.0075020	(94)
Khor	2013		BA12	10.1016/j.meegid.2012.12.017	(114)
Hirano	2014	ON2		10.1016/j.meegid.2014.09.030	(95)
Ren	2015		GB5	10.1002/jmv.23960	(204)
Schobel	2016	TN1 – TN2		10.1038/srep26311	(96)
Gimferrer	2016		BA13	10.1016/j.cmi.2015.09.013	(205)
Gaymard	2018		BA-Ly	10.1016/j.jcv.2018.02.004	(206)

The current genotypes for RSV A are GA1-GA7, SAA1-SAA2, NA1-NA4, ON1-ON2, CB-A and TN1-TN2 and the current genotypes for RSV B are GB1-GB5, SAB1-SAB4, BA1-BA13, URU1-URU2, CB-B, BA-C, CB1 and BA-Ly (Table 3.1). These were described in 20 different papers over 20 years (Figure 3.1). The most common genotypes of recent years are GA2 and GA5 for RSV A and several types of BA genotypes for RSV B.

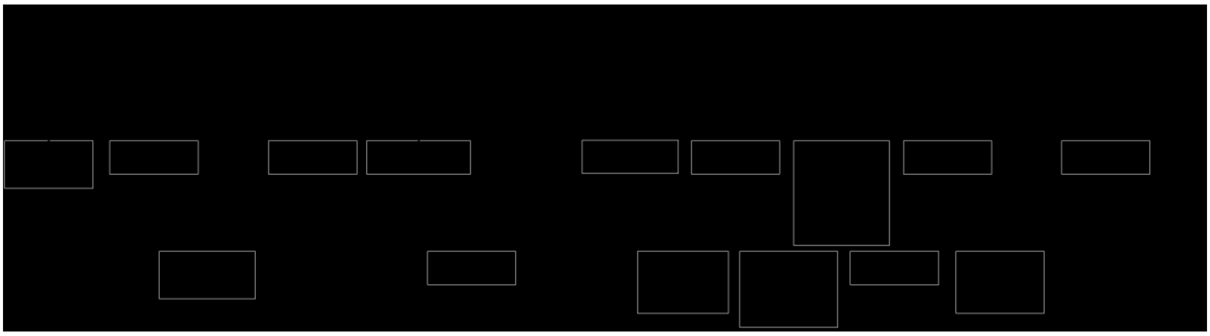
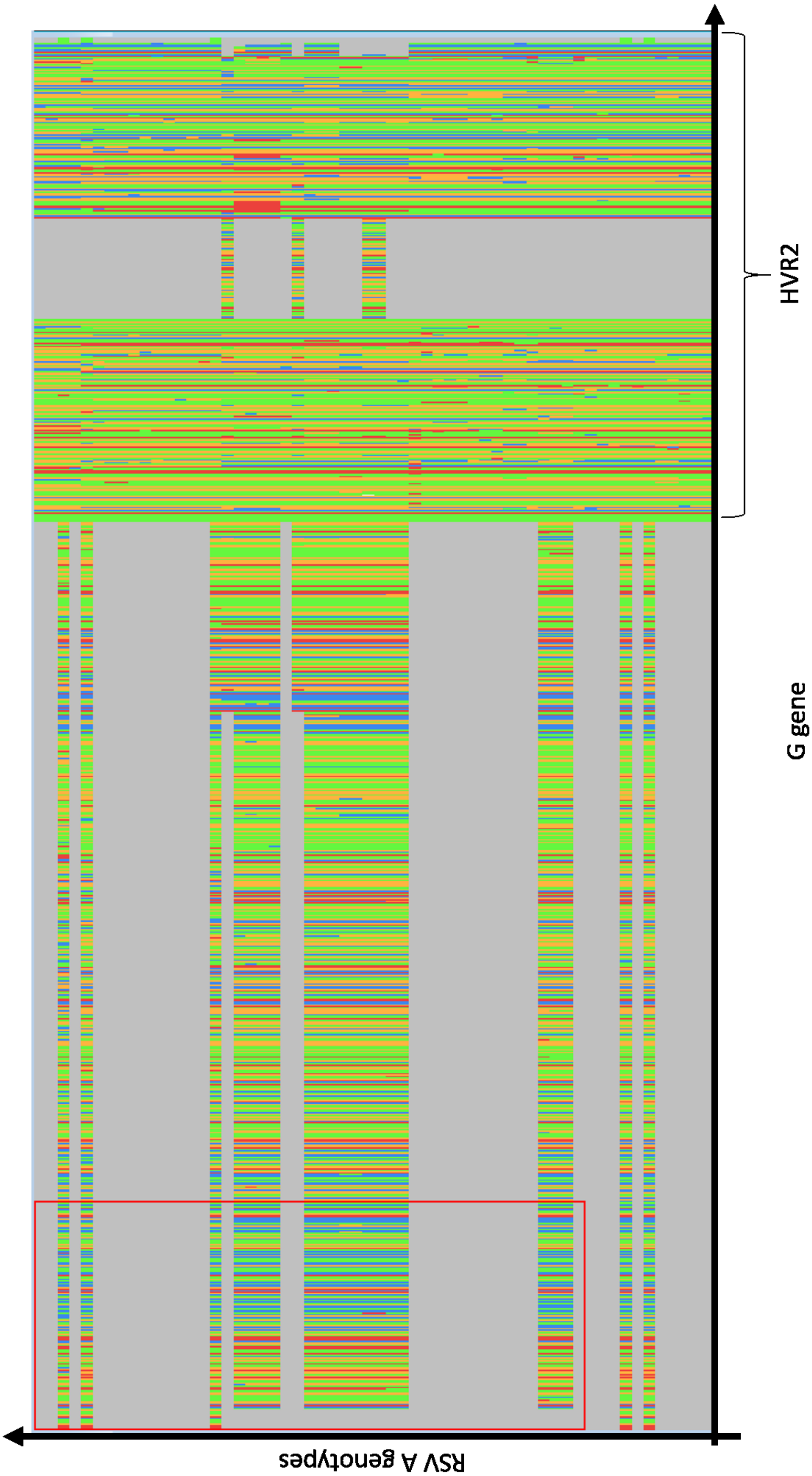


Figure 3.1: Timeline of genotypes discovered since 1998 with Peret et al. describing the first genotypes of the current genotyping system of RSV A and RSV B.

3.2.2. Usefulness of original sequences for genotype purposes

To be able to differentiate all genotypes from each other, including the genotypes that are no longer found in patients today, the original sequences from these papers have been pulled from the online, international database 'National Center for Biotechnology Information' (NCBI). Almost all papers mentioned the accession numbers of the sequences they used for their analysis, which made it possible to go back to most of the original sequences that were the first published sequences to describe that genotype. Sequences for some genotypes of RSV A (NA2 and SAA2) and RSV B (GB2, BA6 and BA11) were not found, either because they were not submitted or because sequence labelling in the paper and in the database did not match and these sequences can therefore no longer be traced.

The part of the genome used to determine new genotypes was not the same for all described genotypes (Figure 3.2). In earlier years, only a small part of the G gene was sequenced, while nowadays full G genes or full genomes are often determined. The length of the sequence of G used for genotyping varied from minimum 270 and 243 base pairs for RSV A and RSV B respectively to full G genes, which were maximum 967 and 984 base pairs long for RSV A and RSV B respectively depending on the genotype.


HVR2
G gene
RSV A genotypes

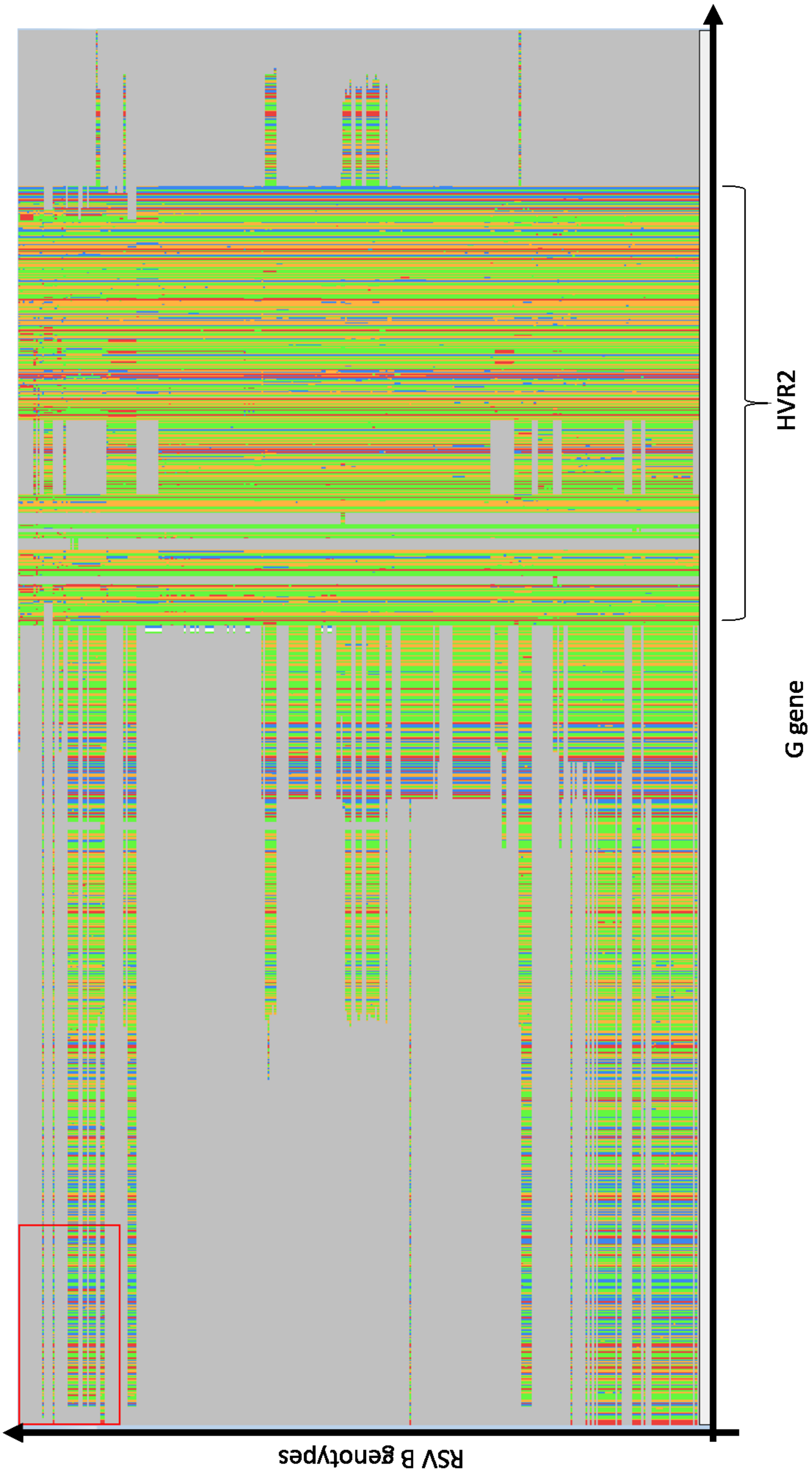


Figure 3.2: Overview in JalView of original sequences of RSV A (top, n=58) and RSV B (bottom, n=317) genotypes. The region used for genotyping ranged from 270 bp to 967 bp for RSV A and 243 bp to 984 bp for RSV B. Green = adenine (A); yellow = cytosine (C); red = guanine (G) and blue = thymine (T).

Completeness scores (C-scores) were calculated for each dataset separately to determine how complete and unambiguous the dataset was and these C-scores were calculated using AliStat v1.7 (https://github.com/thomaskf/AliStat). Different kinds of C-scores were calculated to check whether sequences in this dataset were roughly equally as long, contained gaps and if they were of good quality, meaning they contained few ambiguous positions. The RSV A dataset contained 58 sequences and the overall completeness score (Ca) for the RSV A alignment was 0.512 (Table 3.2). Minimum C-scores for individual sequences (Cr), individual sites (Cc) and pairs of sequences (Cij) in this dataset were 0.272, 0.069 and 0.263 respectively. This suggested short or gapped sequences and some extremely different sequences that only shared 26.3% of unambiguous positions in their genome as indicated by the minimum Cij score (Table 3.2). An overview of all Cij scores was plotted on a triangular heat map, which shows that more than 50% of sequences have completeness scores lower than 0.3 (Figure 3.3).

Table 3.2: Completeness scores of datasets containing original sequences of RSV A and RSV B genotypes.

	Completeness scores original RSV A sequences (n = 58)	**Completeness scores original RSV B sequences (n = 317)**
Completeness score for alignment (Ca)	0.512	0.445
Maximum completeness score for individual sequences (Cr_max)	0.973	0.866
Minimum completeness score for individual sequences (Cr_min)	0.272	0.214
Maximum completeness score for individual sites (Cc_max)	1	1
Minimum completeness score for individual sites (Cc_min)	0.069	0.003
Maximum completeness score for pairs of sequences (Cij_max)	0.972	0.866
Minimum completeness score for pairs of sequences (Cij_min)	0.263	0.198

A similar trend was seen in the RSV B dataset, which contained 317 sequences. Ca was even lower, *i.e.* 0.445, for RSV B sequences and all minimum and maximum C-scores were lower as well for this dataset (Table 3.2) compared to the RSV A dataset. Minimum Cr was 0.214, minimum Cc was 0.003 and minimum Cij was 0.198, indicating that the least similar sequences only shared 19.8% of unambiguous positions in their genomes. A triangular heat map of Cij scores showed that more than 75% of sequences had Cij scores of less than 0.5 (Figure 3.3).

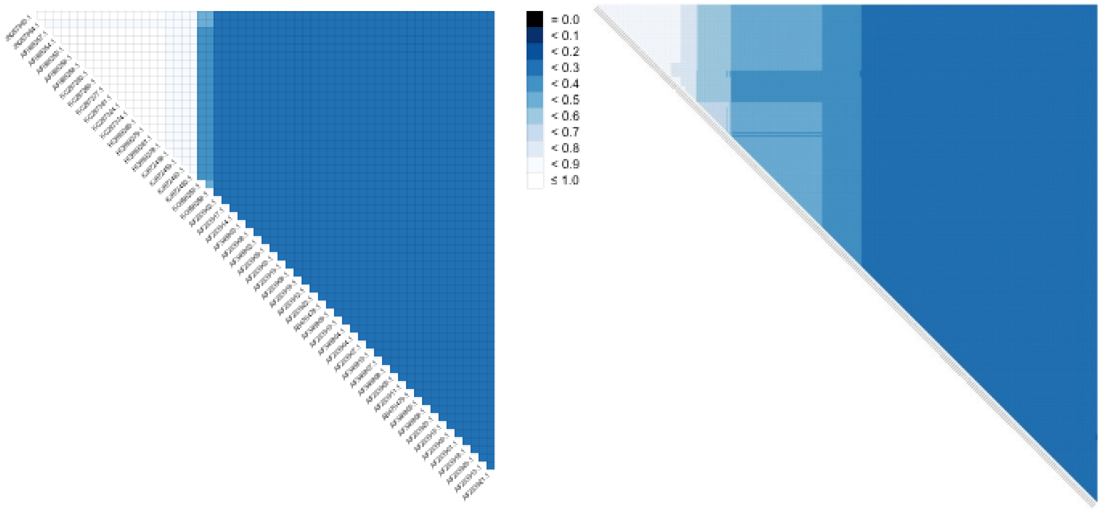


Figure 3.3: Triangle heat maps of RSV A (left, n=58) and RSV B (right, n=317) datasets showed completeness scores comparing all sequences to each other (Cij-scores).

These C-scores indicated that the datasets contained lots of ambiguous base pairs, so IQ-Tree (v1.6.6) software (193, 207) was employed to investigate the usefulness of this dataset for phylogenetic analysis. The best fitting models for the dataset were calculated from 286 DNA models and were TPM3u+F+G4 and GTR+F+G4 for RSV A and RSV B respectively according to the Bayesian Information Criterion (BIC). '+F' indicated that default empirical base frequencies were best for this dataset (no equalising or optimising necessary) and '+G4' indicated that the rate heterogeneity across sites was best based on the discrete Gamma model (208) with rate categories. There was no proportion of invariable sites ('+I') detected which would benefit the model for this dataset.

The RSV A dataset contained 175 parsimony-informative sites, 120 singleton sites and 699 constant sites. The RSV B dataset contained 284 parsimony-informative sites, 142 singleton sites and 710 constant sites. Likelihood mapping showed that random analysis of quartets from this dataset returned 26.0% and 29.0% of uninformative quartets for RSV A and RSV B datasets respectively (Figure 3.4). Only 71.4% quartets from the RSV A dataset and 66.4% of quartets the RSV B dataset were informative. This indicated that phylogenetic analysis would not be able to return a truthful consensus tree as there was not enough information to base tree-branching decisions on with confidence.

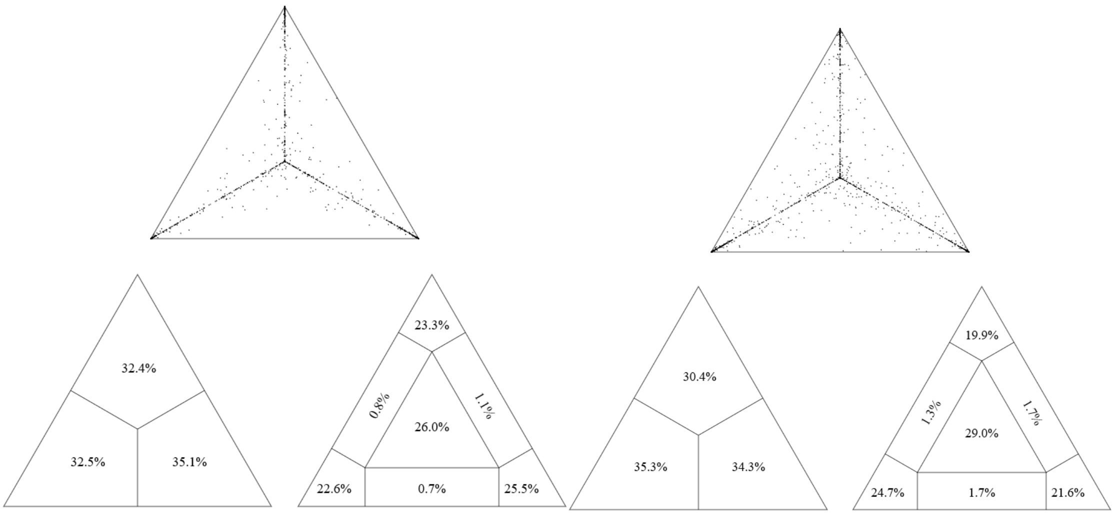


Figure 3.4: Likelihood mapping of quartets in this dataset shows that 26.0% of quartets of the RSV A dataset (left) and 29.0% of quartets of the RSV B dataset (right) are uninformative as shown in the middle section of the triangle graph and phylogenetic analysis of these datasets is therefore not useful.

This makes it impossible to perform accurate phylogenetic analysis on these datasets. To be able to genotype sequences, these partial sequences of 270 or 243 base pairs for RSV A or RSV B respectively do not give enough information to distinguish them from other genotypes. Larger parts of unambiguous positions of the genome are necessary, although it is unclear how much of the genome should be sequenced routinely to be able to classify strains into the correct (sub-)genotype category.

For this purpose, complete genomes from the NCBI database were explored and used as 'alternative genotype references' in the next part of this chapter.

3.3. Phylogenetic analysis to determine necessary and sufficient part of RSV genome for genotyping

In the previous part, the original sequences for RSV genotyping were gathered and studied. They were too short to employ for genotyping using phylogenetic analysis, as likelihood mapping showed too many uninformative quartets and therefore a phylogenetic tree would not be able to confidently show new information. So, complete genomes were gathered and studied to find full genomes that contained the exact or most similar sequence of each genotype. These sequences were used as alternatives to the original sequences to be able to determine the part of the genome that is necessary and sufficient to determine genotypes in clinical samples, while staying as closely as possible to the original sequences.

3.3.1. Selection of alternative sequences for different genotypes

Complete genomes for all genotypes were considered regardless of time and place of sample collection. The search for alternative sequences showed that some genotypes are no longer around and there were no complete genomes available that could serve as alternatives. The closest complete genome that was found by searching in online databases, like NCBI, was used and for genotypes GA2 and ON1 for RSV A, and GB1 and BA9 for RSV B, the exact same sequence as the original was found in a complete genome (Table 3.3A and Table 3.3B for RSV A and RSV B respectively). For genotypes TN1 and TN2, the original sequences were full genomes already, so they are a 100% match as well.

However, not all alternatives had complete genomes specific to that genotype. GA2, GA3, GA6 and NA4 genotypes of RSV A alternatives resemble each other and some of the available full genomes that were most similar to the original sequence were the same for these genotypes. A similar trend was spotted for RSV B alternatives where the search for SAB1-SAB4 genotypes, URU1 and URU2 genotypes, and CB1 genotype returned the same complete genomes for the available original sequences. Furthermore, all BA genotypes were so similar to each other, they all resemble the same available full genomes. Besides that, there were no good alternatives with similarities of at least 96% available for a number of original sequences, *i.e.* GB5, BA12, BA13, CB1, SAB4, and URU1 and URU2.

Table 3.3A: Overview of original RSV A sequences of each genotype and the alternative, full genomes used for phylogenetic analysis. Highlighted in yellow are selected references sequences.

RSV A genotype	Reference	Ref ID	RefSeq length	Alt (Complete seq)	Similarity	Identities
GA1	Peret. 1998.	AF065257	G-gene	KP258717.1	99.24%	915/922
	Peret.2000.	AF233902	G-gene (partial)	KU316164.1	99.26%	268/270
	Peret.2000.	AF233914	G-gene (partial)	KP258744.1	99.63%	269/270
	Peret.2000.	AF233917	G-gene (partial)	KU316123.1	98.89%	267/270
GA2	Peret. 1998.	AF065258	G-gene	KU316131.1	100.00%	922/922
	Peret.2000.	AF233900	G-gene (partial)	MG642080.1	99.63%	269/270
	Peret.2000.	AF233915	G-gene (partial)	MG642050.1	99.26%	268/270
	Peret.2000.	AF233923	G-gene (partial)	MG642033.1	97.78%	264/270
GA3	Peret. 1998.	AF065256	G-gene	KU316139.1	99.02%	913/922
	Peret.2000.	AF233913	G-gene (partial)	MG642081.1	98.89%	267/270
	Peret.2000.	AF233920	G-gene (partial)	MG642056.1	97.78%	264/270
	Peret.2000.	AF233921	G-gene (partial)	MG642026.1	98.52%	266/270
GA4	Peret. 1998.	AF065254	G-gene	KP258704.1	99.57%	918/922
GA5	Peret. 1998.	AF065255	G-gene	MG642061.1	99.78%	920/922
	Peret.2000.	AF233908	G-gene (partial)	MG642055.1	98.15%	265/270
	Peret.2000.	AF233909	G-gene (partial)	MG642052.1	98.15%	265/270
	Peret.2000.	AF233922	G-gene (partial)	MG642048.1	99.63%	269/270
GA6	Peret. 2000.	AF233901	G-gene (partial)	KP258723.1	98.52%	266/270
		AF233905	G-gene (partial)	KU316126.1	95.93%	259/270
		AF233911	G-gene (partial)	KU316166.1	95.17%	256/269
		AF233918	G-gene (partial)	MG642071.1	98.52%	266/270
GA7	Peret. 2000.	AF233904	G-gene (partial)	JX069800.1	99.63%	269/270
		AF233907	G-gene (partial)	JF920065.1	98.89%	267/270

		AF233910	G-gene (partial)	MG642030.1	98.15%	265/270
SAA1	Venter.2001.	AF348807	G-gene (partial)	KP258696.1	97.78%	264/270
		AF348808	G-gene (partial)	KU316093.1	95.56%	258/270
		AF348809	G-gene (partial)	KP258715.1	94.81%	256/270
		AF348810	G-gene (partial)	KJ723468.2	97.78%	264/270
SAA2	Pretorius. 2013.	-	-	-	-	-
NA1	Shobugawa. 2009.	AB470478	G-gene (partial)	JX015495.1	99.63%	269/270
NA2	Shobugawa. 2009.	AB470479	G-gene (partial)	KJ627361.1	99.26%	268/270
NA3	Cui.2013.	KC297292	G-gene	MK109773.1	98.34%	888/903
		KC297260	G-gene	KJ627318.1	98.34%	888/903
		KC297277	G-gene	KJ627306.1	98.23%	887/903
NA4	Cui.2013.	KC297381	G-gene	KU950573.1	96.68%	873/903
		KC297324	G-gene	KP218910.1	96.90%	875/903
		KC297374	G-gene	KJ627648.1	96.79%	874/903
ON1	Eshagi.2012.	JN257693	G-gene	MH447952.1	100.00%	967/967
		JN257694	G-gene	MH447951.1	99.90%	966/967
ON2	Hirano.2014.	KC858255	G-gene (partial)	KJ672440.1	97.01%	454/468
		KC858256	G-gene (partial)	KC731482.1	97.44%	456/468
CB-A	Baek.2012.	HQ699278	G-gene	KJ627320.1	98.32%	876/891
		HQ699279	G-gene	KJ627312.1	98.23%	886/902
		HQ699280	G-gene	KJ627252.1	97.86%	869/888
		HQ699281	G-gene	KJ627355.1	98.31%	873/888
TN1	Schobel.2016.	KJ672443	FULL	-	-	-
		KJ672482	FULL	-	-	-
TN2	Schobel.2016.	KJ672436	FULL	-	-	-
		KJ672459	FULL	-	-	-

Table 3.3B: Overview of original RSV B sequences of each genotype and the alternative, full genomes used for phylogenetic analysis. Highlighted in yellow are selected references sequences.

RSV B genotype	Reference	Ref ID	RefSeq Length	Alt (Complete seq)	Similarity	Identities
GB1	Peret.1998.	AF065250	G-gene (partial)	KJ723480.2	100.00%	921/921
GB2	Peret.1998.	-	-	-	-	-
GB3	Peret.1998.	AF065252	G-gene (partial)	KP258713.1	99.67%	918/921
GB4	Peret.1998.	AF065251	G-gene (partial)	KU316163.1	99.89%	920/921
GB5	Ren.2015.	KC461289	G-gene (partial)	KU316156.1	92.92%	223/240
		KC461291	G-gene (partial)	KU316094.1	93.25%	235/252
		KC461293	G-gene (partial)	KU316111.1	93.12%	230/247
		KC461294	G-gene (partial)	-	-	-
BA1	Trento.2006.	DQ227363	G-gene	MF185752.1	99.49%	976/981
		DQ227365	G-gene	JX576762.1	98.98%	971/981
		DQ227380	G-gene	KP856966.1	98.47%	966/981
		DQ227368	G-gene	JQ582844.1	98.17%	966/984
BA2	Trento.2006.	DQ227377	G-gene	JX576758.1	98.68%	971/984
		DQ227389	G-gene	KF826825.1	98.27%	964/981
		DQ227391	G-gene	JX576761.1	97.87%	963/984

		DQ227393	G-gene	KF826829.1	97.56%	960/984
BA3	Trento.2006.	DQ227381	G-gene	KP258739.1	97.76%	959/981
		DQ227376	G-gene	KU316172.1	98.06%	962/981
		DQ227388	G-gene	KJ939920.1	97.76%	958/980
		DQ227394	G-gene	JX576759.1	98.16%	962/980
BA4	Trento.2006.	DQ227395	G-gene	KF826822.1	99.59%	970/974
		DQ227396	G-gene	KP317925.1	97.86%	959/980
		DQ227407	G-gene	KU950458.1	97.73%	430/440
		DQ227408	G-gene	KJ627254.1	97.73%	430/440
BA5	Trento.2010.	AB175819	G-gene (partial)	JX576757.1	96.96%	319/329
		AB175820	G-gene (partial)	KP317941.1	98.18%	323/329
BA6	Trento.2010.	AY751110	G-gene (partial)	JX576756.1	97.45%	764/784
		AY751116	G-gene (partial)	KJ627317.1	97.58%	765/784
		DQ985148	G-gene (partial)	KJ939934.1	96.05%	754/785
		DQ985147	G-gene (partial)	JX576750.1	96.94%	760/784
BA7	Dapat.2010.	HM459864	G-gene (partial)	KJ627310.1	98.79%	326/330
		HM459867	G-gene (partial)	KJ627280.1	99.09%	327/330
		HM459868	G-gene (partial)	JX576755.1	98.18%	324/330
		HM459870	G-gene (partial)	KX765899.1	98.79%	326/330
BA8	Dapat.2010.	HM459871	G-gene (partial)	LC474522.1	97.27%	321/330
		HM459872	G-gene (partial)	KF826842.1	97.27%	321/330
		HM459873	G-gene (partial)	KU950574.1	96.36%	318/330
		HM459875	G-gene (partial)	KU950569.1	96.67%	319/330
BA9	Dapat.2010.	HM459876	G-gene	KF640637.1	99.70%	329/330
		HM459877	G-gene	LC474532.1	99.39%	328/330
		HM459878	G-gene	LC474531.1	100.00%	330/330
		HM459880	G-gene	KF826859.1	99.39%	328/330
BA10	Dapat.2010.	HM459883	G-gene	KX765893.1	99.39%	328/330
		HM459886	G-gene	KX765900.1	97.88%	323/330
		HM459891	G-gene	KF826844.1	97.88%	323/330
		HM459892	G-gene	KJ939926.1	98.79%	326/330
BA11	Baek.2012.	-	-	-	-	-
BA12	Khor.2013.	JX256973	G-gene (partial)	MK109789.1	95.08%	309/325
		JX256974	G-gene (partial)	JX576743.1	95.99%	311/324
		JX256976	G-gene (partial)	KP317917.1	95.68%	310/324
		JX256979	G-gene (partial)	JX576744.1	95.99%	311/324
BA13	Gimferrer.2016.	KX262621	G-gene (partial)	MK109767.1	95.56%	280/293
		KX262638	G-gene (partial)	MG431251.1	95.56%	280/293
		KX262625	G-gene (partial)	KX765945.1	95.90%	281/293
		KX262619	G-gene (partial)	KX655674.1	95.22%	279/293
CB1	Cui.2013.	KC297466	G-gene	KJ723466.2	96.13%	870/905
		KC297471	G-gene	MG642049.1	95.36%	863/905
		KC297446	G-gene	KP856961.1	94.20%	861/914
		KC297428	G-gene	KP258731.1	94.99%	854/899

CB-B	Baek.2012.	HQ699297	G-gene	KF826845.1	98.82%	922/933
		HQ699300	G-gene	KU950657.1	98.39%	918/933
		HQ699304	G-gene	KJ627348.1	97.21%	907/933
		HQ699289	G-gene	JX576739.1	97.64%	911/933
BA-C	Cui.2013.	KC297486	G-gene	KJ939922.1	98.24%	949/966
		KC297429	G-gene	KJ939919.1	99.69%	963/966
		KC297456	G-gene	MF185753.1	96.58%	932/965
		KC297477	G-gene	KU950614.1	96.79%	934/965
BA-Ly	Gaymard.2018.	MF510895	G-gene (partial)	LC474529.1	96.70%	410/424
		MF510897	G-gene (partial)	KX765963.1	97.30%	540/555
SAB1	Venter.2001.	AF348825	G-gene (partial)	MF185751.1	98.15%	265/270
		AF348826	G-gene (partial)	KU316156.1	94.44%	255/270
SAB2	Venter.2001.	AF348821	G-gene (partial)	KU316179.1	98.89%	267/270
		AF348822	G-gene (partial)	KP258742.1	90.98%	242/266
SAB3	Venter.2001.	AF348811	G-gene (partial)	KP258724.1	98.89%	267/270
		AF348812	G-gene (partial)	MF185754.1	97.78%	264/270
		AF348813	G-gene (partial)	KU316134.1	97.78%	264/270
SAB4	Arnott.2011.	JN119969	G-gene (partial)	KP258708.1	95.93%	259/270
		JN119976	G-gene (partial)	KP258745.1	95.93%	259/270
		JN119979	G-gene (partial)	KP258713.1	96.30%	260/270
		JN119987	G-gene (partial)	KU316100.1	94.07%	254/270
URU1	Blanc.2005.	AY488794	G-gene (partial)	KP856965.1	96.46%	436/452
		AY488804	G-gene (partial)	KU316094.1	94.20%	357/379
		AY488805	G-gene (partial)	KP258720.1	93.93%	356/379
URU2	Blanc.2005.	AY488802	G-gene (partial)	KU316111.1	94.02%	330/351
		AY488803	G-gene	KU316163.1	97.83%	361/369

3.3.2. Determine variability of each RSV gene

Next, the variability of each gene was determined to find the best gene for genotyping. To be able to use phylogenetic analysis on a dataset, there has to be enough variation to determine differences between sequences, but there has to be enough similarity to be able to find genetic relatedness. If there is not a good balance, tree-branching decisions cannot confidently be made and phylogenetic analysis is not useful.

3.3.2.1. Alternative reference sequence datasets

To determine the variability of each RSV gene, the Shannon Entropy (SE) was calculated. The Shannon entropy was determined at each position of each gene separately (by using the Shannon Entropy-One tool, www.hiv.lanl.gov). This is a way of determining the uncertainty of a nucleotide at a certain position in the genome. If every sequence carries the same nucleotide at position x, the uncertainty is zero, hence the SE is zero for position x. On the contrary, if there are multiple possible nucleotides at that position in an alignment, the SE will increase depending on how many different nucleotides are detected and what their frequency is.

The SE of each position for each gene was calculated from the datasets containing alternative reference sequences, which is shown in Figure 3.5A for RSV A and Figure 3.5B for RSV B. Then, the total SE (SE_t) for each gene was calculated, which is the sum of SE from each position of that gene. For RSV A and RSV B respectively, the minimum SE_t ranged from 10.018 and 4.256 for the SH gene to a maximum SE_t of 227.640 and 139.074 for the L gene (Table 3.4). This was not unexpected as the SH gene is very short and can therefore not accumulate much variation and uncertainty, while the L gene is by far the largest gene and can therefore easily accumulate more variation. Proteins G, F and L have accumulated the most variation and the second hypervariable region of G (with or without duplication) was clearly visible on the graphs as particularly prone to variation for both RSV A and B (Figure 3.5A and B). For RSV A, both the first and second hypervariable region were visible, while the RSV B dataset showed an exceptional variable HVR2 compared to the rest of the G gene.

To eliminate the factor of gene length, the SE_t was normalised *e.g.* the SE_t of each gene was divided by the number of nucleotides of that gene (SE_{nt}) (Figure 3.6). This clearly showed that the G gene is the most variable of all RSV genes for both RSV A and RSV B.

Table 3.4: Total Shannon entropy (SE_t) and Shannon entropy per nucleotide (SE_{nt}) for RSV A and RSV B alignments.

	RSVA_SE_t	RSVA_SE_{nt}	RSVB_SE_t	RSVB_SE_{nt}
NS1	14.038	0.033	7.626	0.018
NS2	15.289	0.041	12.039	0.032
N	41.081	0.035	24.461	0.021
P	23.072	0.032	22.293	0.031
M	26.556	0.034	17.217	0.022
SH	10.018	0.051	4.256	0.021
G	99.558	0.103	91.505	0.095
F	73.910	0.043	41.277	0.024
M2-1	20.836	0.036	14.417	0.025
M2-2	14.227	0.053	6.895	0.025
L	227.640	0.035	139.074	0.021

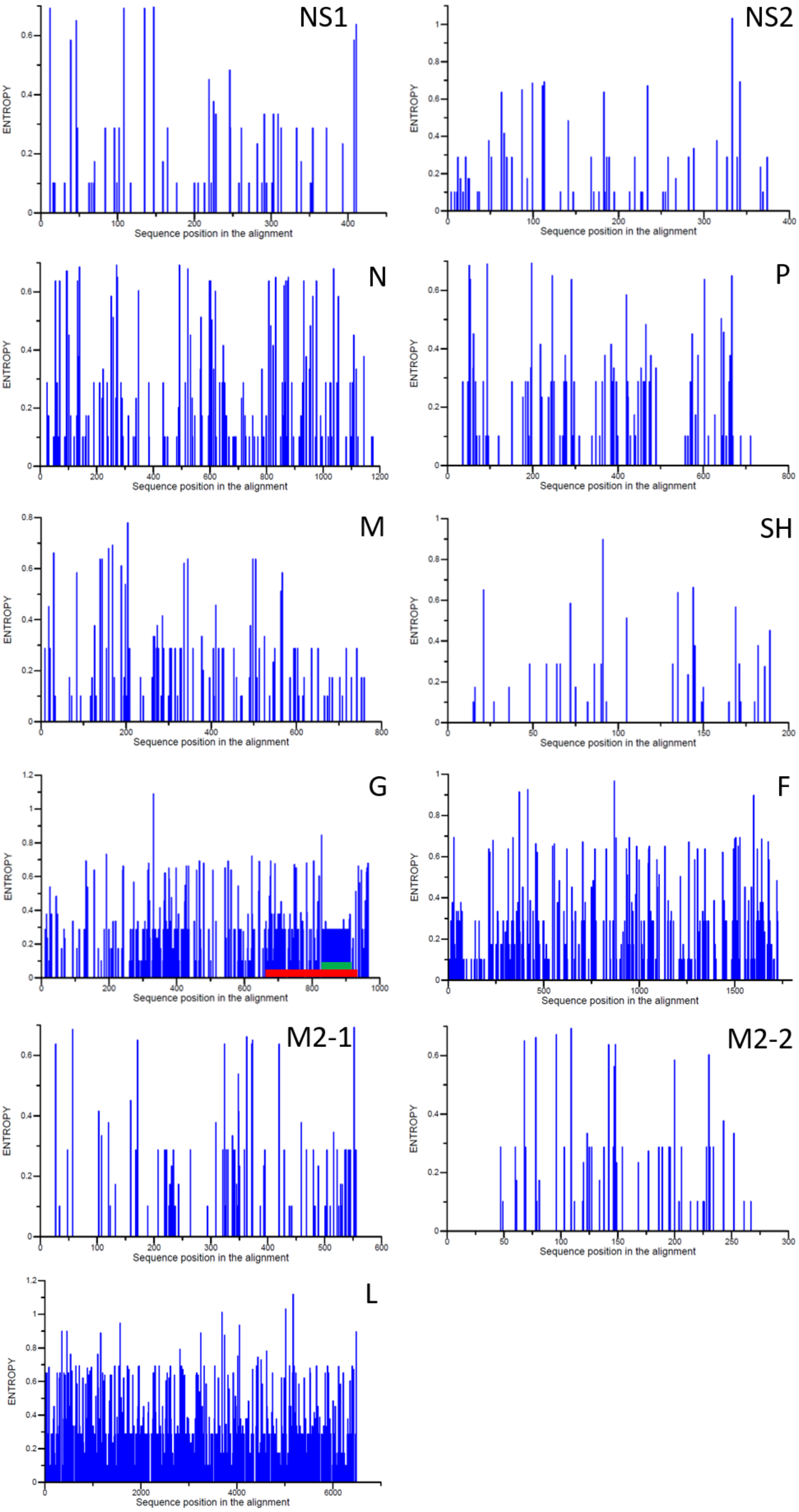


Figure 3.5A: Shannon Entropy values were calculated for each position of each gene for RSV A. The values are mapped in graphs per gene. Note that the scale of the x- and y-axis is not constant. Red = HVR2 of the G gene; green = 72-nucleotide duplication.

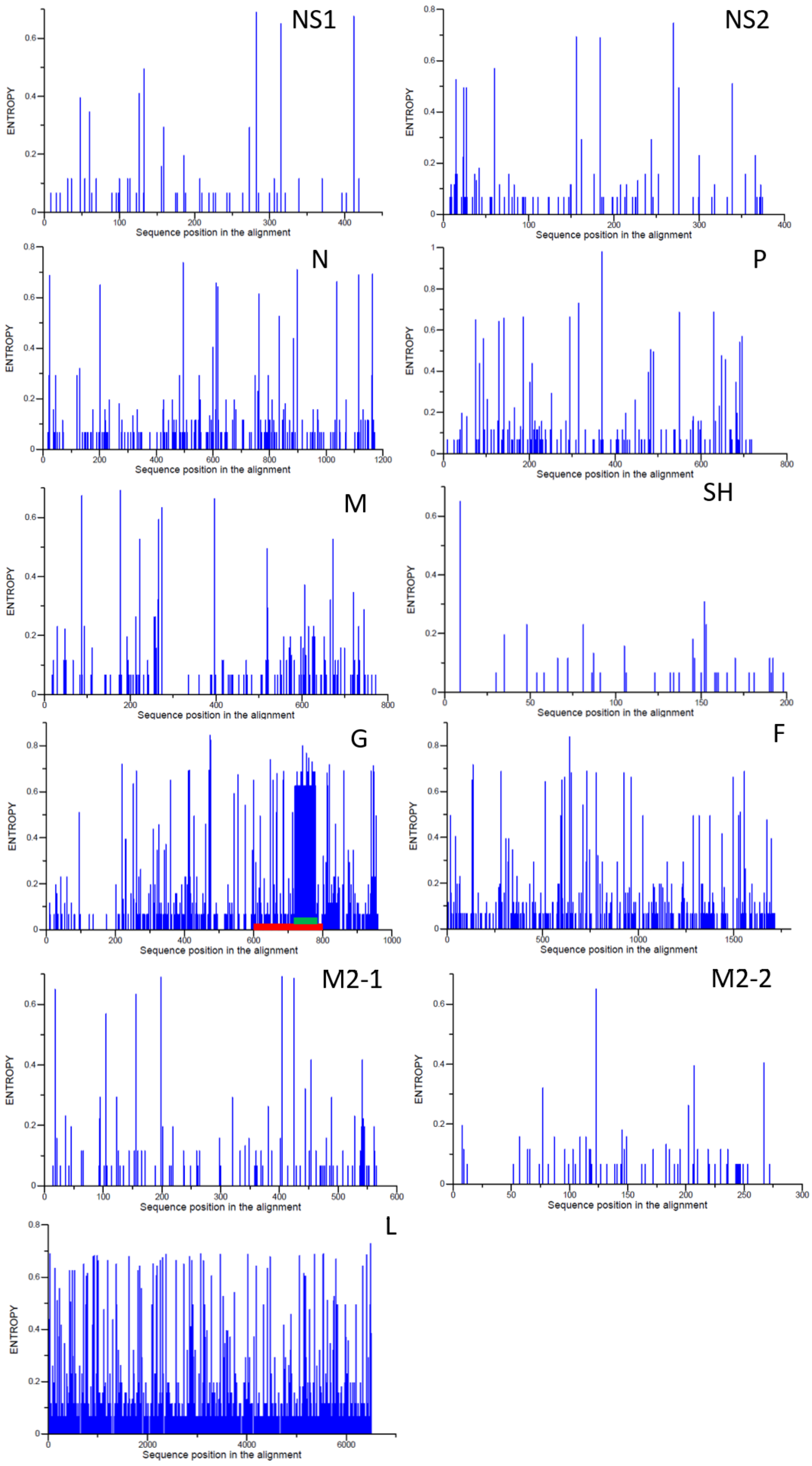


Figure 3.5B: Shannon Entropy values were calculated for each position of each gene for RSV B. The values are mapped in graphs per gene. Note that the scale of the x- and y-axis is not constant. Red = HVR2 of the G gene; green = 60-nucleotide duplication.

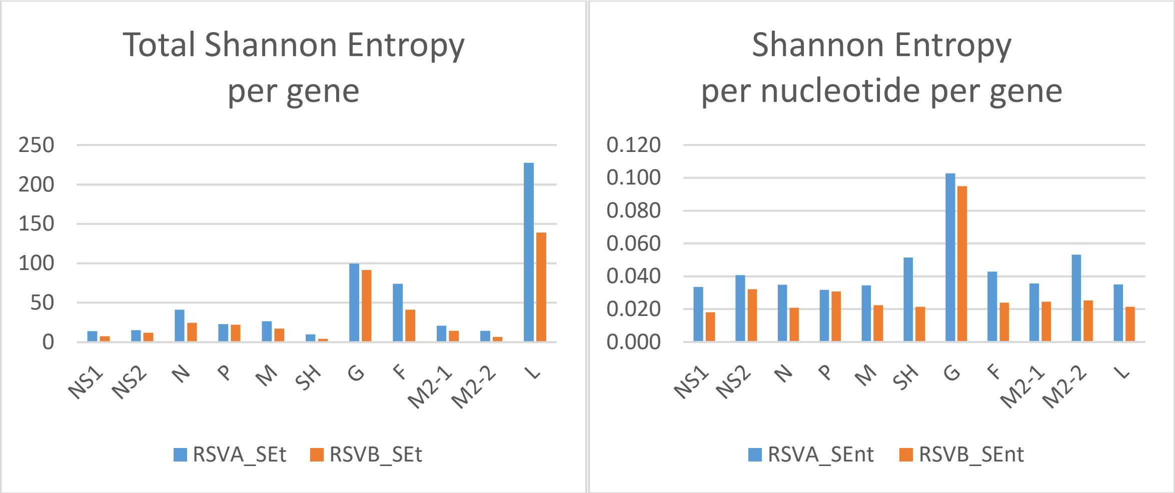


Figure 3.6: The total Shannon entropy (SEt) values per gene showed that L, G and F accumulated the most variability (left). G had the highest Shannon entropy per nucleotide (SEnt) and was therefore the most variable per nucleotide per gene compared to other genes.

The usefulness for phylogenetic analysis of these datasets was calculated for each gene separately by calculating the number of informative quartets in the alignment. It showed that L was the most useful gene for confident phylogenetic analysis with 92.8% and 84.9% of informative quartets for RSV A and RSV B respectively (Table 3.5). This could be explained by its high number of nucleotides to base phylogenetic decisions on. However, the second most useful gene for phylogenetic analysis is not G, as would be expected, but F with 85.1% for RSV A and 84.5% for RSV B of informative quartets. It has a higher number of nucleotides than G and also carries variation, although the SE per nucleotide is only half that of G.

Table 3.5: The percentage of informative quartets (Quqrtets$_{inf}$) for RSV A and RSV B per gene shows that several genes have more informative quartets than G in this database and would therefore produce a more informative phylogenetic tree. F and L would be more informative for both the RSV A and RSV B dataset.

	RSV A Quartets$_{inf}$	RSV B Quartets$_{inf}$
NS1	77.4	72.5
NS2	80.6	63.2
N	83.7	65.3
P	67.4	58.5
M	77.6	60.3
SH	65.5	31.1
G	82.6	74.3
F	85.1	84.5
M2-1	73.7	74.3
M2-2	77.2	42.6
L	92.8	84.9

These data were based on quite small datasets, so a larger dataset was compiled and similar analyses were performed to confirm or disprove the results of this analysis.

3.3.2.2. Gene variability from comprehensive dataset

To determine the variability in a larger population, all complete G genes and complete genomes from the international NCBI database have been downloaded on September 13th, 2018. Several filters were put in place to select only useful sequences. The search terms 'Respiratory syncytial virus' and 'Complete G' were combined with a sequence length of at least 800 bp and maximum 15300 bp. All sequences were filtered to be genomic DNA or RNA and be from published articles. From this selection all strains from bovine (5), ovine (1) and mouse (4) origins were removed. There were 5 lab grown mutant viruses that were discarded and one Pneumonia Virus of Mice strain was taken out as well. From 1064 RSV strains downloaded, 1048 were left to be allocated to the RSV A or RSV B dataset. At a later date one more sequence was found to be more than 15300 bp long, *e.g.* 15333 bp, which was not included in this analysis.

All sequences were allocated to either the RSV A or RSV B dataset and aligned using MUSCLE software. The two datasets were divided into smaller datasets containing only one gene for further analysis of gene variation. Not all genomes were complete genomes and therefore not all datasets contained an equal number of sequences. Incomplete sequences in a gene dataset were removed. For RSV A, the complete dataset contained 653 different sequences, while the number of sequences in the datasets for NS1, NS2, N, P, M, SH, G, F, M2-1, M2-2 and L were respectively 153, 153, 153, 153, 153, 154, 653, 163, 153, 153 and 152. For RSV B, the complete dataset contained 398 sequences and the numbers for each dataset in the same order as described above were 76, 76, 76, 76, 76, 76, 398, 86, 42, 76 and 76.

The SE was determined and graphed per nucleotide as in the previous subchapter, which demonstrated that the second hypervariable region of G was even more evident on graphs of SE for both RSV A and RSV B (Figure 3.7A and 3.7B). A strong background variation can be seen in L for both serotypes and for F, although that feature was more distinct in the RSV B dataset for the F gene.

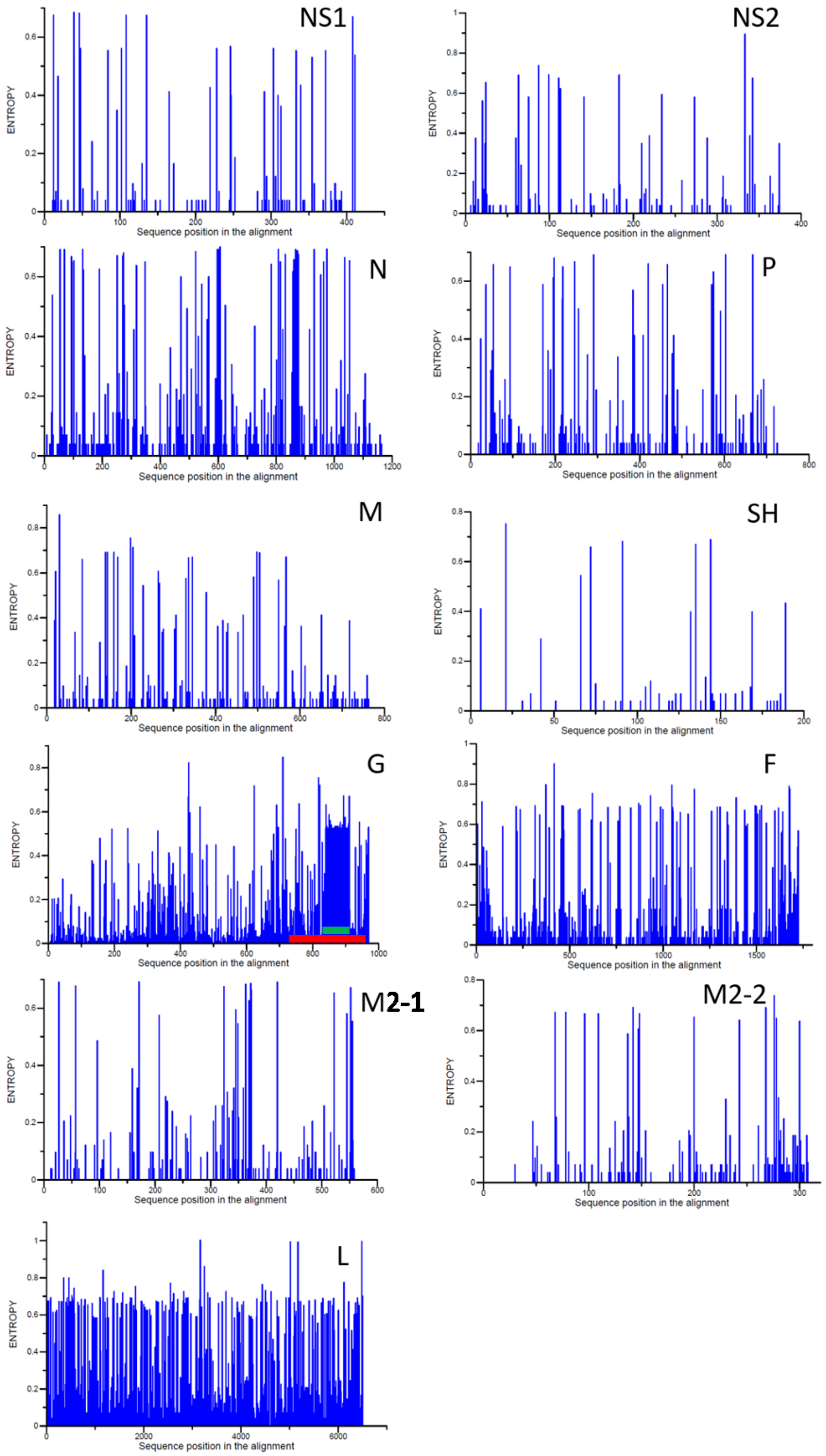


Figure 3.7A: Shannon Entropy values were calculated for each position of each gene for RSV A from a dataset of NCBI sequences. The values are mapped in graphs per gene. Note that the scale of the x- and y-axis is not constant. Red = HVR2 of the G gene; green = 72-nucleotide duplication.

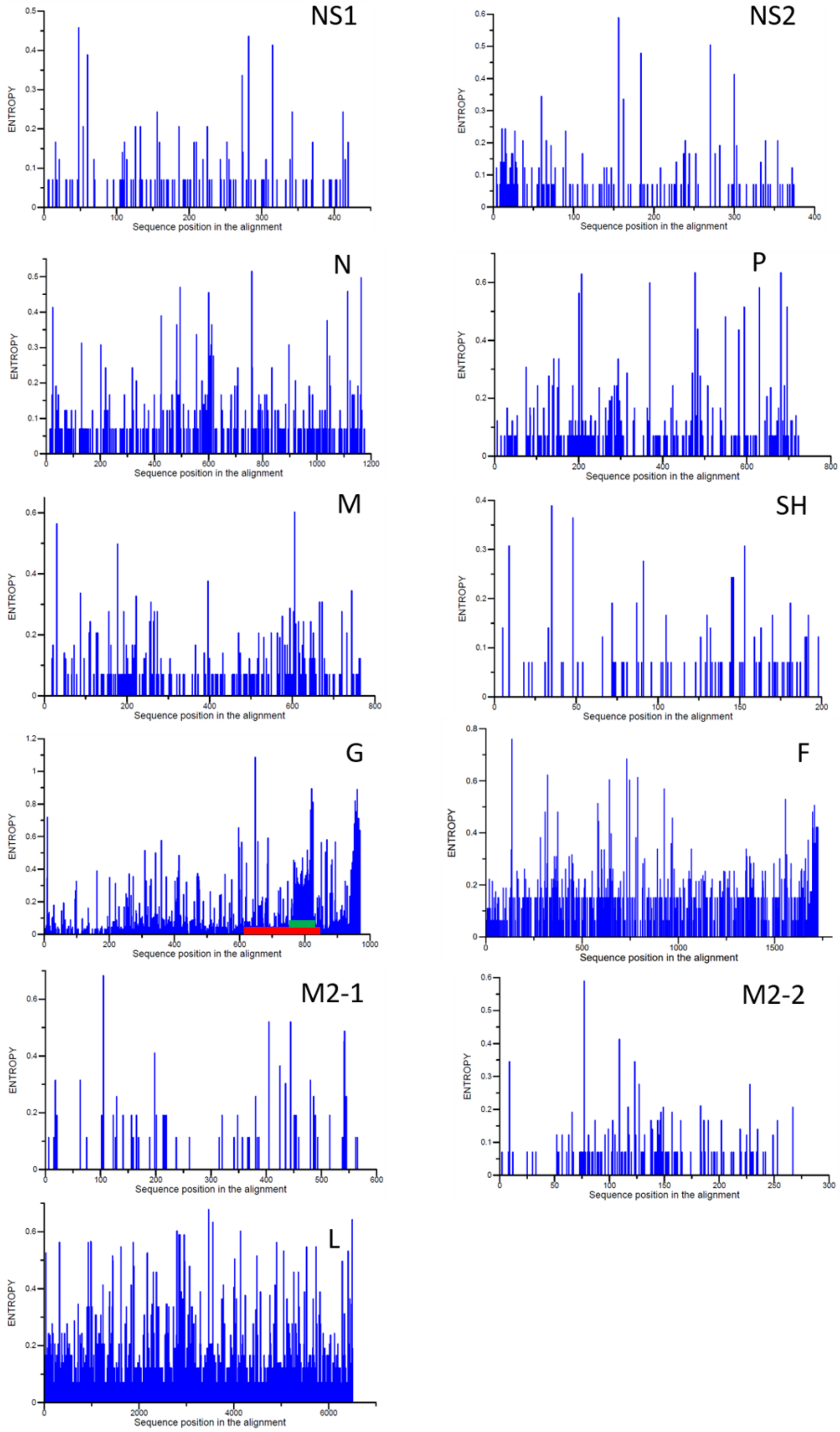


Figure 3.7B: Shannon Entropy values were calculated for each position of each gene for RSV B from a dataset of NCBI sequences. The values are mapped in graphs per gene. Note that the scale of the x- and y-axis is not constant. Red = HVR2 of the G gene; green = 60-nucleotide duplication.

The total SE for both RSV A and B datasets were calculated per gene (Table 3.6). This revealed similar trends as shown in the previous subchapter with the alternative reference datasets. The L gene had a SE_t of 218 and 178 for RSV A and B respectively, which was by far the highest SE_t. The G gene had the second highest SE_t of 100 and 92, and the F gene was a close third with 82 and 84. This was in accordance with earlier findings.

Next, these numbers were normalised for gene length and the G gene showed to be the most variable gene by far (Figure 3.8) with a normalized SE of 0.10 for both RSV A and B datasets compared to 0.05 for the F gene and 0.03 for the L gene (Table 3.6). This demonstrated that G accumulated double the amount of variation compared to the next most variable gene, *e.g.* F (Figure 3.8).

Table 3.6: Total Shannon entropy (SEt) and Shannon entropy per nucleotide (SE_{nt}) for comprehensive RSV A and RSV B alignments.

	RSVA_SE_t	RSVA_SE_{nt}	RSVB_SE_t	RSVB_SE_{nt}
NS1	16.54	0.04	11.05	0.03
NS2	16.17	0.04	13.16	0.04
N	44.03	0.04	29.38	0.02
P	24.90	0.03	24.96	0.03
M	26.05	0.03	22.16	0.03
SH	7.76	0.04	7.81	0.04
G	100.31	0.10	92.49	0.10
F	82.47	0.05	84.49	0.05
M2-1	21.85	0.04	11.56	0.02
M2-2	18.13	0.06	11.81	0.04
L	218.18	0.03	178.18	0.03

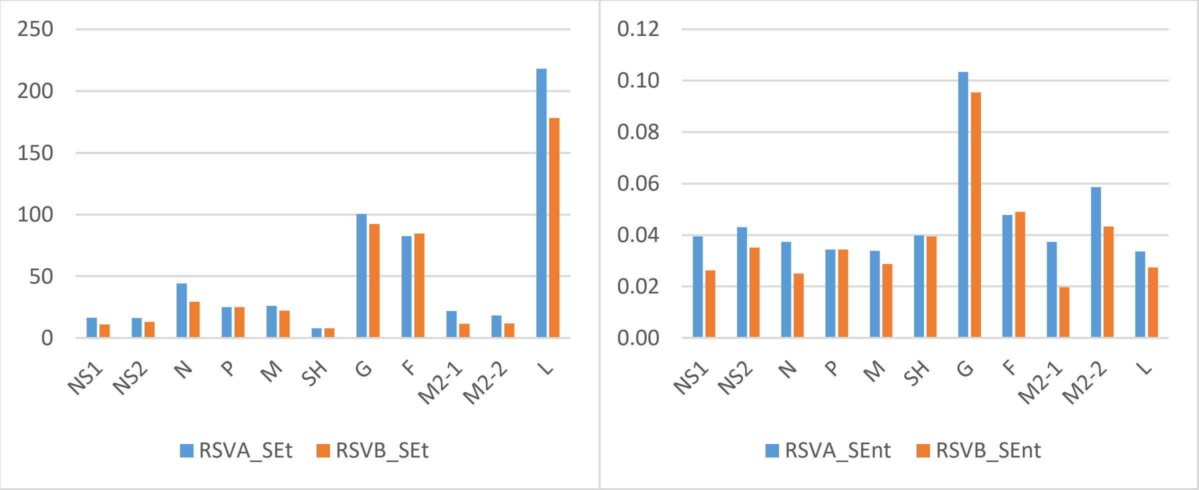


Figure 3.8: Total Shannon Entropy per gene (SEt) and total Shannon Entropy per nucleotide per gene (SEnt) for comprehensive RSV A and B datasets show similar trends as the database with only some strains for each genotype.

To study the usefulness for phylogenetic analysis of each gene from these datasets, the proportion of informative quartets were calculated (Table 3.7). The highest percentage of useful quartets was found in the L gene for the RSV A dataset with an overwhelming 90.5% of useful quartets. Phylogenetic analysis would return a tree with mostly confident branching decisions for this gene. However, the overall variability was only one third of the variability that was seen for the G gene. The second most useful gene was F with 82.1% of informative quartets after which M2-1 (76.8), M2-2 (76.3), N (75.6) and M (72.6) were the most useful for phylogenetic analysis. Surprisingly, the G gene was only the seventh useful gene with 71.3% of informative quartets. This was remarkable as this was not noted for the previous dataset of reference strains for RSV A.

For the RSV B dataset, the G gene was the most useful for phylogenetic analysis with 72.9% of informative quartets. This contrasted with the previous dataset and with the comprehensive RSV A dataset. It was noted that the proportion of informative quartets dropped for each gene in both RSV A and B datasets, apart for M2-2 of the RSV B dataset, which remained the same.

Table 3.7: The percentage of informative quartets for comprehensive datasets of RSV A and RSV B per gene show similar trends to the database with some reference strains. Several other genes would produce a more informative phylogenetic tree than G.

	RSVA_Quartets$_{inf}$	RSVB_Quartets$_{inf}$
NS1	64.0	48.9
NS2	66.4	48.3
N	75.6	34.9
P	70.5	48.7
M	72.6	36.9
SH	66.6	58.9
G	71.3	72.9
F	82.1	64.9
M2-1	76.8	66.0
M2-2	76.3	42.6
L	90.5	67.5

The combination of variability and usefulness of quartets indicated that G carried the most variability per nucleotide. The usefulness for phylogenetic analysis was greatly dependent on the dataset.

3.3.3. Determining the necessary and sufficient part of RSV genome for genotyping

In the previous sub-chapter, it was shown that L could be an informative gene for phylogenetic analysis. However, the Shannon entropy, and therefore the variability, per nucleotide of the L gene was three times lower than the variability of the G gene. The percentage of informative quartets for F was higher than for G in three out of four datasets, but the SE for F was much lower than for G. Therefore, it was decided to investigate the G gene to use for genotyping.

The next question was to study the amount of nucleotides necessary to be able to genotype strains. The previous datasets were used to try and determine the necessary and sufficient part of the genome necessary for confident genotyping. The alternative complete genomes that were gathered earlier were used for this purpose. This set of complete genomes was aligned with MUSCLE (v3.8.31) and manually improved. The dataset was split up in RSV A and RSV B strains, containing 48 sequences and 82 sequences respectively.

First, the completeness of the datasets was determined by calculating the C-scores for the complete genomes of the RSV A and RSV B datasets. This showed good numbers compared to the datasets containing the original sequences, with an overall alignment score of 0.99 for both RSV A and RSV B alignments (Table 3.8). This was confirmed by comparing all sequences to each other, which all showed higher values than 0.9 (Figure 3.9).

Table 3.8: C-scores of the dataset for RSV A and RSV B containing alternative, full genomes indicate good, unambiguous sequences are present in this dataset.

	Completeness scores alternative RSV A sequences (n = 48)	**Completeness scores alternative RSV B sequences (n = 82)**
Completeness score for alignment (Ca)	0.99	0.99
Maximum completeness score for individual sequences (Cr_max)	0.99	1.00
Minimum completeness score for individual sequences (Cr_min)	0.98	0.98
Maximum completeness score for individual sites (Cc_max)	1.00	1.00
Minimum completeness score for individual sites (Cc_min)	0.02	0.01
Maximum completeness score for pairs of sequences (Cij_max)	0.99	1.00
Minimum completeness score for pairs of sequences (Cij_min)	0.97	0.97

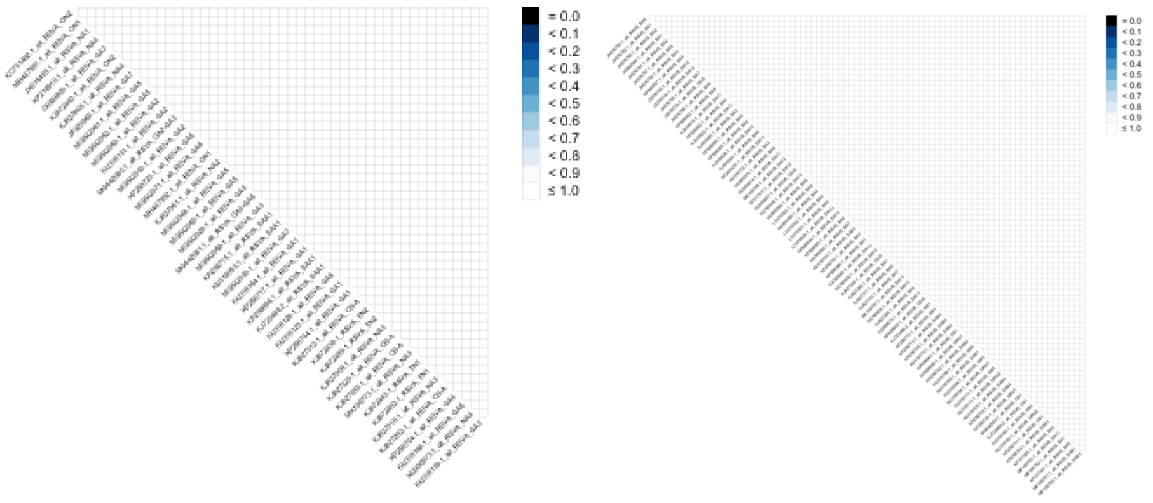


Figure 3.9: Triangle heatmaps of RSV A (left) and RSV B (right) datasets showed completeness scores comparing all sequences to each other (Cij-scores). Blue = incomplete sequences; white = complete sequences.

In the next step, the usefullness of the alignment was estimated by calculcating the percentage of informative quartets for these datasets. The RSV A dataset contained 95.0% of informative quartets and the RSV B dataset contained 92.6% of informative quartets (Figure 3.10). This indicated that confident phylogenetic trees could be built from these datasets with complete genomes.

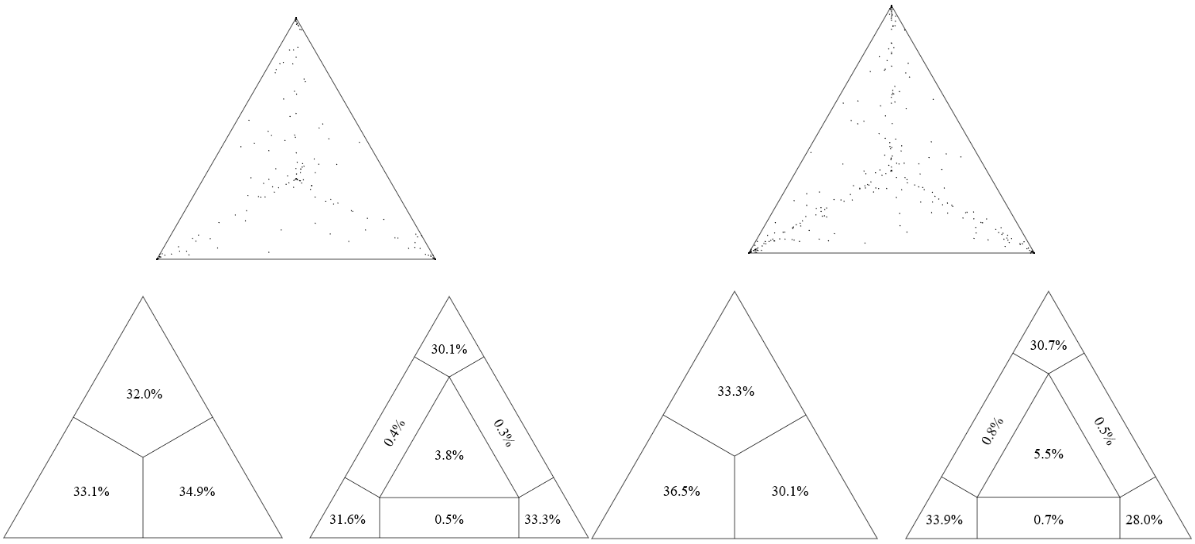


Figure 3.10: Triangular likelihood mapping tests whether a quartet (four randomly selected sequences from the dataset) produces an informative phylogenetic tree. Each dot in the top triangle is a quartet. The three parts of the left triangle indicate whether any subsets (if they were selected) have a bigger number of quartets. The right triangle shows uninformative quartets in the middle and informative quartets in the corners. Partial informative quartets are shown on the sides of the triangle. Likelihood mapping of quartets in the dataset with complete genomes of alternative sequences show that 3.8% of quartets of the RSV A dataset (left) and 5.5% of quartets of the RSV B dataset (right) are uninformative.

Similar analyses were done for the same dataset which was cut down to contain only the full G gene and again for partial G containing only the second hypervariable region (Table 3.9 and Figure 3.11).

The results showed slightly lower C-scores for the complete G gene analyses compared to full genomes and much lower values for the second hypervariable region analyses. This suggested that HVR2 might not be long enough to confidently determine genotypes.

Table 3.9: C-scores of the dataset for RSV A and RSV B containing alternative, full G genes and second hypervariable region of G indicate the dataset with full G genes contains enough unambiguous sites and the dataset with HVR2 sequences contains a lot more ambiguous sites.

C-scores for full G gene	**Completeness scores alternative RSV A sequences (n = 48)**	**Completeness scores alternative RSV B sequences (n = 82)**
Completeness score for alignment (Ca)	0.93	0.97
Maximum completeness score for individual sequences (Cr_max)	1.00	1.00
Minimum completeness score for individual sequences (Cr_min)	0.92	0.93
Maximum completeness score for individual sites (Cc_max)	1.00	1.00
Minimum completeness score for individual sites (Cc_min)	0.08	0.01
Maximum completeness score for pairs of sequences (Cij_max)	1.00	1.00
Minimum completeness score for pairs of sequences (Cij_min)	0.92	0.93
C-scores of HVR2 of G (342 sites for RSV A; 393 sites for RSV B)	**Completeness scores alternative RSV A sequences (n = 48)**	**Completeness scores alternative RSV B sequences (n = 82)**
Completeness score for alignment (Ca)	0.81	0.94
Maximum completeness score for individual sequences (Cr_max)	1.00	1.00
Minimum completeness score for individual sequences (Cr_min)	0.79	0.84
Maximum completeness score for individual sites (Cc_max)	1.00	1.00
Minimum completeness score for individual sites (Cc_min)	0.08	0.01
Maximum completeness score for pairs of sequences (Cij_max)	1.00	0.99
Minimum completeness score for pairs of sequences (Cij_min)	0.79	0.84

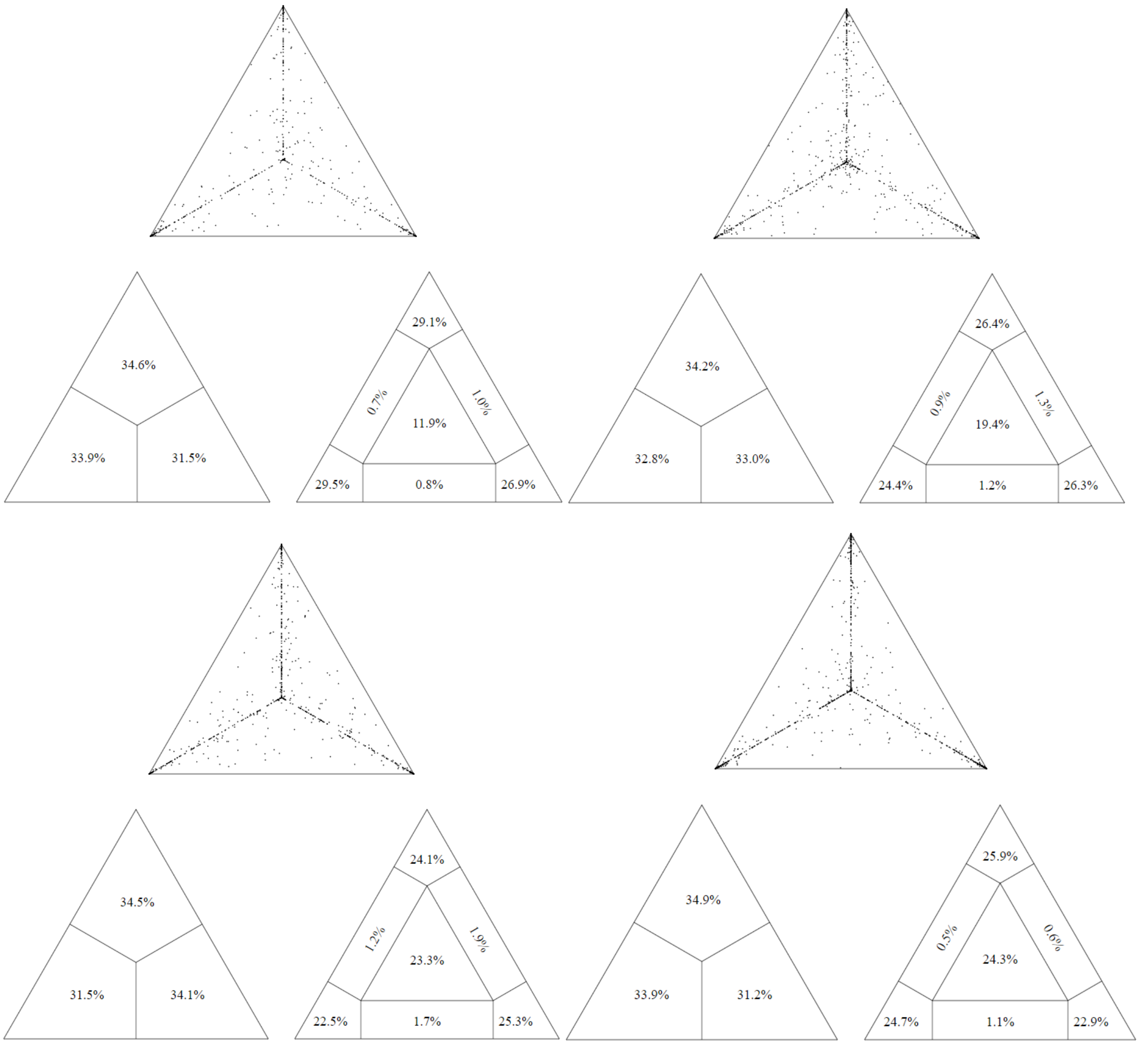


Figure 3.11: Triangular likelihood mapping tests whether a quartet (four randomly selected sequences from the dataset) produces an informative phylogenetic tree. Each dot in the top triangle is a quartet. The three parts of the left triangle indicate whether any subsets (if they were selected) have a bigger number of quartets. The right triangle shows uninformative quartets in the middle and informative quartets in the corners. Partial informative quartets are shown on the sides of the triangle. Likelihood mapping of quartets of RSV A (left) and RSV B (right) datasets of full G genes (top) or only second hypervariable regions (bottom) of alternative sequences.

In the next step, the phylogenetic trees were built. These were maximum likelihood trees that were built based on the best fitting model selected for each dataset individually. For the RSV A datasets, the best fitting models were GTR+F+R2 for full genomes and TN+F+G4 for both full G genes and partial G genes (HVR2) according to the Bayesian Information Criterion (BIC). For the RSV B datasets, GTR+F+R3 was the best fitting model for complete genomes, TPM2u+F+R3 for complete G genes and TN+F+G4 for partial G genes (HVR2).

These models were used to build phylogenetic trees. Both the approximate likelihood ratio test (aLRT) and bootstrap (run 1000 times) were calculated for each tree to distinguish confident branching and uncertain branching.

The first phylogenetic tree was constructed from the complete genomes from the RSV A dataset (Figure 3.12). Most branches were built confidently as only some bootstrap values were below 80 and lower values were mostly noticed within genotype clusters. Genotypes GA2, GA3, GA6 and NA4 did fall into combined clusters, which was to be expected as these alternative sequences lied very close together when looking for them in the NCBI database.

Tree topology of trees from full G genes showed only minor differences and less detailed clusters compared to the tree built from full genomes (Figure 3.13). The bootstrap values were lower as well, illustrating less confident branching. The topology from the tree built from partial G genes was not as confident. The topology showed obvious differences and the branching was not confident, illustrating that this dataset did not contain enough information to base phylogenetic analysis on (Figure 3.13 a and b). It is therefore also not suitable to determine genotypes.

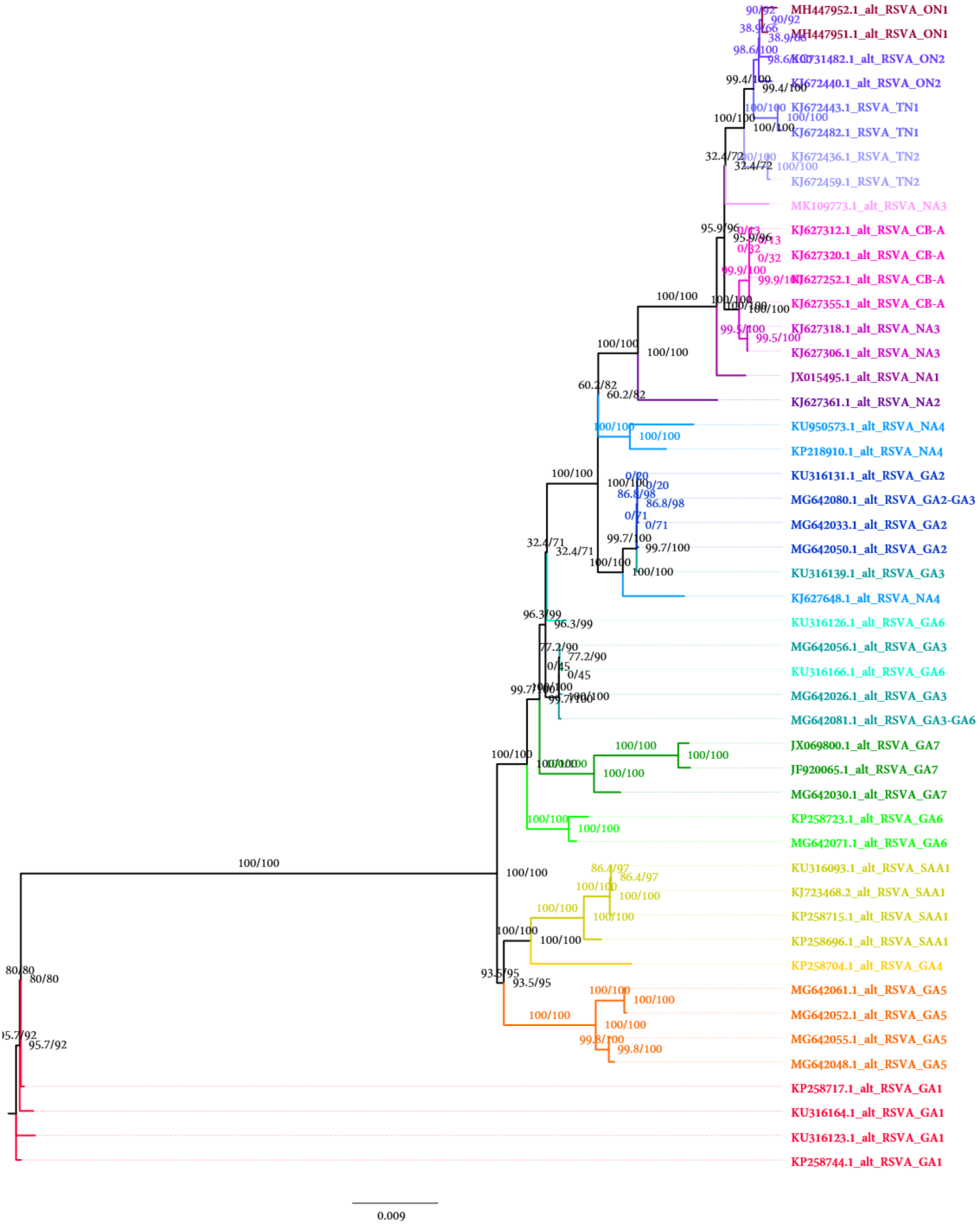


Figure 3.12: Maximum likelihood tree of dataset containing alternative, full genomes of RSV A genotypes shows that most alternative reference sequences do cluster according to their genotype. GA3, GA6 and NA4 are the exception where no clear clustering is to be found. (approximate likelihood ratio/ bootstrap)

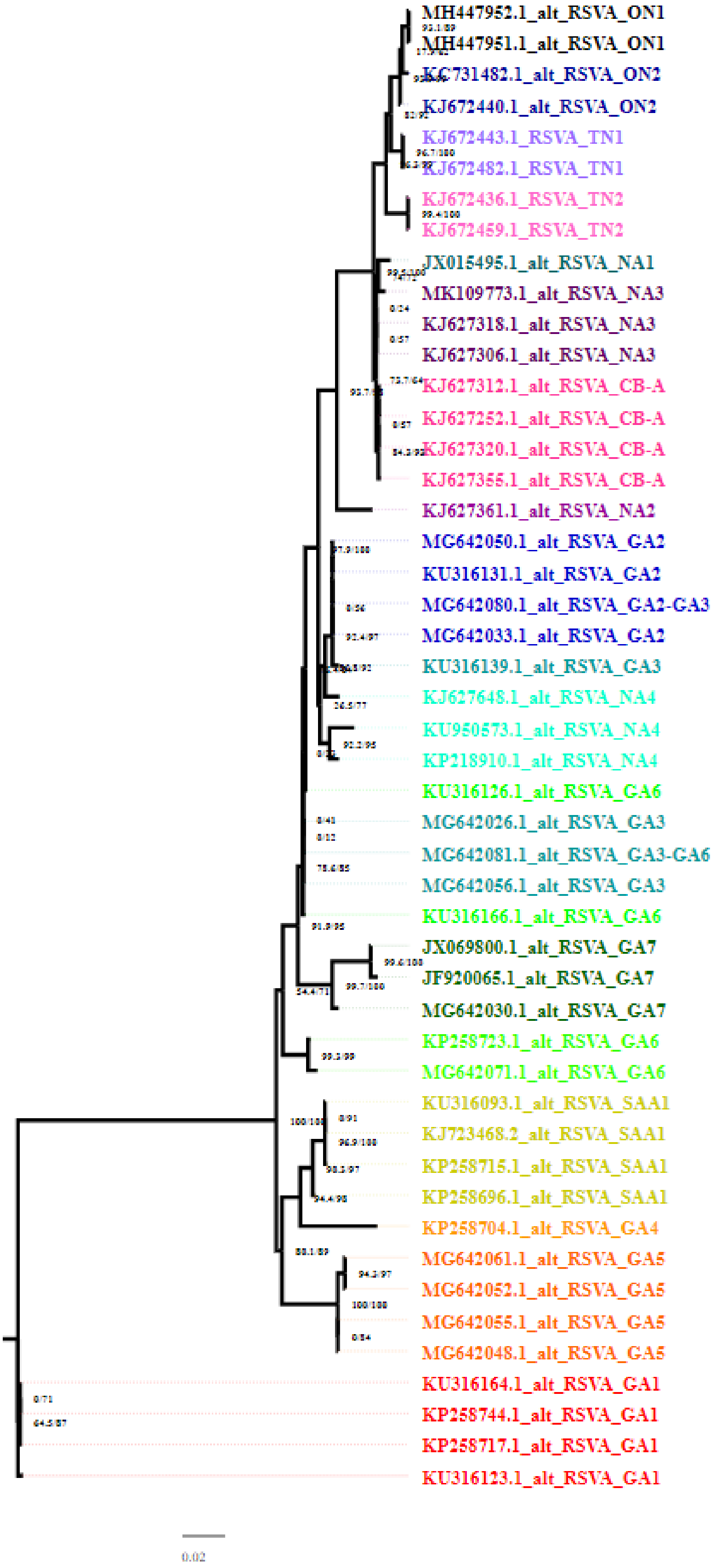


Figure 3.13 (a): Maximum likelihood tree of dataset containing full G genes of alternative RSV A genotypes. These sequences show similar results to the phylogenetic tree based on full genomes, although the bootstrap values and approximate likelihood ratio test values are not as high.

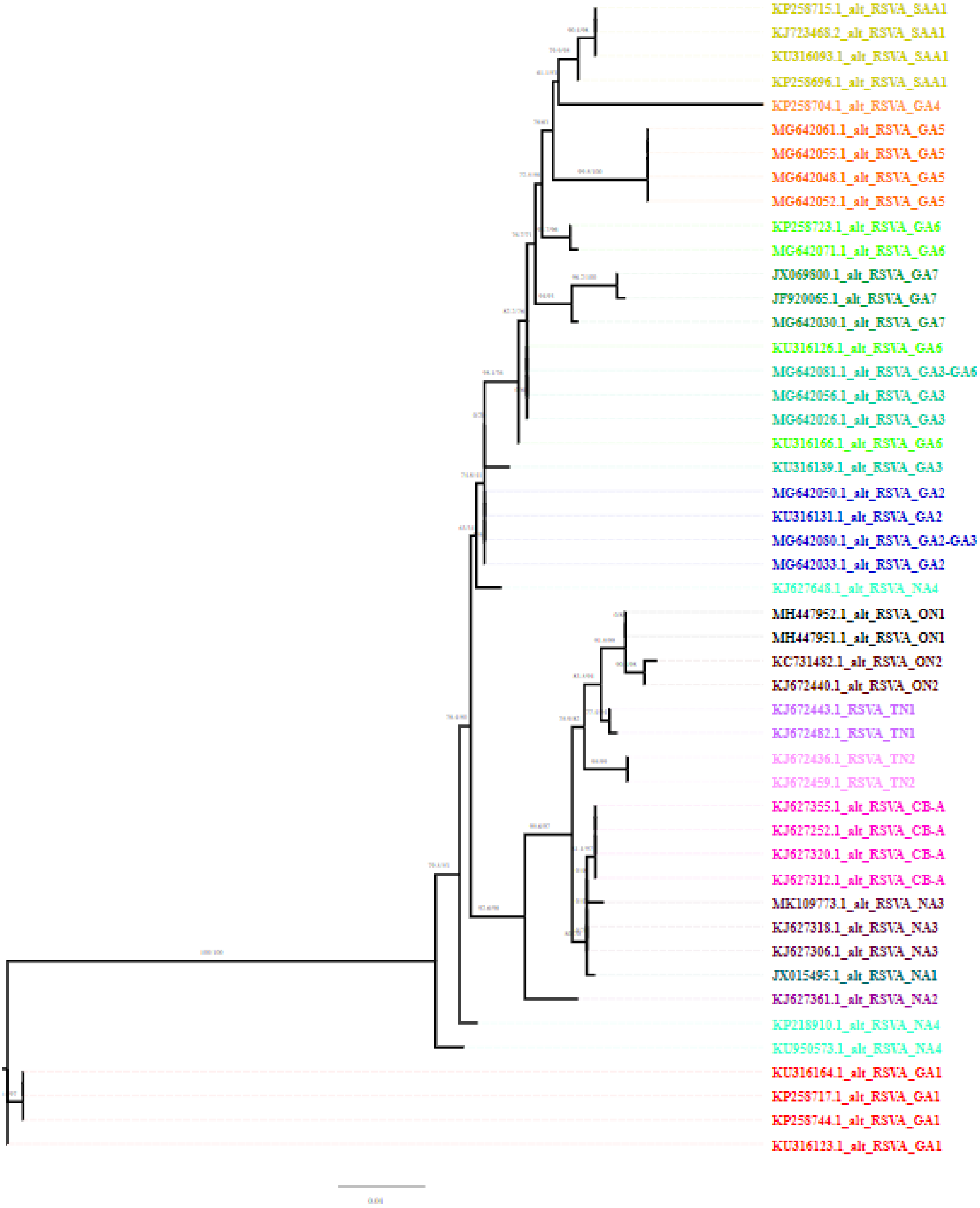


Figure 3.14 (b): Maximum likelihood tree of dataset containing the second hypervariable region of alternative sequences of RSV A genotypes. This phylogenetic tree shows more mixing of genotypes and has low bootstrap values and approximate likelihood ratio test values. (approximate likelihood ratio test/ bootstrap)

The dataset of RSV B full genomes was more complex. The tree based on full genomes did not produce clear clusters of genotypes, which was to be expected as genotypes SAB1 to SAB4, URU1 to URU2 and CB1 were all related to the same complete genomes from the international NCBI database (Figure 3.14). Besides those genotypes clustering together, all BA genotypes clustered together as well. Keeping this information in mind, it became somewhat clearer, but there were still some genotypes found all over the tree, *e.g.* BA-C. The reliability of this tree was generally lower than for the RSV A tree of complete genomes (Figure 3.15 a and b). This diminished even further for the trees based on full G and partial G. It is possible the selected sequences were not good representatives of these genotypes. It is also possible that determining new genotypes was not structured enough to be left with clear genotypes.

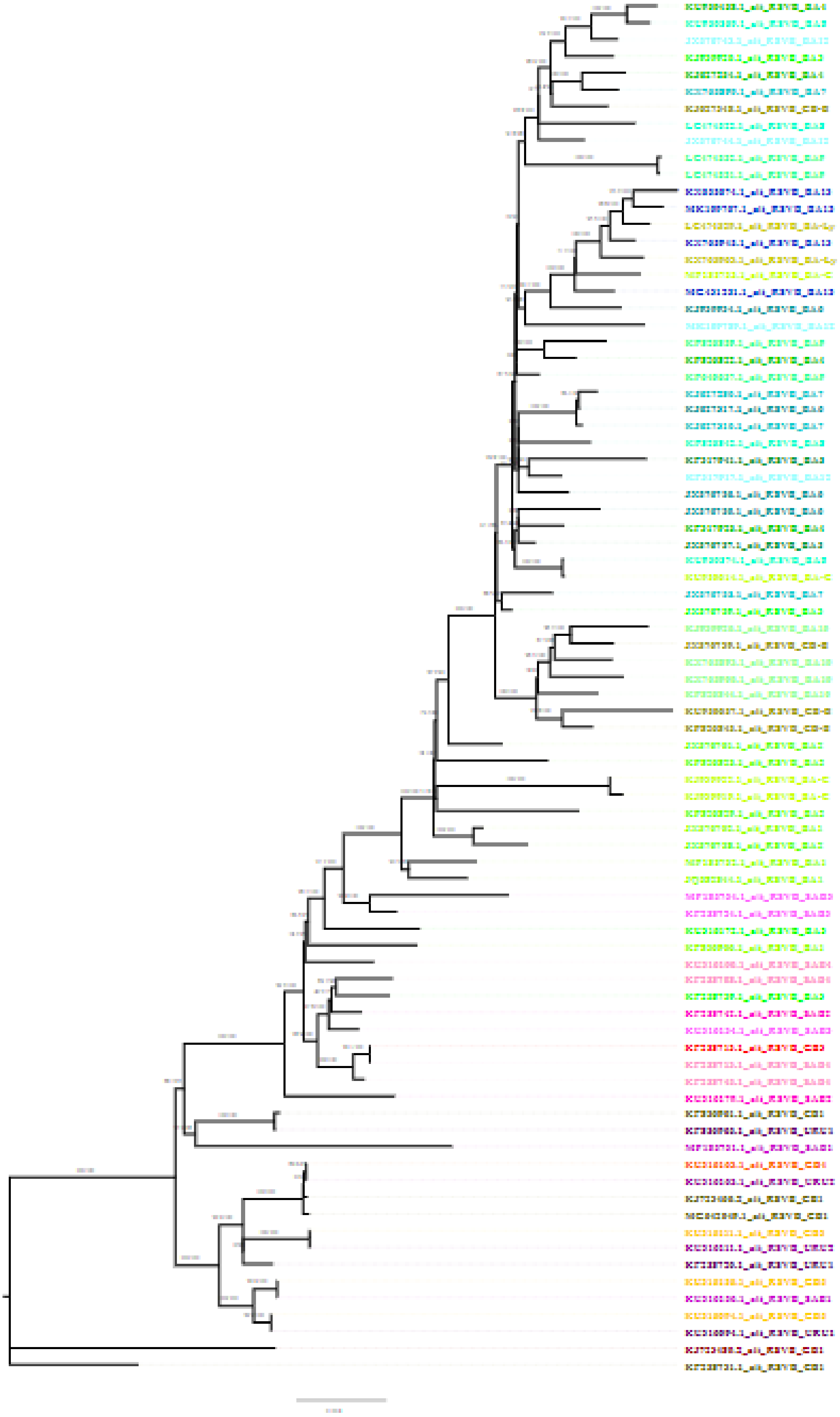


Figure 3.15: Maximum likelihood tree of RSV B full genomes of alternative sequences shows that most alternative reference sequences do cluster according to their genotype. BA strains are less clearly distinguishable. (approximate likelihood ratio/ bootstrap).

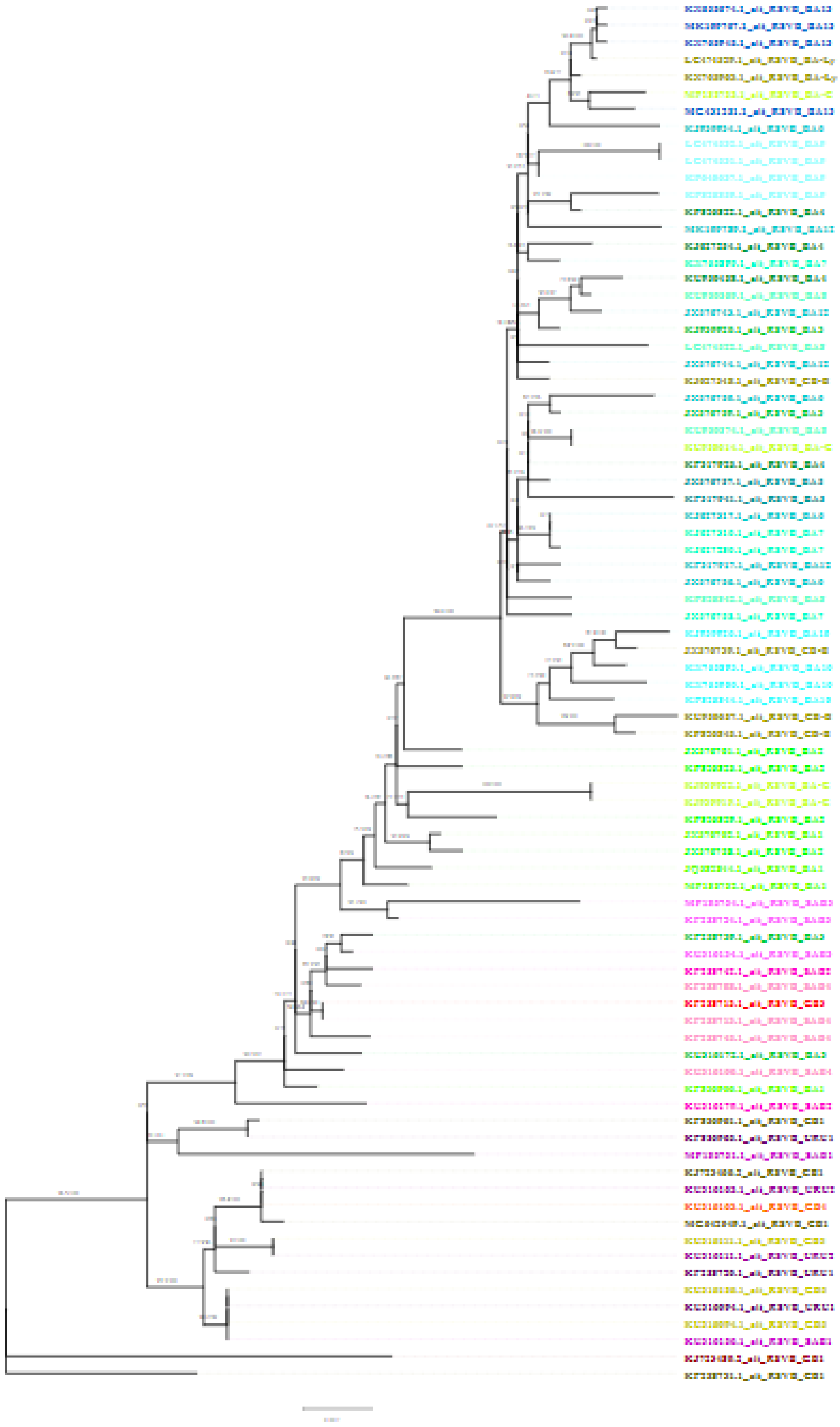


Figure 3.16 (a): Maximum likelihood trees of full G genes of alternative sequences of RSV B genotypes. This phylogenetic tree shows similar results to the phylogenetic tree based on full genomes. (approximate likelihood ratio/ bootstrap)

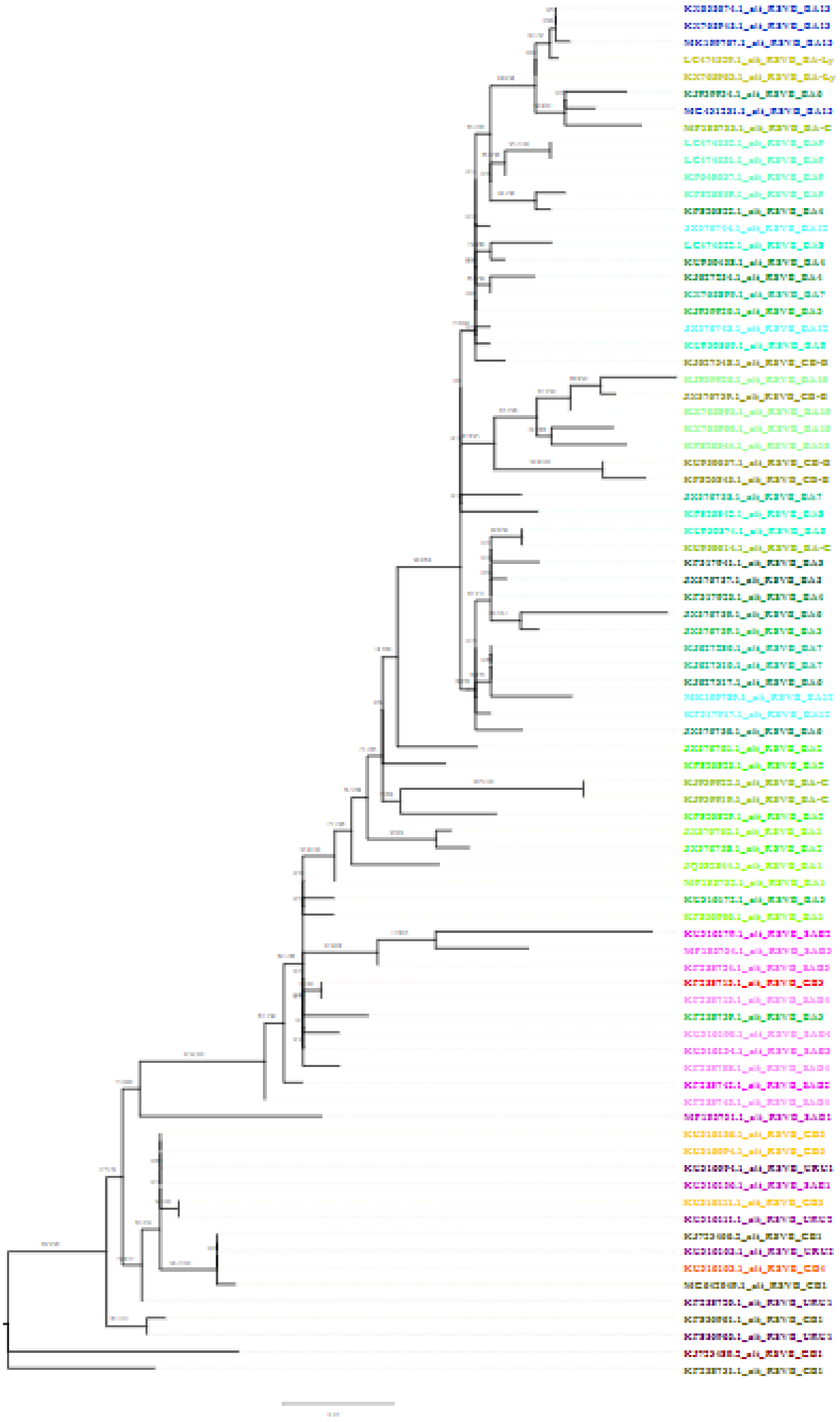


Figure 3.17 (b): Maximum likelihood trees of the second hypervariable region of G of alternative sequences of RSV B genotypes. This tree shows lower bootstrap values and approximate likelihood ratio values compared to phylogenetic trees based on full G genes or full genomes. (approximate likelihood ratio test/ bootstrap)

3.4. Determine reference dataset for genotyping RSV strains

3.4.1. Select references and test quality and usefulness of datasets for genotyping

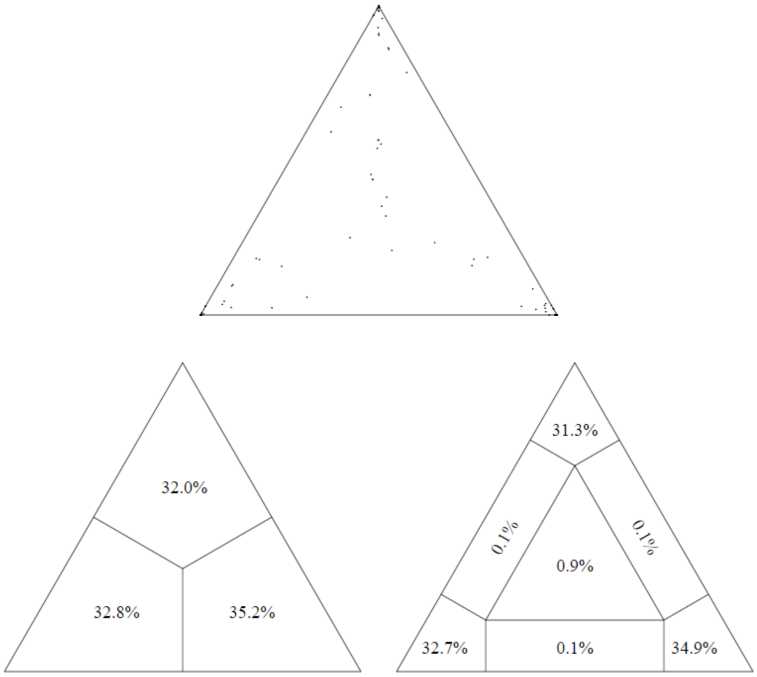
A set of reference sequences was selected based on the best complete genome alternatives that were available for each genotype. Alternatives with a similarity rate lower than 97.00% were considered too dissimilar to be used as a reference sequence. Therefore, genotype NA4 for RSV A and genotypes GB5, BA12, BA13, CB1, SAB4 and URU1 for RSV B did not have references in these datasets. This left 16 and 20 different genotypes to be represented in the selection of reference sequences for RSV A and RSV B respectively. The selected references are highlighted in yellow in Tables 3.3A and 3.3B.

The C-scores for each dataset were calculated, which showed little ambiguous positions in the RSV A and RSV B datasets (Table 3.10). The dataset containing both RSV A and B references was tested as well. The C-score for the overall alignment was 0.99 for both RSV A and B datasets separately and 0.98 for the combined dataset. The maximum C-score for individual sequences was only 0.99 for the combined dataset compared to 1.00 for both separate datasets. This was to be expected as the duplicated region for RSV A and RSV B strains is not the same and therefore there cannot be a sequence that carries both duplications and is “complete” in that sense. There will always be a gap in the sequences compared to each other. Overall, the completeness scores of these datasets were very high and the datasets could be used for further analysis.

Table 3.10: C-scores of reference sequences in datasets containing only RSV A, only RSV B or both RSV A and B references.

	Completeness scores references RSV A (n = 16)	**Completeness scores references RSV B (n = 20)**	**Completeness scores references RSV A and B (n=36)**
Completeness score for alignment (Ca)	0.99	0.99	0.98
Maximum completeness score for individual sequences (Cr_max)	1.00	1.00	0.99
Minimum completeness score for individual sequences (Cr_min)	0.98	0.98	0.97
Maximum completeness score for individual sites (Cc_max)	1.00	1.00	1.00
Minimum completeness score for individual sites (Cc_min)	0.06	0.05	0.03
Maximum completeness score for pairs of sequences (Cij_max)	0.99	1.00	0.99
Minimum completeness score for pairs of sequences (Cij_min)	0.98	0.98	0.96

In the next step, the proportion of informative quartets was tested for the separate datasets and the combined dataset. Figure 3.16 showed very low numbers of uninformative quartets, so the branching of phylogenetic trees would be confident in both RSV A and RSV B reference trees. The RSV A dataset only contained 0.9% of uninformative quartets, the RSV B dataset contained 3.6% of uninformative quartets and the combined dataset contained 3.1% of uninformative quartets. This meant that phylogenetic trees would produce confident trees with high bootstrap values and could be used for meaningful phylogenetic analysis.



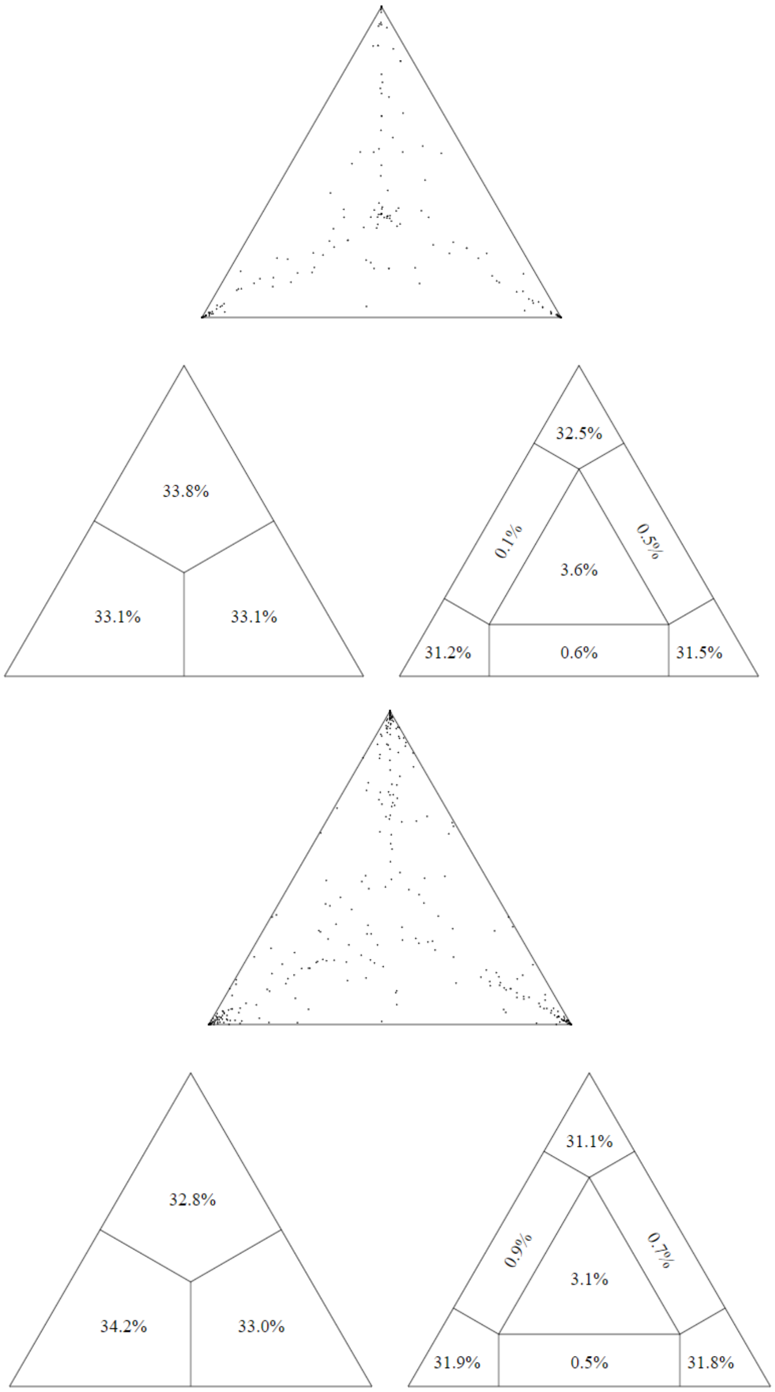


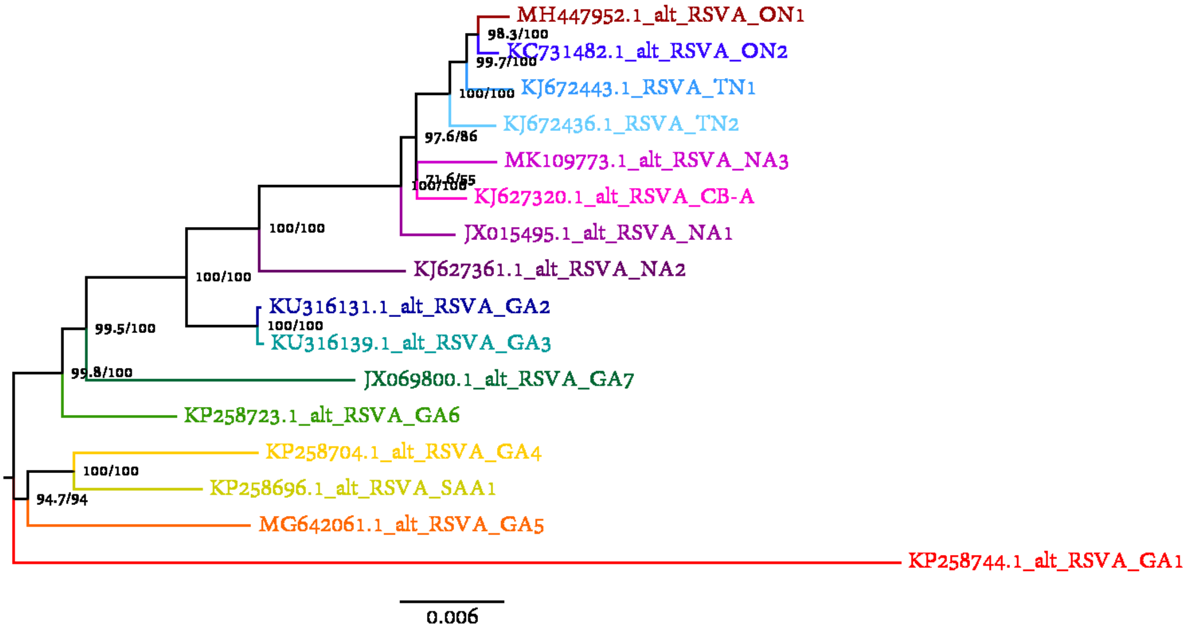
Figure 3.18: Triangular likelihood mapping tests whether a quartet (four randomly selected sequences from the dataset) produces an informative phylogenetic tree. Each dot in the top triangle is a quartet. The three parts of the left triangle indicate whether any subsets (if they were selected) have a bigger number of quartets. The right triangle shows uninformative quartets in the middle and informative quartets in the corners. Partial informative quartets are shown on the sides of the triangle. Informative quartets for RSV A (top), RSV B (middle) and combined datasets (bottom) are shown.

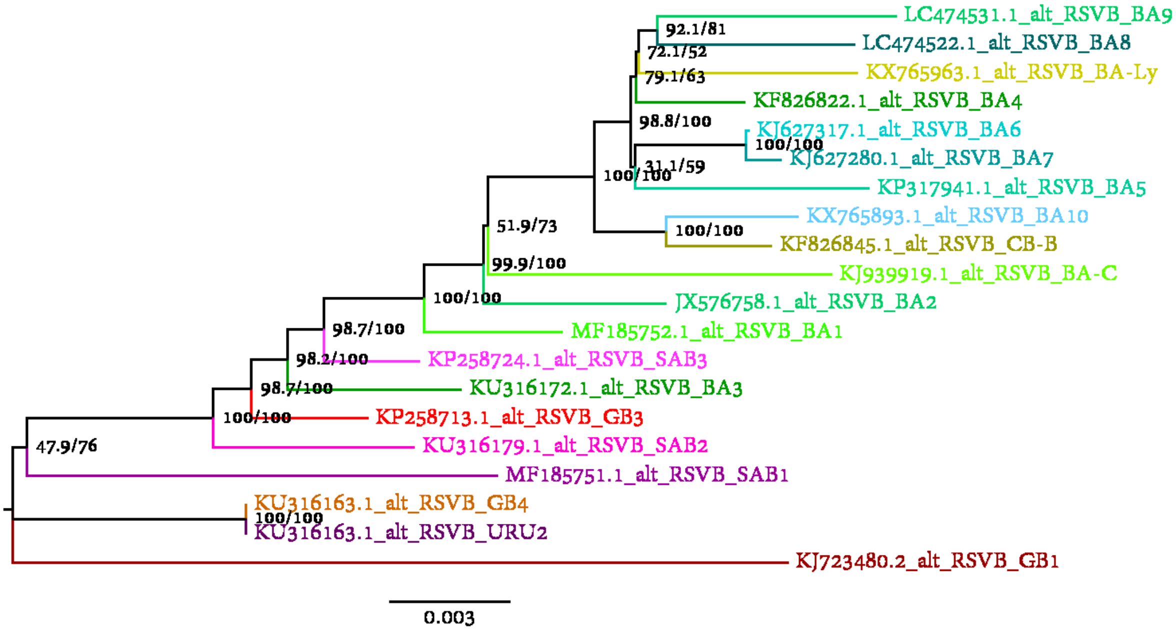
The last step to test the quality of these datasets for genotyping was running maximum likelihood trees with a bootstrap of 1000 and consulting the bootstrap values on the branches. The resulting trees showed good results for both separate and combined datasets (Figure 3.17). The lowest bootstrap value for the ML tree from the RSV A dataset was 55%, while all other bootstrap values were above 85%. Overall, this tree has good support as indicated by the high bootstrap values and aLRT values. The lowest aLRT value in this tree was 71.6, which was on the same node as the lowest bootstrap value.

A second ML tree was run from the RSV B dataset. There were more low values in this tree, although most branching still had high support (Figure 3.17). The lowest bootstrap value was 52%, followed by a node with 59%, 63%, 73%, 76% and 81%. All other nodes had bootstrap values of 100%. The same nodes that had bootstrap values under 85% also showed lower aLRT values: 72.1, 31.1, 79.1, 51.9, 47.9 and 92.1 respectively. All but one of those nodes were part of a cluster containing all BA strains, which were already known to be very similar.

The combined dataset produced an ML tree which contained the best bootstrap values and aLRT values of all trees. The lowest bootstrap value in this tree was 72% and the lowest aLRT value was 35.7, which was the lowest out of the three trees.

However, there were two clusters in this tree, RSV A and RSV B, and one outlier strain from the RSV B dataset. Using this dataset for genotyping would not improve the information retrieved from this analysis. The next sub-chapter will still compare the two separate reference datasets with the combined dataset nonetheless. This subchapter tests the use of these datasets to genotype unknown clinical samples. Only the G gene of these samples was determined and one RSV A and one RSV B dataset was constructed to genotype 116 clinical samples that were collected between 2014 and 2018 as well as a combined dataset.


MH447952.1_alt_RSVA_ON1
98.3/100
KC731482.1_alt_RSVA_ON2
99.7/100
KJ672443.1_RSVA_TN1
100/100
KJ672436.1_RSVA_TN2
97.6/86
MK109773.1_alt_RSVA_NA3
71.6/55
100/100
KJ627320.1_alt_RSVA_CB-A
JX015495.1_alt_RSVA_NA1
100/100
KJ627361.1_alt_RSVA_NA2
100/100
KU316131.1_alt_RSVA_GA2
100/100
KU316139.1_alt_RSVA_GA3
99.5/100
99.8/100
JX069800.1_alt_RSVA_GA7
KP258723.1_alt_RSVA_GA6
KP258704.1_alt_RSVA_GA4
100/100
KP258696.1_alt_RSVA_SAA1
94.7/94
MG642061.1_alt_RSVA_GA5
KP258744.1_alt_RSVA_GA1
0.006


LC474531.1_alt_RSVB_BA9
92.1/81
LC474522.1_alt_RSVB_BA8
72.1/52
KX765963.1_alt_RSVB_BA-Ly
79.1/63
KF826822.1_alt_RSVB_BA4
98.8/100
KJ627317.1_alt_RSVB_BA6
100/100
KJ627280.1_alt_RSVB_BA7
31.1/59
100/100
KP317941.1_alt_RSVB_BA5
KX765893.1_alt_RSVB_BA10
100/100
KF826845.1_alt_RSVB_CB-B
51.9/73
99.9/100
KJ939919.1_alt_RSVB_BA-C
100/100
JX576758.1_alt_RSVB_BA2
MF185752.1_alt_RSVB_BA1
98.7/100
98.2/100
KP258724.1_alt_RSVB_SAB3
98.7/100
KU316172.1_alt_RSVB_BA3
100/100
KP258713.1_alt_RSVB_GB3
KU316179.1_alt_RSVB_SAB2
47.9/76
MF185751.1_alt_RSVB_SAB1
KU316163.1_alt_RSVB_GB4
100/100
KU316163.1_alt_RSVB_URU2
KJ723480.2_alt_RSVB_GB1
0.003

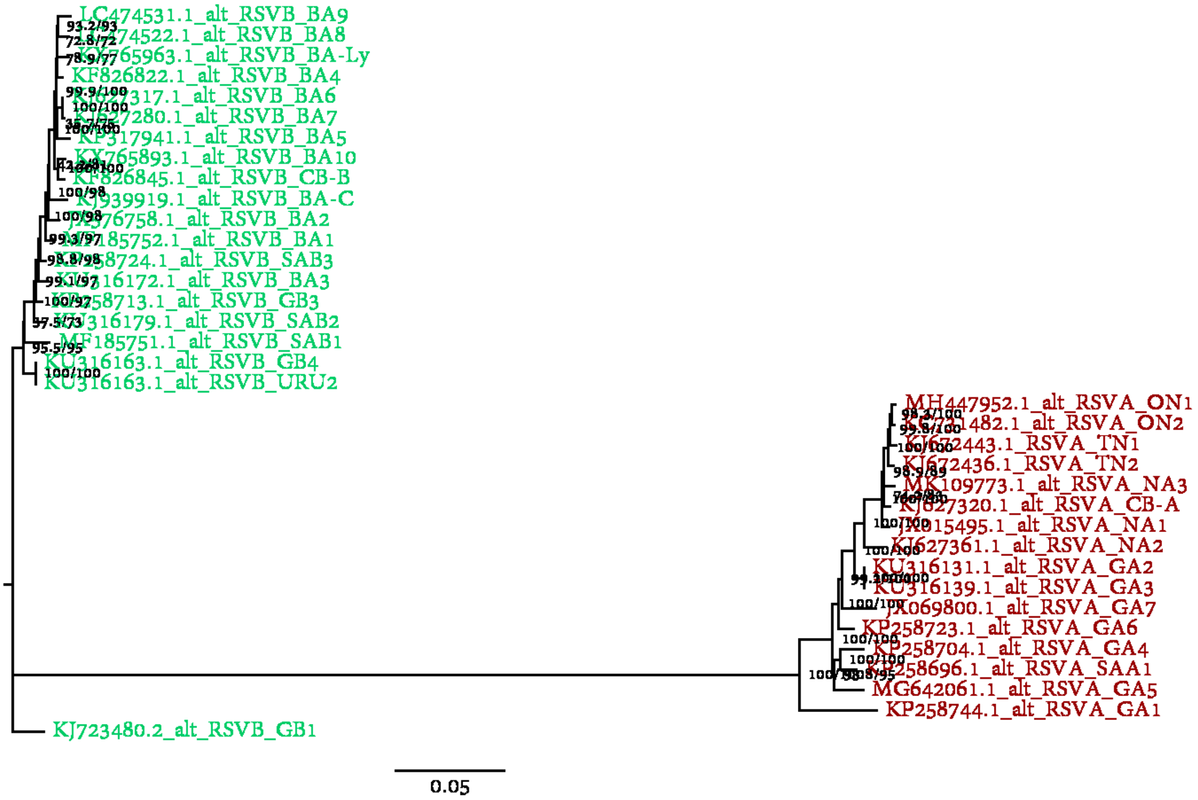


Figure 3.19: Phylogenetic trees of RSV A only (top), RSV B only (middle) and combined datasets of RSV A in bordeaux and RSV B in green (bottom) showed that separate datasets are more useful than one combined dataset for RSV A and B. The approximate likelihood ratio test values and bootstrap replicate values (aLRT/BB) were calculated as well and indicated in black at each node.

3.4.2. Testing the reference datasets with 116 clinical samples

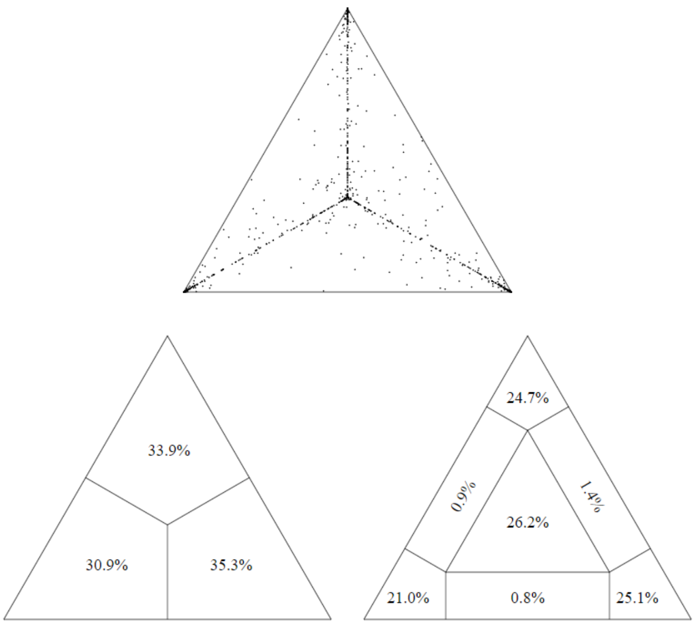
In the final part of this chapter, the selected reference datasets were tested in function of genotyping unknown strains. Samples were collected via Public Health England in collaboration with the Royal College of General Practitioners. 116 samples were selected from all over England and during season 2014-2015 to season 2017-2018. The G gene was sequenced from 52 RSV A and 60 RSV B strains by Sanger sequencing. These sequences were combined in three different datasets: one with RSV A clinical strains and references, one with RSV B clinical strains and references, and one with both RSV A and B clinical and reference sequences.

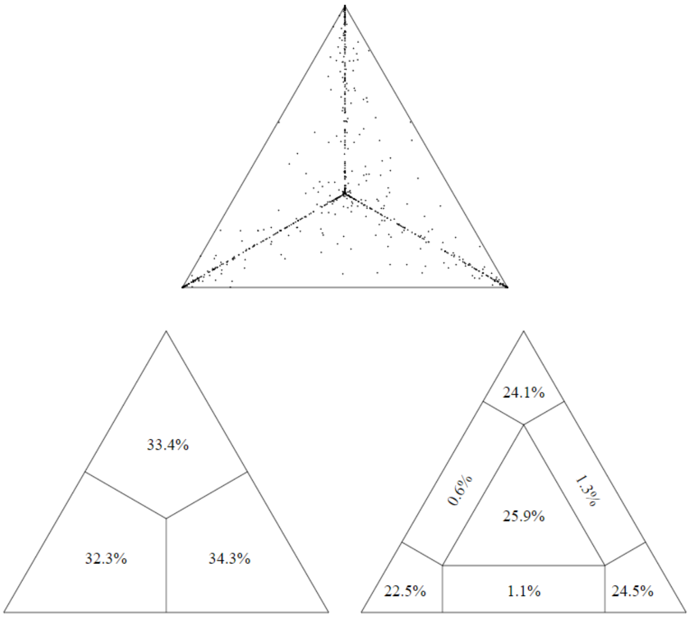
After alignment of the datasets, the C-scores were determined (Table 3.11). These were not as high as for the datasets with only reference sequences; the overall Ca was 0.98 for the RSV A dataset, 0.95 for the RSV B dataset and 0.94 for the combined dataset.

Table 3.11: C-scores of RSV A dataset with 52 clinical strains and 12 reference sequences, RSV B dataset with 60 clinical strains and 20 reference sequences and a combined dataset.

	Completeness scores references RSV A (n = 68)	**Completeness scores references RSV B (n = 80)**	**Completeness scores references RSV A and B (n = 148)**
Completeness score for alignment (Ca)	0.98	0.95	0.94
Maximum completeness score for individual sequences (Cr_max)	1.00	0.99	0.98
Minimum completeness score for individual sequences (Cr_min)	0.80	0.54	0.53
Maximum completeness score for individual sites (Cc_max)	1.00	1.00	1.00
Minimum completeness score for individual sites (Cc_min)	0.74	0.01	0.01
Maximum completeness score for pairs of sequences (Cij_max)	1.00	0.99	0.98
Minimum completeness score for pairs of sequences (Cij_min)	0.80	0.12	0.12

Next, the portion of informative quartets was calculated. This showed that a high percentage is uninformative for RSV A (26.2%), RSV B (25.9%) and the combined dataset (34.4%) as indicated in Figure 3.18. These percentages were a lot higher than for the reference sequence datasets, which indicated that branching decisions were not all as highly supported. This was due to the fact that some of these clinical strains were very similar either to each other or to the selected references.


33.9%
30.9%
35.3%
24.7%
0.9%
26.2%
1.4%
21.0%
0.8%
25.1%


33.4%
32.3%
34.3%
24.1%
0.6%
25.9%
1.3%
22.5%
1.1%
24.5%

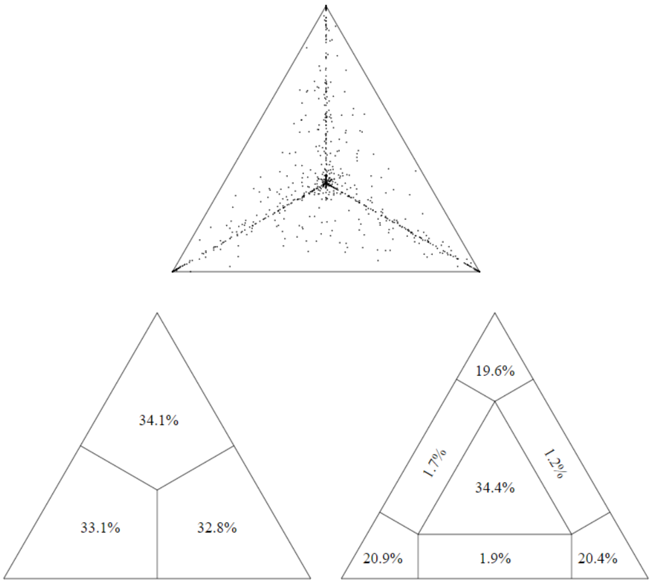


Figure 3.20: Triangular likelihood mapping tests whether a quartet (four randomly selected sequences from the dataset) produces an informative phylogenetic tree. Each dot in the top triangle is a quartet. The three parts of the left triangle indicate whether any subsets (if they were selected) have a bigger number of quartets. The right triangle shows uninformative quartets in the middle and informative quartets in the corners. Partial informative quartets are shown on the sides of the triangle. Informative quartets of RSV A (top), RSV B (middle) and combined (bottom) datasets with clinical strains and reference sequences.

The last step was to test the datasets to be able to genotype the unknown strains via building a maximum likelihood tree. The separate trees had better support, because of higher numbers of informative quartets and therefore higher numbers of confident branching. However, this did not influence genotypes of unknown strains compared to the combined tree (Figure 3.19).

Depending on the references used in the datasets, the outcome might be slightly different. Certain strains might fall in between genotypes and therefore cluster more with one genotype or another depending on the reference sequence used. This issue should be investigated and ideally, there would be an internationally recognised, standard reference dataset which should be used for genotyping. However, the best references need to be selected for that and right now, these might not have been determined yet.

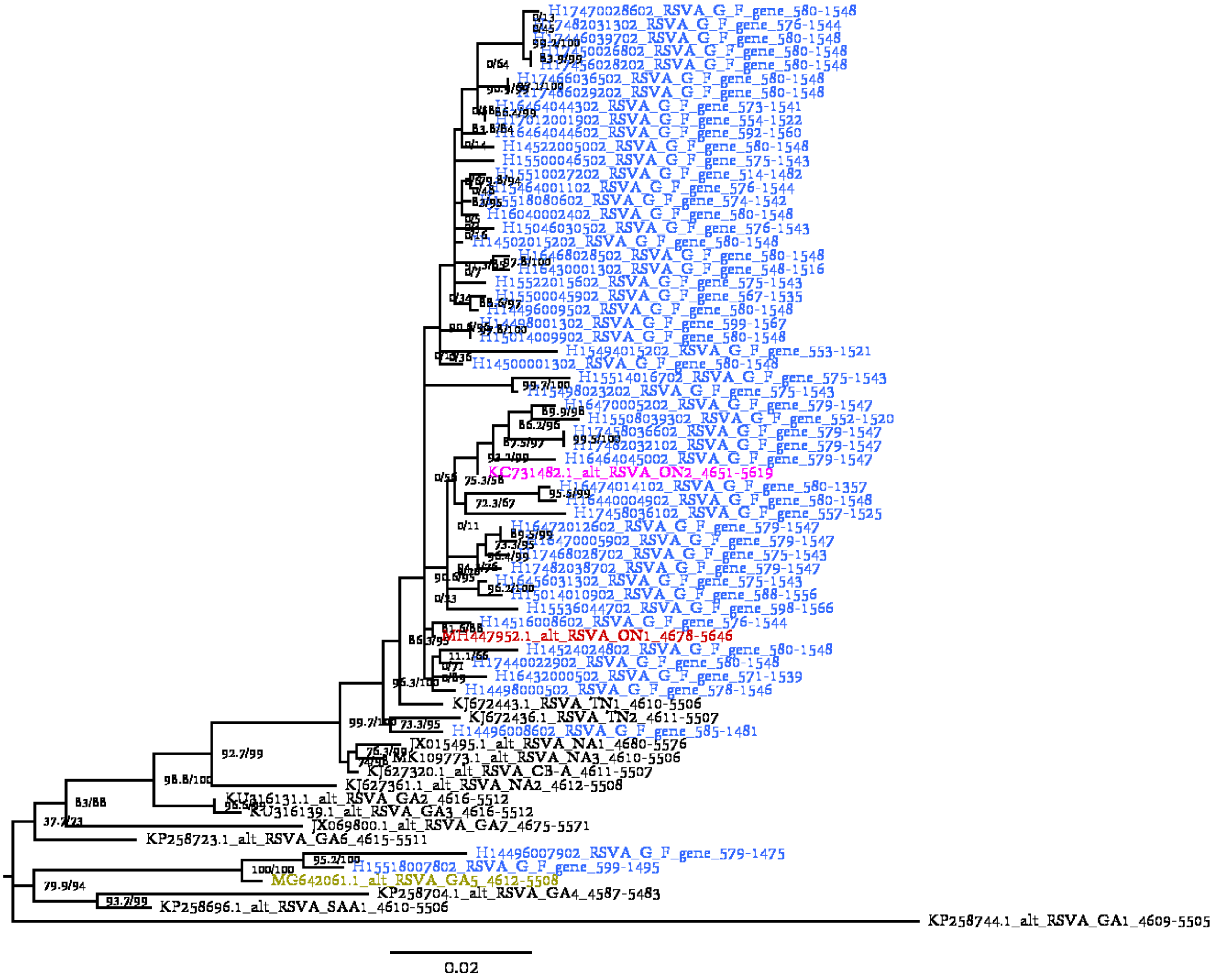

H17470028602_RSVA_G_F_gene_580-1548
H17482031302_RSVA_G_F_gene_576-1544
H17446039702_RSVA_G_F_gene_580-1548
H17450026802_RSVA_G_F_gene_580-1548
H17456028202_RSVA_G_F_gene_580-1548
H17466036502_RSVA_G_F_gene_580-1548
H17486029202_RSVA_G_F_gene_580-1548
H16464044302_RSVA_G_F_gene_573-1541
H17012001902_RSVA_G_F_gene_554-1522
H16464044602_RSVA_G_F_gene_592-1560
H14522005002_RSVA_G_F_gene_580-1548
H15500046502_RSVA_G_F_gene_575-1543
H15510027202_RSVA_G_F_gene_514-1482
H15464001102_RSVA_G_F_gene_576-1544
H15518080602_RSVA_G_F_gene_574-1542
H16040002402_RSVA_G_F_gene_580-1548
H15046030502_RSVA_G_F_gene_576-1543
H14502015202_RSVA_G_F_gene_580-1548
H16468028502_RSVA_G_F_gene_580-1548
H16430001302_RSVA_G_F_gene_548-1516
H15522015602_RSVA_G_F_gene_575-1543
H15500045902_RSVA_G_F_gene_567-1535
H14496009502_RSVA_G_F_gene_580-1548
H14498001302_RSVA_G_F_gene_599-1567
H15014009902_RSVA_G_F_gene_580-1548
H15494015202_RSVA_G_F_gene_553-1521
H14500001302_RSVA_G_F_gene_580-1548
H15514016702_RSVA_G_F_gene_575-1543
H15498023202_RSVA_G_F_gene_575-1543
H16470005202_RSVA_G_F_gene_579-1547
H15508039302_RSVA_G_F_gene_552-1520
H17458036602_RSVA_G_F_gene_579-1547
H17482032102_RSVA_G_F_gene_579-1547
H16464045002_RSVA_G_F_gene_579-1547
KC731482.1_alt_RSVA_ON2_4651-5619
H16474014102_RSVA_G_F_gene_580-1357
H16440004902_RSVA_G_F_gene_580-1548
H17458036102_RSVA_G_F_gene_557-1525
H16472012602_RSVA_G_F_gene_579-1547
H16470005902_RSVA_G_F_gene_579-1547
H17468028702_RSVA_G_F_gene_575-1543
H17482038702_RSVA_G_F_gene_579-1547
H16456031302_RSVA_G_F_gene_575-1543
H15014010902_RSVA_G_F_gene_588-1556
H15536044702_RSVA_G_F_gene_598-1566
H14516008602_RSVA_G_F_gene_576-1544
MH447952.1_alt_RSVA_ON1_4678-5646
H14524024802_RSVA_G_F_gene_580-1548
H17440022902_RSVA_G_F_gene_580-1548
H16432000502_RSVA_G_F_gene_571-1539
H14498000502_RSVA_G_F_gene_578-1546
KJ672443.1_RSVA_TN1_4610-5506
KJ672436.1_RSVA_TN2_4611-5507
H14496008602_RSVA_G_F_gene_585-1481
JX015495.1_alt_RSVA_NA1_4680-5576
MK109773.1_alt_RSVA_NA3_4610-5506
KJ627320.1_alt_RSVA_CB-A_4611-5507
KJ627361.1_alt_RSVA_NA2_4612-5508
KU316131.1_alt_RSVA_GA2_4616-5512
KU316139.1_alt_RSVA_GA3_4616-5512
JX069800.1_alt_RSVA_GA7_4675-5571
KP258723.1_alt_RSVA_GA6_4615-5511
H14496007902_RSVA_G_F_gene_579-1475
H15518007802_RSVA_G_F_gene_599-1495
MG642061.1_alt_RSVA_GA5_4612-5508
KP258704.1_alt_RSVA_GA4_4587-5483
KP258696.1_alt_RSVA_SAA1_4610-5506
KP258744.1_alt_RSVA_GA1_4609-5505
0/13
0/45
99.2/100
83.9/99
0/64
97.1/100
86.4/99
83.8/64
0/14
79.8/94
0/48
83/95
0/5
0/16
97.8/100
0/7
0/34
88.6/97
97.8/100
99.7/100
89.9/98
86.2/96
99.5/100
87.5/97
0/56
75.3/58
95.5/99
72.3/67
0/11
89.5/99
73.3/95
96.4/99
90.6/95
96.2/100
0/13
81.6/88
86.3/95
11.1/66
0/71
0/89
96.3/100
99.7/100
73.3/95
76.3/99
92.7/99
98.8/100
83/88
37.7/73
95.2/100
100/100
79.9/94
93.7/99
0.02

Figure 3.21 (a): Phylogenetic tree of RSV A dataset with clinical strains (blue) and reference sequences (pink = ON2, red = ON1, yellow = GA5). (approximate likelihood ratio test/ bootstrap)

A total of 52 clinical strains were used to build a phylogenetic tree with 12 reference sequence in the dataset of RSV A strains (Figure 3.19 (a)). Four different genotypes were detected in this group of clinical samples: two strains were part of GA5, one strain was part of TN2, 8 sequences clustered with the ON2 reference strain and 41 sequences were ON1 strains. Only three clinical RSV A strains, namely the GA5 and TN2 strains, did not contain the 72-nucleotide duplication, while 49 strains, namely ON1 and ON2 strains, did.

The most common genotype during this season is ON1, which was confirmed by these data (110). It is thought to have a fitness advantage, but there is no evidence to date to suggest altered pathogenicity (111).

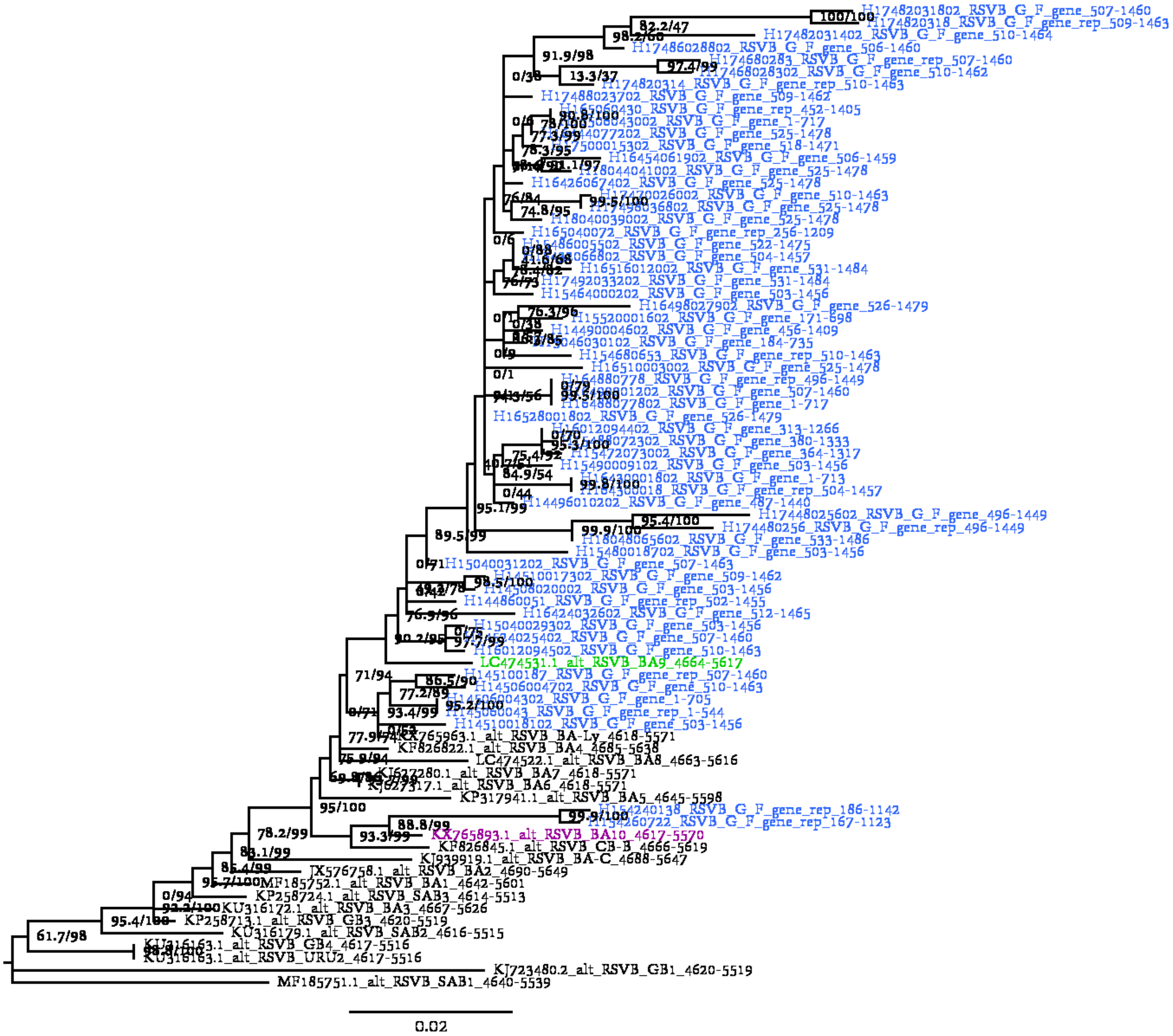

H17482031802_RSVB_G_F_gene_507-1460
H174820318_RSVB_G_F_gene_rep_509-1463
H17482031402_RSVB_G_F_gene_510-1464
H17486028802_RSVB_G_F_gene_506-1460
H174680283_RSVB_G_F_gene_rep_507-1460
H17468028302_RSVB_G_F_gene_510-1462
H174820314_RSVB_G_F_gene_rep_510-1463
H17488023702_RSVB_G_F_gene_509-1462
H165060430_RSVB_G_F_gene_rep_452-1405
H16508043002_RSVB_G_F_gene_1-717
H16544077202_RSVB_G_F_gene_525-1478
H17500015302_RSVB_G_F_gene_518-1471
H16454061902_RSVB_G_F_gene_506-1459
H18044041002_RSVB_G_F_gene_525-1478
H16426067402_RSVB_G_F_gene_525-1478
H17470026002_RSVB_G_F_gene_510-1463
H17498036802_RSVB_G_F_gene_525-1478
H18040039002_RSVB_G_F_gene_525-1478
H165040072_RSVB_G_F_gene_rep_256-1209
H15486005502_RSVB_G_F_gene_522-1475
H16506066802_RSVB_G_F_gene_504-1457
H16516012002_RSVB_G_F_gene_531-1484
H17492033202_RSVB_G_F_gene_531-1484
H15464000202_RSVB_G_F_gene_503-1456
H16498027902_RSVB_G_F_gene_526-1479
H15520001602_RSVB_G_F_gene_171-698
H14490004602_RSVB_G_F_gene_456-1409
H17504603010_RSVB_G_F_gene_184-735
H154680653_RSVB_G_F_gene_rep_510-1463
H16510003002_RSVB_G_F_gene_525-1478
H164880778_RSVB_G_F_gene_rep_496-1449
H15490001202_RSVB_G_F_gene_507-1460
H16488077802_RSVB_G_F_gene_1-717
H16528001802_RSVB_G_F_gene_526-1479
H16012094402_RSVB_G_F_gene_313-1266
H15488072302_RSVB_G_F_gene_380-1333
H15472073002_RSVB_G_F_gene_364-1317
H15490009102_RSVB_G_F_gene_503-1456
H16430001802_RSVB_G_F_gene_1-713
H164300018_RSVB_G_F_gene_rep_504-1457
H14496010202_RSVB_G_F_gene_487-1440
H17448025602_RSVB_G_F_gene_496-1449
H174480256_RSVB_G_F_gene_rep_496-1449
H18048065602_RSVB_G_F_gene_533-1486
H15480018702_RSVB_G_F_gene_503-1456
H15040031202_RSVB_G_F_gene_507-1463
H14510017302_RSVB_G_F_gene_509-1462
H14508020002_RSVB_G_F_gene_503-1456
H144860051_RSVB_G_F_gene_rep_502-1455
H16424032602_RSVB_G_F_gene_512-1465
H15040029302_RSVB_G_F_gene_503-1456
H15524025402_RSVB_G_F_gene_507-1460
H16012094502_RSVB_G_F_gene_510-1463
LC474531.1_alt_RSVB_BA9_4664-5617
H145100187_RSVB_G_F_gene_rep_507-1460
H14506004702_RSVB_G_F_gene_510-1463
H14506004302_RSVB_G_F_gene_1-705
H145060043_RSVB_G_F_gene_rep_1-544
H14510018102_RSVB_G_F_gene_503-1456
KX765963.1_alt_RSVB_BA-Ly_4618-5571
KF826822.1_alt_RSVB_BA4_4685-5638
LC474522.1_alt_RSVB_BA8_4663-5616
KJ627280.1_alt_RSVB_BA7_4618-5571
KJ627317.1_alt_RSVB_BA6_4618-5571
KP317941.1_alt_RSVB_BA5_4645-5598
H154240138_RSVB_G_F_gene_rep_186-1142
H154260722_RSVB_G_F_gene_rep_167-1123
KX765893.1_alt_RSVB_BA10_4617-5570
KF826845.1_alt_RSVB_CB-B_4666-5619
KJ939919.1_alt_RSVB_BA-C_4688-5647
JX576758.1_alt_RSVB_BA2_4690-5649
MF185752.1_alt_RSVB_BA1_4642-5601
KP258724.1_alt_RSVB_SAB3_4614-5513
KU316172.1_alt_RSVB_BA3_4667-5626
KP258713.1_alt_RSVB_GB3_4620-5519
KU316179.1_alt_RSVB_SAB2_4616-5515
KU316163.1_alt_RSVB_GB4_4617-5516
KU316183.1_alt_RSVB_URU2_4617-5516
KJ723480.2_alt_RSVB_GB1_4620-5519
MF185751.1_alt_RSVB_SAB1_4640-5539
100/100
82.2/47
98.2/60
91.9/98
97.4/99
0/38
13.3/37
90.8/100
78/100
77.3/99
78.3/95
91.1/97
76/84
99.5/100
74.8/95
0/6
0/88
41.6/68
78.4/82
76/73
76.3/96
0/1
0/38
0/9
0/1
0/72
99.5/100
0/70
95.3/100
75.4/92
84.9/54
99.8/100
0/44
95.1/99
95.4/100
99.9/100
89.5/99
0/71
98.5/100
76.9/96
0/75
97.7/99
90.2/95
71/94
86.5/90
77.2/89
95.2/100
93.4/99
0/62
77.9/74
75.9/94
95/100
99.9/100
88.8/99
93.3/99
78.2/99
83.1/99
85.4/99
95.7/100
0/94
92.2/100
95.4/100
61.7/98
0.02

Figure 3.22 (b): Phylogenetic tree of RSV B dataset with clinical strains (blue) and reference sequences (green = BA9 and purple = BA10). (approximate likelihood ratio test/ bootstrap)

In the RSV B dataset, there were 60 clinical samples and 20 reference strains. This dataset was used to build a phylogenetic tree to study the genotypes of these clinical samples (Figure 3.19 (b)). Three different genotypes were detected in this dataset: 2 sequences belonged to BA10, 5 sequences clustered most closely with BA-Ly and 53 sequences were part of BA9. All RSV B clinical strains contained the 60-nucleotide duplication in the G gene.

Similarly to RSV A, RSV B has evolved with a 60-nucleotide duplication in its genome. These strains are grouped under the BA genotypes (83). Furthermore, BA genotypes have also replaced most strains without the duplication its genome in our current epidemics and BA9 is the most common of the BA genotypes (92, 114, 115). These data mirror that trend and therefore agree with the literature to date.

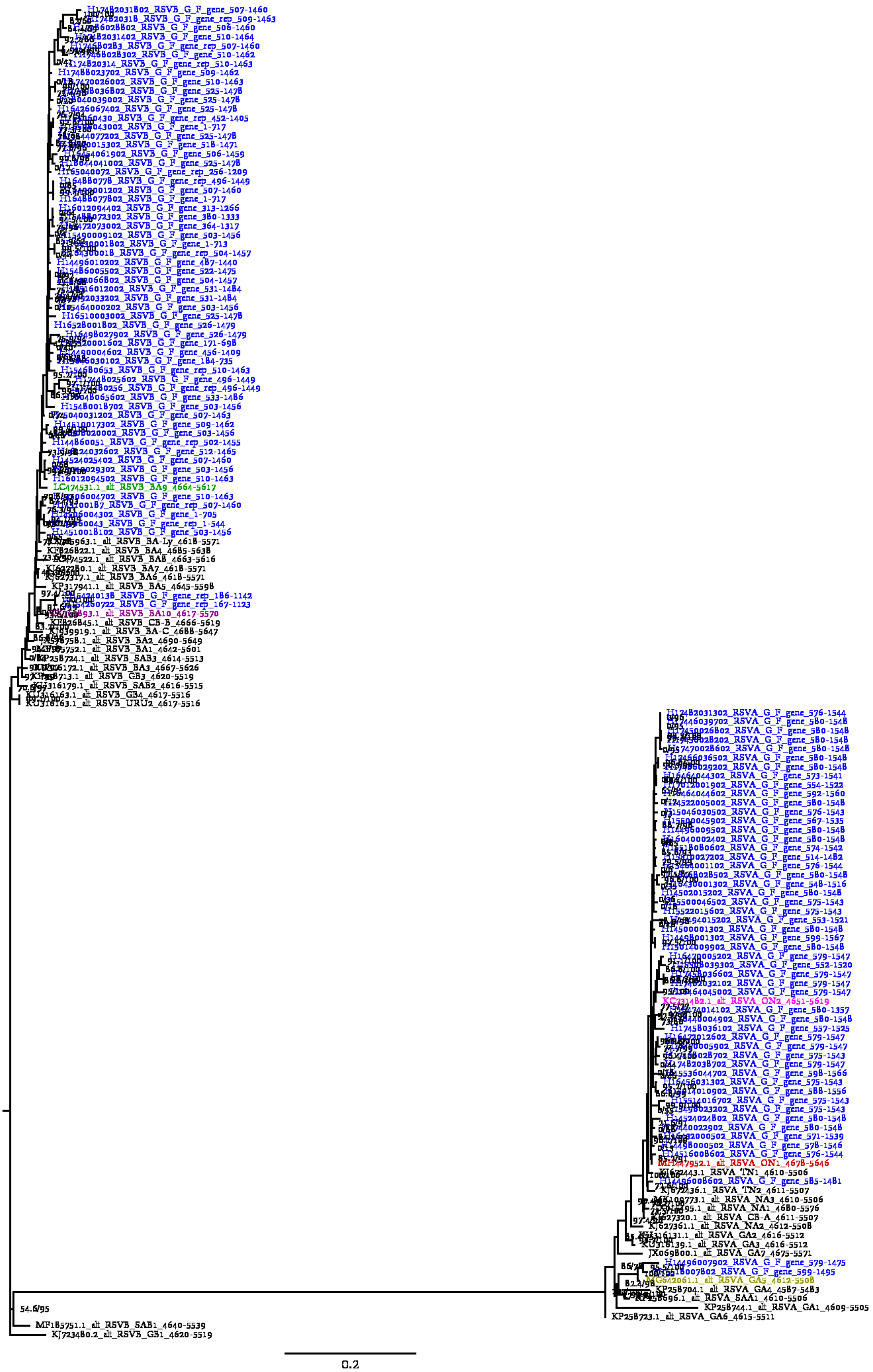

H17482031802_RSVB_G_F_gene_507-1460
H174820318_RSVB_G_F_gene_rep_509-1463
H17486028802_RSVB_G_F_gene_506-1460
H17482031402_RSVB_G_F_gene_510-1464
H174680283_RSVB_G_F_gene_rep_507-1460
H17468028302_RSVB_G_F_gene_510-1462
H174820314_RSVB_G_F_gene_rep_510-1463
H17488023702_RSVB_G_F_gene_509-1462
H17470026002_RSVB_G_F_gene_510-1463
8036802_RSVB_G_F_gene_525-1478
5040039002_RSVB_G_F_gene_525-1478
H16426067402_RSVB_G_F_gene_525-1478
60430_RSVB_G_F_gene_rep_452-1405
043002_RSVB_G_F_gene_1-717
44077202_RSVB_G_F_gene_525-1478
0015302_RSVB_G_F_gene_518-1471
H16454061902_RSVB_G_F_gene_506-1459
H18044041002_RSVB_G_F_gene_525-1478
H165040072_RSVB_G_F_gene_rep_256-1209
H164880778_RSVB_G_F_gene_rep_496-1449
0001202_RSVB_G_F_gene_507-1460
H16488077802_RSVB_G_F_gene_1-717
H16012094402_RSVB_G_F_gene_313-1266
88072302_RSVB_G_F_gene_380-1333
472073002_RSVB_G_F_gene_364-1317
H15490009102_RSVB_G_F_gene_503-1456
0001802_RSVB_G_F_gene_1-713
4300018_RSVB_G_F_gene_rep_504-1457
H14496010202_RSVB_G_F_gene_487-1440
H15486005502_RSVB_G_F_gene_522-1475
2066802_RSVB_G_F_gene_504-1457
16012002_RSVB_G_F_gene_531-1484
92033202_RSVB_G_F_gene_531-1484
5464000202_RSVB_G_F_gene_503-1456
H16510003002_RSVB_G_F_gene_525-1478
H16528001802_RSVB_G_F_gene_526-1479
H16498027902_RSVB_G_F_gene_526-1479
20001602_RSVB_G_F_gene_171-698
490004602_RSVB_G_F_gene_456-1409
5046030102_RSVB_G_F_gene_184-735
H154680653_RSVB_G_F_gene_rep_510-1463
H17448025602_RSVB_G_F_gene_496-1449
480256_RSVB_G_F_gene_rep_496-1449
048065602_RSVB_G_F_gene_533-1486
H15480018702_RSVB_G_F_gene_503-1456
5040031202_RSVB_G_F_gene_507-1463
H14510017302_RSVB_G_F_gene_509-1462
8020002_RSVB_G_F_gene_503-1456
H144860051_RSVB_G_F_gene_rep_502-1455
24032602_RSVB_G_F_gene_512-1465
H14524025402_RSVB_G_F_gene_507-1460
029302_RSVB_G_F_gene_503-1456
H16012094502_RSVB_G_F_gene_510-1463
LC474531.1_alt_RSVB_BA9_4664-5617
06004702_RSVB_G_F_gene_510-1463
00187_RSVB_G_F_gene_rep_507-1460
06004302_RSVB_G_F_gene_1-705
60043_RSVB_G_F_gene_rep_1-544
H14510018102_RSVB_G_F_gene_503-1456
963.1_alt_RSVB_BA-Ly_4618-5571
KF826822.1_alt_RSVB_BA4_4685-5638
474522.1_alt_RSVB_BA8_4663-5616
KJ627280.1_alt_RSVB_BA7_4618-5571
KJ627317.1_alt_RSVB_BA6_4618-5571
KP317941.1_alt_RSVB_BA5_4645-5598
4240138_RSVB_G_F_gene_rep_186-1142
4260722_RSVB_G_F_gene_rep_167-1123
765893.1_alt_RSVB_BA10_4617-5570
KF826845.1_alt_RSVB_CB-B_4666-5619
KJ939919.1_alt_RSVB_BA-C_4688-5647
59758.1_alt_RSVB_BA2_4690-5649
05752.1_alt_RSVB_BA1_4642-5601
KP258724.1_alt_RSVB_SAB3_4614-5513
6172.1_alt_RSVB_BA3_4667-5626
8713.1_alt_RSVB_GB3_4620-5519
KU316179.1_alt_RSVB_SAB2_4616-5515
KU316163.1_alt_RSVB_GB4_4617-5516
KU316163.1_alt_RSVB_URU2_4617-5516
54.6/95
MF185751.1_alt_RSVB_SAB1_4640-5539
KJ723480.2_alt_RSVB_GB1_4620-5519
H17482031302_RSVA_G_F_gene_576-1544
7446039702_RSVA_G_F_gene_580-1548
7450026802_RSVA_G_F_gene_580-1548
028202_RSVA_G_F_gene_580-1548
H17470028602_RSVA_G_F_gene_580-1548
H17466036502_RSVA_G_F_gene_580-1548
029202_RSVA_G_F_gene_580-1548
H16464044302_RSVA_G_F_gene_573-1541
7012001902_RSVA_G_F_gene_554-1522
6464044602_RSVA_G_F_gene_592-1560
4522005002_RSVA_G_F_gene_580-1548
15046030502_RSVA_G_F_gene_576-1543
H15500045902_RSVA_G_F_gene_567-1535
H14496009502_RSVA_G_F_gene_580-1548
H16040002402_RSVA_G_F_gene_580-1548
H15518080602_RSVA_G_F_gene_574-1542
0027202_RSVA_G_F_gene_514-1482
464001102_RSVA_G_F_gene_576-1544
8028502_RSVA_G_F_gene_580-1548
430001302_RSVA_G_F_gene_548-1516
H14502015202_RSVA_G_F_gene_580-1548
5500046502_RSVA_G_F_gene_575-1543
H15522015602_RSVA_G_F_gene_575-1543
94015202_RSVA_G_F_gene_553-1521
H14500001302_RSVA_G_F_gene_580-1548
H14498001302_RSVA_G_F_gene_599-1567
H15014009902_RSVA_G_F_gene_580-1548
H16470005202_RSVA_G_F_gene_579-1547
5508039302_RSVA_G_F_gene_552-1520
7458036602_RSVA_G_F_gene_579-1547
7482032102_RSVA_G_F_gene_579-1547
464045002_RSVA_G_F_gene_579-1547
KC731482.1_alt_RSVA_ON2_4651-5619
474014102_RSVA_G_F_gene_580-1357
440004902_RSVA_G_F_gene_580-1548
H17458036102_RSVA_G_F_gene_557-1525
H16472012602_RSVA_G_F_gene_579-1547
0005902_RSVA_G_F_gene_579-1547
028702_RSVA_G_F_gene_575-1543
H17482038702_RSVA_G_F_gene_579-1547
5536044702_RSVA_G_F_gene_598-1566
H16456031302_RSVA_G_F_gene_575-1543
014010902_RSVA_G_F_gene_588-1556
H15514016702_RSVA_G_F_gene_575-1543
1549823202_RSVA_G_F_gene_575-1543
H14524024802_RSVA_G_F_gene_580-1548
7440022902_RSVA_G_F_gene_580-1548
2000502_RSVA_G_F_gene_571-1539
H14498000502_RSVA_G_F_gene_578-1546
H14516008602_RSVA_G_F_gene_576-1544
MH447952.1_alt_RSVA_ON1_4678-5646
2443.1_RSVA_TN1_4610-5506
H14496008602_RSVA_G_F_gene_585-1481
KJ672436.1_RSVA_TN2_4611-5507
109773.1_alt_RSVA_NA3_4610-5506
95.1_alt_RSVA_NA1_4680-5576
627320.1_alt_RSVA_CB-A_4611-5507
KJ627361.1_alt_RSVA_NA2_4612-5508
16131.1_alt_RSVA_GA2_4616-5512
KU316139.1_alt_RSVA_GA3_4616-5512
JX069800.1_alt_RSVA_GA7_4675-5571
H14496007902_RSVA_G_F_gene_579-1475
8007802_RSVA_G_F_gene_599-1495
MG642061.1_alt_RSVA_GA5_4612-5508
KP258704.1_alt_RSVA_GA4_4587-5483
696.1_alt_RSVA_SAA1_4610-5506
KP258744.1_alt_RSVA_GA1_4609-5505
KP258723.1_alt_RSVA_GA6_4615-5511
0.2

Figure 3.23 (c): Phylogenetic tree of RSV A and RSV B combined dataset with clinical strains (blue) and reference sequences (pink = ON2, red = ON1, yellow = GA5, green = BA9 and purple = BA10). (approximate likelihood ratio test/ bootstrap)

3.5. Discussion

When RSV was discovered in 1956, there was no way to divide strains into different groups. The first way to group strains was based on serotyping techniques and in 1966 Coates *et al.* showed antigenic differences based on which RSV could be divided into two different groups (71). Once more detailed sequence determination methods were used, it became clear that RSV accumulates amino acid changes and that this is likely due to immune driven natural selection (84, 85). Some divergent strains survived over time with some strains showing some fitness advantage compared to strains from different genotypes. In 1998, the first paper on the currently used genotyping system appeared detailing the division of strains into groups based on the sequence of the virus. Currently, the most common strains carry a duplication in their G gene for both RSV A and RSV B genotypes (83, 110).

In this chapter, the current selection of genotypes was investigated and it showed that genotype names were appointed quite randomly over time. The basis for naming a new genotype is very flexible. All papers declaring a new genotype, used a (partial) sequence and compared it to some selected other (partial) sequences and if they clustered separate from the rest of the tree in phylogenetic analysis, a new genotype was "discovered". All these papers on new genotypes have one thing in common; they describe genotypes that are different from each other and previous genotypes by selecting a limited amount of reference sequences. The choice of references to compare new strains to determines whether it clusters separately or not. Even when a great amount of sequences is used, it is possible that the selection is too narrow in location or time to determine the correct cluster the new strains belong to. The current genotyping system seems to have developed quite organically by a number of different groups sequencing some parts of the RSV genome and publishing about it. A clear system or a distinct, rational structure is not incorporated.

This raises the question whether we should continue using the original sequences to keep determining the genotypes of RSV. Perhaps, all current genotypes are not all true separate genotypes or some might be too closely related to one another to all be full-fledged genotypes. The original genotypes that have been used so far are mostly based on partial sequences of the G gene. Analysis in this chapter showed that such a small piece of the sequence does not carry enough information to determine clusters confidently in phylogenetic analysis. However, old samples are not as readily available for sequencing as current samples. Furthermore, RSV is an RNA virus which is very labile. That makes it extra difficult to extract full sequences from older samples. There are not enough full genomes from samples from the 1960's, 70's and 80' available to determine full genomes from older genotypes. It might not be possible to ever do so. Perhaps, we should stop looking back and start with a new genotype system based on current and future genomes as we cannot go back in time to collect the necessary information to include old genotypes?

It is also possible there are just not enough strains sequenced yet to see the evolution of one strain over time. It might just look as if there are different genotypes, while in fact, it might just be different snapshots of the same virus strain over time with space between clusters where there were not enough strains sequenced to see the evolution of said strain. Perhaps, there are a very limited number of genotypes which all evolve into slightly different strains which disappear over time. It is hard to investigate this hypothesis at the moment as there is just not enough sequencing information available to do so. Looking to the future, it would be interesting to investigate much more samples as the virus strains will travel around the world and come back mutated to the same places one year later. Perhaps, sub-genotypes might be called to life to solve this issue. The same strain will then appear as different sub-genotypes, although they all descended from the same strain originally and are still closely related to one another and are not actually new genotypes just yet. That idea circles back to the question “What makes a new genotype a new genotype?”.

For measles, genotyping is based on a 450 nucleotide region of the nucleotide gene (N), although, for all representative strains, the complete haemagglutinin gene (H) should be sequenced as well (209). These are the most variable genes of the measles virus. Representative strains from specific countries or strains causing large outbreaks should also be isolated and H should be sequenced. So far, 24 genotypes have been detected, although not all of them are still active (210). Each clade in phylogenetic analysis is assigned with a letter and clusters within a clade are numbered and represent genotypes. The genotyping system that was adopted in 1998 contains a set of these representative strains against which all new strains should be compared (209). New genotypes can be suggested (211), but can only be accepted after virus isolation and approval by WHO (212, 213). The purpose of this measles genotype system is extensive. It is a means to track genotypes by surveillance and epidemiological investigations, but it also allows identification of sources and transmission pathways. It helps to record the global distribution of measles genotypes and measures the effectiveness of control and elimination programs. A genotyping system that could do the same for RSV would be the most useful.

Human rhinoviruses (HRV) have been classified based on serology, antiviral susceptibility, nucleotide sequence relatedness and receptor usage. In the current system, more than 100 immunologically distinct serotypes have been described. Phylogenetic analysis of VP1 has shown a correlation with these serotypes (214). So far, three major groups of rhinoviruses have been described, namely HRV-A with 74 serotypes, HRV-B with 25 serotypes and the most recent group, HRV-C, which was only discovered in 2007 (215). However, sequence analysis of VP4/VP2 has highlighted some complexities in this virus group. One of the rhinovirus strains, HRV87, was found to belong to another species, human enterovirus D (HEV-D). This species has been shown to carry both rhinovirus and enterovirus characteristics (216, 217). The relatedness between the species was exposed by a study in 2007 that showed that HRV-B and HEV both diverged from HRV-A and share a common ancestor (218). Besides the intertwining of species, extensive intraspecies antigenic variation has been discovered as well. Genotype A21 contains strains that are not neutralised by antiserum of the prototype strain of genotype A21, which might be due to newly discovered coding polymorphisms in viral proteins VP1, VP2 and VP3 (219).

It would be impossible to determine full length genomes from all strains worldwide, so it might help to determine the smallest part of the genome that is necessary for (sub)-genotyping. To investigate which part of the genome that is, the variability of different genes was determined. Obviously, full length genomes carry the most information. However, lots of laboratories just do not have enough funding and possibilities to determine the full-length genomes of all strains that can be collected. The variability of each gene of the RSV genome was determined. It showed that some genes carry a lot of background variability (*e.g.* L and F genes), but G carries the most variability overall and a lot of that variability is extraordinarily high in the HVR2 region. Even though L and F are most useful for confident phylogenetic analysis, they both carry a lot less variation than G, which would make G the best gene to use for genotyping using phylogenetic analysis. On the other hand, F is currently the most targeted gene for vaccines, so perhaps a new genotyping system based on the F gene would be most useful in the current climate. A large disadvantage would be the lack of F gene sequences available in online databases and from older strains at the moment. G has always been the most sequenced part of the genome, hence there are over 4 times more G gene sequences available compared to F gene or full genome sequences. Unless a big change is made and the entire RSV community stands behind the idea of switching from the G gene to the F gene for genotyping, it will not be possible to collect enough information that is needed for transitioning to genotyping based on F. There are both advantages and disadvantages to using F as the region to base sequencing on, so for now, in this chapter, the G gene was still used for genotyping new strains. First, the minimum necessary part for correct genotyping was determined.

The clustering of maximum likelihood trees based on full length genomes was therefore determined and compared to clustering of trees based on full G genes and the original partial G gene sequences. This analysis was complicated because of the lack of full-length genomes of each genotype. The clustering of full G genes is similar to the clustering of full-length genomes, although the bootstrap values are lower and branching was less confident. The full-length genome clustering was also compared to the clustering based on the HVR2 region of the G gene. This showed a very different picture. The ML tree had very low support for most branches and showed different clustering compared to both full length and full G trees. It strengthens the hypothesis that the HVR2 region does not carry enough information for confident genotyping of new strains. It is possible it changes too quickly and too frequently to confidently track the evolution of this part of the genome.

These analyses were performed on a selected dataset of sequences which served as full genome alternatives to the partial sequences that are currently available for RSV. A selection of representatives for each genotype was made and 116 clinical strains were genotyped based on this selection. The genotyping was performed on full G sequences to test if this carried enough information for genotyping unknown strains. This showed that most RSV strains from England in the seasons 2014-2015 to 2017-2018 were part of the GA2/TN1/TN2/ON1/ON2 genotype cluster and GA5 genotype cluster for RSV A strains. For RSV B, all strains were part of the BA genotypes cluster.

ML trees based solely on reference sequences had high support. The confidence in the ML trees with clinical strains was lower, however, this was to be expected as more similar sequences were found in this dataset. The clustering of the strains was quite clear, however, some genotypes seemed to flow over into one another. This supports the theory that sub-genotypes might be more useful to label these strains. However, these selected alternatives are not necessarily good representatives of all genotypes and for some genotypes, there were indeed no good alternatives available at all.

This analysis also poses the same issues that have been pointed out earlier. This was only a selection of reference sequences and this selection determines the outcome of the genotyping efforts. When genotyping clinical strains, ideally all studies should use the same dataset of genotype reference strains and the same part of the genome for genotyping. An internationally recognised, standard reference dataset could be set up which all researchers use for genotyping to overcome this issue. However, not all current genotypes have full length references available. It might not be possible yet to compose these datasets completely and new (sub-)genotypes might evolve, hence this hypothetical dataset should be revised with regular intervals or after big impact discoveries (like ON1). This is a big job that would require multiple research groups to sit together and determine the best strategy for investigation and to streamline and test the process. Currently, there have been working groups set up to do just that.

Many genotypes have been established in the past and it is clear a standardised way of appointing genotypes should be agreed upon, as well as a reference dataset to determine genotypes of clinical samples. It should also be revised regularly, because RSV is a virus that evolves at a rate that would be expected from an RNA virus, although the HVR2 might evolve quicker than the rest of the genome and the evolutionary rate depends on the genotype as well (113, 116, 220, 221). However, with this degree of variability in the virus, a very specific question comes to mind. How and when does the virus accumulate so many mutations? This virus has no other known host than the human population, so it seems reasonable to hypothesize that variations arise during acute infections in one person, which can then be transmitted to another person. If these variations survive the bottleneck of transmission and do not succumb when establishing a new infection, the mutations will stay in the population and contribute to strain variation and possibly even evolve to become a new (sub-)genotype.

To be able to test this hypothesis, determining the origin of new mutations and seeing them arise in clinical samples, a new method had to be set up. Establishing the consensus sequence is no longer good enough, so a method had to be established where all minor variants could be detected as well. The set-up, optimisation and use of this method in different clinical cohorts will be discussed in the following chapters of this thesis.

4. Optimisation of a deep-sequencing method to obtain full length genomes of clinical RSV strains

4.1. Introduction

Methods to determine consensus genomes from RSV exist, however, deep-sequencing is not common practice yet and to determine minority variants, this would be the best way forward. In this chapter, a method is optimised to deep-sequence samples to be able to detect minority variants from clinical samples. To deep-sequence viruses, enough starting material needs to be available to extract, convert to cDNA and amplify to enhance the signal to the background signal of bacterial or host DNA. In acute infections with high viral loads, this can be achieved easily. However, even in severe disease, viral loads are not always high enough and to investigate less severe disease caused by RSV, lower viral loads need to be sequenced as well.

Since the goal is to research in-host variation and look for minority variants, it is also of interest to increase cDNA yield from a sample as much as possible. Therefore, optimisation of cDNA yield was performed by using known test viruses, testing different enzymes, two different sets of primers and different protocols. The optimised protocol was then used in further experiments in the following chapters of this thesis.

4.2. Characterisation of test viruses

First, lab grown RSV aliquots were characterised with diagnostic tests to determine viral load and test the first protocol with high viral titres. Two RSV strains were tested: RSV A Long and RSV B 9320. These are known lab strains and are grown frequently in culture. This test ensured the availability of a large batch of RSV with a known viral load that could be used to test different conditions. Extraction was performed using the EasyMag® instrument as described in 2.2.4.2. A diagnostic qPCR was performed to determine viral titre for all four lab strain aliquots using multiplex PCR as described in 2.2.6.2 after RT conversion as described in 2.2.5.2. This showed the multiplex PCR was able to detected the internal control in all samples and RSV A was detected in all samples of the RSV A dilution series of dilutions up to 1:10^6 (Figure 4.1). No RSV B was detected in RSV A samples. Similarly, RSV B was detected in all samples of the RSV B dilution series up to 1:10^5 (Figure 4.2). No RSV A was detected in RSV B samples. C_T values ranged from 16.82 in undiluted RSV A Long to 37.34 in a 1:10^6 dilution of RSV A Long and 13.70 in undiluted RSV B 9320 aliquots to 32.48 in a 1:10^5 dilution of RSV B 9320. Any further dilutions did not contain enough virus to measure viral loads.

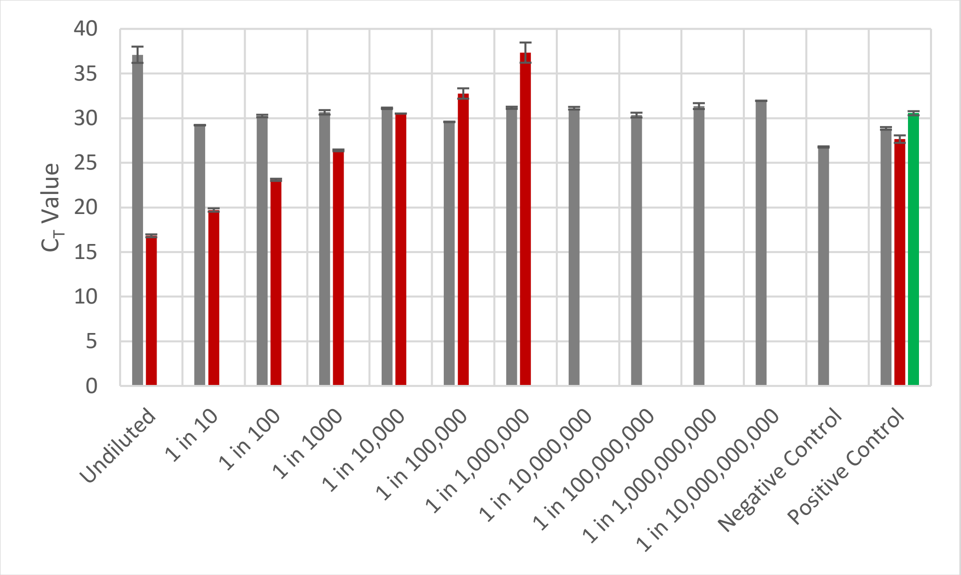


Figure 4.1: Characterisation of RSV A Long C_T value in dilution series with multiplex PCR. Red = RSV A, Green = RSV B, Grey = Internal Control.

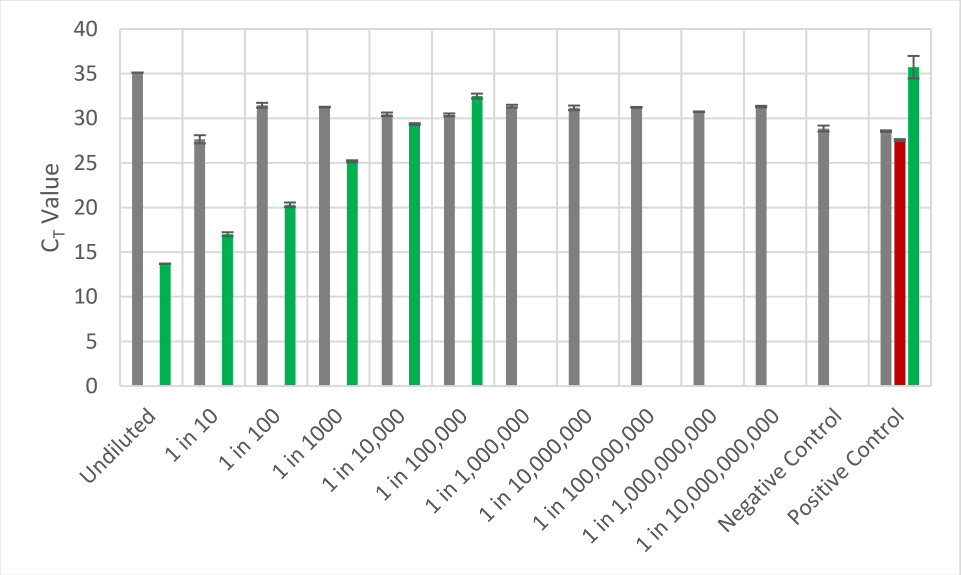


Figure 4.2: Characterisation of RSV B 9320 C_T value in dilution series with multiplex PCR. Red = RSV A, Green = RSV B, Grey = Internal Control.

4.3. Testing different enzymes to increase cDNA yield

The next step was to test the enzymes used in the standard protocol. For all samples, extraction and cDNA conversion were performed as described in 2.2.4.2 and 2.2.6.2 respectively. Next, the amplification PCR was performed using Pfu Ultra II fusion HS DNA polymerase to test the protocol with 16 primer pairs covering the entire RSV genome in 15 fragments. The first fragment was tested by two different primer pairs.

Amplification was performed as described in 2.2.7.1 with RSV A Long samples diluted 1:10 to 1:10^6 and this showed that all bands were visible when the virus was diluted 1:10 to 1:10^3 with C_T values of

19.71, 23.09 and 26.39 respectively, but dilution 1:10^4 with C_T values of 30.5 started showing missing bands on the gel (Figure 4.3). Hardly any bands were visible on the gels from dilutions 1:10^5 and 1:10^6 in which C_T values reached 33.17 and 36.54 respectively. Therefore, we used RSV A Long 1:10^4 dilutions to test which enzyme produced more and/or stronger bands and therefore increased the cDNA yield.

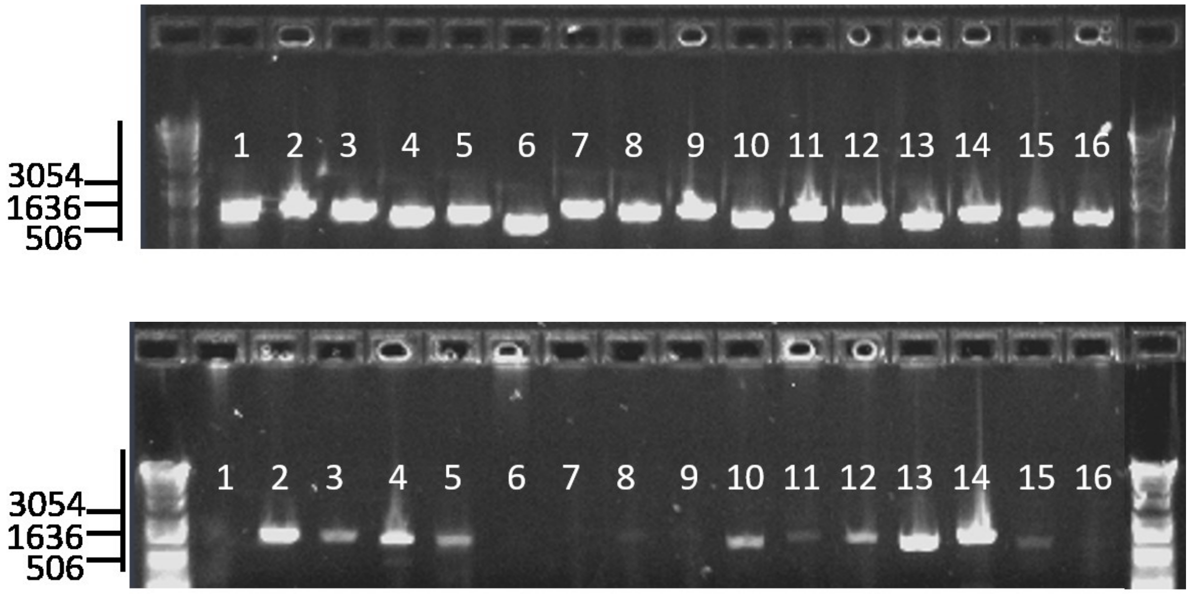
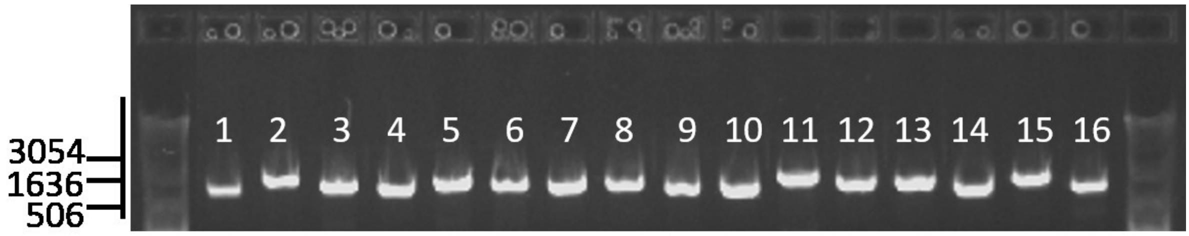


Figure 4.3: RSV A Long sample were diluted in a 1:10 series and amplified with Pfu Ultra II enzyme. Dilution 1:10 (top) showed good viral yield for each amplicon, while some amplicons started showing reduced viral yields from the 1:10^4 dilution onwards (bottom).

A similar test was done with RSV B 9320. In this case, dilution 1:10 to 1:10^4 all produced strong bands, which was in accordance with C_T values of 13.70, 17.01, 20.30 and 25.18 respectively. When amplifying dilution 1:10^5 which had C_T value 29.34, some bands became weaker or were no longer visible. Dilution 1:10^6 had a C_T value of 32.48 and only produced 5 visible bands (Figure 4.4).



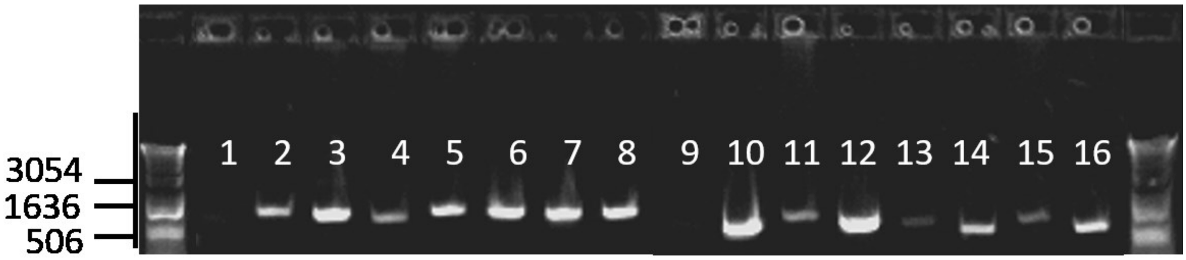


Figure 4.4: RSV B 9320 samples were diluted in a 1:10 series and amplified with Pfu Ultra II enzyme. Dilution 1:10 (top) showed good viral yield for each amplicon, while some amplicons started showing reduced viral yields from the 1:10^4 dilution onwards (bottom).

Two dilutions were selected producing borderline quantities of amplified material, which were 1:10^4 and 1:10^5 for RSV A and RSV B, to test two different protocols. Current use of PfuUltra II Fusion HS DNA polymerase (Pfu) protocol was compared to Platinum^{TM} Taq DNA polymerase High Fidelity (HIFI) protocol. The reagents used only slightly differed from each other as can be seen in Table 4.1. The protocols used for each enzyme differed slightly as well as detailed in Table 4.2.

Table 4.1: Composition of amplification mix for protocol 1 with Pfu and protocol 2 with HIFI.

	Protocol 1	**(µl)**	**Protocol 2**	**(µl)**
Water		37.75		35.8
$MgSO_4$		0		2
Buffer	10X Pfu buffer	5	10X HIFI buffer	5
dNTPs		1.25		1
Enzyme	PfuUltra II Fusion (1U/ml)	1	Platinum Taq HIFI (5U/ml)	0.2
Primer mix		2		3
cDNA		3		3
TOTAL		50		50

Table 4.2: PCR program used for protocol 1 with Pfu and protocol 2 with HIFI.

Protocol 1			**Protocol 2**		
95°C	1 minute		95°C	10 minutes	
95°C	20 seconds	40 cycles	94°C	30 seconds	35 cycles
55°C	20 seconds		55°C	30 seconds	
72°C	45 seconds		68°C	2 minutes	
72°C	5 minutes		68°C	5 minutes	

When comparing each fragment separately, it showed that the Pfu enzyme, buffer and protocol combination returned more and brighter bands than the HIFI enzyme, buffer and protocol combination. The Pfu combination returned 13 bands from 16 fragments for RSV A Long, while HIFI only resulted in 9 bands. For RSV B, a similar result was seen with 16 out of 16 bands with Pfu and 7 bands with HIFI. It was decided to use the Pfu enzyme, buffer and protocol for further optimisation of the protocol (Figure 4.5).

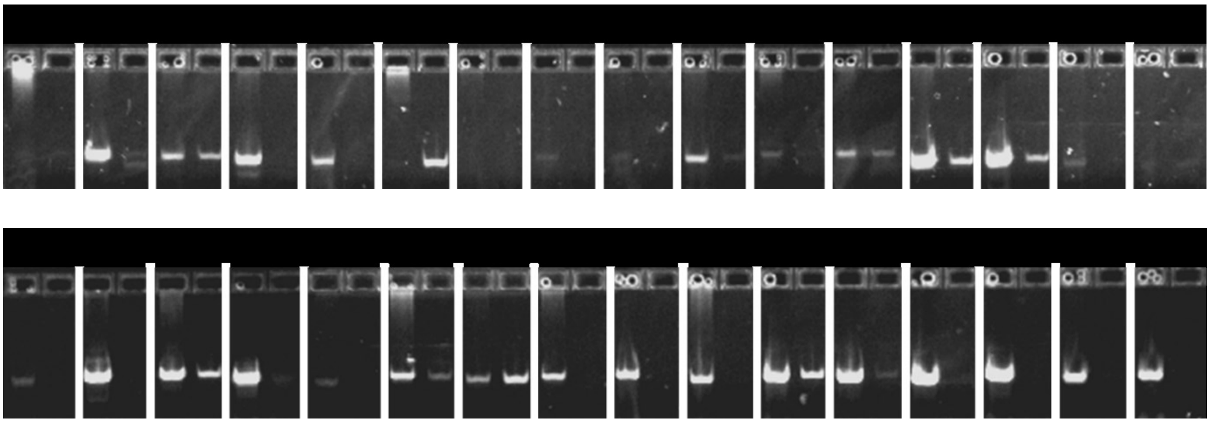


Figure 4.5: Comparison of viral yield after the amplification reaction for each fragment (F1 to F15 and $F1_{alt}$) using PfuUltra II Fusion HS DNA polymerase (left) versus PlatinumTM Taq DNA polymerase High Fidelity (right) reagents and protocol. RSV A Long (top) and RSV B 9320 (bottom) were tested.

4.4. Testing different volumes in the protocol to increase cDNA yield

Next, the volumes of sample in the different steps were tested. The idea was that by eluting nucleic acids in 25 µl instead of 100 µl, the nucleic acids are more concentrated and by using more eluted RNA for reverse transcriptase, more RNA will be converted to cDNA and using more cDNA for amplification will give more amplified PCR product. This was tested according to Table 4.3. The combination of elution of nucleic acids in 25 µl and using 42 µl was not practically possible though and was not tested as generally not enough volume would be available to perform this protocol on clinical samples.

Table 4.3: Different protocols were tested based on different volumes in each step. The best protocol for high cDNA yield is highlighted in yellow. B was the standard protocol.

Protocol	Elute volume	µl of RNA used	cDNA used for aPCR
A	100	15	3
B	100	15	10
C	100	42	3
D	100	42	10
E	25	15	3
F	25	15	10

This resulted in 6 different possibilities for each fragment and these were compared to find the highest cDNA yielding protocol. Intuitively, it may seem that eluting in 25 µl and using 42 µl of RNA and using

10 µl of cDNA will give the best results, however, primer and enzymes are optimised for set quantities of RNA/cDNA. When the quantity is exceeded, this does not necessarily give better results.

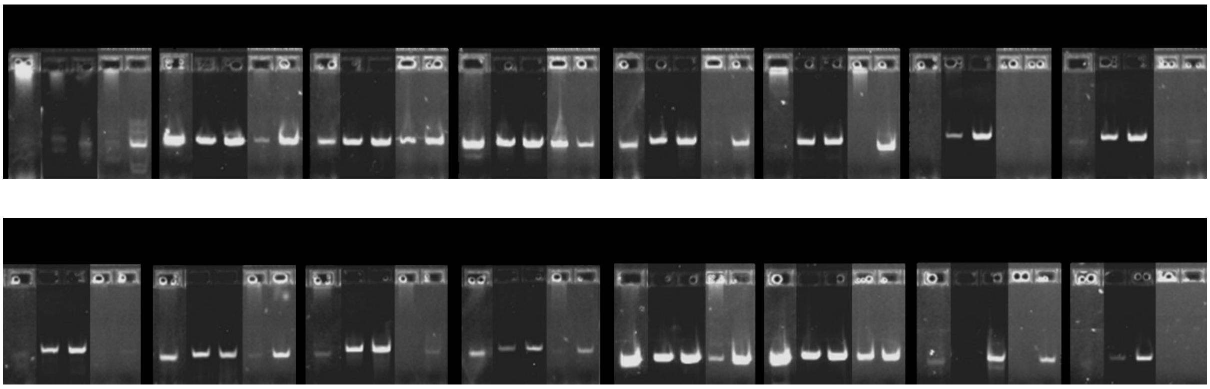


Figure 4.6: Protocols A, C, D, E and F (left to right) were run for all 16 fragments F of the RSV A Long strain and are shown here in lanes side by side per fragment.

The best results seemed to be produced by protocol C (Figure 4.6). This involved eluting nucleic acids in 100 µl and using 42 µl of that RNA for reverse transcriptase. Then, 3 µl of cDNA was used for amplification. This was more practical for clinical samples with low volumes and the increased amount of cDNA did not improve cDNA yield enough to justify the extra enzyme use.

4.5. Testing different primer sets

The previous protocols were all tested by amplifying the RSV genome in 16 separate reactions to produce 15 fragments (and fragment 1 twice). Another set of primers was tested where 2 or 3 forward and reverse primers were used per fragment so that at least one forward and reverse primer would bind the genome, even when variation was seen in the binding site of one of the primers. This was done in 6 separate reactions that produced 6 fragments that covered the entire genome as shown in Figure 4.7. These primers were based on the publication of complete RSV sequencing by Agoti *et al*. in 2015 (222) as described in 2.2.7.2.

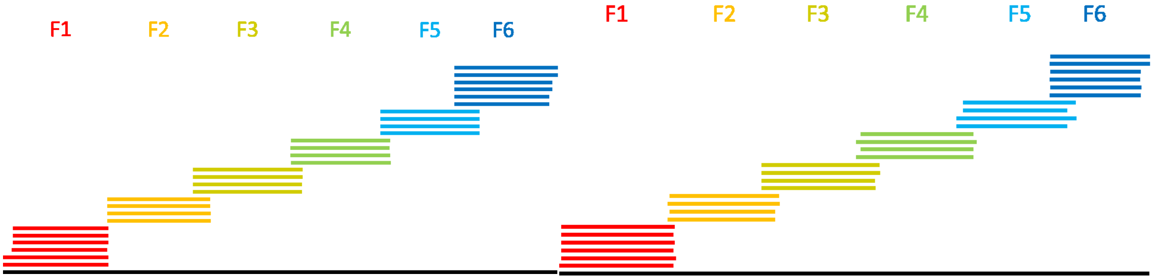


Figure 4.7: Diagram of primer sets used for 6 reaction covering the entire RSV A (left) and B (right) genome. Multiple primers are used at the start and end of each fragment to reduce bias by variation in primer binding site.

The protocol used for these primers was not entirely clear from the publication, so the first try was based on the manufacturer's recommendations for Superscript III RT enzyme and is shown in Table

4.4 for reverse transcriptase. This protocol used forward primers in the reverse transcription process as described in the publication rather than random primers that were used for reverse transcriptase in the test conditions from the previous chapter.

Table 4.4: Reverse transcriptase using 6 fragment primers for targeted conversion of viral RNA to cDNA in the second set of testing conditions.

Reagents RT mix	**Volume (µl)**
5X First Strand Buffer	4.0
0.1 M DTT	2.0
10 mM dNTPs mix	2.0
1 µM Primer forward	1.0
RNAsin 40 U/ul	1.0
Superscript III RT 200 U/ul	1.0
RNA	4.0
RNase-free water	
Final volume	20

The program for reverse transcriptase was 50ºC for 60 minutes and 70ºC for 15 minutes to end the reverse transcription. The amplification PCR was then performed with the reagents mix as described in Table 4.5 for each of the 6 fragments separately.

Table 4.5: Amplification PCR using 6 reactions with each a set of primers specific for one fragment to cover the RSV genome in the second set of testing conditions.

Reagents aPCR mix	**Volume (µl)**
10X High Fidelity PCR buffer	5.0
50mM $MgSO_4$	2.0
10 mM dNTPs mix	1.0
Primer mix	4.0
Platinum® Taq DNA polymerase (5U/µl)	0.2
cDNA	5.0
RNase-free water	32.8
Final volume	50

The program for amplification consisted of a first step at 98ºC for 30 seconds, then 40 cycles of 98ºC for 10 seconds, 53ºC for 30 seconds and 72ºC for 3 minutes and the last step was at 72ºC for 10 minutes. However, Figure 4.8 shows that bands were not (clearly) visible on E-gels with this protocol and adapted testing conditions had to be designed to improve the results.

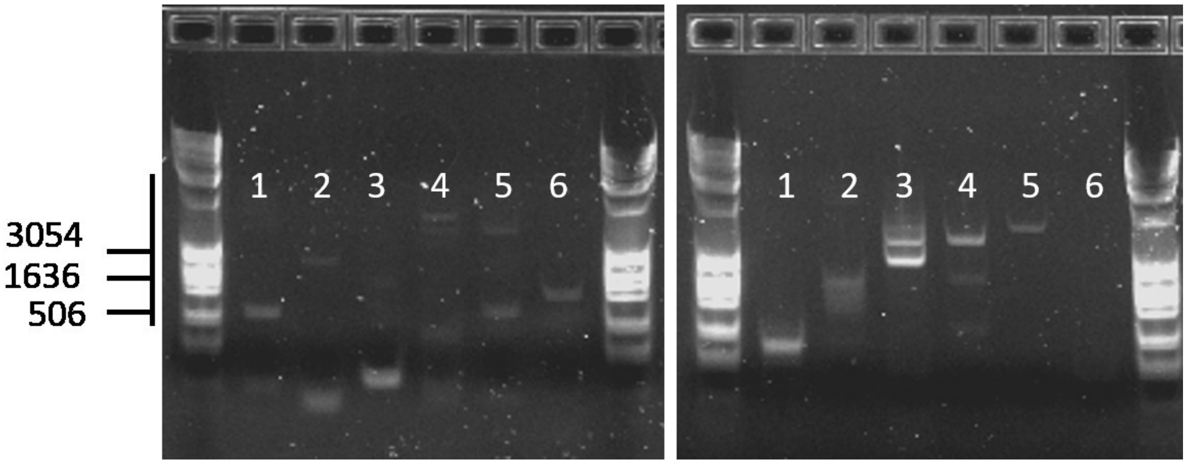


Figure 4.8: E-gels show little replication of the desired fragments for RSV A Long (left) and RSV B 9320 (right) with the first set of testing conditions using 6 sets of primers as described above.

The next set of conditions, were based on an in-house one-step protocol used for Influenza A and B virus amplification. This protocol combined the reverse transcription step with amplification PCR using the SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity DNA Polymerase kit. The reaction mix is detailed in Table 4.6 and the program ran on a Thermal cycler, which is detailed in Table 4.7.

Table 4.6: Reagent mix for reverse transcription and amplification of RSV genome in 6 reactions with fragment specific primers.

Reagents One-Step reaction mix	**Volume (µl) second test**
2X reaction mix	25
Primer mix	4
Enzyme mix (SuperScript III + TaqHIFI)	1
RNA	10
RNase-free water	10
Total	50

Table 4.7: Program run for one-step reaction PCR.

Temperature	**Time**	
42°C	60 minutes	Reverse transcription
94°C	2 minutes	
94°C	30 seconds	5 cycles
44°C	30 seconds	
68°C	3 minutes	
94°C	30 seconds	31 cycles
57°C	30 seconds	
68°C	3 minutes	
68°C	5 minutes	

PCR fragments were run on agarose gels, to show clearer and more detailed bands of cDNA produced by these methods. This set of conditions showed more promising results, although fragments 3 and 5 of RSV A and fragments 2 and 3 of RSV B were replicated less abundantly (Figure 4.9).

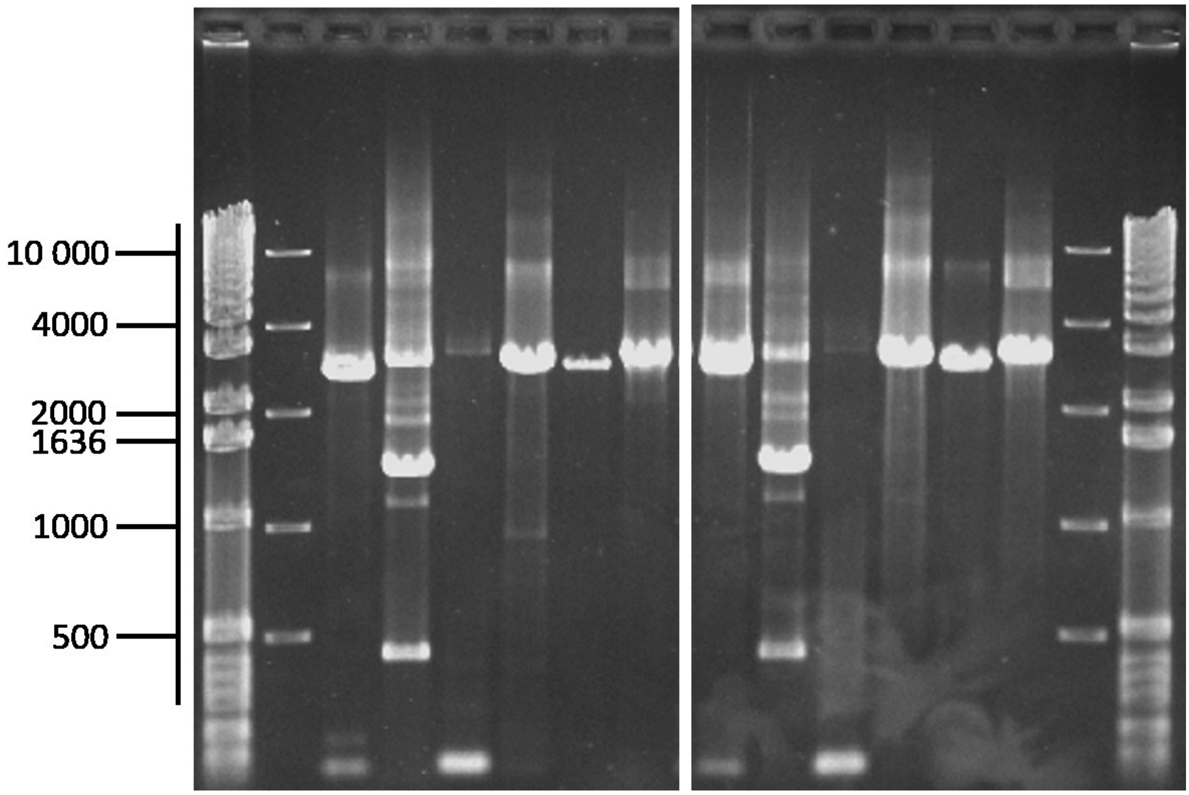


Figure 4.9: Agarose gels of one-step protocol as described in Table 4.7 with the reagents mixture as described in Table 4.6 for RSV A Long (left) and RSV B 9320 (right). The first two lanes contain ladders (1kb DNA ladder by Invitrogen and FastRuler High Range DNA Ladder by Thermo Fisher Scientific), the following 6 lanes contain 6 different RSV A fragments. On the next gel, 6 lanes show 6 different RSV B fragments and 2 ladders at the end of the gel.

However, a 1:10 dilution of RSV A Long and RSV B 9320 showed no visible bands whatsoever with the same protocol (Figure 4.10). Further dilutions did not show any bands either.

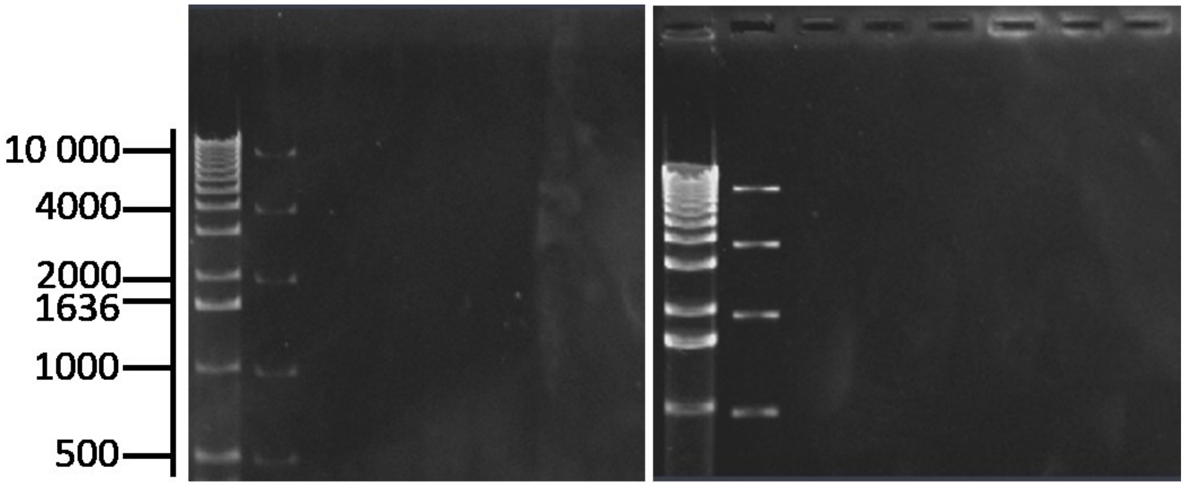


Figure 4.10: Agarose gels of one-step protocol for RSV A Long (left) and RSV B 9320 (right) after 1:10 dilution compared to previous agarose gel in Figure 4.9 shows no visible bands. The first two lanes contain ladders (1kb DNA ladder by Invitrogen and FastRuler High Range DNA Ladder by Thermo Fisher Scientific), the following 6 lanes contain 6 different RSV fragments.

The last attempt to optimising the conditions, was performing a gradient PCR. RSV A Long's fragment 5 was selected to be tested on a gradient PCR in undiluted form and 1:10 dilution. This band was clearly visible in the previous testing conditions, but was not the most abundant fragment. The gradient would run from 45ºC to 60ºC in step 2 of the 31 cycles of the PCR. Other conditions were kept identical.

The gradient PCR showed that a lower temperature would be beneficial to the reaction, but in the 1:10 dilution of RSV A Long (with a C_T value of 20), this still showed a rather weak band (Figure 4.11). Therefore, this method was deemed unsuitable for clinical samples which reach far lower viral loads than the ones tested here.

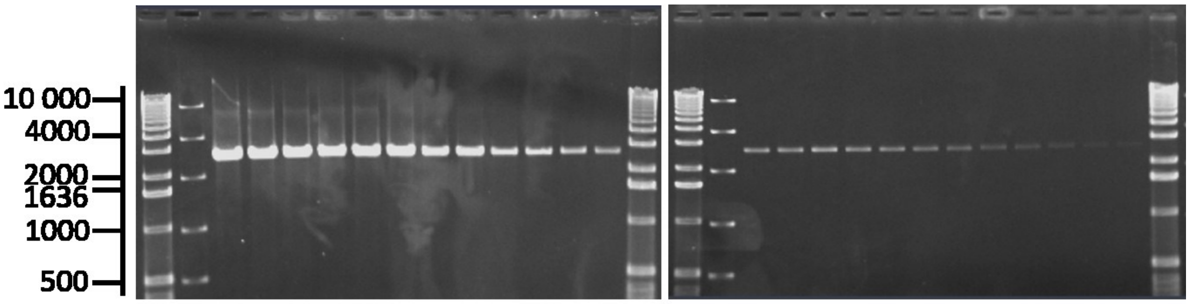


Figure 4.11: Agarose gel of RSV A Long fragment 5 was tested in a gradient PCR spanning 45ºC to 60ºC in 12 steps with an undiluted sample (left) and a 1:10 diluted sample (right). The first two lanes contain ladders (1kb DNA ladder by Invitrogen and FastRuler High Range DNA Ladder by Thermo Fisher Scientific), the following lanes contain the RSV fragment.

4.6. Discussion

There are several factors that can influence the use and range of an experiment. Each step should be optimised to end up with the best possible protocol. However, practicalities should not be left out of consideration and a perfect protocol is not of any use if it cannot be performed on clinical samples.

In this subchapter, several testing conditions were evaluated: total nucleic acid extraction, enzymes for reverse transcriptase and volumes, and primers and primer pairs. Other papers from experts in the field were used as inspiration for protocol optimisation (222).

A standard protocol was adapted for the specific purpose of increasing cDNA yield to be able to identify minor variants. The following protocol was derived from the experiments: elute nucleic acids in 100 µl, use 42 µl of those nucleic acids for reverse transcriptase with PfuUltra II Fusion HS DNA polymerase and buffer, following the complementary protocol instructions. Then, use 3 µl of cDNA for 16 separate PCR reactions, each with their own primer pair. This resulted in the best practical protocol for the purpose of deep-sequencing clinical samples and looking for minor variants in these samples.

With more funding, further optimisation would have been done by re-testing all primer pairs or by using elements from the two primer strategies together. Combining some of the primers from the set of 16 with new primers to amplify larger parts of the genome (as was done in the protocol based on the paper by Agoti *et al.* (222)) could have resulted in good outcomes too.

This new, optimised protocol made it possible to try and deep-sequence clinical samples. This protocol was used in combination with Illumina MiSeq standard protocols in the following chapters of this thesis.

5. Optimisation of cell culture experiments to grow RSV from clinical samples

5.1. Introduction

In previous chapters, the variability of each RSV gene and between strains was investigated. RSV infection can cause severe disease, although asymptomatic infections occur as well. Cell culture experiments were set up to link the genetic information to cytopathology. Clinical samples are valuable and scarce, so experiments were optimised using viral laboratory strains and then with clinical-like samples.

The aim of the first experiment was to learn to work with cell culture, recognise RSV cytopathology and master viral seeding and harvest. The experiment was set up with a strain used for human challenge studies with RSV, *i.e.* RSV Memphis-37 (M37).

In this experiment, the best time for harvesting RSV from cell culture was investigated. Microscopy was used to examine the cytopathic effect (CPE) and photos were taken to document these findings. Viral titres were verified by conducting plaque assays of samples at all different time points and viral RNA loads were determined by performing qPCR. The viral aliquots that were characterised with these methods were used for further optimisation experiments.

In the second experiment, more clinically relevant samples were produced to test if it was possible to grow RSV from clinical samples, which contain much more than just virus. To know how much virus had to be present to be able to successfully infect cell culture, “clean” nasal lavages from an uninfected donor were spiked with a known concentration of virus. The RSV M37 that was characterised in the previous experiment was used here to spike the nasal lavage aliquots. Microscopy was used to document CPE and to determine the presence of bacterial infection. Uncontaminated virus was harvested and viral titre determined by plaque assay and qPCR.

This chapter will discuss the optimisation and results of these experiments.

5.2. Optimisation of culture and characterisation of RSV M37

The amount of virus needed to infect cell culture and successfully replicate the virus is well established. The lab strain chosen for these experiments was RSV M37, a strain used for human challenge studies and therefore well characterised. The time of harvest is one of the factors that differs between laboratories; a Standard Operating Protocol (SOP) previously set up in our lab states that virus should be harvested when the cytopathic effect is 50%. To test this SOP, an experiment with different harvesting times was set up and CPE was documented through microscopy photos to compare CPE at different times. Viral titres were then verified with plaque assays to find the best time to harvest RSV and attain the highest yield.

5.2.1. Temporal optimisation of RSV harvest from cell culture

Seven T175 flasks were seeded with 10 million Human Epithelial type 2 (HEp2, ATCC) cells. Once they were 80% confluent, one of the flasks was mock infected with cell medium, while the other six flasks were infected with RSV M37 from our virus stock at a multiplicity of infection (MOI) of 0.1. After 24 hours, the cytopathic effect was checked twice daily. At 31, 55, 72, 74, 76, 78 and 80.5 hours post infection (hpi), the CPE was recorded by microscopy photos and at all time points except 31 hpi, one flask was harvested as described in 2.2.2. Cell medium was not changed during this time. The harvested virus was snap frozen and kept in liquid nitrogen in 1 ml aliquots. The harvesting process was performed in the same way until all flasks were harvested and ready for further processing.

At 31 hpi, the first syncytia and dying cells were visible under the microscope in the infected flasks, while the mock infected flask only showed cells growing to 100% confluence. At 55 hpi, syncytia were bigger and more dying cells were detected. These look like white, round cells that are floating on the surface (see arrows on Figure 5.1). The mock infected flasks had grown to 100% confluence and some cells started to die as well, although not as abundantly as seen in infected flasks.

At this point, a first flask was harvested and virus was snap frozen and stored in liquid nitrogen (Figure 5.2). The next day, at 72 hours post infection, the next flask was assessed and harvested. By this point, the mock infected flask started showing more dying cells and the infected flask already showed 40-50% cell death with gaps in the cell layer. Cells were either part of syncytia or dying and filaments connecting several syncytia were seen as well (see circled in Figure 5.1).

These steps were repeated every 2 hours from then onwards, assessing and harvesting flasks at 74 hours, 76 hours, 78 hours and 80.5 hours post infection (Figure 5.2). In the mock infected flask, increasing numbers of dying cells were seen and the number of dead cells in the infected flasks increased as well. Syncytia broke up and became less prevalent and medium became acidic, as shown by the media changing colour from red to orange and yellow in both infected and mock infected flasks.

[redacted]

Figure 5.1: Timeline of photos and harvest of infected flasks. Square = photo, star = harvest of one flask.

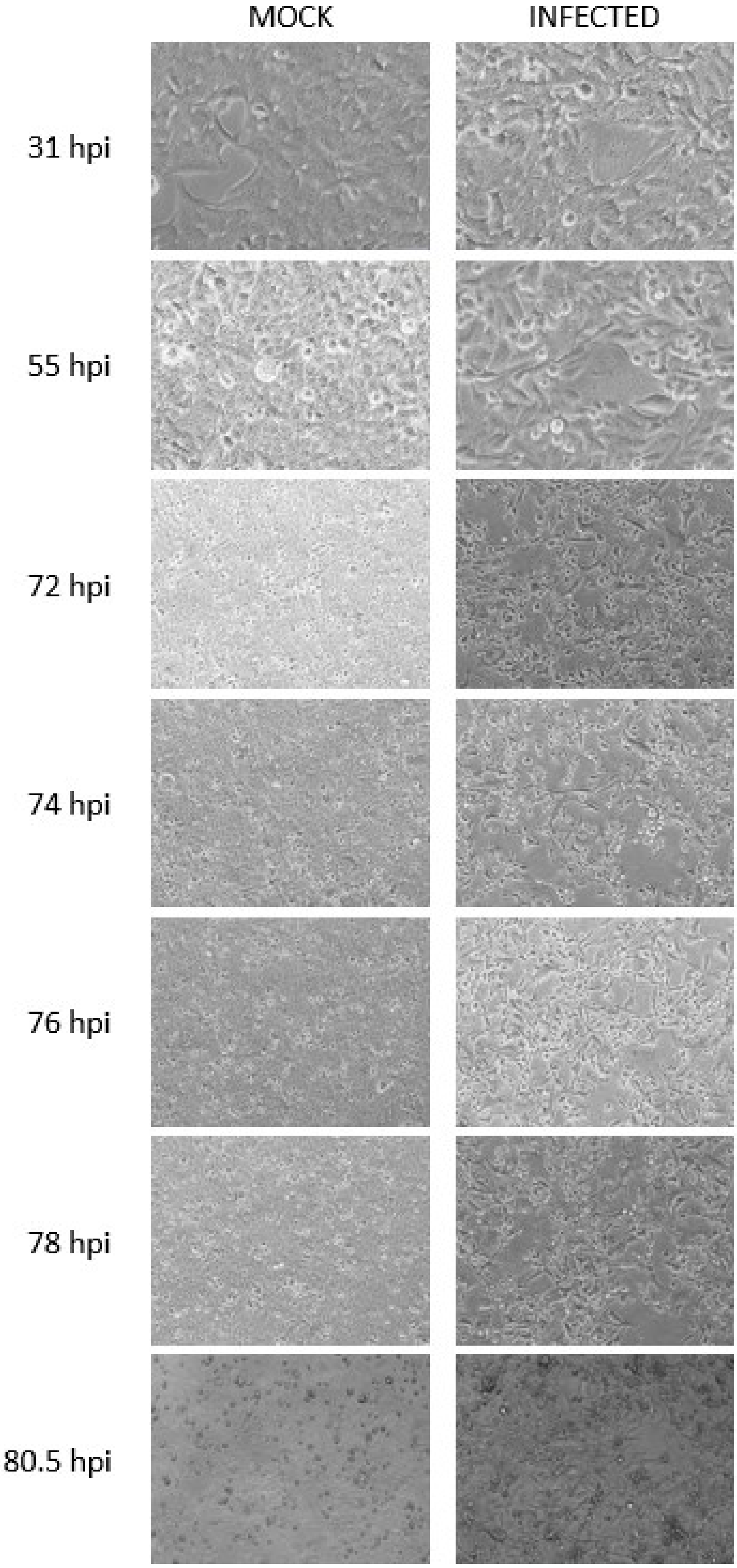


Figure 5.2: Microscopy photos of HEp2 cells infected with RSV M37 and mock infected at different time points. Dead cells looked like white spots and were floating in the medium (circled in black).

5.2.2. Characterisation of RSV M37 samples grown in cell culture

5.2.2.1. Characterisation via plaque assays

The next step was to quantify the amount of viable virus and characterise the harvested virus further. This was managed by conducting plaque assays as described in 2.2.3 followed by qPCR to determine the amount of viral RNA in the sample. If many defective particles were produced during cell culture, the viral load determined by qPCR might be high, while a plaque assay might indicate less infectious virus than expected. Plaque assays were performed on aliquots from all different time points, while qPCR was carried out on the sample with the highest plaque-forming unit (PFU) as this was considered to be the sample harvested at the best time point and used for further experiments.

First, the amount of infectious virus harvested at each time point was determined by plaque assays. HEp2 cells were seeded in 96 well plates and aliquots from each time point were thawed. The virus was titrated out and a starting dilution of 1 in 10 was prepared in Phosphate Buffered Saline (PBS). From there, 1 in 2 dilution series were set up until 12 dilutions were prepared from each aliquot. 50 μl of virus was added to the wells in duplicates and incubated. After 24 hours, the cells were fixed and stained. Plaques were manually counted for each time point and in three different wells per row: one well where the dilution showed about 50 plaques and one well with a lower and one well with a higher dilution. The PFUs were calculated using the average of the three wells per time point in Equation 5.1.

$$\mathrm{PFU} = \text{ number of plaques} * \text{ dilution} * \text{ } 20/\mathrm{ml}$$

Equation 5.1: Equation to calculate the number of plaque forming units per ml of viral sample.

PFUs of all six time points were determined. The average PFUs ranged from 3.1E6 at the first time point to 8.3E5 at the last time point (Table 5.1). For each aliquot, the duplicate rows showed PFUs in the same order of magnitude. The results showed that the viral titre was a lot higher when harvested at the time that infected cells were showing CPE, but did not die yet (Figure 5.3). Once cells had died and detached, the number of viable viral particles decreased. The ideal time of harvest seems to be the moment right before cells start to die and detach.

Table 5.1: Viral load of samples at different times of harvest in PFU/ml. Equation 5.1 was used to calculated the PFU of the first and second dilution series. The average was calculated from these two rows.

	First dilution	**Second dilution**	**Average**

55 hpi	2.9E6	3.3E6	3.1E6
72 hpi	1.3E6	1.1E6	1.2E6
74 hpi	1E6	8.7E5	9.3E5
76 hpi	1.1E6	1.3E6	1.2E6
78 hpi	1E6	6.7E5	8.6E5
80.5 hpi	8.9E5	7.7E5	8.3E5

5.2.2.2. Statistical analysis on effects on viral yield of sonication during virus harvest

There were two samples that were not sonicated during harvest and these seemed to show slightly lower viral loads than expected. There are not a lot of statistical tests that can be done on such little data, but a prediction model based on the sonicated sample data could show whether the data points from unsonicated samples fell within the predicted range. Therefore, a linear regression model was calculated with R to further investigate that hypothesis. However, it is unsure that viral load follows a linear decline over time after cells start to die rather than an exponential decline as most studies focus on viral growth or decline after admission of interfering substances. Therefore, this analysis may not be correct and is just to explore the idea that sonication does improve viral yield.

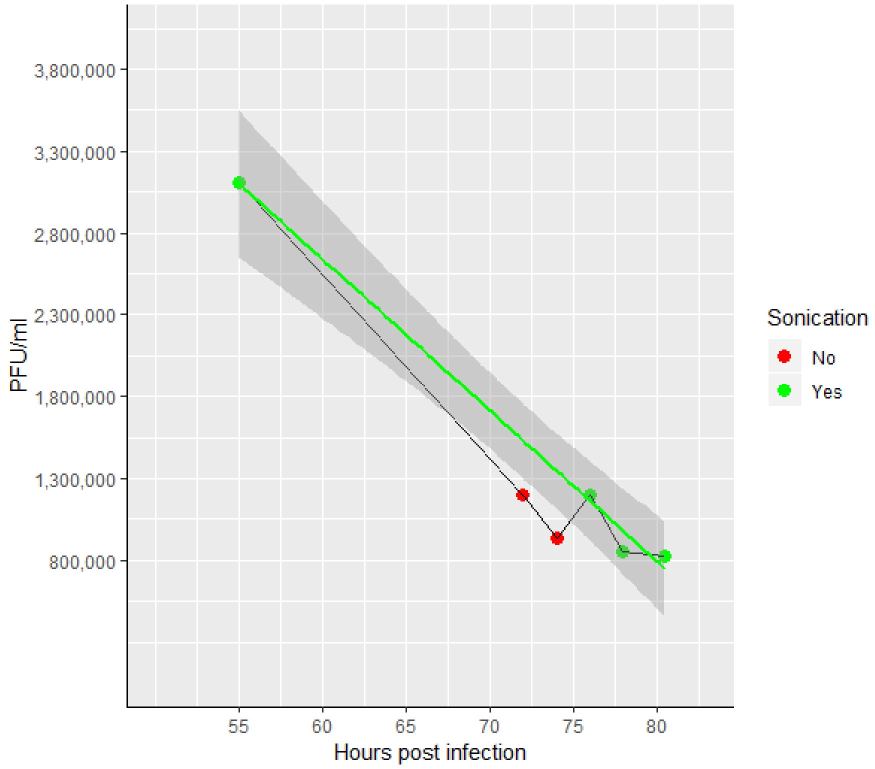


Figure 5.3: RSV M37 viral load from HEp2 cells harvested at different time points with a linear regression line in green calculated from sonicated samples only and 95% confidence intervals in grey. Green points = sonicated during harvest, red points = not sonicated during harvest.

The linear regression model was calculated using R and using data from sonicated samples only (Figure 5.3). The estimated coefficient showed that the viral load dropped with 92,437 ± 5,206 PFUs with every hour after 55 hours post infection. The p-value was 0.00316 which meant the null hypothesis (no change in viral load over time) could be rejected and the chance of this correlation being coincidental was less than 0.316%.

Since this was a prediction model, the quality should be checked. The residual standard error showed that this linear regression might be deviating 105,800 PFUs from the true fit. The multiple R^2 is 0.9937 indicating that this model fit the experimental data very well and the F-statistic confirms this with a value of 315.3 (Table 5.2). The points from the unsonicated samples fell just outside of the 95% confidence interval (Figure 5.3). This suggests sonication might increase viral yield during harvest.

When comparing this data to a linear regression model including all data, it was noted that the estimated coefficients were similar and indicated a significant correlation, but there was a tighter fit when using only sonicated data as seen from the higher residual standard error and lower R^2 data and F-statistic (Table 5.2).

Table 5.2: Summary table of linear regression calculation in R of sonicated samples compared to all samples taken together.

	Sonicated samples	**All samples**
R formula	lm(Average ~ HPI, SonT)	lm(Average ~ HPI, CellCul)
Intercept	8,188,819.42	8,133,664.75
Slope	-92,437.35	-93,386.27
n	2	2
Slope of Standard Error	5,205.74	11,247.87
R2	0.99	0.95
Adjusted R2	0.99	0.93
F-statistic	315.30	68.93
p-value	0.003	0.001

5.2.2.3. Characterisation via qPCR

Once the PFU was determined for all time points, the sample with the highest PFU was chosen to investigate the amount of viral RNA via qPCR. An aliquot was thawed and complete nucleic acid extraction was performed as described in 2.2.4.1. Immediately after, cDNA conversion was performed as described in 2.2.5.1. The amount of cDNA was quantified using qPCR with primers recognising a sequence of 70 base pairs in the beginning of the N gene and a FAM- and TAMRA-tagged RSV pan-A probe (see 2.2.6.1.).

10-fold dilutions were set up starting at 1 in 10 and going to 1 in 1,000,000 to determine the viral load (in copies/ml) and C_T value. All samples for qPCR were tested in duplicate to ensure correct quantification of viral cDNA for this virus stock that was used in later experiments.

The standard used for this qPCR was a series of 10-fold dilutions set up specifically for this purpose and the quantities ranged from 1,000,000/ml to < 1/ml. This standard curve was plotted on a graph comparing the cDNA quantity and C_T values and the sample dilution series was plotted on this same graph to check if the qPCR worked well and if the dilution series was set up correctly (Figure 5.4). The graph showed a linear relation for both the standard series and sample dilution series confirming that the qPCR was successful.

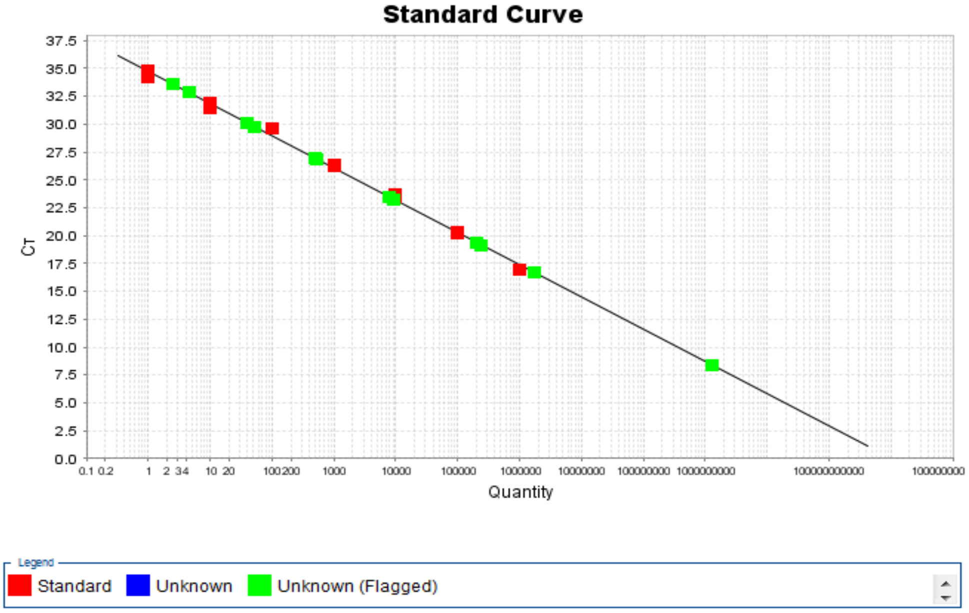


Figure 5.4: Samples align along the standard curve indicating a successful dilution series. Red = standard, green = samples.

Figure 5.4: Samples align along the standard curve. Red = standard, green = samples.

The viral cDNA load in the 1 in 10 dilution duplicates is 1,303,490 and 1,454,671. This corresponded to C_T values of 18.1 and 18.0 respectively. Lower dilutions were all with similar orders of magnitude for each duplicate and the lowest dilution, 1 in 1 million, had quantities of 4.79 and 2.77 with corresponding C_T values of 33.9 and 34.6 respectively (Table 5.3).

Table 5.3: qPCR results of 10-fold dilution series of RSV M37 grown in cell culture.

Well	Dilution	C_T value	Quantity (/ml)
A1	1:10	18.11	1303490
A2	1:10	17.97	1454671
A11	Standard	18.06	1000000
A12	Standard	18.06	1000000
B1	1:100	20.84	150688

B2	1:100	20.48	199352
B11	Standard	21.28	100000
B12	Standard	21.30	100000
C1	1:1000	24.51	8209
C2	1:1000	24.35	9370
C11	Standard	24.44	10000
C12	Standard	24.91	10000
D1	1:10,000	27.94	544
D2	1:10,000	27.83	593
D11	Standard	27.23	1000
D12	Standard	27.39	1000
E1	1:100,000	31.19	42
E2	1:100,000	30.93	51
E11	Standard	30.96	100
E12	Standard		
F1	1:1,000,000	33.92	5
F2	1:1,000,000	34.61	3
F11	Standard	32.40	10
F12	Standard	33.05	10
G11	Standard	35.36	1
G12	Standard	35.90	1
H11	Standard	Undetermined	
H12	Standard	Undetermined	

Another check of duplicate results was performed by drawing an amplification plot of the qPCR. This graph plotted the number of cycles in the PCR and the ΔRn value, which is the difference in fluorescent signal between the reaction signal and background signal. Figure 5.5 shows that all duplicated samples showed the same threshold detection level of fluorescence in the same PCR cycle. This indicated that the qPCR ran successfully.

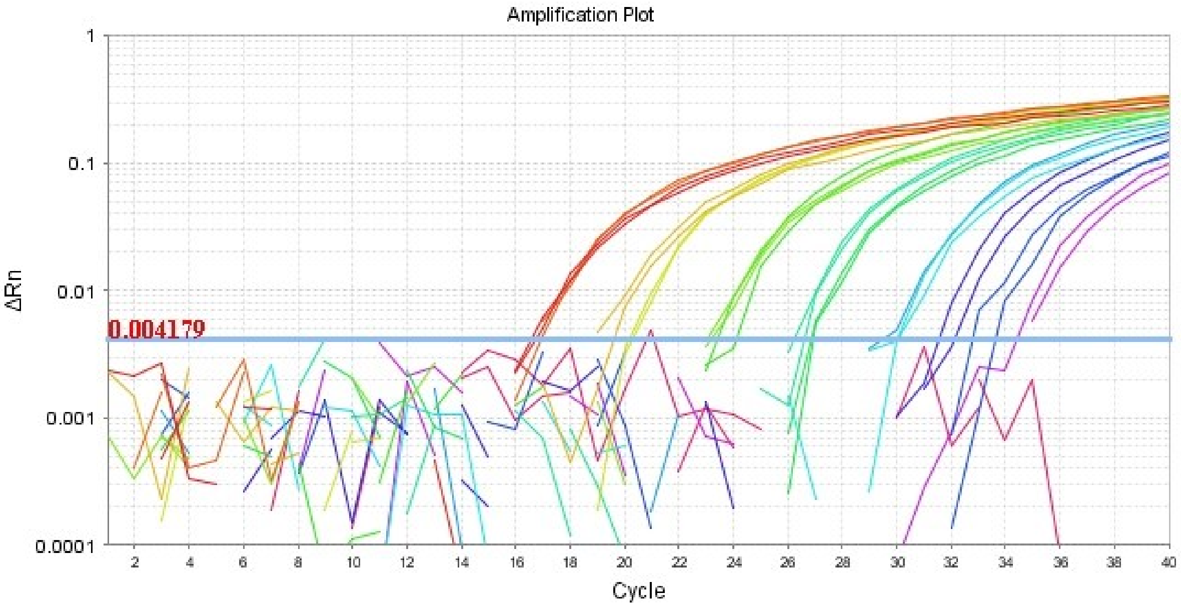


Figure 5.5: Amplification plot of viral stock showing amount of fluorescence detected per PCR cycle of duplicates of sample dilution series.

5.3. Optimisation of viral culture from clinical samples

In this chapter I have described the way that viral stock was grown in cell culture and characterised by plaque assays and qPCR. The virus collected in this way was used for comparison with clinical samples. The lowest C_T value encountered in children in hospital with bronchiolitis at ICCRU was 18 and the data from the qPCR showed the viral load ranged from a C_T of 17.97 in a 1 in 10 dilution to a C_T of 34.6 in a 1 in 1 million dilution. This covered the range of viral load that were seen in clinical samples and which could be sequenced easily.

C_T values higher than 30 were seen as well, but were too high to sequence. The range that was tested with qPCR was ideal to try and test if they could be grown *in vitro*. Clinical samples contain material other than virus, but using precious clinical samples for optimisation was not an option, so samples were constructed to be a clinically relevant as possible without wasting actual clinical samples.

The next experiment was therefore to investigate the minimum viral load necessary for growing clinical samples in cell culture. To make sure the viral load put in culture was known, 900 µl “clean” nasal lavage from a healthy donor was used and this was spiked with 100 µl of virus of a known quantity for each sample. A 10-fold dilution series set up starting from 1:10, which was done by 100 µl of virus stock in 900 µl of “clean” nasal lavage. Six more 1 in 10 dilutions were set up until a range of 1 in 10 to 1 in 10,000,000 was reached. This was similar to an infection of MOI 0.05 to 0.00000005. Photos were taken with a microscope to document CPE and virus was harvested when the cytopathic effect was greatest, but cell death had not taken over yet.

Photos in Figure 5.6 showed big syncytia in flasks with 1:10 to 1:1000 dilutions and smaller syncytia in 1:10,000 dilution. The flasks with dilutions of 1:100,000 to 1:1,000,000 showed some attempt at syncytium formation, but these were rare and there was no wide spread CPE in these flasks.

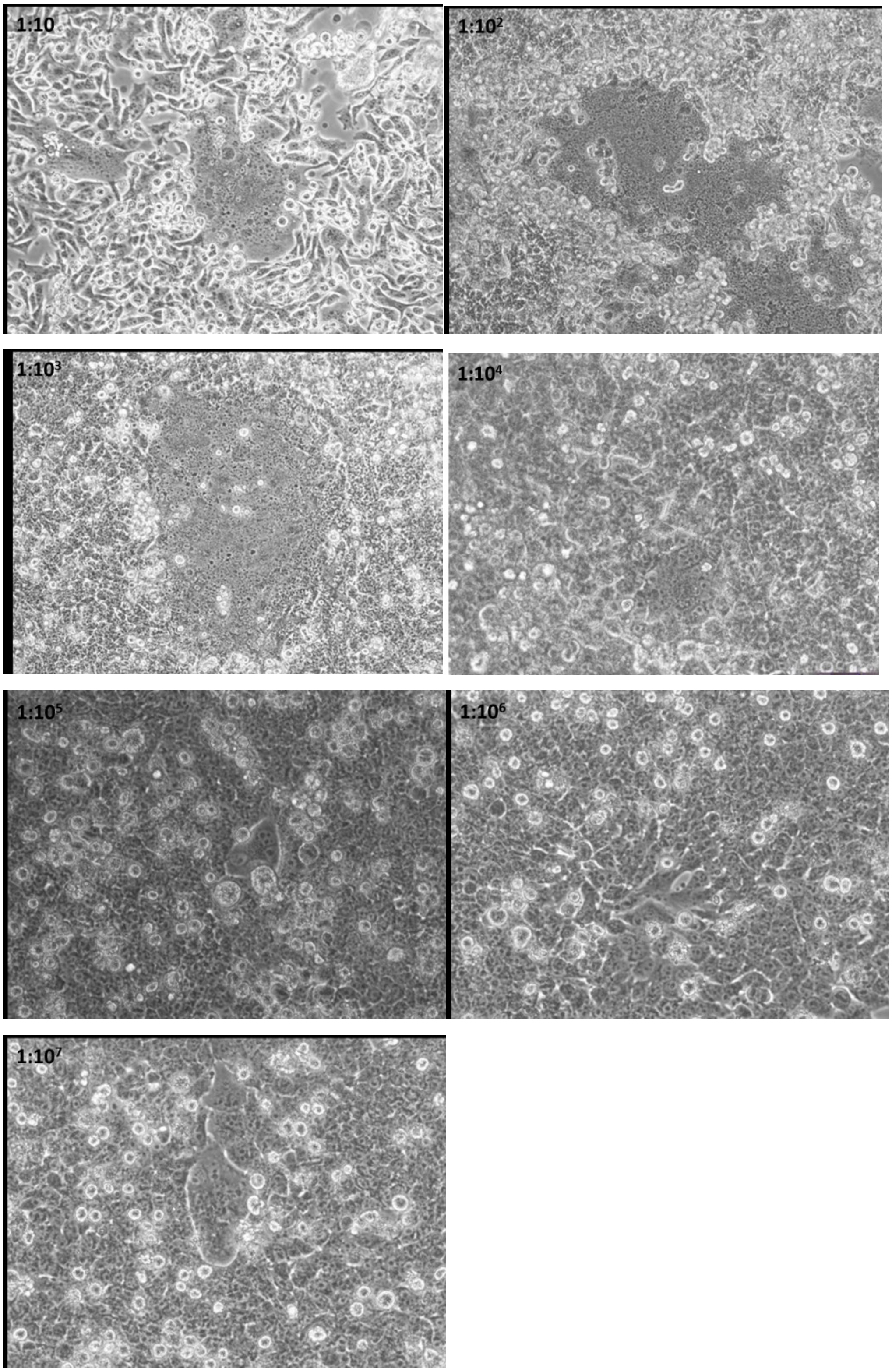


Figure 5.6: Microscopy photos of HEp2 viral cultures of clinical-like samples in 10-fold dilutions.

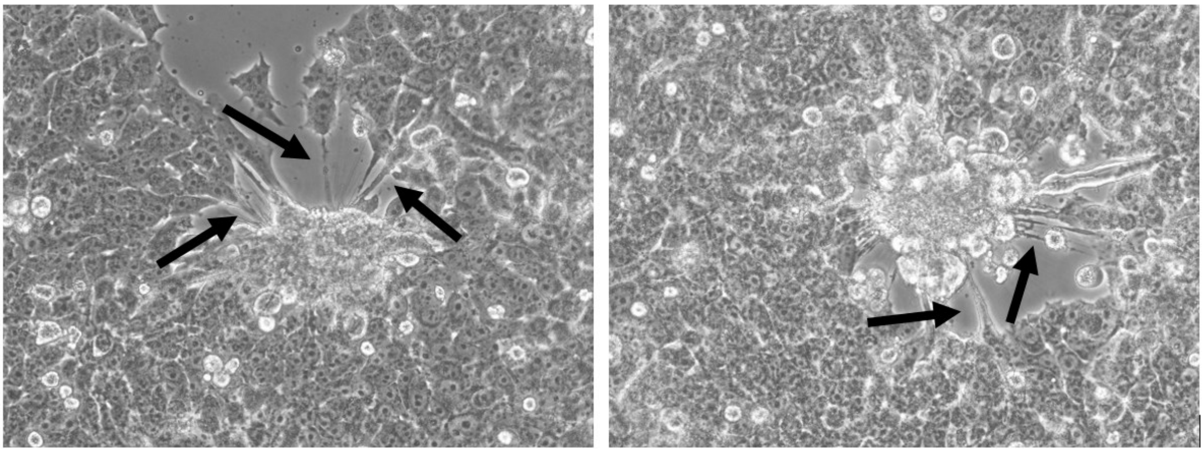


Figure 5.7: Big syncytia show threads when surrounding cells are dying (black arrows).

Bigger syncytia showed filaments reaching out to cells (Figure 5.7), although it is not clear whether these are from dead cells, retraction of cells that were being pulled into syncytia or filaments growing towards other cells. This might be a way of transporting viral particles to other cells without having budding viral particles and it could be a way of evading the immune system. However, this is experiment was unfit to investigate that hypothesis.

After growing these clinical-like samples *in vitro*, plaque assays were performed to check the amount of viable virus after harvesting and if culturing of samples *in vitro* was successful. The amount of viable virus used to spike the nasal lavage was known as plaque assays were performed on the aliquots in the previous part of the chapter. The plaque assays showed that the viral load harvested from culture of the clinical-like samples had PFUs ranging from 311,040 in a 1 in 10 dilution to undetectably low amounts in dilutions of 1 in 100,000 and 1 in 1 million. All detectable PFUs were in the same order of magnitude as expected from the viral load that was used for infection (Table 5.4).

Table 5.4: PFU/ml of seeding virus for each dilution and of harvested virus.

Dilution	**PFU/ml of seeding virus**	**PFU/ml of harvested virus**
1:10	3.1E5	2.3E5
1:100	3.1E4	2E4
1:1000	3.1E3	3.1E3
1:10,000	Estimated at 311	600
1:100,000	Estimated at 30	Unable to determine
1:1,000,000	Estimated at 3	Unable to determine

5.4. Discussion

I learned several techniques during the experiments in this chapter, like cell culture, growing and harvesting virus, plaque assays, microscopy and RT-qPCR.

This chapter discusses the set up and optimisation of growing clinical samples *in vitro* and what is necessary to be able to do that. The results indicate that the best time of harvest is the point where the cytopathic effect is greatest and the number of dead cells is lowest. Once a cell dies, the virus cannot replicate and the number of viral particles declines. In these experiments, that time was 55 hours post-infection. However, there are no data from earlier time points and the gap with the next time point is too big to conclude this time as the best time for harvest. This time might differ between various RSV strains and it would have been an interesting fact to analyse if there was more time.

Characterisation of the first time point indicated that the viral load of the N gene was over 13 million per millilitre. Multiple mRNA strands of the N virus are produced, however, it is generally known that some RNA is lost during the freeze-thaw cycle and during extraction and it is impossible to know how much, so even more copies might be produced. Further experiments with engineered RSV carrying fluorescently-labelled N proteins, for example by adding the sequence for Green Fluorescent Protein (GFP) to the sequence of the N gene, which then infects cell culture could provide us with more information on the number of N proteins that are produced during infection. Combined with fluorescent photography, the number of N proteins per viral particle could be investigated and other proteins might be investigated in a similar way.

After testing whether clinical-like samples can be grown in culture, it is clear that a sample needs a C_T value under 30 to grow with any visible cytopathic effect and any detectable viral titre with plaque assays. It has proven possible to culture clinical samples with higher viral loads and with visible syncytia.

If clinical samples have been sequenced and show distinctive characteristics, these could be selected to be grown in culture to determine the mutation rate, but also replication rate and ability to induce cytopathic effect and the speed at which this happens. This would be the best way to start investigating the effect of mutations on clinical disease. Comparing different strains to each other can show whether these characteristics have any effect on the pathology induced by the virus or not.

If certain characteristics seem to influence the pathology induced *in vitro*, further experiments, like gene expression studies, can shed light on the cellular pathways that might be differentially activated by different strains. Certain mutations in the F protein have already shown to influence mucin induction and cytokine production (143). Besides overall differential gene expression, single-cell sequencing can be used as well to find differences between infected and neighbouring uninfected cells. Via flow cytometry, cells are sorted in *e.g.* infected and uninfected neighbour, but also epithelial or immune cell, or other categories. Then, all RNA of individual cells is sequenced and can show the differential expression of cytokines and pathways that are known to be up or down regulated, but unknown factors can be identified this way as well. By comparing this to mock infected cells, a clear light can be shed on the workings of the cell while infected with RSV with specific characteristics, like a duplication in the G gene or specific mutations in the F gene.

6. Minority variant detection in clinical samples of community and hospitalised patients

6.1. Introduction

In this chapter, clinical samples from patients infected with RSV were sequenced using the newly established methods from the previous chapter. The clinical samples were from two different severity settings, namely community and hospital samples, and whole-length genomes were determined. The samples were spatiotemporally matched and ages from the infected infants were matched as well between the two groups. Consensus genomes were used for phylogenetic analysis and to establish the genotype of each clinical sample. Furthermore, the amount of variation in community versus hospital samples was investigated and the variants found in each sample were probed in more detail.

The aims of this chapter were to apply the optimised methods from the previous chapter to clinical samples, assemble genomes without having a reference genome to start from, perform quality control, analyse the sequencing results by genotyping the samples and performing phylogenetic analysis, and in-depth variant calling analysis. Moreover, the reason why some samples produced lower quality data was explored and the results from two disease severity groups were compared.

First, the samples were selected based on several characteristics that were discussed and pinned down beforehand. The characterisation of the selected samples is detailed in the first part of this chapter. In the second part, the sequencing and quality control of the data is described. This is followed by a subchapter on phylogenetic analysis and genotyping. This part depicts the sequences that were recovered, their genotypes and where they could be placed in the general RSV population. Finally, the last part spells out each variant that was detected in these samples, what their prevalence was and how that differs between the disease severity groups.

6.2. Characterisation of samples

The samples used to test the newly established methods were from two different cohorts. Community samples were obtained from a collaboration of PHE with the Royal College of General Practitioners (RCGP). General Practitioners sent nasal and/or throat swabs of patients suspected of having respiratory viral infections to PHE where samples were tested for RSV amongst other viruses, and viral loads were determined via multiplex qPCR as described in 2.2.6.2. These samples were aliquoted and kept at -80°C at PHE.

Samples from hospitalised patients were retrieved via the Imperial College Healthcare NHS Trust hospitals and stored at -80°C at the Tissue Bank in the Charing Cross hospital. These were nasopharyngeal aspirates and nasal swabs from infants with severe RSV infection needing hospitalisation and supportive care or intubation.

6.2.1. Describing the clinical data from two cohorts

This part of chapter 6 will discuss the selection process of the samples and the clinical data of the two cohorts that will be compared. Eight samples were selected from each cohort to test the NGS methods and to compare the prevalence of genetic variance. The goal of this part of the study was to select samples in such a way that it would control for any differences in clinical data that could affect RSV sequence variance between community samples and hospital samples. Therefore, the samples were spatiotemporally matched to avoid bias caused by local strains that might emerge in different regions or during different seasons. It is difficult to determine which infants had a primary or secondary infection, so patient samples were age-matched to keep that bias as low as possible. The viral load (C_T value) had to be lower than 30 to be selected, because a higher C_T value indicates there is too little qualitative RNA for sequencing.

Table 6.1: Clinical data of 16 selected clinical samples from community and hospital cohorts.

	Cohort	Viral Load (C_T)	Date of collection	Place of collection	Age of patient (months)
Sample1	Community	26.64	07/01/2016	Bristol	3m
Sample2	Community	29.14	12/11/2015	Kent	6m
Sample3	Community	19.04	07/12/2015	Cambridge	6m
Sample4	Community	23.93	07/11/2015	Bristol	7m
Sample5	Community	19.08	19/01/2016	Bristol	7m
Sample6	Community	19.94	09/11/2015	Portsmouth	9m
Sample7	Community	22.73	09/12/2015	Kent	10m
Sample8	Community	25.96	10/11/2015	Kent	11m
Sample9	Hospital	27.47	15/01/2016	London	3m
Sample10	Hospital	18.30	30/11/2015	London	6m
Sample11	Hospital	23.52	19/11/2015	London	6m
Sample12	Hospital	29.05	15/12/2015	London	7m
Sample13	Hospital	23.09	15/12/2015	London	7m
Sample14	Hospital	25.02	16/11/2015	London	9m
Sample15	Hospital	28.67	17/11/2015	London	10m
Sample16	Hospital	25.76	10/12/2015	London	11m

The first selection contained eight samples from community patients and eight samples from hospital patients that were matched (Table 6.1). All samples were from the 2015-2016 season (Figure 6.1) and were age-matched between community and hospital samples. The samples were collected over a period of 10 weeks and 4 days. The first sample was collected on the 7th of November 2015, while the last samples was collected on the 19th of January 2016. The samples in each cohort were from one patient aged 3 months, two of 6 months, two of 7 months, one of 9 months, one of 10 months and one of 11 months old. These ages were matched between the two groups to reduce age-related bias, like primary or secondary infection (Figure 6.1).

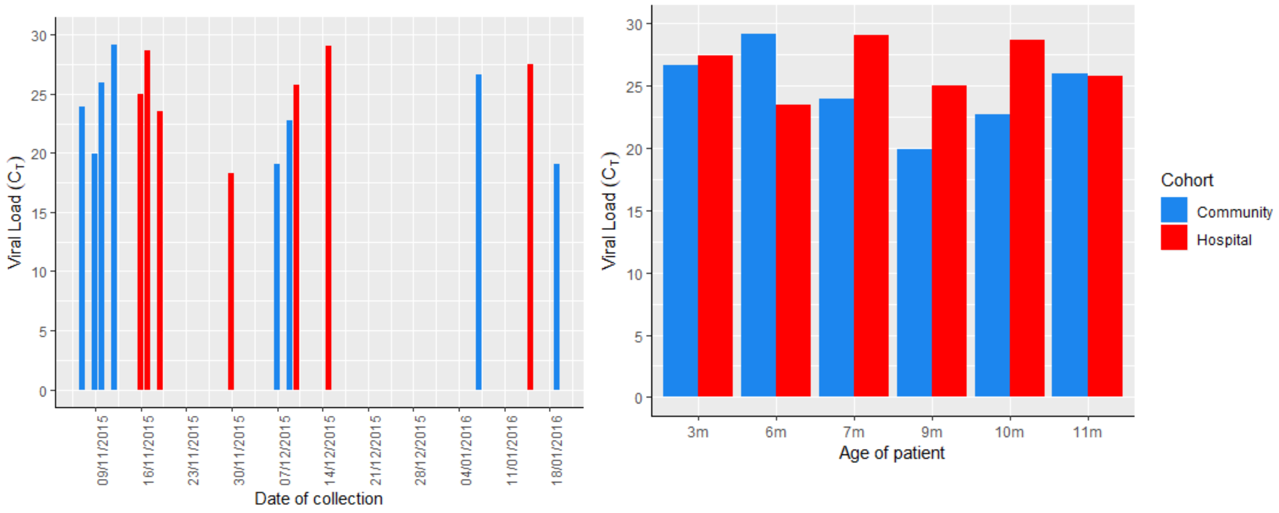


Figure 6.1: Date of collection and viral load were not associated with each other. Disease severity did not influence viral load in these selected samples (left). Age and viral load were plotted against each other and were not correlated (right).

Besides temporal and age matching, samples were collected in the same region, namely the south of England. Community samples ranged from Cambridge (the most northern point of collection) to Portsmouth in the south of England, Kent area in the east to Bristol as the most western point of collection. All hospital samples came from London, which lies in the middle of the quadrangle that covers the area from which the community samples were collected (Figure 6.2).

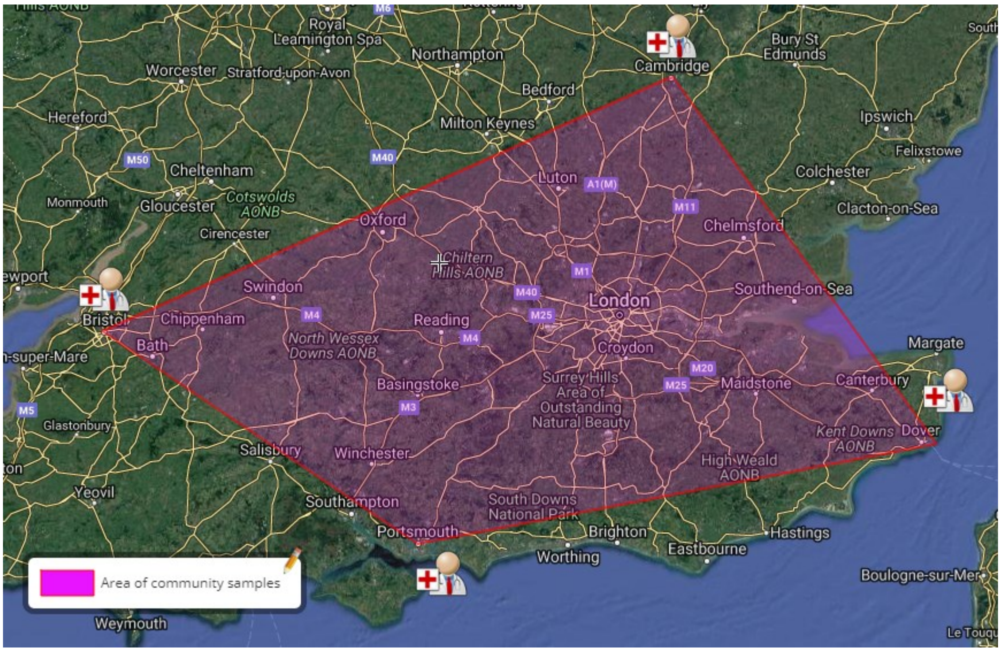


Figure 6.2: Map of the south of England detailing the farthest north, east, south and west points of the area from which community samples were collected. Hospital samples all came from London which is located in the middle of this area. Made via www.scribblemaps.com.

6.2.2. Statistical testing to identify viral load variance in two cohorts

The aim of this chapter is to study the difference in genetic variance between community and hospital samples from patients with RSV infections. In the previous part, clinical data was employed to create two similar cohorts with different disease severity. This part will discuss the investigation into the viral load and whether there is a difference between the selected samples from each cohort, since this can also affect the genetic variance that will be studied later. Viral load can affect the amount genetic variance in a sample. If there is a higher viral load and therefore more genetic material, the possibility of developing variants is higher as well. This would not be related to disease severity, so it has to be checked before further experiments are performed. Besides that, viral load could affect disease severity regardless of the genetic variation in the sample and therefore this bias was tested for with statistical tests.

First, the cohorts were inspected by summarising the data for community and hospital samples separately (Table 6.2). The minimum C_T value for community and hospital samples was 19.04 and 18.30 respectively, while the maximum C_T value was 29.14 for community samples and 29.05 for hospital samples. The median differed with 2.07 and the mean values were 23.30 and 25.11 for community and hospital samples respectively.

Table 6.2: Summary table of viral loads from community samples (top) and hospital samples (bottom) and the results from the Shapiro-Wilk test.

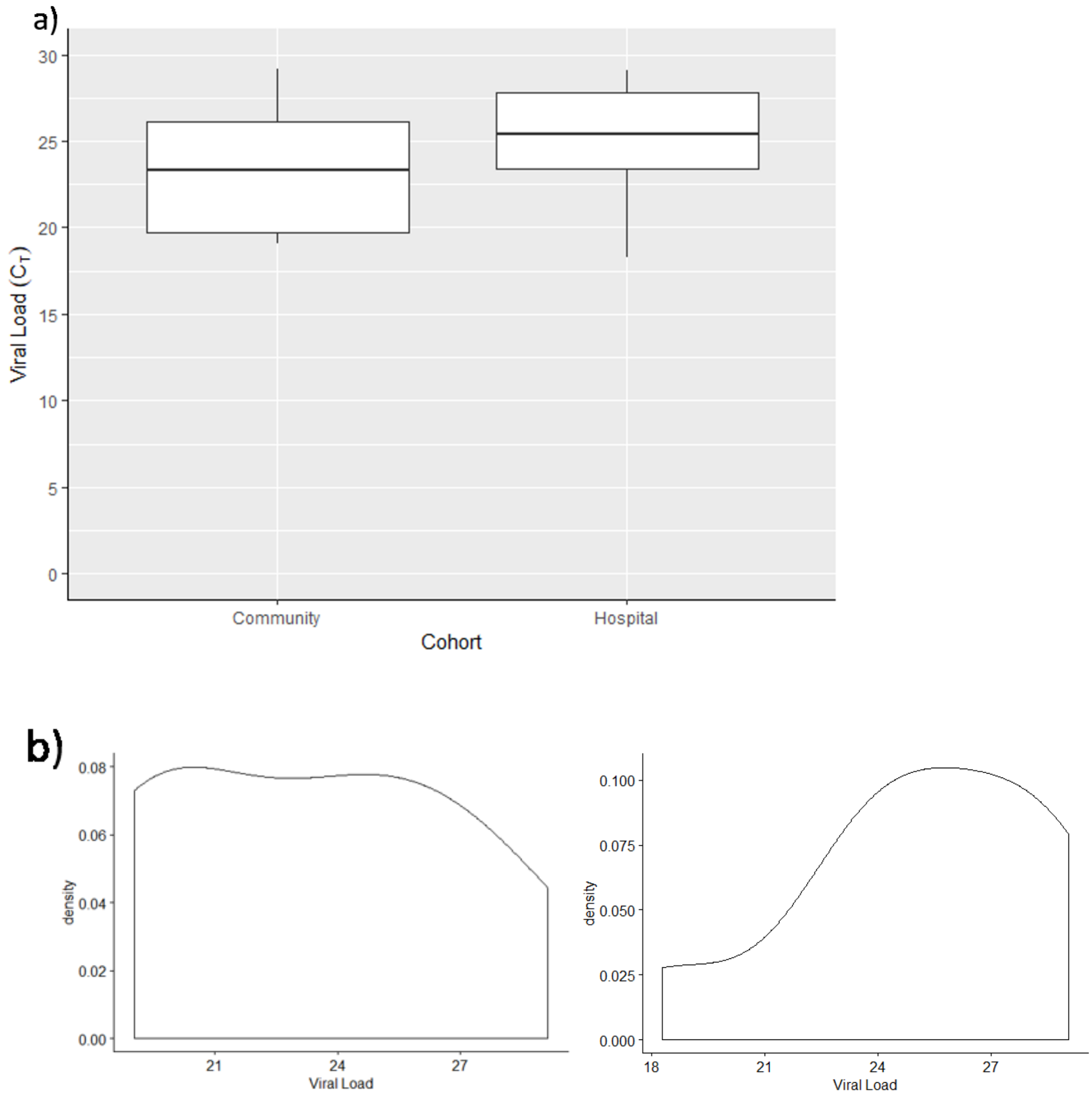
Summary of community samples						**Shapiro-Wilk test**	
Minimum	1st Quadrant	Median	Mean	3rd Quadrant	Maximum	W	p-value
19.04	19.72	23.32	23.30	26.12	29.14	0.92361	0.4598

Summary of hospital samples						**Shapiro-Wilk test**	
Minimum	1st Quadrant	Median	Mean	3rd Quadrant	Maximum	W	p-value
18.30	23.41	25.39	25.11	27.77	29.05	0.92968	0.5132

Besides a numerical summary, three plots were produced to further probe the data. First, a boxplot was created to compare the cohorts, then a density plot and Q-Q plot for each cohort were created to assess normality of the data of each group (Figure 6.3). The normality of the data was also tested statistically with the Shapiro-Wilk test. As a caveat, note that the sample groups are very small and not ideal to run statistics on. Therefore, these statistics must be interpreted with great care. The boxplot showed that the viral load was not entirely equally distributed. The density plots of both cohorts failed to show a bell curve as would be expected from normally distributed data. This might be due to the low number of samples in this test. However, the Q-Q plots did show that the data stayed within the confidence intervals as would be expected from normally distributed data. The Shapiro-Wilk normality test showed no significant p-value for community (p-value of 0.4598) or hospital (p-value of 0.5132) cohorts, hence normality could be assumed. The W value of the Shapiro-Wilk normality test did hint at non-normal distribution of the data.

Keeping all of this in mind, three tests were run to check if there was any significant difference in viral load between the two cohorts. First, a two-sample t-test was run on the viral load data comparing the community and hospital groups with a confidence level of 95%. This test assumes normality and the Shapiro-Wilk test did suggest normally distributed data (although this was not entirely convincing in general). The p-value returned from this test was 0.3402 and indicates there is no significant difference in the viral load.

As is it possible the data was not normally distributed, but it was not picked up due to the small sample size, the data was log-transformed and the test was run again. This returned a p-value of 0.3434 and confirmed the findings of the first data. Transforming the data by exponent, square or square root did not result in a different outcome for this test. Therefore, a third test was run. The Mann-Whitney-Wilcoxon test (MWW) was used as this is the alternative for non-normally distributed data. This returned a value of 0.5054 confirming once again what was calculated by the two-sample t-test. The conclusion was that there was no significant difference in viral load between the two cohorts and it could be assumed that differences in variation between the groups were not due to viral load differences.



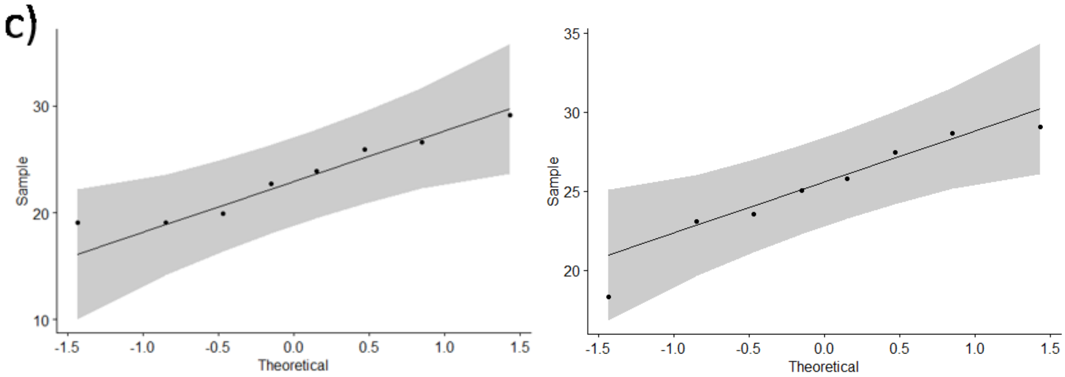


Figure 6.3: (a) Boxplot of community and hospital samples comparing the distribution of viral load. (b) Density plots of viral loads for community (left) and hospital (right) cohorts. (c) Q-Q plots of viral load data for community (left) and hospital (right) cohorts.

6.3. Sequencing and quality control

In the previous chapter, the optimisation of sequencing of RSV samples is described and this newly established workflow was used for deep-sequencing of the samples from both the community and hospital cohort. All samples were extracted with NucliSENS® easyMag®, converted to cDNA, amplified in 16 fragments, which was checked on E-gels, cDNA quantified and sequenced with Illumina MiSeq (see 2.2.4.2, 2.2.5.3, 2.2.7.1, 2.2.8.2, 2.2.1 and 2.2.3 respectively). The output files were Fastq files and were first assembled with SPAdes as described in 2.3.2.1. Then, a reference sequence was determined by BLASTing contigs and looking through the available online databases. Secondly, the reads were mapped against the reference sequence that was returned by BLAST as described in 2.3.2.2. The output file of BWA assembly was used to create a consensus sequence from the reads and the reads were then mapped against the obtained consensus sequence of the sample to look for variants. The genome coverage was plotted in R to inform about the read depth and therefore the success of whole-genome retrieval (Figure 6.4). Quality checks were performed using general statistics generated by samtools and FastQC as described in 2.3.2.3.

Sample 1 from the community cohort failed to deliver a sequence. There was not enough PCR product generated from the sample to be sequenced. Most bands that were distinguishable for this sample were too low on the gel indicating the cDNA fragments were too small for what we expected to find (fragments ranging from 1168 to 1520 bp). Two partial sequences were acquired, namely from sample 2 and sample 8. The E-gels showed the presence of clear bands for some of the fragments and were absent for other fragments. These missing bands indicating missing fragments were mirrored in the plots of the read depth. Five full RSV genomes were recovered, namely sample 3, 4, 5, 6 and 7. These samples showed much clearer E-gels with brighter bands and bands that were not smeared out. The partial or failed samples all had C_T values higher than 26.0, while all complete sequences samples had

C_T values lower than 23.9 (Table 6.1). An overview of the output data from the quality control is collected in Table 6.3.

Out of eight hospital samples, two failed to provide sequences, namely sample 11 and 14. The E-gels showed few clear bands of the right size and lots of smearing on the gel. Sample 10, 12 and 16 resulted in a partial sequence. The E-gels showed this might happen with some bands of the right size and some missing bands for other fragments of the genome. Samples 9, 13 and 15 showed the best E-gels out of all hospital samples and generated complete sequences. All samples that produced smeared gels with RSV A primers were run again with RSV B primers to confirm their subtype. All samples produced better gels with RSV A primers apart from sample 10. NGS was performed on the cDNA which showed the best bands on their gel.

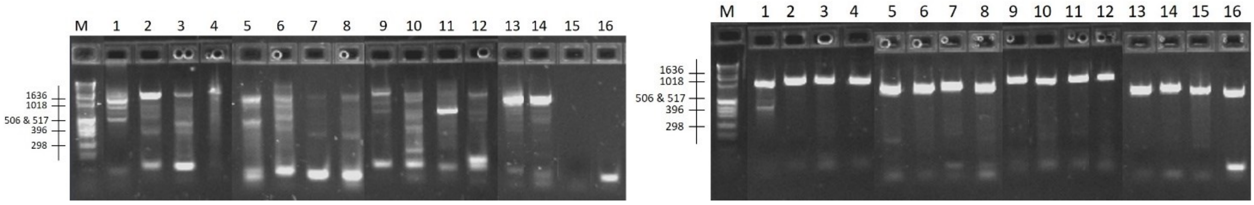

M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
1636
1018
506 & 517
396
298
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
1636
1018
506 & 517
396
298

Figure 6.4: Sample 2 data is shown on the left, sample 5 data is shown on the right. E-gels show bands if fragment contains good amounts of PCR product for sequencing. Marker (M) size is explained at the left of the E-gel picture.

Table 6.3: Summary output of FastQC quality control from bam file produced by mapping of the reads against the consensus sequence. The accession number refers to the strain used as reference during initial mapping of the reads to produce the first bam file. %GC = overall GC content of reads.

	Accession number	Retrieved sequence	Data file	Total number of mapped reads	Reads flagged as poor quality	Read length	%GC
Sample 1		Fail	Raw data forward reads	75511	0	50-151	48
			Raw data reverse reads	75511	0	50-151	48
Sample 2	KU950592.1	Partial	Raw data forward reads	75667	0	50-151	45
			Raw data reverse reads	75667	0	50-151	45
			Bam file	151374	0	30-151	45
Sample 3	KU950523.1	Complete	Raw data forward reads	135033	0	50-151	36
			Raw data reverse reads	135033	0	50-151	36
			Bam file	271973	0	30-151	36
Sample 4	KX765948.1	Complete	Raw data forward reads	158490	0	50-151	45
			Raw data reverse reads	158490	0	50-151	44

			Bam file	317242	0	30-151	45
Sample 5	KX765948.1	Complete	Raw data forward reads	227909	0	50-151	34
			Raw data reverse reads	227909	0	50-151	34
			Bam file	456005	0	30-151	34
Sample 6	KU950540.1	Complete	Raw data forward reads	109795	0	50-151	35
			Raw data reverse reads	109795	0	50-151	35
			Bam file	219845	0	30-151	35
Sample 7	KC731482.1	Complete	Raw data forward reads	136986	0	50-151	37
			Raw data reverse reads	136986	0	50-151	37
			Bam file	274401	0	31-151	37
Sample 8	JX627336.1	Partial	Raw data forward reads	157733	0	50-151	40
			Raw data reverse reads	157733	0	50-151	40
			Bam file	315526	0	31-151	40
Sample 9	KM360090.1	Complete	Raw data forward reads	158655	0	50-151	39
			Raw data reverse reads	158655	0	50-151	39
			Bam file	317569	0	30-151	39
Sample 10	KM360090.1	Partial	Raw data forward reads	207626	0	50-151	33
			Raw data reverse reads	207626	0	50-151	33
			Bam file	419316	0	30-151	33
Sample 11		Fail	Raw data forward reads	112612	0	50-151	44
			Raw data reverse reads	112612	0	50-151	44
			Bam file	349347	0	30-151	40
Sample 12	KU950459.1	Partial	Raw data forward reads	121540	0	50-151	43
			Raw data reverse reads	121540	0	50-151	43
			Bam file	243144	0	30-151	43
Sample 13	KU950502.1	Complete	Raw data forward reads	153409	0	50-151	39
			Raw data reverse reads	153409	0	50-151	39
			Bam file	306867	0	31-151	39
Sample 14		Fail	Raw data forward reads (A)	159307	0	50-151	43
			Raw data reverse reads (A)	159307	0	50-151	43
			Bam file (RSV A primers)	318617	0	38-151	43
		Fail	Raw data forward reads (B)	105441	0	50-151	41
			Raw data reverse reads (B)	105441	0	50-151	41

			Bam file (RSV B primers)	210894	0	131-151	41
Sample 15	KM360090.1	Complete	Raw data forward reads	185047	0	50-151	34
			Raw data reverse reads	185047	0	50-151	34
			Bam file	370837	0	30-151	34
Sample 16	KX765970.1	Partial	Raw data forward reads	122129	0	50-151	43
			Raw data reverse reads	122129	0	50-151	43
			Bam file	244274	0	30-151	43

6.4. Genotyping using phylogenetic inference analysis

This subchapter focuses on retrieving information from the RSV genomes that were sequenced in the previous subchapter. Once the genomes were quality checked and assembled, they were used to determine the viral genotype. In chapter 3, a set of reference sequences was selected to determine genotypes based on specific parts of the genome. This analysis showed that the G gene of RSV is necessary and sufficient to determine the RSV genotype of a sample. The 12 clinical samples that had a good quality sequence of the G gene were added to the reference dataset containing 48 reference strains. This dataset was aligned using Clustal Omega v1.2.2 and MAFFT v7.409 algorithms via AliView v1.23. Using manual editing, the two alignments were compared and optimised to retrieve the best alignment. Two other datasets were created by supplementing the clinical sequences with only RSV A or only RSV B reference sequences respectively. Quality was tested in three different ways and a phylogenetic tree was calculated and annotated to study the genotypes of the clinical samples.

6.4.1. Quality control of nucleotide sequences from three datasets

First, completeness scores for the alignments were calculated using AliStat v1.7. This gives an indication of the general quality of the sequences, interrogates the number of sequences that have indels compared to the alignment, the number of indels at each position in the alignment and the number of ambiguous characters in each sequence. The formulas for the completeness scores (C-scores) that are calculated are shown in Table 6.4.

Table 6.4: Formulas used by AliStat v1.7 to calculate completeness scores (C-scores) for the sequences in an alignment.

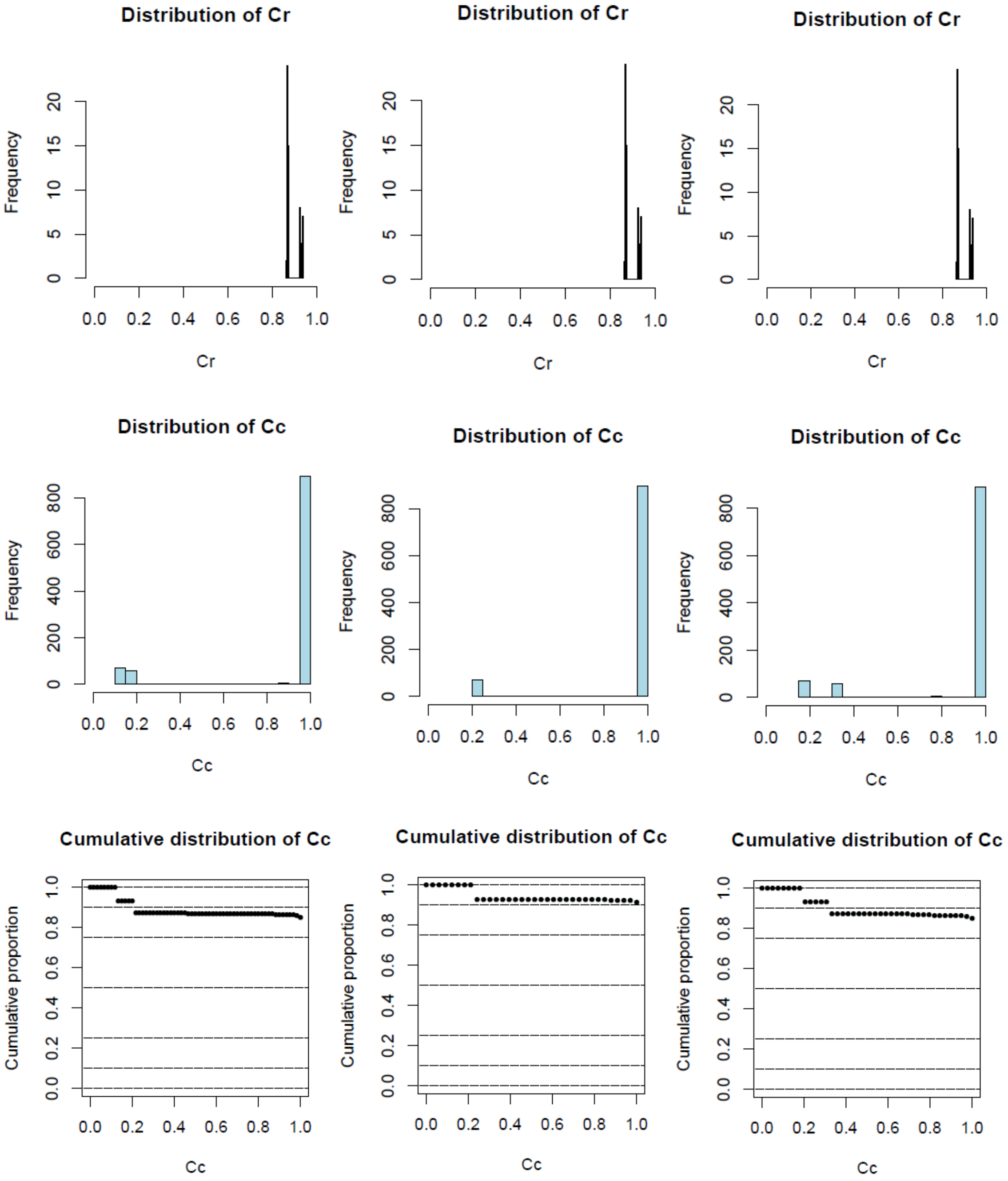
C-score for the alignment	$Ca = \frac{total\ number\ of\ unambiguous\ characters}{number\ of\ sequences * length\ of\ alignment}$
C-scores for individual sequences (C_r_min and C_r_max)	$Cr = \frac{number\ of\ unambiguous\ characters\ in\ the\ sequence}{alignment\ length}$
C-scores for individual sites (C_c_min and C_c_max)	$Cc = \frac{number\ of\ unambiguous\ characters\ in\ the\ column}{number\ of\ sequences}$
C-scores for pairs of sequences (C_{ij}_min and C_{ij}_max)	$Cij = \frac{number\ of\ columns\ with\ unambiguous\ corresponding\ characters\ in\ two}{length\ of\ alignment}$

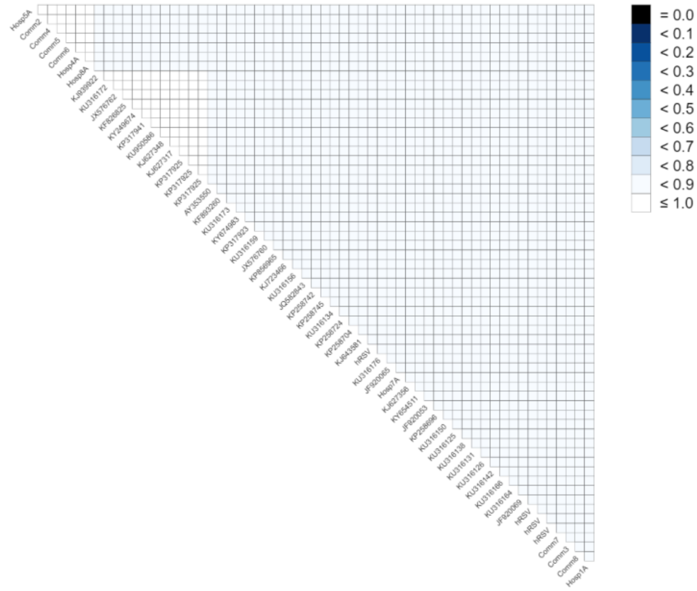
The C-scores were calculated for three datasets: Reference strains for RSV A and RSV B plus sequences from all clinical samples or RSV A reference strains plus sequences from all clinical samples or RSV B reference strains plus sequences from all clinical samples (Table 6.5). There were 1032 nucleotides in all alignments that contains RSV B sequences and 969 nucleotides in the alignment that only contained RSV A sequences. The difference in sequence length was mostly due to a longer non-coding tail at the end of the reference strain for RSV B BA4 genotype. The complete alignment C-scores ranged from 0.89 to 0.94 which means most characters were unambiguous.

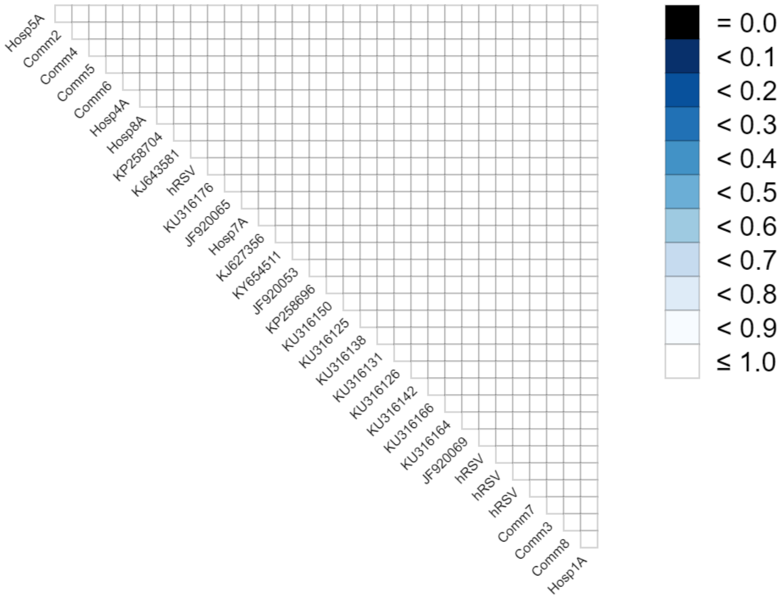
Table 6.5: Table with completeness scores from three datasets: Community and hospital sequences with references strains for both A and B, only A or only B. Ca = Completeness score for the alignment; Cr = completeness scores for individual sequences; Cc = completeness scores for individual sites; Cij = completeness scores for pairs of sequences; Pij = p-distance for pairs of sequences.

		Reference A and B	Reference A	Reference B
Number of sequences in alignment		60	33	39
Ca		0.89	0.94	0.90
Cr	Minimum	0.86	0.92	0.86
	Maximum	0.94	1.00	0.94
Cc	Minimum	0.12	0.21	0.18
	Maximum	1.00	1.00	1.00
Cij	Minimum	0.86	0.92	0.86
	Maximum	0.94	1.00	0.94
Pij	Minimum	0.00	0.00	0.00
	Maximum	0.35	0.13	0.35

The C-scores for individual sequences (Cr) ranged from 0.86 to 1 (Figure 6.5). The C-scores for individual sites (Cc) ranged from 0.12 to 1 (Figure 6.5), meaning at least one position had unambiguous nucleotides across all sequences, although not necessarily the same nucleotide. The cumulative Cc showed that the dataset with only RSV A references had the best overall Cc-scores (Figure 6.5). When looking at the triangular heat maps based on C-scores for paired sequences (Cij), the higher quality in the dataset with only RSV A references compared to the other datasets is confirmed (Figure 6.5).




Hosp5A
Comm2
Comm4
Comm5
Comm6
Hosp4A
Hosp8A
KJ939922
KU316172
JX576762
KF826825
KY249674
KP317941
KU950586
KJ627348
KJ627317
KP317925
KP317925
KP317925
AY353550
KF893260
KU316173
KY674983
KP317923
KU316159
JX576760
KP856965
KJ723466
KU316156
JQ582843
KP258742
KP258745
KU316134
KP258724
KP258704
KJ643581
hRSV
KU316176
JF920065
Hosp7A
KJ627356
KY654511
JF920053
KP258696
KU316150
KU316125
KU316138
KU316131
KU316126
KU316142
KU316166
KU316164
JF920069
hRSV
hRSV
hRSV
Comm7
Comm3
Comm8
Hosp1A
= 0.0
< 0.1
< 0.2
< 0.3
< 0.4
< 0.5
< 0.6
< 0.7
< 0.8
< 0.9
≤ 1.0


Hosp5A
Comm2
Comm4
Comm5
Comm6
Hosp4A
Hosp8A
KP258704
KJ643581
hRSV
KU316176
JF920065
Hosp7A
KJ627356
KY654511
JF920053
KP258696
KU316150
KU316125
KU316138
KU316131
KU316126
KU316142
KU316166
KU316164
JF920069
hRSV
hRSV
Comm7
Comm3
Comm8
Hosp1A
= 0.0
< 0.1
< 0.2
< 0.3
< 0.4
< 0.5
< 0.6
< 0.7
< 0.8
< 0.9
≤ 1.0

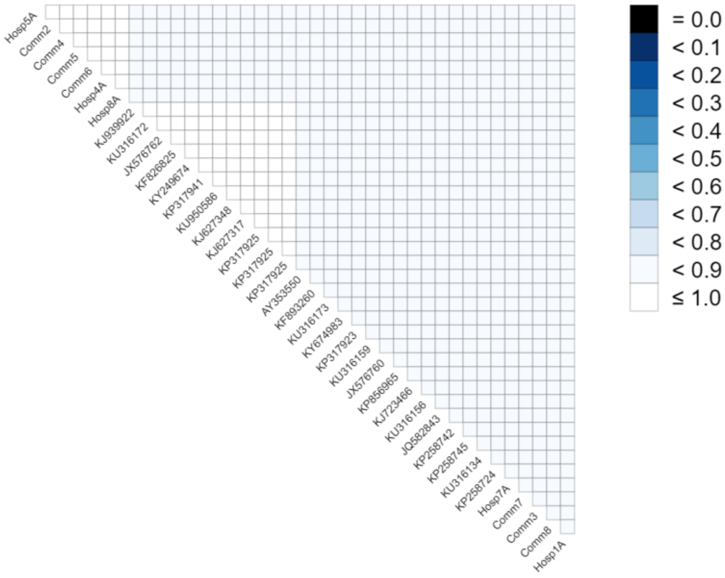

Hosp5A
Comm2
Comm4
Comm5
Comm6
Hosp4A
Hosp8A
KJ939922
KU316172
JX576762
KF826825
KY249674
KP317941
KU950586
KJ627348
KJ627317
KP317925
KP317925
KP317925
AY353550
KF893260
KU316173
KY674983
KP317923
KU316159
JX576760
KP856965
KJ723466
KU316156
JQ582843
KP258742
KP258745
KU316134
KP258724
Hosp7A
Comm7
Comm3
Comm8
Hosp1A
= 0.0
< 0.1
< 0.2
< 0.3
< 0.4
< 0.5
< 0.6
< 0.7
< 0.8
< 0.9
≤ 1.0

Figure 6.5: Histogram of distribution of completeness scores for individual sequences (top row), individual sites (second row) and cumulative distribution of completeness scores for individual sites (third row) from datasets containing clinical samples and both RSV A and B reference sequences (left), only A reference sequences (middle) or only B reference sequences (right). Bottom row: Triangular heatmaps with completeness scores for pairs of sequences (Cij) displayed. Sequences with duplications are more complete compared to sequences with a gap at that position, which is visualised by colouring the reduced Cij completeness score in light blue.

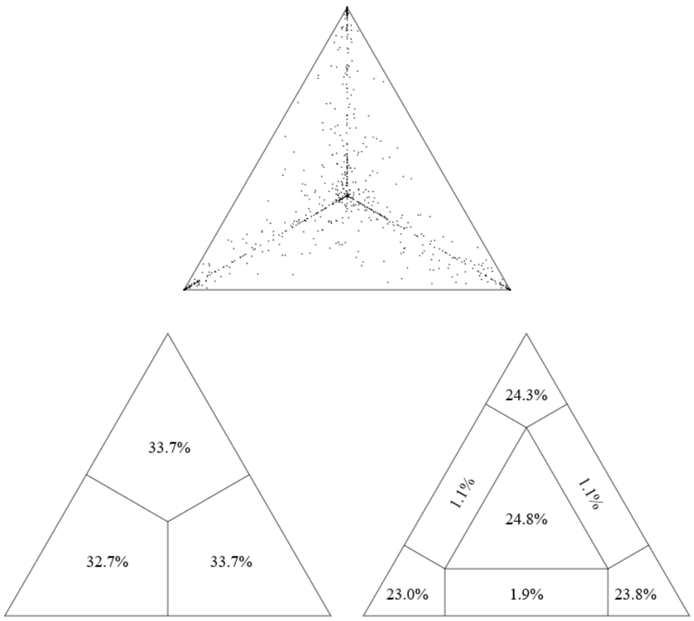
All the quality results taken together suggest that the quality of sequences in these datasets was high and that the dataset with only RSV A references was qualitatively the best dataset out of three. The suggested that most clinical samples are RSV A strains and therefore show fewer gaps (and thus higher C-scores) compared to RSV A references than compared to RSV B references.

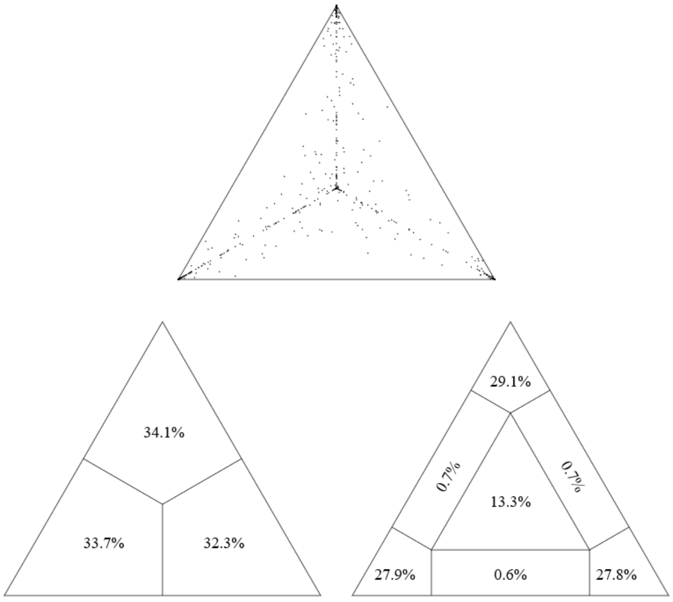
However, these results only describe the quality of nucleotides and the alignment. This does not equal a good dataset for phylogenetic analysis. In the next step, the quality of the alignment for phylogenetic analysis is assessed.

6.4.2. Quality control of alignments for phylogenetic analysis

The quality of the sequences is important, but it is only the first step for analysis. Not all datasets are optimal for phylogenetic analysis. A dataset could contain sequences that are all very similar and cannot be solved for a definitive phylogenetic tree. To check the quality of the datasets used in the previous step, IQ-Tree v1.6.6. was employed and likelihood mapping was visualised with triangle plots. This type of analysis explores if definitive trees can be constructed with phylogenetic analysis. In this process, an indicative Parsimony tree was built and used for initial review of the clinical sequence clustering in the reference tree.

Likelihood mapping analysis is based on the random selection of quartets. The number of quartets used for likelihood mapping should be at least 25 times the number of sequences in the alignment to cover each sequence approximately 100 times. The largest alignment contained 60 sequences, so each dataset was analysed with 1500 quartets to cover the alignments sufficiently. Uninformative quartets collect in the middle of the triangle and indicate that a phylogenetic tree cannot be calculated for this quartet. Partially informative quartets gather at the sides of the triangle and in these quartets only a partial tree can be constructed. Fully informative quartets are most desirable and cluster in the corners of the triangle. Unclustered mappings should have equally distributed quartets in the three corners of the triangles. Clustered mappings usually show a preference for one of the corners depending on the provided clusters and their sizes.


33.7%
32.7%
33.7%
24.3%
1.1%
1.1%
24.8%
23.0%
1.9%
23.8%


34.1%
33.7%
32.3%
29.1%
0.7%
0.7%
13.3%
27.9%
0.6%
27.8%

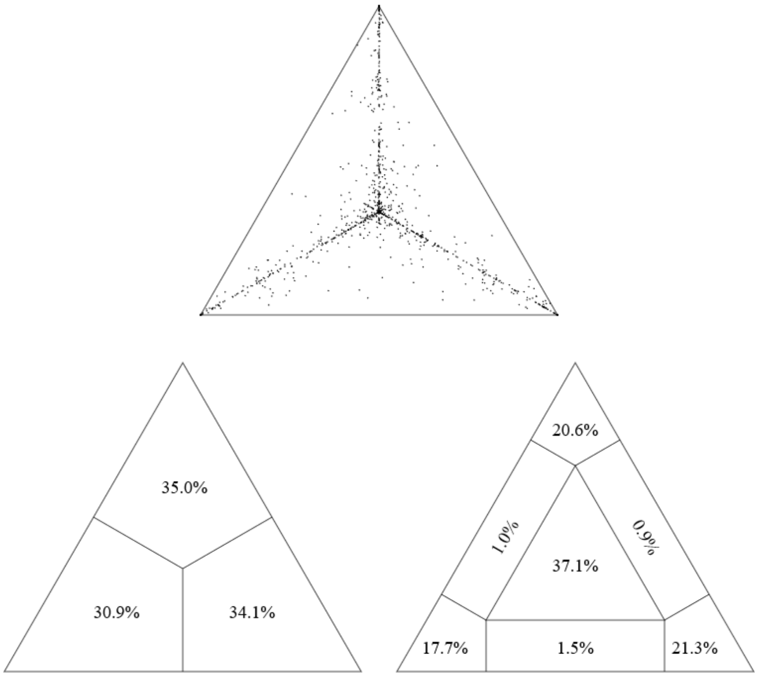


Figure 6.6: Triangle plots for datasets with all sequences top), only RSV A sequences (middle) and only RSV B sequences (bottom) were run with 1500 quartets. Corners show informative quartets, side bars show partially informative quartets and the centre shows uninformative quartets.

The triangle plots for the dataset containing RSV A and RSV B references showed a high number of uninformative quartets (24.8%), while the dataset with only RSV A reference sequences contained fewer uninformative quartets (13.3%) and the dataset with only RSV B reference sequence contained the highest number (37.1%) of uninformative quartets. This suggested that the dataset with only RSV A reference strains is the most useful for phylogenetic analysis (Figure 6.6). An indicative parsimony tree was constructed as well showing what earlier data also suggested: the clinical strains from both community and hospital patients were all part of the RSV A subgroup (Figure 6.7).

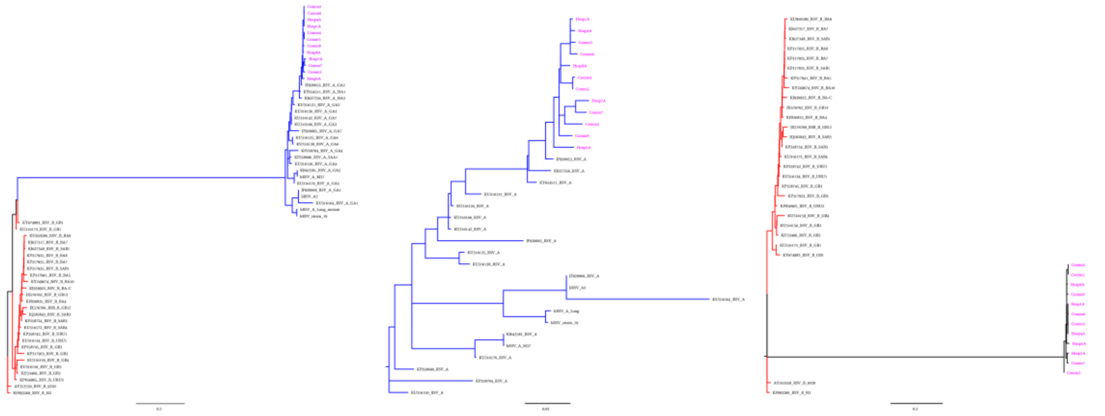


Figure 6.7: The indicative parsimony trees that were produced by likelihood mapping analysis for datasets with both RSV A and RSV B references strain (left), only RSV A reference strains (middle) or only RSV B reference strains (right). Blue = RSV A subtypes; red= RSV B subtypes; purple = clinical samples.

The alignment containing only RSV A reference strains was qualitatively the best dataset and the parsimony tree showed that all clinical samples were from subgroup A and more specifically GA2 genotypes. Some of those contained the 72-nucleotide duplication as was seen in the alignments and as indicated in the triangular heat maps produced by AliStat. The dataset containing only RSV A reference strains was used from this point onwards for further analysis.

6.4.3. Phylogenetic analysis to determine genotypes of community and hospital samples

The next step in phylogenetic analysis was determining the best fitting evolutionary model. The best fitting model of substitution for this dataset was determined by running Model Finder which returned TIM3+F+R2 according to the Bayesian information criterion (BIC) scores. Base frequencies were calculated empirically (indicated by "F" in the model) and the number of the rate type was R with two categories (as indicated by "R2" in the model). This rate type is a generalisation of the G rate type in that it relaxes the gamma distribution rates assumption.

The final step in the phylogenetic analysis was to calculate a Maximum Likelihood tree based on the best fitting model. The tree was run with 1000 UltraFast Bootstrap replicates and 1000 approximate Likelihood Ratio Test replicates. This phylogenetic tree showed that all clinical samples from both cohorts were indeed part of subtype A and genotype GA2 (Figure 6.8). Genotype ON1 is part of the GA2 genotype cluster and contains a 72-nucleotide duplication in the G gene. Sample 2, 4, 5 and 6 from the community cohort and sample 12, 13 and 16 from the hospital cohort carried this 72-nucleotide duplication and were therefore part of genotype ON1. This analysis showed that there was no difference in genotype between the two cohorts.

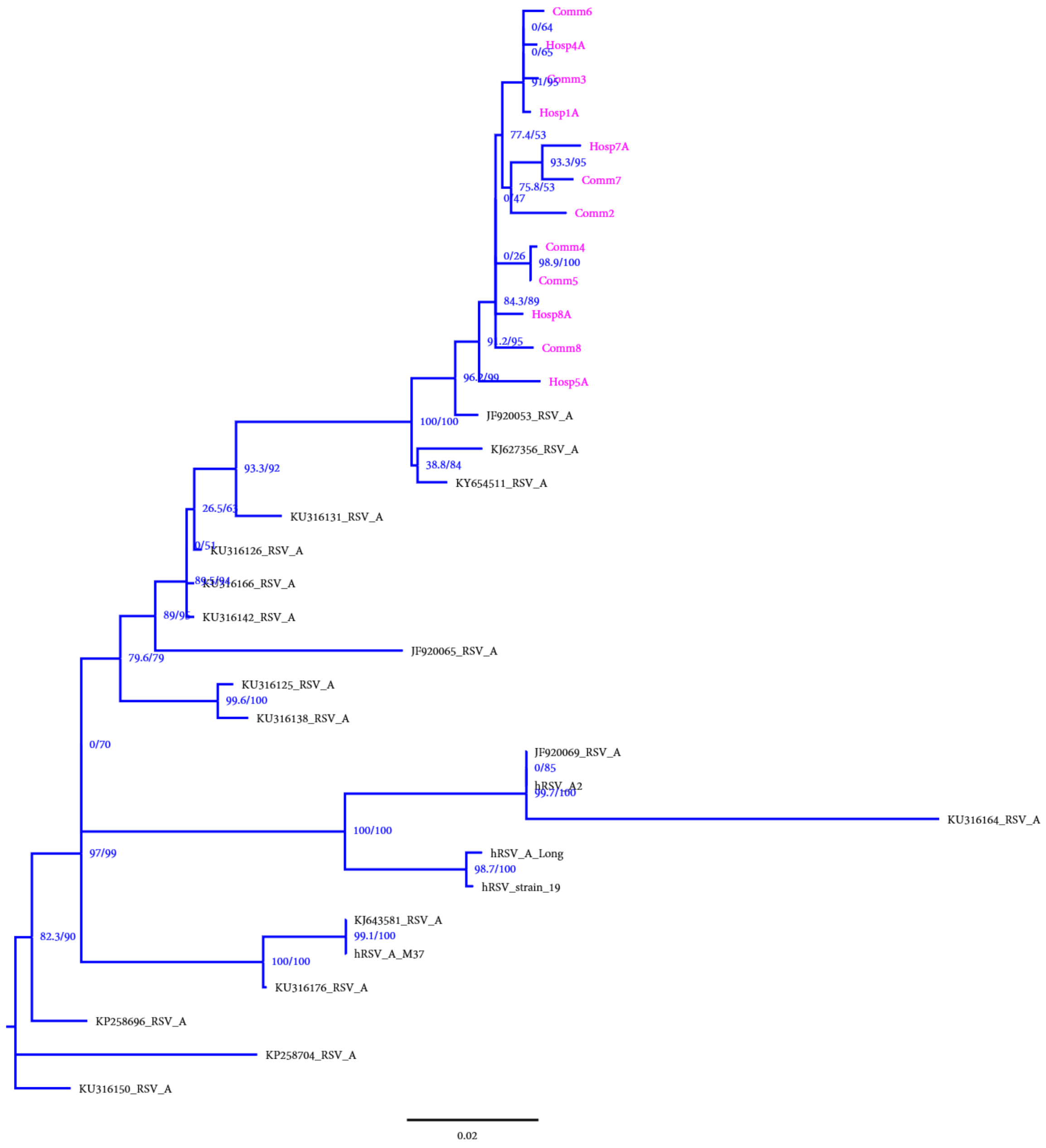


Figure 6.8: Maximum Likelihood tree of dataset containing clinical samples and RSV A reference strains. Branch nodes show approximate Likelihood Ratio Test/Bootstrap values. Pink = clinical samples.

6.5. In-depth analysis of variants

After genotyping the clinical samples, further inspection was carried out on the number and type of individual variants seen in each sample to identify specific differences that might affect disease severity. The number of variants and the prevalence of each variant was determined for each sample by variant calling analysis using samtools and bcftools v1.8. Then, the effect on protein sequence and the localisation of variants was studied. Lastly, the number and prevalence of variants were used to calculate the Shannon Entropy, which reflects the overall abundance of variation present in each sample.

Each method in the sequencing workflows has specific error rates, meaning they can introduce errors in the sequence. Extraction has negligible error rate, but reverse transcription has an error rate of 1/2000. The following step is PCR amplification with a Taq polymerase which has an error rate of 1.8/10000. During library preparation another PCR step follows, although it is not clear which enzyme is used in the Illumina Nextera kits. Assuming it is Phusion enzyme, this error rate is only 4.4/1E7. Lastly, sequence determination by Illumina MiSeq machines has an error rate too. This is specific for each read and each read gets an overall and position specific quality value. All reads with a value of Q=35 or higher were considered, meaning the error rate in these reads/positions is about 1/3125 or less. This implies that about 0.1% of positions can contain errors. To be sure, a threshold of 10 times that number was handled during analysis. Variants with a frequency under 1% were deemed possible errors. Each variant was also screened for position in the reads where it occurs. If a variant is found only at the beginning or end of a read, it was deemed a possible error.

6.5.1. Description of variants, their location and effect on protein sequence

The number of variants ranged from 2 to 9 in a complete sequence in community samples and from 3 or 4 (in a partial or complete sequence respectively) to 5 in a complete sequence in hospital samples. In total, 24 unique variants were encountered, 20 in the community cohort and 22 in the hospital cohort (Table 6.6).

Table 6.6: Genotypes, variants, their frequency and location in the genome, and the overall Shannon Entropy for each sample. Orange = partial sequences.

	Genotype	**Variants**	**Frequency**	**Location**	**Shannon Entropy**	**Codon Ref**	**Codon Alt**	**AA Sub**	**Synonimity**
Sample 1									
Sample 2	ON1	A5419C	0.55	G	1.4	CAC	CCC	H258P	Nsyn
		G9507A	0.14	L		GAG	GAA	E346	Syn
		A14400G	0.39	L		AAA	AAG	K1958	Syn
Sample 3	GA2	A5445G	0.30	G	1.87	AAG	GAG	K262E	Nsyn
		A12159C	0.37	L		ATA	ATC	I1230	Syn
		T12168A	0.15	L		ATC	ACA	I1233T	Nsyn
		T12486C	0.15	L		CGT	CGC	R1339	Syn
Sample 4	ON1	C5383G	0.53	G	2.91	CCC	CGC	P258R	Nsyn
		A8509G	0.26	M2-1		TAG	TGA	Stop303	Syn
		G8510A	0.25	M2-1					

		A12183C	0.18	L		ATA	ATC	I1230	Syn
		T12510C	0.18	L		CGT	CGC	R1339	Syn
		A14367G	0.32	L		AAA	AAG	K1958	Syn
Sample 5	ON1	C2013T	0.13	N	2.18	TAC	TAT	Y301	Syn
		A5423C	0.24	G		CAC	CCC	H258P	Nsyn
		A8546G	0.11	M2-2		AGT	GGT	S116G	Nsyn
		A12220C	0.38	L		ATA	ATC	I1230	Syn
		A14404G	0.16	L		AAA	AAG	K1958	Syn
Sample 6	ON1	A5036G	0.18	G	0.84	CCA	CCG	P126	Syn
		G9519A	0.14	L		GAG	GAA	E327	Syn
Sample 7	GA2	T2342G	0.30	Inbetween N and P	4.35				
		G2998A	0.34	P		CTG	CTA	L226	Syn
		A5429G	0.41	G		CAC	CGC	H259R	Nsyn
		A5445G	0.34	G		GAA	GAG	E264	Syn
		A8484G	0.20	M2-1		TAR	TGR	Stop305	Syn
		G9452C	0.15	L		GGG	GGC	G330	Syn
		A12158C	0.31	L		ATA	ATC	I1232	Syn
		T12485C	0.10	L		CGT	CGC	R1341	Syn
		A14342G	0.46	L		AAA	AAG	K1960	Syn
Sample 8	GA2	G893A	0.14	NS2	1.28	GAT	AAT	D274N	Nsyn
		G9452C	0.15	L		GGG	GGC	G330	Syn
		A14342G	0.55	L		AAA	AAG	K1960	Syn
Sample 9	GA2	C4082A	0.13	M	1.76	CCC	CAC	P285H	Nsyn
		G7081A	0.22	F		GTG	GTA	V486	Syn
		A8486G	0.28	M2-1		TAR	TGR	Stop307	Syn
		G9454C	0.13	L		GGG	GGC	G332	Syn
Sample 10	RSV B	-	-	-	0				
Sample 11									
Sample 12	ON1	A2320T	0.23	Inbetween N and P	2.15				

		T2329C	0.22	Inbetween N and P					
		G2992A	0.26	P		CTG	CTA	L226	Syn
		G4890A	0.09	G		ACG	ACA	T80	Syn
		A14404G	0.12	L		AAA	AAG	K1958	Syn
Sample 13	ON1	G2989A	0.22	P	2.45	CTG	CTA	L226	Syn
		A5418C	0.33	G		CAC	CCC	H258P	Nsyn
		G9506A	0.16	L		GAG	GAA	E328	Syn
		A12215C	0.38	L		ATA	ATC	I1231	Syn
		A14399G	0.52	L		AAA	AAG	K1959	Syn
Sample 14									
Sample 15	GA2	A5430C	0.28	G	2.62	CAC	CCC	H258P	Nsyn
		A5479T	0.45	G		CCA	CCT	P275	Syn
		C5495T	0.33	G		CAT	TAC	H280Y	Nsyn
		T5497C	0.33	G					
		A12159C	0.35	L		ATA	ATC	I281	Syn
Sample 16	ON1	G9117T	0.15	L	1.46	GCA	TCA	A196S	Nsyn
		A12221C	0.40	L		ATA	ATC	I1230	Syn
		A14405G	0.45	L		AAA	AAG	K1958	Syn

A total of 54 variants were detected in all samples combined. In general, variants were most often seen in the G gene and L gene. NS1 and SH were the only genes were no variants were detected. The number of variants detected for each gene was 0 for NS1, 1 for NS2, 2 for N, 3 for P, 1 for M, 0 for SH, 12 for G, 1 for F, 4 for M2-1, 1 for M2-2 and 26 for L (Figure 6.9). Thirteen non-synonymous variants were detected in all samples together. Eight of those were located in the G gene of which five in the community cohort and three in the hospital cohort. The other non-synonymous variants were distributed in the NS2 (1), M (1), M2-2 (1) and L (2) genes. The remaining 41 variants were synonymous variants and did not affect the protein sequence. The L gene showed the most variants, which was expected since it spans over a third of the genome. When looking at the number of variants per nucleotide for each gene, it was clear that G contains the most variants, followed by M2-2 (Figure 6.9).

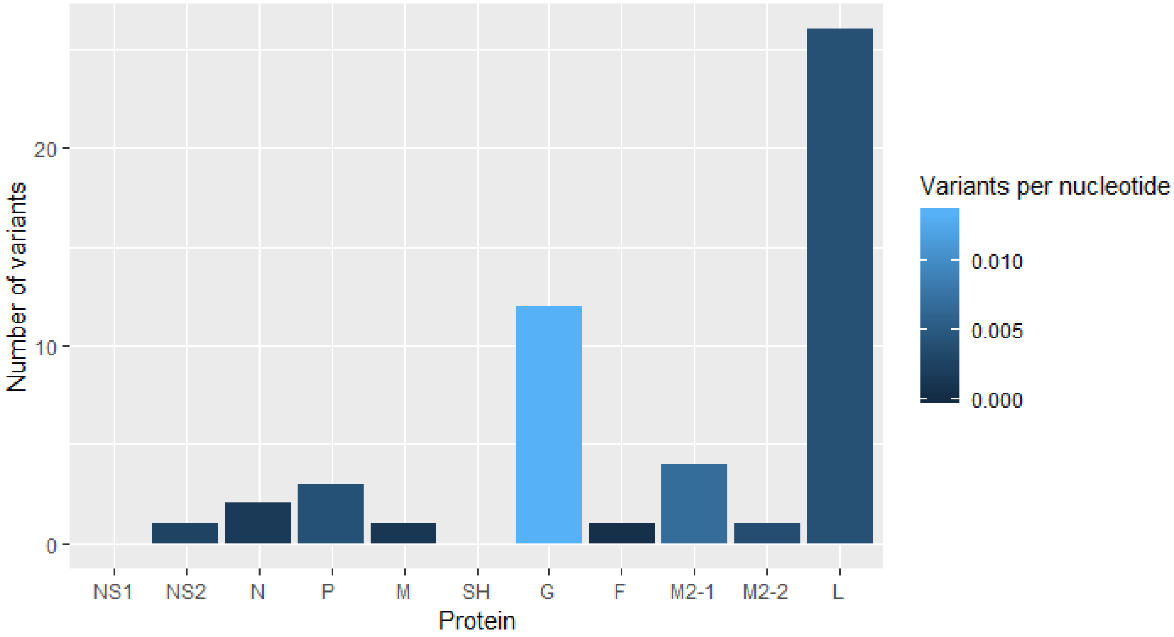


Figure 6.9: Bar plot to show number of variants per gene and number of variants per nucleotide for each gene indicating G has the highest substitution rate with M2-1 having the second highest substitution rate. L has accumulated the highest absolute number of variants.

When comparing the two cohorts, there is a difference in number of variants overall with 32 variants for the community samples and 22 variants for the hospital samples. However, since only 1 sequence failed and two are incomplete in the community cohort compared to two failed and three incomplete sequences in the hospital cohort, this was unreliable information. The relative number of non-synonymous variants was similar with 25.0% out of 32 being non-synonymous in the community cohort compared to 27.3% out of 22 variants in the hospital cohort. The variants that did not affect the protein sequence were therefore also similarly in relative numbers: 75.0% in the community cohort and 72.7% in the hospital cohort.

The variants seen in the clinical samples often show non-synonymous variations at sites which are known to be under positive selection. Non-synonymous variants in G are often seen at position 258 in these samples, but both position 262 and position 280 show non-synonymous variation in this cohort as well. Non-synonymous variants were also found in the NS2 gene (D274N), the M gene (P285H), the M2-2 gene (S116G) and in the L gene (A196S and I1233T).

6.5.2. Comparison of Shannon Entropy between cohorts to study variance

The Shannon Entropy (SE) is a quantity that is also known as information entropy as described in information theory by Claude Shannon. It is used to describe the amount of information that is carried by data. In this thesis, it describes the amount of genetic variance in a sample. It takes both the number of variants and their abundance into account. The SE ranged from 0.84 to 4.35 for community samples and from 0 to 2.62 for hospital samples. The overall mean SE for community samples was 2.119

compared to 1.740 for hospital samples. A boxplot of the SE for both cohorts showed that there seems to be little difference between hospital and community samples, although community samples seemed to have a broader range than hospital samples (Figure 6.10).

To test whether there was a difference in variance between the community and hospital cohorts, statistics were brought into the equation. As a caveat, note once again that the sample groups are very small and not ideal to run statistics on. Therefore, these statistics must be interpreted with great care. First, the normality of the data had to be checked. The density plots and Q-Q plots already showed that the data did not seem to be normally distributed (Figure 6.11)and the Shapiro-Wilk test confirmed that the null hypothesis of could not be rejected with p-values of 0.4225 and 0.2504 for the SE of community and hospital samples respectively. Therefore, the Mann-Whitney-Wilcoxon test was used for statistical testing and it returned a p-value of 0.9452. This confirmed that there was no significant difference. The overall conclusion is that there was no significant difference in genetic variance between the two cohorts.

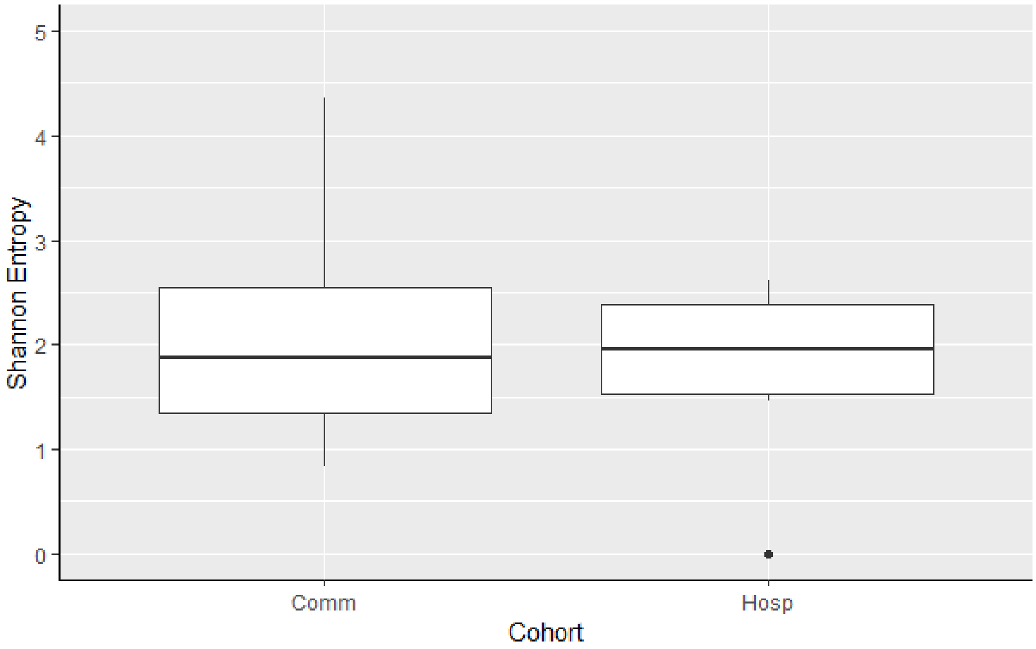


Figure 6.10: Boxplot of Shannon Entropy for each cohort shows similar average diversity of samples between both cohorts, although the community cohort has a higher standard deviation.

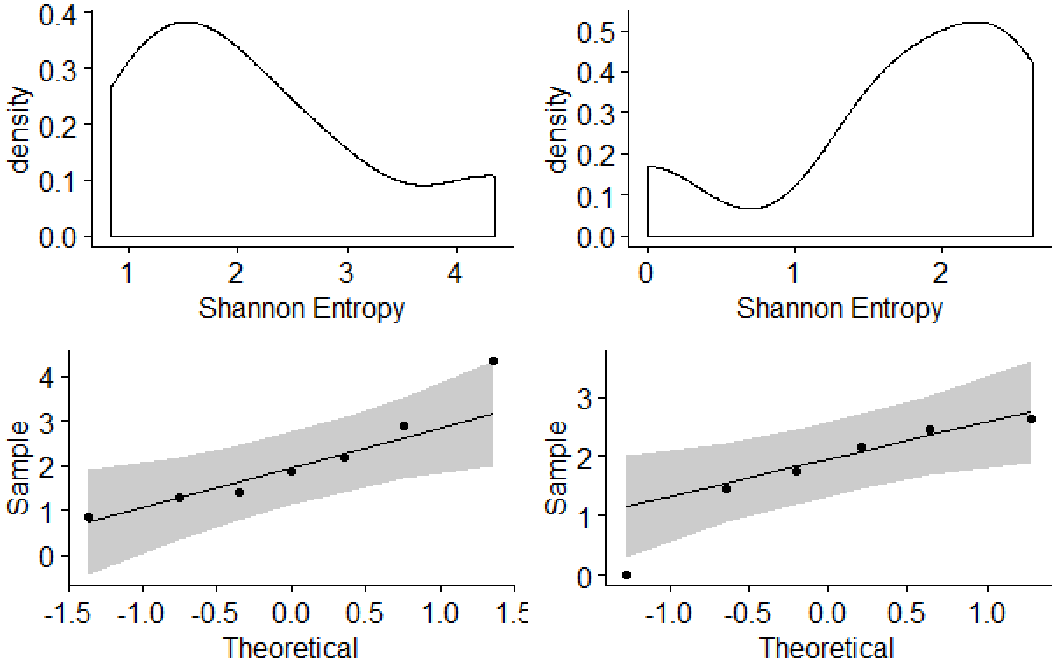


Figure 6.11: Normality plots for the Shannon entropy for community (left) and hospital (right) cohorts. Top = Density plots; botton = Q-Q plots.

6.6. Discussion

This chapter focused on using the next-generation sequencing methods that were optimised in chapter 2. The clinical samples were selected from two disease severity groups, namely community and hospital patients. The hypothesis was that disease severity could be explained by a higher genetic variance in the viral genome or that disease severity might drive increased viral variation in the hospital patients or by specific variants in the RSV genome.

The first part of this chapter was about the selection and characterisation of the samples from each cohort. These samples were selected based on several characteristics: viral load high enough to sequence, geographically located in the south of England, from the 2015-2016 season and patients' ages were matched. All these measures were taken to reduce any bias in the results. Before any further analysis was done, the difference in viral load was investigated as this could influence the amount of genetic viral variation detected in the two cohorts.

Statistical testing showed that there was no difference in the viral load between the two cohorts. The numbers were low in these tests, so type II statistical errors are possible in this analysis. When more samples are tested and added to each cohort, more subtle differences could be detected. It should be kept in mind that these samples were selected to have high enough viral loads to be sequenced by NGS. Thus, samples with lower viral loads were excluded from these cohorts and an overall difference in viral load might be present, but missed because of the selection criteria. This was not the main point of study in this chapter and was therefore not studied in more detail.

The second part of this chapter dealt with testing the NGS methods by sequencing clinical samples without knowing which genotype or subtype was present. The methods to retrieve whole-length genomes were successful in 50% of the samples. In an additional 31% of samples, partial sequences were retrieved and in 19% of samples, the NGS methods failed to deliver RSV genomes. There are several reasons why this could have happened.

External factors could have decreased the quality of the sample. RSV is a labile virus that breaks down very quickly. If the sample is not frozen within 30 minutes, the RNA starts to break down and quality decreases. Furthermore, once the sample is frozen, it should not be freeze-thawed if this can be avoided as this process also decreases quality. Therefore, it is best to freeze aliquots of one sample rather than freezing everything in one container. Unfortunately, not all samples were aliquoted or not enough aliquots were available for all experiments. After thawing and freezing to test the presence and quantity of RSV, the sample was thawed again and used for sequencing at which point it had been frozen twice before. Adding RNA-stabilising agents like RNAlater could somewhat prevent breakdown of RNA and snap freezing samples could reduce RNA breakdown even further. Unfortunately, this is a limitation of working with clinical samples taken for diagnosis as not every facility has the time and means to treat samples like this as their time goes to treating patients, which obviously takes precedent.

Biological factors could have caused the decreased sequence quality as well. Since there are several genotypes in RSV and the G gene is very variable as shown in chapter 3, the primers might bind with different specificity to different strains. Considering the genotypes are not known beforehand, it is impossible to adapt the set primers to all genotypes. Primer design based on retrieved partial sequences would create the possibility to recover a complete genome if there are enough aliquots available.

A more technical factor that makes it more difficult to assemble a whole genome is the fact that there are no reference strains set out to be the best match. The genotype is unknown, so assembly of the reads against a reference genome is impossible as a lot of reads will be left out if the wrong reference genome is selected. Hence, *de novo* assembly was performed first which will give bad results if there are gaps in the sequenced genome as the reads cannot connect form one long contig. The longest contig produced by *de novo* assembly was used to compare against public databases and find the genome that looks most alike. This genome was then used as a reference against which the reads were assembled again. However, if there are gaps in the genome, this method only uses part of the genome, namely the region contained in the largest contig. This loss of information on the rest of the genome might mean that the genome looking most like the largest contig does not necessarily look most like the entire genome from the sample. Considering only the reads from the sample that look most like the reference are used in the assembly, this newly assembled sequence looks most like the reference genome that was used when compared against the public database later.

The next part of this chapter was based on the consensus sequences determined from the assembly to perform phylogenetic analysis. In the first chapter, it was determined that the G gene is necessary and sufficient to determine the genotype and thus only the G gene was used for phylogenetic analysis. It would be better to use complete genomes to investigate this, however, there were few complete genomes available and it is less computationally demanding to compare part of the RSV genome. Before starting to build a tree, the quality of the sequences were determined, *e.g.* if there were a lot of ambiguous characters in the G genes or not. These completeness scores are calculated on databases and since the genotypes were not known and even the subtype was not known for sure for all samples, there were three datasets set up to be tested: a dataset containing both RSV A and RSV B genotypes references, a dataset containing only RSV A references and a dataset containing only RSV B references.

The completeness score for the overall alignment was best for the dataset based on only RSV A genotypes. When looking at the completeness scores for individual sequences (Cr), it shows that the datasets containing RSV B reference sequences do not contain any completely unambiguous sequences. Having the knowledge afterwards that all clinical strains were RSV A genotypes, this makes sense. The 60-nucleotide duplication in the G gene of RSV B genotypes is never seen in RSV A genotypes and the 72-nucleotide duplication in the G gene of RSV A genotypes is never seen in RSV B genotypes either. This explains why there will always be a gap in an individual sequence when both genotypes are present. The completeness score can therefore never be 1. However, in the RSV A genotype, the maximum Cr was 1 and this was a first indication that all clinical samples are RSV A genotypes. The minimum Cr will never be 1 as both genotypes with and without the 72-nucleotide duplication are present and therefore there is always at least one sequence in the alignment with a gap compared to the other sequences.

The completeness score for individual sites (Cc) showed a maximum of 1 for all datasets, so there is at least one position which was unambiguous for all sequences (but not necessarily the same nucleotide in all sequences). The minimum Cc is very low, but knowing about the 72 nt duplication, this is to be expected as well. If only 21% of the sequences had an unambiguous character at a specific position, this indicated that 1/5th of the sequences have the 72 nt duplication in this dataset and 4/5th do not have the duplication and have therefore a gap at this position. The completeness scores for pairs of sequences (Cij) confirmed that the dataset with only RSV A references contained sequences that are more complete compared to each other as both the minimum and maximum Cij are higher in this dataset.

After checking the quality of the sequences in the dataset, the phylogenetic information was probed. This indicates whether there is enough (but not too much) variation in the dataset to calculate a phylogenetic tree with enough confidence to trust its branching. Likelihood mapping consists of selecting random quartets and checking if they are informative enough to calculate a tree. Sequences that are completely the same are very uninformative, while sequences with several different nucleotides compared to each other are more informative. When sequences are too different, the quartets become uninformative again. The three datasets were compared again and it clearly showed that the dataset with RSV A reference sequences only is the most informative and will thus return the tree with the highest confidence. To check all sequences at least 100 times, the number of tested quartets should be 25 times the number of sequences in the dataset. The resulting triangle plots showed that the most informative dataset for phylogenetic analysis is the one with only RSV A reference sequences.

Based on these data, the best fitting model was searched for and the tree was built on this substitution model with the dataset containing only RSV A references. The model describes the frequency of each possible nucleotide substitution in this specific dataset.

The resulting tree was a maximum likelihood tree, meaning that a number of trees have been built and compared to each other with the most likely one as the output. In this tree, all clinical samples gather in the GA2 cluster, which also contains ON1 genotypes. This is to be expected since this is by far the most common genotype of RSV A in the recent years. The distinction between GA2 and ON1 is based on the presence or lack of the 72 nt duplication in the G gene.

Finally, the variants were investigated in detail. All variants were localised in the genome and their effect on protein sequence was determined. The number of variants was highest in the L gene and that makes sense since this gene entails more than a third of the genome. After normalising for gene length, the G gene contained the most variants and that is to be expected too. This gene is incredibly variable and differs immensely between subtypes, genotypes and even within the same genotype. It makes sense that lots of variation is seen in this gene, both synonymous and non-synonymous. The number of non-synonymous variants was highest in G as well, which is also to be expected. The second most variable gene after normalisation is M2-1, which functions as a transcription factor. These variants are all synonymous though and do not affect the protein sequence. The fact that G is the most variable protein of RSV is confirmed in these data.

Most of the variants was observed in the G gene, which was not unexpected. It has been observed in the past that certain positions have higher numbers of non-synonymous changes occurring, indicating positive selection. Some of these positions have been shown to carry escape mutants suggesting these positions are under immune driven positive selection (223). Several studies have investigated which exact positions have positive selection and found that these are not the same for RSV A and RSV B. RSV A positions under positive selection in the G gene are 97,101, 104, 106, 111, 115, 117, 121, 122, 123, 126, 127, 131, 132, 142, 146, 161, 206, 215, 217, 225, 226, 230, 233, 247, 250, 258, 262, 274, 276, 280, 286, 290, 291 and 297 (224, 225). For RSV B, the sites under positive selection in G are 98, 99, 142, 152, 219, 223, 224, 237, 247, 251, 257, 258, 259, 267/287 and 297/317 (depending on the presence of the duplication) (225, 226). Since non-synonymous variants were often found in known sites of positive selection, like position 258 of the G gene, this study confirms earlier work.

Neutralizing epitopes in G are found at the end of the conserved region and at the beginning of the HVR2, which covers amino acid positions 145-160 and 187-218 of G (227). The duplicated region of G also carries a neutralizing region, which has been shown to evolve as well (228).

A lot more is known about neutralizing epitopes in F as this protein is more important for infection and therefore the main focus of vaccine development (229, 230). Changes in neutralizing epitopes can also affect the only prophylactic currently available, palivizumab. Single mutations at positions 268, 272 and 276 have shown to induce partial or complete palivizumab resistance (231-233). Surveillance studies have not been able to find these mutations occurring naturally (234).

The number of variants and their prevalence was determined and based on that, the Shannon Entropy was calculated. This number demonstrates the amount of variation present in one sample. If the Shannon Entropy is different between the two cohorts, this could be an explanation for the differences in disease severity seen in children infected with RSV. However, after statistical testing, there was no measured difference. The genetic variance of a sample is not correlated to disease severity in these cohorts.

Bigger groups could enhance detection of statistically significant differences. More spatiotemporally and age-matched samples were already selected, but funding was not sufficient to sequence more samples. With more resources other questions could be examined as well. There might be differences in quasispecies development in different age groups like babies compared to young children, older adults, healthy adults or immunocompromised patients. Perhaps different RSV genes will evolve at different rates when the immune system is at different stages of maturation.

There are lots of different host variants that might affect disease severity as well. Plenty of human genes are involved in immune responses and each of those genes can carry variants that influence the response to RSV infection. It is even possible that the combination of certain RSV variants combined with specific human gene variants causes advantageous or disadvantageous reactions during infection. To study the effect of viral variants only, a genetically similar cohort of people should be studied, however, such studies are nearly impossible.

In this study, there was no detectable difference in strains between the community cohort and hospital cohort. Furthermore, there was no observable difference in frequency of certain variant frequencies in either cohort. Overall, there seems to be no clear association between RSV variants or strains and disease severity.

However, it was shown here that most clinical samples do carry minority variants. It presents the question of whether these variants arose during acute infection or whether they were transmitted at the moment of infection. There are many variants know to occur in the RSV genome, but it is unclear where these originate from. In the next chapter, the optimised methods from the previous chapter will be used to investigate the abundance, origination and evolution of minority variants during acute infection.

7. Within-host minority variant dynamics during acute RSV infection

7.1. Introduction

In chapter 3, genome variation was studied in clinical samples from two cohorts. In this chapter, another population was investigated to find variants in the RSV genome over time during acute infection. To be able to study infected patients during the acute phase of RSV infection, healthy, adult volunteers were infected with a known and well-studied RSV strain, *i.e.* RSV A M37. Nasal lavage samples were taken every day for 10 consecutive days after inoculation and again at day 14 and day 28 post-inoculation. These samples were sequenced with the optimised methods from chapter 4 that were also used in chapter 6. The amount of variation in each sample was followed up over time. The second part of this chapter will focus on an infant that contracted RSV while receiving palivizumab. Several consecutive samples were taken from this infant over the course of three weeks. Three of these samples were sequenced and analysed in detail, and variants were characterised.

The aim of this chapter was to investigate RSV genome variant frequencies over time in a controlled population to study bottleneck events and in natural infection. The hypothesis was that RSV genome variance seen on a population level is caused by quasispecies variants developing during acute infection and surviving the bottleneck of transmission. Therefore, the inoculum was sequenced first and then variants seen in the inoculum were searched for and investigated over time in the infected volunteers. After that, the genome was also scanned for *de novo* variants and these were characterised as well.

First, the samples from infected volunteers were investigated to ensure sufficient quality and high enough viral loads. The selected samples were characterised and so were the symptom scores from the volunteers for each day the samples were taken. Then, the variants that were present in the inoculum were identified and described in detail. After that, the variants seen in subsequent clinical samples from volunteers were detailed, identifying those that survived the bottleneck of transmission, and whether frequencies increased or decreased over time. In the following part, the development of *de novo* variants is outlined and these variants are probed in more detail. In the second subchapter, a similar study was performed on three clinical samples from an infant naturally infected with RSV while on palivizumab. All variants present in these samples are described over time.

7.2. Analysis of samples from volunteer cohort

7.2.1. Characterisation of samples from infected volunteers

The samples used to investigate RSV genome variants from a known strain over time were from a cohort of volunteers previously infected in 2011, 2012 and 2013 (235). These healthy, adult volunteers had nasal lavage samples taken daily for 10 consecutive days and follow up samples were taken at day 14 and day 28 post-inoculation. Samples at time points later than day 10 were not investigated as RSV was no longer detected. Samples from the acute phase of infection with a C_T value of less than 30 (viral load of more than 500 copies/ml) were selected for deep-sequencing.

Fifty-two samples from 10 different volunteers met the criteria and were sequenced using the Illumina MiSeq platform with the optimised methods as described in chapter 2. Samples from six men and four women were used. The volunteer's ages ranged between 19 and 34 years old and the mean age was 22.3 years. On average, 3.2 samples were sequenced per volunteer. The earliest sample post-infection with viral loads high enough to sequence was from day 3 and the latest sample was from day 10. None of the volunteers had high enough viral loads for sequencing on day 14 post-inoculation or later. Thirty-two samples returned good quality sequences. Of those, 19 partial sequences and 13 complete viral genomes were assembled (Table 7.1). Assembly was performed as described in 2.3.2.2. and quality checks were carried out as described in 2.3.2.3.

Table 7.1: Characteristics of volunteers infected with RSV M37 and samples that were sequenced.

Subject	Sex	Age	Sample day	Log_{10}(viral load)	Viral load (C_T)	Symptom scores
#1	F	24	3	3.25	28.76	5
			5	2.75	30.39	13
			6	6.45	18.30	20
			8	4.61	24.33	19
			9	5.98	19.86	14
#2	M	20	7	4.20	25.66	1
			10	3.95	26.46	0
#3	F	19	4	5.08	22.78	3
			5	5.68	20.82	6
			6	5.00	23.06	8
			9	4.27	25.44	2
#4	M	22	4	3.10	29.27	1
			6	3.17	29.02	0
			8	3.00	29.58	0
#5	M	21	6	4.43	24.90	2
			8	4.42	24.93	2
#6	F	20	5	2.12	32.46	2
			6	3.58	27.70	14
			7	3.26	28.72	10
			8	2.52	31.16	5
#7	M	19	7	4.74	29.42	16
#8	M	20	7	2.55	30.72	10
			8	3.76	27.27	4
#9	M	24	6	2.70	30.54	1

			7	3.95	26.46	1
			8	4.82	23.65	0
			9	3.41	28.23	0
#10	F	34	6	3.80	26.96	7
			7	5.26	22.20	11
			8	5.47	21.51	10
			9	5.00	23.04	6
			10	4.38	25.08	6

Volunteers were asked to score their symptoms during the first 10 days post-inoculation. Symptoms scores quantified severity of sneezing, nasal discharge, nasal obstruction, sore throat and cough as well as general symptoms like headache, fatigue and fever. These were graded on a scale of 0 to 3 by the volunteers and varied greatly between people. Some of the infected individuals showed symptoms of infection (with some scores as high as 20), while other remained asymptomatic (scores of 0), but had detectable amounts for RSV and were therefore infected nonetheless (Table 7.1).

To examine the correlation between viral load and symptom scores, statistical testing was performed. First, the normality of the data was checked by producing a data summary (Table 7.2) and graphing a density plot and a Q-Q plot (Figure 7.1). Then, the normality was tested with the Shapiro-Wilk test (Table 7.2). Since the data were not normally distributed ($p<0.05$ in SWT), the log(viral load), $\log_{10}$(viral load), exponent, square and square root of the viral load were calculated and checked for normality with the Shapiro-Wilk test. This showed that $\log_{10}$-transformed viral load data was normally distributed; a summary of the data is shown in Table 7.2. The Shapiro-Wilk test resulted in a p-value of 0.89 and therefore the $\log_{10}$-transformed data were used for further testing. Density plots and Q-Q plots were graphed as a second visual check (Figure 7.1) and showed that $\log_{10}$(VL) data was normally distributed, but symptom scores were not.

Table 7.2: Summary of viral load data and symptom score data (left) and Shapiro-Wilk test results (right).

Summary of VL of Volunteers						**Shapiro-Wilk test**	
Minimum	1st Quadrant	Median	Mean	3rd Quadrant	Maximum	W	p-value
130.90	1710.80	12430.50	168271.00	73800.10	2838835.00	0.35	0.00

Summary of Symptom score of Volunteers						**Shapiro-Wilk test**	
Minimum	1st Quadrant	Median	Mean	3rd Quadrant	Maximum	W	p-value
0.00	1.00	5.00	6.22	10.00	20.00	0.89	0.00

Summary of $\log_{10}$(VL) of Volunteers						**Shapiro-Wilk test**	
Minimum	1st Quadrant	Median	Mean	3rd Quadrant	Maximum	W	p-value
2.12	3.23	1.08	4.08	4.86	6.45	0.98	0.89

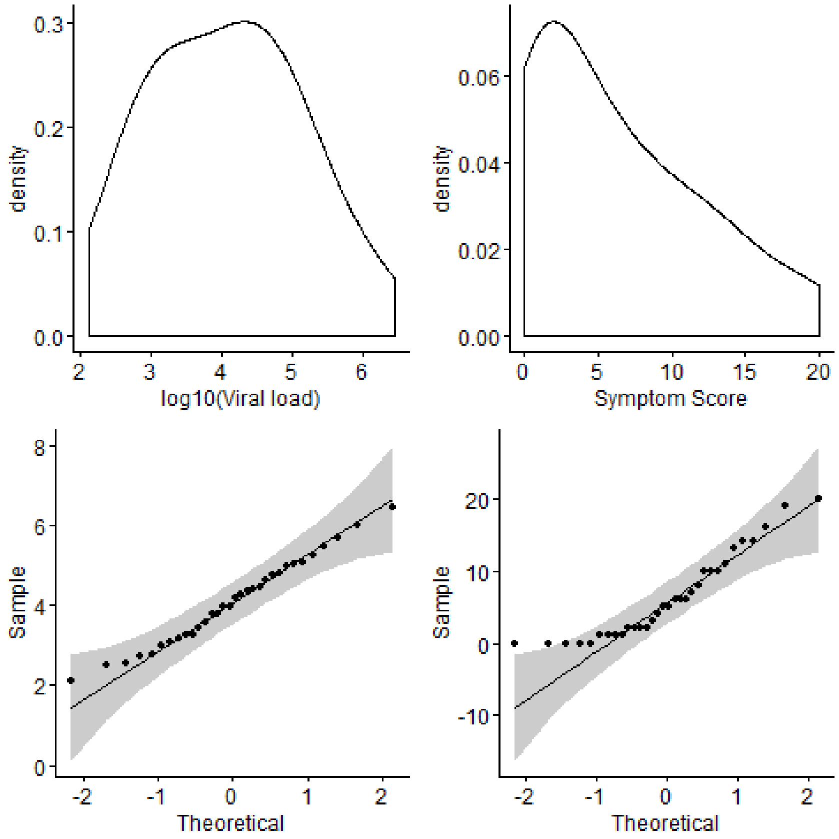


Figure 7.1: Density plots (top) and Q-Q plots (bottom) for $\log_{10}$(VL) and symptom scores on the left and right respectively showed that the transformed viral load data were normally distributed, but the symptom scores were not.

The $\log_{10}$-transformed data were used to calculate the correlation using the Spearman's Rank Correlation test. The Spearman's Rank Correlation test was selected since the symptom score data were not normally distributed and therefore the Pearson Correlation test could not be used. A significant (p = 0.036), positive correlation of 0.37 was observed between viral load and symptom scores after the Spearman's Rank Correlation test in the cohort used for these experiments (Figure 7.2). This implied a link between the symptom scores and higher viral loads.

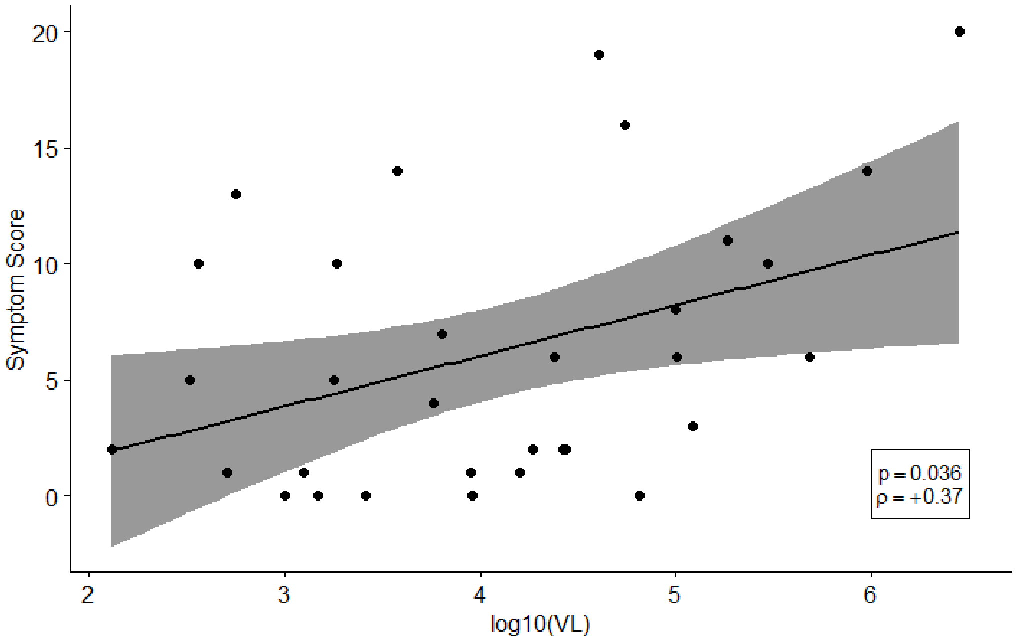


Figure 7.2: The log_{10}-transformed viral loads and symptom scores were significantly correlated. P = p-value of Spearman's Rank Correlation test; ρ = correlation coefficient.

7.2.2. Description of variants present in the inoculum

The inoculum was sequenced to get the exact consensus sequence and to find all variants present in the inoculum before volunteers were infected. The inoculum was processed twice to exclude errors caused by PCR and sequencing. Reads were assembled by BWA as described in 2.3.2.2. and quality was checked as described in 2.3.2.3. This process was carried out for each experiment separately and assembly and quality checks were run again for both fastq files combined. Statistical analysis shows that 986,346 out of 1,046,996 reads (94%) were mapped and paired. The average read quality was 37.7. This was considered a good quality of the sequence. A read quality of 30 translates to 1 error in 1000 nucleotides and a read quality of 40 translates to 1 error in 10,000 nucleotides. Nucleotides with a quality above 35 are considered good in the community. This translates to approximately 1 error in 3000, while 37.7 translates to 1 error in 6000.

There were 9 variations found in the inoculum of which the prevalence ranged from 16.3% to 42.5% (Table 7.3). The variations were seen in three different proteins: P, G and L (Figure 7.3). One of the variations was intergenic: A7485T lies between the F gene and M2 gene and therefore did not affect amino acid sequence of any protein. Three variations were found in coding regions and were synonymous: G2998A, A3007G and A12163T. These variants had no effect on the protein sequence either. The first two (G2998A and A3007G) were located in the P gene and the third synonymous variant was located in the L gene. The amino acids were conserved in their respective proteins: leucine at position 226 (L226), glycine at position 229 (G229) and threonine at position 1233 (T1233).

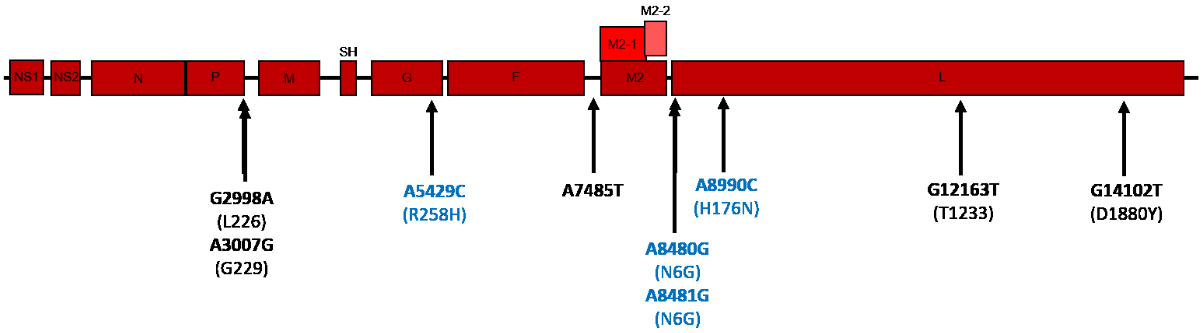


Figure 7.3: Location of variations of inoculum in the RSV genome and their position in their respective proteins. A7485T is an intergenic variant and therefore has no ORF to read the possible amino acid change from. Blue = non-synonymous variation.

The other five variations were non-synonymous. A5429C was located in the attachment gene (G gene) at position 258 and changed from histidine to proline (H258P). Histidine is a positively charged amino acid and proline is a non-polar amino acid. This substitution might therefore affect the protein formation or function, as also suggested by previous studies indicating that this site is under positive selection (224-226).

The four other non-synonymous variations were located in the polymerase (L) gene. A8480G and A8481G together cause a substitution of asparagine to glycine (N6G) with loss of polarity at position 6. The combination of these two variants will be discussed in detail below. A8990C is a substitution that caused the exchange of a positively charged amino acid (histidine) with a neutral, polar amino acid (asparagine) at position 176 in the L gene (H176N). G14102T caused a substitution of the negatively charged amino acid aspartic acid with the neutral, polar amino acid tyrosine at position 1880 in the L gene (D1880Y). This variation was detected in the inoculum at 31.6% prevalence and was not observed in clinical samples from volunteers.

Minority variants with a prevalence of less than 1% were left out of analysis as described in the previous chapter. The minority variants, by definition, never reached 50% or more in the inoculum. However, in later samples, some of these variants did overtake the original consensus bases. The consensus sequence of this inoculum is the same as the published sequence of RSV M37 (35).

Table 7.3: Variations present in the RSV M37 inoculum, their prevalence and the effect on the amino acid sequence.

Position	Reference base	Alternative base	Variant presence (%)	Read depth	Protein	Amino acid substitution
2998	G	A	42.54	4500	P	L226
3007	A	G	27.22	3969	P	G229
5429	A	C	17.27	3331	G	R258H
7485	A	T	27.76	1934	Intergenic	
8480	A	G	24.77	3012	L	N6G
8481	A	G	19.99	2828	L	N6G
8990	A	C	16.34	1386	L	H176N
12163	G	T	26.09	5704	L	T1233
14102	G	T	31.61	3995	L	D1880Y

7.2.3. Characterisation of variations from inoculum present in clinical samples

Out of 9 original variations present in the inoculum, 8 were seen in later samples of the volunteers with their prevalence ranging from 14.6% to 52.5%. G14102T (D1880Y) was the only variation in the inoculum that was not detected in clinical samples from volunteers after inoculation.

G2998A (L226) was seen in four clinical samples of three different volunteers. The prevalence dropped from 42.5% in the inoculum to ~15% in the latest time point of each volunteer (Figure 7.4a). To test if there was any correlation between the prevalence of L226 and time after inoculation, the normality of the data was first tested with the Shapiro-Wilk test. It showed that the time points (day of sampling) and frequency of L226 were not normally distributed. Therefore, the Pearson Correlation test could not be used and the Spearman's Rank Correlation test was used instead. We observed a significant (p = $1.08*10^{-9}$) negative correlation with rho (ρ) = -0.98. It is important to note that there are not many data points for this variant and therefore statistical analysis does not necessarily reflect a correct calculation. Nevertheless, the correlation is strong and it is likely that this variant is reducing in time.

A3007G (G229) was identified in all 10 volunteers after inoculation and in a total of 28 clinical samples spread out over 7 different time points. In the inoculum, this variation was present in 27.2% of reads. SWT showed the data was not normally distributed and therefore the Spearman's Rank Correlation test was used. This resulted in a significant (p=0.007) negative correlation of -0.43 over time (Figure 7.4b).

A5429C (R258H) was detected in a total of 21 clinical samples from 9 different volunteers. It was only undetectable in volunteer #7 of which only one quality sequence was available. The general trend for this variation was an increase in frequency; however, there were also several drops in prevalence between consecutive samples of the same volunteer. In the inoculum, the variation had a prevalence of 17.3% and this increased in all samples to a range of 21.7 to 78.7% (Figure 7.4c). The SWT showed data were not normally distributed and the Spearman's Rank Correlation test revealed that there was a significant ($p = 5.5*10^{-8}$) positive correlation ($\rho = 0.80$) of variant frequency over time.

A7485T is a variation that was seen in 8 out of 10 volunteers in 16 clinical samples, but it is intergenic. Its prevalence was 27.8% in the inoculum and it ranged from 9.2% to 100% in post-inoculation samples. It is located between the F gene and M2 gene. There was no significant correlation over time according to the Spearman's Rank Correlation test (Figure 7.4d).

A8990C(H176N) was seen in 8 samples from 6 different volunteers and ranged from 16.3% in the inoculum to 17.9% to 100% in post-inoculation samples. The data were not normally distributed according to the Shapiro-Wilk test and Spearman's Rank Correlation test showed a significant ($p = 3.782*10^{-7}$) positive correlation ($\rho = 0.90$) over time (Figure 7.4g). It is located in the L gene and changes the amino acid from a histidine to an asparagine. This means the amino acid changed from a positively charged amino acid to a polar, uncharged amino acid.

G12163T (T1233) was prevalent in the inoculum at 26.1% and was detected in 24 clinical samples. Data were not normally distributed, so Spearman's Rank Correlation test was used and it showed a significant negative correlation ($\rho = -0.64$) of frequency over time with a p-value of $3.96*10^{-5}$ (Figure 7.4h). The frequency ranged from 17.4 to 29.9% in post-inoculation samples. This was a synonymous variation.

A8480G and A8481G were located in the same codon and were seen at similar frequencies (Figure 7.4 e and f). They caused a non-synonymous mutation N6G. In the inoculum, A8480G and A8481G were prevalent at 24.8% and 20.0% respectively. Their frequency in post-inoculation samples ranged from 8.8% to 36.4% for A8480G and 11.4% to 39.9% for A8481G. There was no significant correlation of frequency over time according to statistical testing with the Spearman's Rank Correlation test.

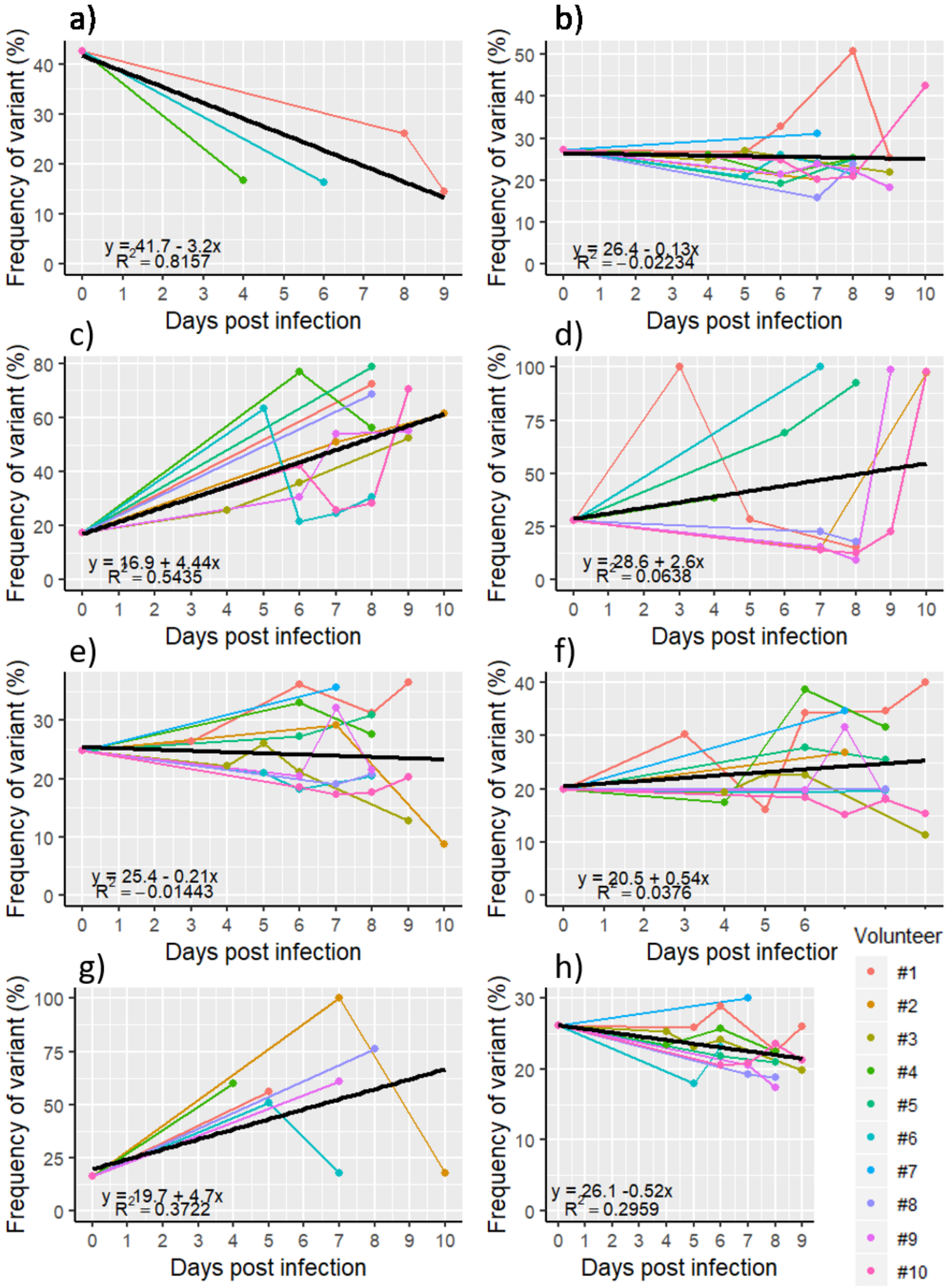


Figure 7.4: Evolution of variant prevalence and frequency over time. The frequency of variants over time with correlation coefficient and p-value added to the graph. (a) G2998A, (b) A3007G, (c) A5429C, (d) A7485T, (e) A8480G, (f) A8481G, (g) A8990C, (h) G12163. Variant G14102T from the inoculum was not seen in later time points. Black line = trend line.

These two variants were found at very similar levels and if both were detected at the same time point, they followed similar trends (Figure 7.5). The most common codon was AAT which translates to asparagine, an uncharged, polar amino acid. If the variations at position 8480 and 8481 both occurred in the same sequence, the codon changed to GGT and switched the amino acid to glycine, which is a neutral, aliphatic amino acid. However, it is also possible that only one variation was present in the same sequence. If the variation at position 8480 is present, but the variation at position 8481 is not, this would change the codon to GAT and switch the amino acid to aspartate, a negatively charged amino acid. If the variation at position 8481 is present, but the variation at position 8480 is not, then the codon would change to AGT, which would switch the amino acid to serine, which is an uncharged, polar amino acid.

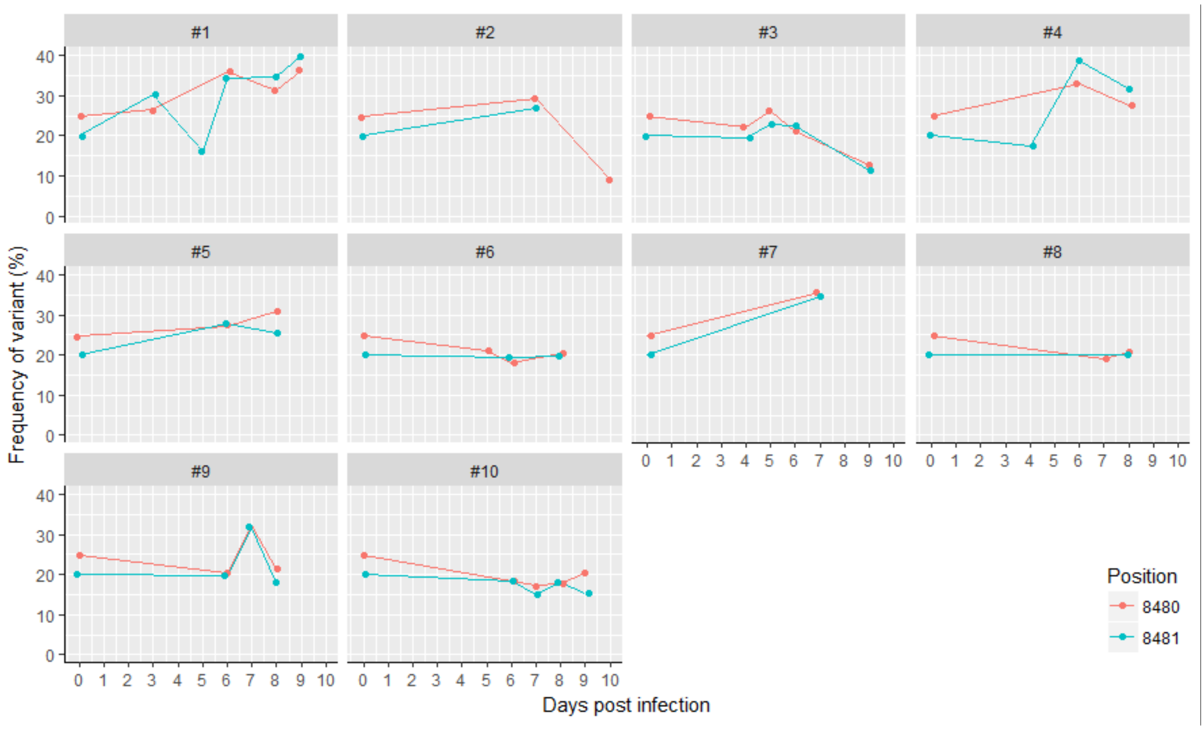


Figure 7.5: Trends of variant frequency of A8480G and A8481G at the sixth codon of the L gene per volunteer show clear resemblance and suggest these variants in the same codon occur together.

Two variants that caused amino acid substitutions, R258H and H176N, seemed to increase in frequency, while other variants either remained present in equal quantities or decreased in frequency.

7.2.4. Characterisation of new variations in clinical samples

Fifty-one *de novo* variations were detected in clinical samples from the volunteers. On average, there were 9.1 variations per sample (ranging from 5 to 23). Of those, on average 4.3 (47.8%) were synonymous variations (ranging from 1 to 18 variants) and 3.6 (39.9%) non-synonymous variations (ranging from 2 to 8 variants) per sample, and 1.1 (12.4%) was intergenic (ranging from 0 to 3 variants).

Thirty out of 51 *de novo* variations appeared only once (one volunteer at one time point). These unique variations were detected at a prevalence of 11.9% to 88.8%. Of those 30 unique variations, 6 were intergenic, 10 were synonymous and 14 were non-synonymous. The non-synonymous variations were seen in NS1, P, M, four times in SH, three times in G and four times in L (Table 7.4).

Table 7.4: Newly emerged variations detected in clinical samples during experimental RSV M37 infection of healthy adults.

Variant	Gene	Amino acid	Volunteer	Days post infection	Frequency
A492C	NS1	*140Y	#3	5	15.75
G765A	NS2	R55K	#1	5	50.00
			#4	4	51.48
C940T	NS2	H113	#1	3	65.84
C1176T	N	S21	#1	3	14.36
			#1	6	24.71
			#7	7	23.84
C1182T	N	Y23	#1	3	41.46
			#1	6	44.68
			#3	5	14.13
			#7	7	45.58
			#10	10	10.20
C1185T	N	T24	#1	3	24.81
			#1	6	34.47
			#7	7	31.88
A2001T	N	G296	#1	5	34.55
			#1	6	44.54
			#4	4	27.27
			#7	7	43.36
G2302A			#3	6	26.09
A2589G	P	D90G	#9	7	33.14
A3099G			#3	5	16.04
			#3	9	25.52
A3218G			#8	8	18.20
A3428G	M	K66	#5	6	16.77
A3863T	M	G211	#1	6	19.63
			#2	7	99.30
			#2	10	21.98
			#7	7	22.01
			#9	7	53.64
C3988T	M	P253L	#8	8	88.80
C4081A			#1	6	15.62
T4280C	SH	N3	#4	4	12.62
T4324C	SH	F18S	#4	4	14.72
T4330C	SH	L20P	#4	4	14.54
T4351C	SH	I27T	#4	4	14.43

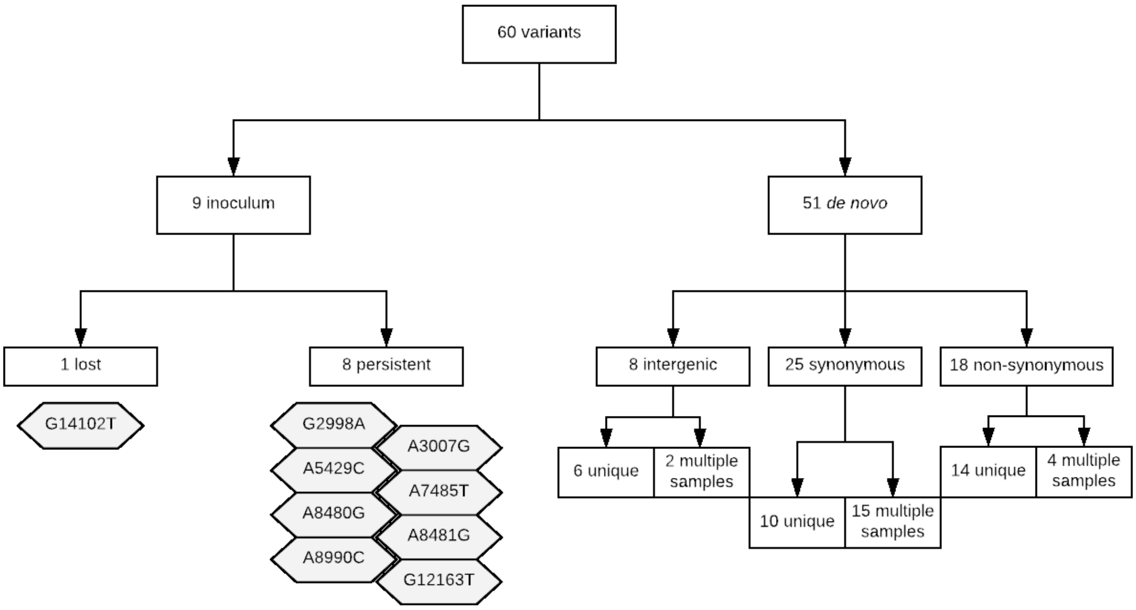
T4363C	SH	L31P	#4	4	16.10
C4628A			#1	6	25.43
			#1	8	29.59
			#7	7	23.62
T4880C	G	I75T	#3	6	15.07
A5129G	G	K158R	#8	7	74.07
T5363C	G	I236T	#10	6	49.28
			#10	7	20.01
			#10	10	36.66
G5440A	G	E262K	#1	9	20.00
A5595G			#10	9	12.44
T6518C	F	L297	#8	7	15.67
C6844T	F	S405	#1	3	15.06
			#1	6	30.18
			#7	7	30.88
G7075A	F	V482	#1	9	25.22
			#4	6	26.16
C7285T	F	A552	#4	6	25.75
A7842T	M2-1	I90	#5	6	62.50
A8031T	M2-1	P153	#6	7	38.36
			#10	10	21.62
C8037T	M2-1	D155	#2	10	42.86
			#5	6	45.00
			#6	7	67.89
			#10	10	52.94
C8230T	M2-2	N33	#1	5	22.73
			#4	4	19.47
C8342T	M2-2	L71	#5	6	25.73
A9102G	L	K213R	#1	6	20.37
			#7	7	14.93
T9184C	L	N240	#1	5	37.56
			#2	10	66.21
			#4	4	42.40
A9445G	L	E327	#1	6	39.22
			#6	6	14.11
			#7	7	24.07
G9448C	L	G328	#1	6	54.17
			#6	6	13.88
			#7	7	61.22
G9826A	L	L454	#2	7	98.82
			#2	10	26.50
T11027C	L	L855	#3	9	11.91
A11140G	L	P892	#5	6	15.11
A11364G	L	N967S	#8	8	85.60
A12313G	L	K1283	#8	8	75.46

C13298A	L	L1612I	#2	7	17.60
A13370G	L	M1636V	#3	4	18.78
A13990C	L	K1842N	#6	6	41.95
			#6	8	21.39
			#9	6	21.39
T14127C	L	I1888T	#3	4	17.22
A15023G			#8	7	34.62
T15049TTTT			#3	6	16.00

The other 21 variations were seen at several time points and in 1 to 4 different volunteers (Table 7.4). Their prevalence ranged from 10.2% to 99.3% and they were spread over 8 genes. 2 of the variations were intergenic (A3099G and C4628A), four variations were non-synonymous (G765A in NS2, T5363C in G, and A9102G and A13990C in L) and 15 were synonymous. Those 15 synonymous variations were located in the N gene (C1176T, C1182T, C1185T and A2001T), M gene (A3863T), F gene (C6844T and G7075A), M2-1 (A8031T, C8037T and A8040G), M2-2 (C8230T) and L gene (T9184C, A9445G, G9448C and G9826A).

The 30 unique variations consisted of 10 synonymous variations (33.3%), 14 non-synonymous variations (46.7%) and 6 intergenic variations (20.0%). The 21 variations that were detected at multiple time points or in multiple volunteers, consisted of 15 synonymous variations (71.4%), 4 non-synonymous variations (19.0%) and 2 intergenic variations (9.5%) (Figure 7.6).

There were 8 genes in which synonymous variations were found (NS2, N, M, SH, F, M2-1, M2-2 and L) and 7 genes in which non-synonymous variations were found (NS1, NS2, P, M, SH, G and L). Non-synonymous variations were most common in SH, G and L, while synonymous variations were most common in N and L (Table 7.4).


60 variants
9 inoculum
51 *de novo*
1 lost
8 persistent
8 intergenic
25 synonymous
18 non-synonymous
G14102T
G2998A
A3007G
A5429C
A7485T
A8480G
A8481G
A8990C
G12163T
6 unique
2 multiple samples
10 unique
15 multiple samples
14 unique
4 multiple samples

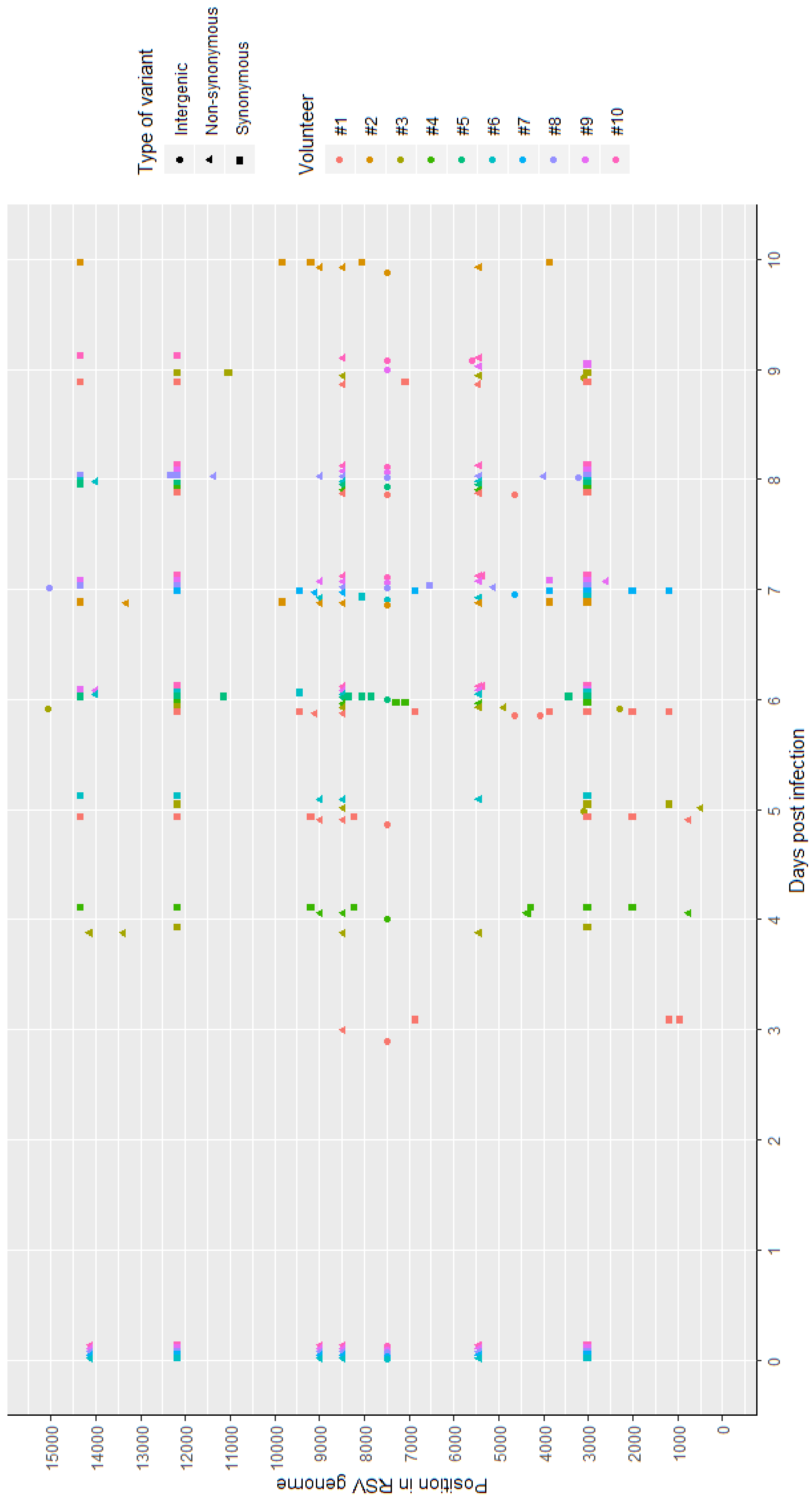


Figure 7.6: Flow chart showing all variants detected in the inoculum and volunteers (top). Overview of all variants over time indicating the type of variant and coloured by volunteer. All variants indicated at day 0 were found in the inoculum (bottom).

In total, most variation were found in the L gene for both synonymous and non-synonymous variants. However, after taking the gene length into account, L is one of the least tolerant proteins towards variations, while SH and G are extremely tolerant towards non-synonymous variants, which was expected based on the literature for the G gene. In the case of the SH gene, this is all based on one volunteer showing several variants and this has to be interpreted with caution. All non-synonymous variations in G were located in the two hypervariable regions of G. Other proteins carrying non-synonymous variations were NS1, P and M (Figure 7.7).

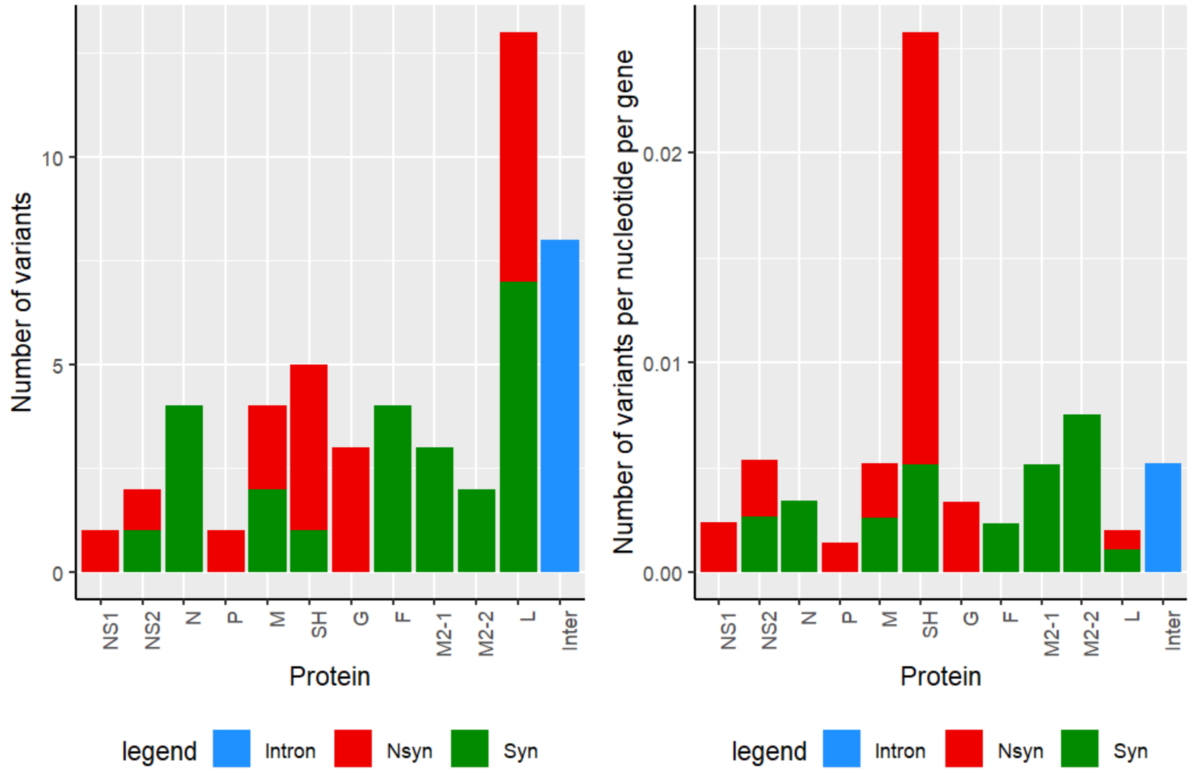


Figure 7.7: Absolute numbers (left) and normalised numbers (right) of variants per gene for both synonymous (Syn) and non-synonymous (Nsyn) variants.

In total, there were 60 distinct variations confidently detected in 32 clinical samples from 10 different volunteers with their prevalence ranging from 8.8% to 100%. Out of 9 variations seen in the inoculum, all but one were common variations in all samples from volunteers at later time points. There were also 51 *de novo* variations detected of which 30 were unique and only seen in one volunteer at one time point. Out of all variations, 9 (15.0%) were intergenic, 28 (46.7%) were synonymous and 23 (38.3%) were non-synonymous variations.

Some of the non-synonymous variants only transiently appeared, while others persisted or increased in frequency.

7.3. Variations in naturally infected infant on Palivizumab

While analysing the samples from infected volunteers over time, samples from infants on prophylactic palivizumab treatment who then got naturally infected with RSV were collected. There were several consecutive samples available from these infants during the same infection. This allowed to study variants in the RSV genome over time in naturally infected patients. Several samples from two infants infected with RSV while receiving prophylactic palivizumab treatment were collected. These samples were sequenced to find out if this was a palivizumab resistant strain or not. The following part of this chapter is the mini-study about one of these two infants.

7.3.1. Characterisation of samples of RSV-infected infants on palivizumab

The first infant was a female between 12 and 18 months of age. Four samples were available over a period of 23 days during January and February of 2018. The C_T values ranged from 18.25 to 24.64 and all samples were nasopharyngeal aspirates. The second infant was a male between 18 and 24 months of age. Two consecutive samples were available which were taken during January of 2018. The C_T values were 35.17 and 34.86 and these samples were also nasopharyngeal aspirates (Table 7.5). These C_T values were too high to sequence, so for this case study, only the samples from the first infant were examined.

Table 7.5: Clinical characteristics of infants on prophylactic palivizumab treatment contracting RSV.

Patient	**Sex**	**Age**	**Sample number**	**Sample date**	**Sample type**	**C_T value**
1	Female	12-18 months	1	20/01/2018	NPA	18.25
			2	22/01/2018	NPA	18.96
			3	29/01/2018	NPA	24.64
			4	11/02/2018	NPA	21.87
2	Male	18-24 months	1	21/01/2018	NPA	35.17
			2	22/01/2018	NPA	34.86

These children would most likely have had some kind of immunosuppressive condition or predisposition to severe disease, since palivizumab is expensive and only given as a prophylactic treatment for children with a heightened risk of RSV infection (236). Both of these infants were part of PERFORM (237), which is a consortium made up of 18 international organisations that aims to study differences between viral and bacterial infections and to improve diagnosis and management of febrile patients struck by infection. More clinical information was not shared about these two patients.

The C_T values of the samples ranged from 18.25 to 35.17 (Figure 7.8). Samples with C_T values higher than 30 were not feasible for whole-genome sequencing. These samples were disregarded and therefore only samples from the first infant with low enough C_T values were requested for sequencing. Of the sample collected on the 29th of January 2018, only a small volume was left.

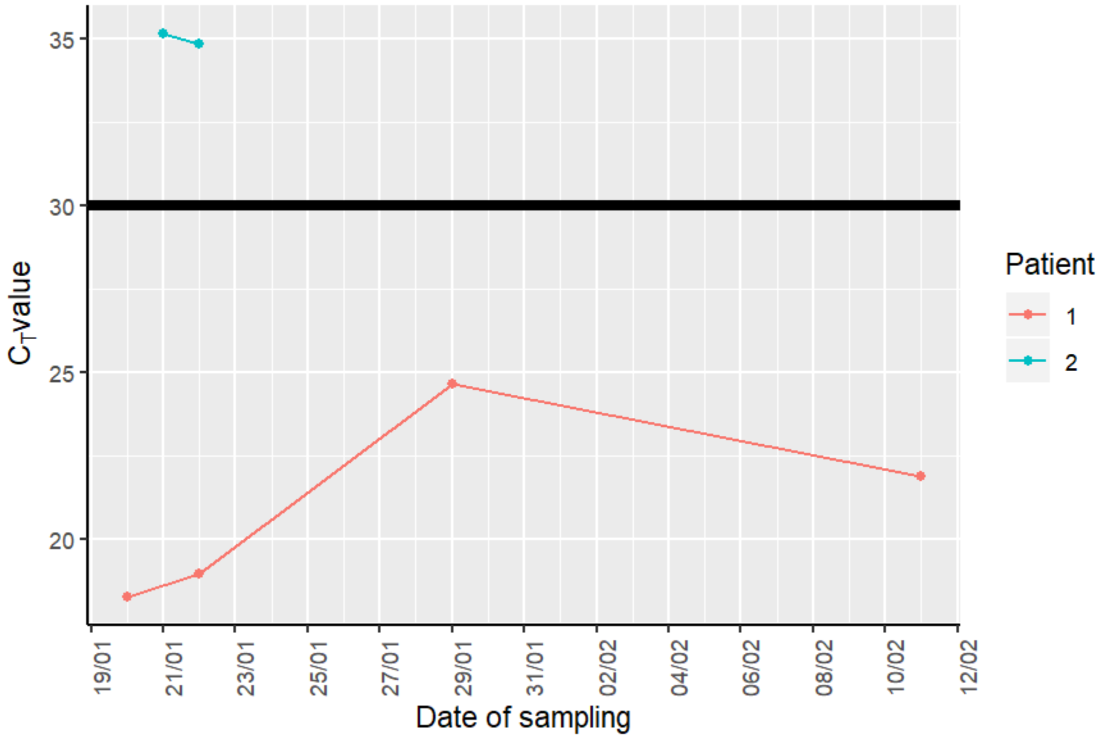


Figure 7.8: C_T values of samples from two patients with RSV infection while on prophylactic palivizumab treatment over a time period of three weeks with the cut-off value for sequencing indicated at $C_T = 30$.

7.3.2. Characterisation and sequencing of RSV strain in samples from infant on palivizumab

The test performed by PICU to detect which viral infection was present did not distinguish between RSV A and B. Therefore, the samples were extracted as described in 2.2.4.2., reverse transcription was run as described in 2.2.5.2 and multiplex PCR was performed on these samples as described in 2.2.6.2. Samples were checked in duplicate and with positive and negative controls. The multiplex analysis showed that only control samples for RSV A were convincing in the red channel. All samples (excluding non-RSV B controls) did reflect a RSV B infection in the yellow channel (Figure 7.9). Human metapneumovirus A and B infection were also excluded.

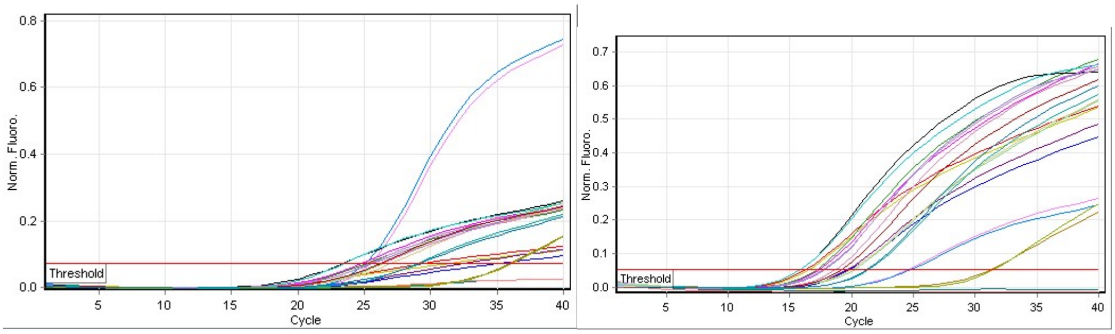


Figure 7.9: Amplification plots of the multiplex PCR to determine the viral subtype of the samples from an infant with RSV infection while on prophylactic palivizumab treatment. RSV A = red channel on the left; RSV B = yellow channel on the right.

Based on the multiplex results, further RSV B analysis was conducted. The reverse transcriptase was repeated, but this time according to the description in 2.2.5.3. and the RSV B genome was amplified in 16 fragments as described in 2.2.7.1. To check the results, 1%-agarose were prepared as described in 2.2.8.1. The bands on the agarose gels showed that for the first two samples all bands were present and visible and this was also the case for all bands but one for the fourth sample. Unfortunately, there was not enough sample left of sample three, which was collected on the 29th of January 2018, to extract enough viral RNA for the entire genome after amplification (Figure 7.10). The part of the genome that contains the palivizumab binding site was not amplified sufficiently and could therefore not be investigated any further. Therefore, three samples were sequenced using the Illumina MiSeq platform according to 2.2.9.2., 2.2.10. and 2.2.12.

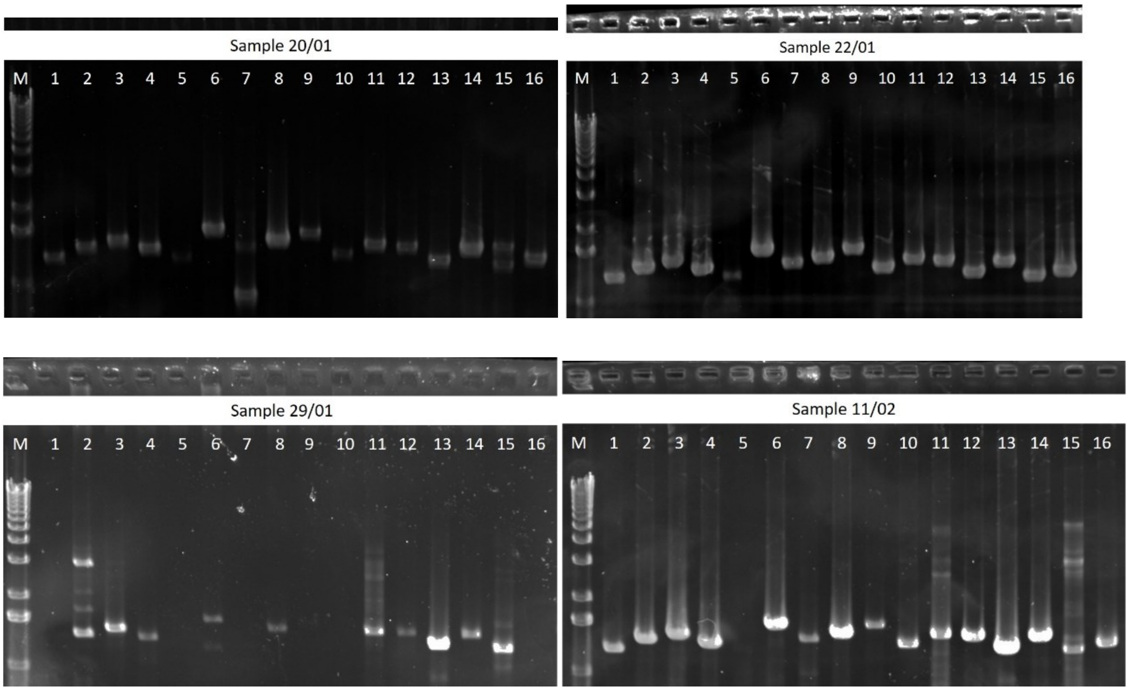

Sample 20/01
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Sample 22/01
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Sample 29/01
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Sample 11/02
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Figure 7.10: Bands per fragment of amplified RSV B genome from infant on palivizumab with persistent RSV infection on 1% agarose gel for four days over a period of one month.

First, *de novo* assembly protocols with SPAdes (2.3.2.1.) were run as the specific strain was unknown and therefore there was no reference strain to assemble reads against. Once a large contig was available, a reference strain (KM517573.1) was found by running BLAST algorithms on the NCBI database. Further assembly against this reference strain was performed as described in 2.3.2.2 and quality checks were performed as described in 2.3.2.3.

The read depth was mapped to check if the whole genome was covered. It ranged between zero and > 8000 (Figure 7.11). Certain parts of the genome were not covered sufficiently for sequence analysis. The insufficiently covered parts were the same in all three sequenced samples.

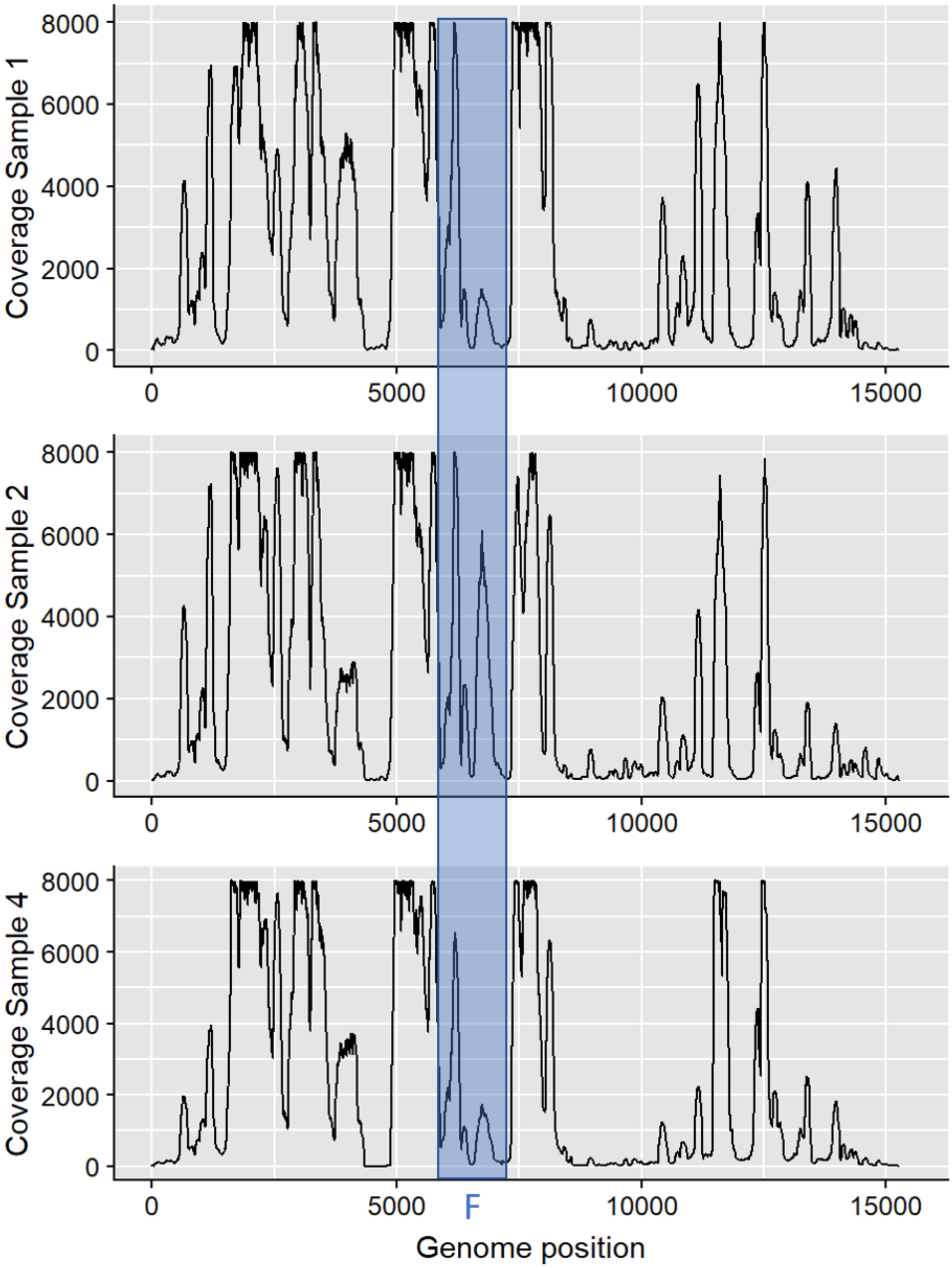


Figure 7.11: Read depth of three samples from an infant that contracted RSV while on prophylactic palivizumab treatment. The read depth of the F gene (highlighted in blue) shows the read depth is not enough extract the sequence of the gene.

7.3.3. In-depth analysis of the RSV genome and its variants

First, the consensus sequence of the sample was used to determine the genotype of this RSV B strain using phylogenetic analysis. BA genotypes were the most prevalent genotype of the RSV B subtype at the time of infection. Therefore, the consensus sequence was aligned against a set of RSV B reference strains that was used in the previous chapter. The alignment was performed using the Clustal Omega algorithm and manually adjusted to obtain the best possible alignment. The quality of the alignment and the number of informative quartets for phylogenetic analysis were assessed as described in the first chapter. The overall completeness score for the alignment was 0.96 and the completeness scores ranged between 0.94 and 1 for individual sequences (Cr), 0.46 and 1 for individual sites (Cc), and 0.93 and 1 for pairs of sequences (Figure 7.12a). The triangle plots for informative quartets showed that 83.5% of quartets were informative and these equally distributed over the three corners of the triangle plot (Figure 7.12b).

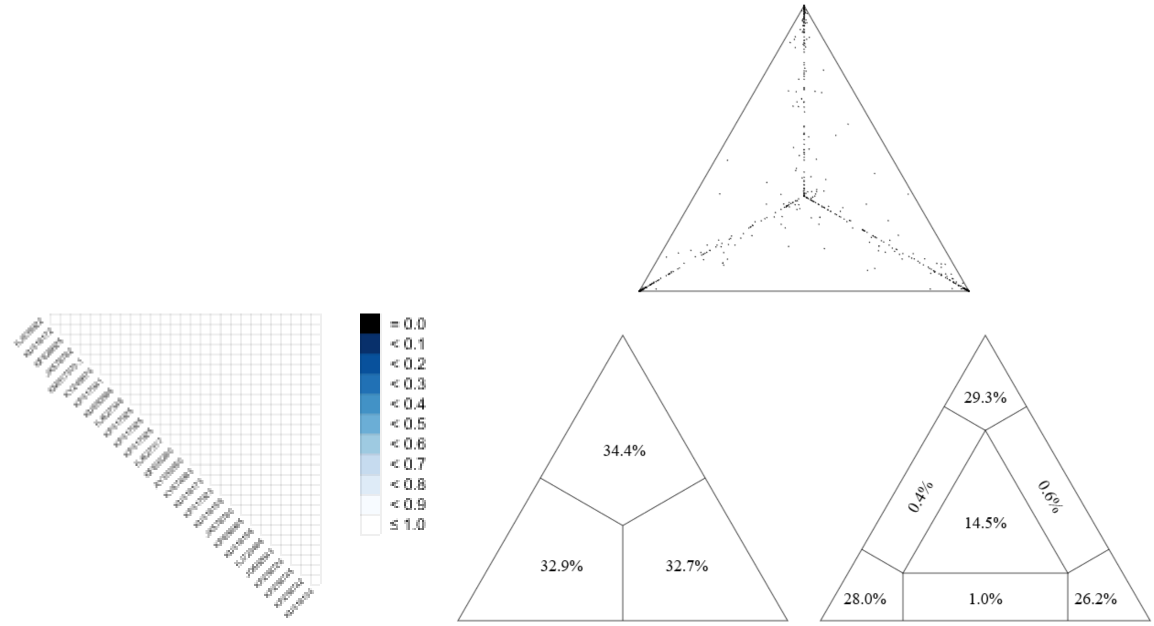


Figure 7.12: Completeness scores for pairs of sequences in triangular heatmap (left) show good quality data. Triangle plots for the alignment run with 1500 quartets show 14.5% of uninformative quartets.

After quality checks of the alignment and usefulness of phylogenetic analysis showed the data sufficed, a maximum likelihood tree was calculated. The best fitting model was determined according to the Bayesian Information Criterion and was HKY+F+R2. The ML tree showed that this strain was part of the BA genotypes, as were most RSV B strains in that season (Figure 7.13).

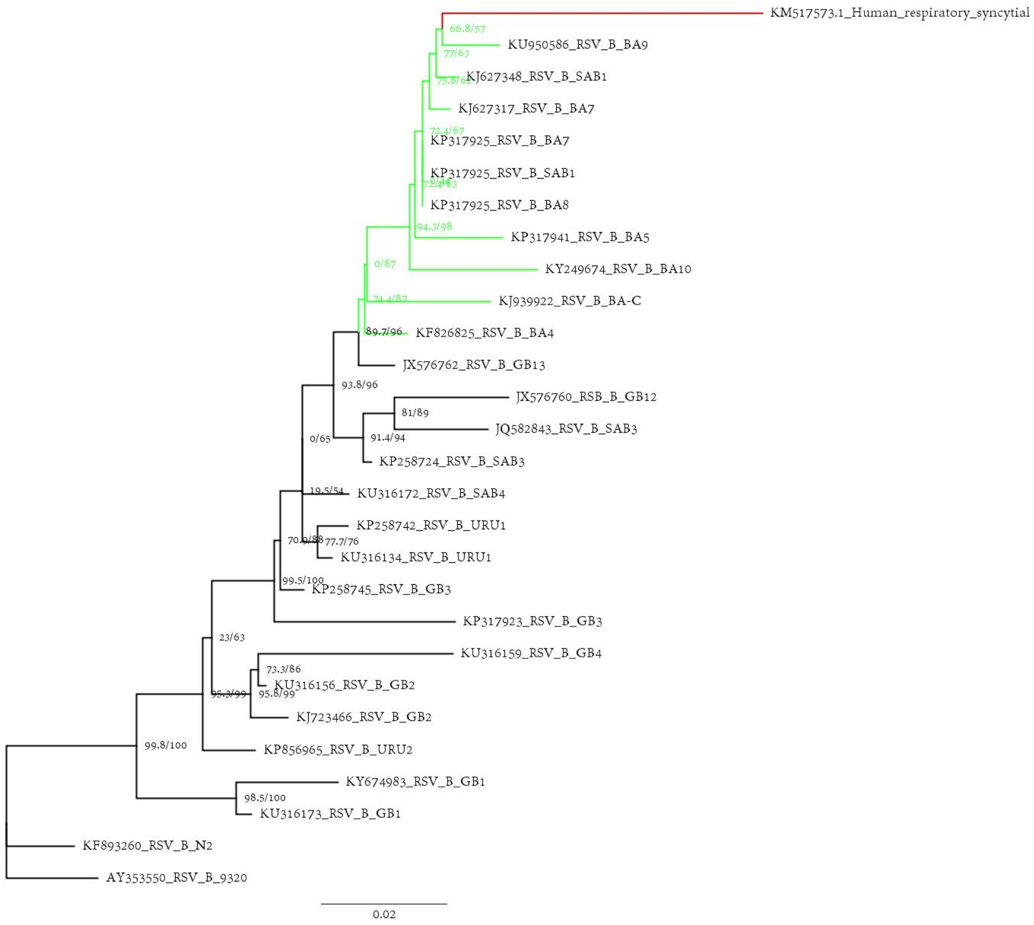


Figure 7.13: Maximum likelihood tree of RSV B genotype references and sample 2 (in red) of the infected infant on palivizumab treatment. Green = BA genotypes.

Secondly, variant calling analysis was performed on the obtained sequences as described in 2.3.2.4. In the first sample, there were four variations and all of them were located in the G gene. In the second sample, there were two variations and they were located in the G gene too, and in the fourth sample, there were 4 variations and two of them were located in the G gene and two of them in the L gene (Table 7.6). Variants with a frequency of less than 1% were left out of analysis.

The positions known to be under positive selection for RSV B are 223 and 224. Position 225 is known as a site under positive selection pressure for RSV A. This study might suggest that position 225 is under positive pressure for RSV B also.

Table 7.6: Variants found in samples from infant infected with RSV while on prophylactic palivizumab treatment.

Sample	Sampling date	Variant	Gene	Amino acid	Frequency (%)
1	20/01/2018	C5306T	G	T206	1.20
		A5362C	G	N225T	1.15
		C5363A	G	N225T	4.51
		C5422T	G	P245L	1.16
2	22/01/2018	A5362C	G	N225T	1.21
		C5363A	G	N225T	3.22
4	11/02/2018	A5362C	G	N225T	2.08
		C5363A	G	N225T	2.06
		C9955T	L	F465	79.17
		A13154G	L	N1532D	76.37

The frequency of these variants was very low throughout the infection, apart from two variants that appeared in the fourth sample and became the most prevalent nucleotide at that position (Table 7.6). Some variants stayed present in all samples over time, while other disappeared or reappeared later. There were two variants in neighbouring positions coding for the same amino acid that were seen in all samples, *e.g.* A5362C and C5363A. Their frequency never reached 5% or more (Figure 7.14).

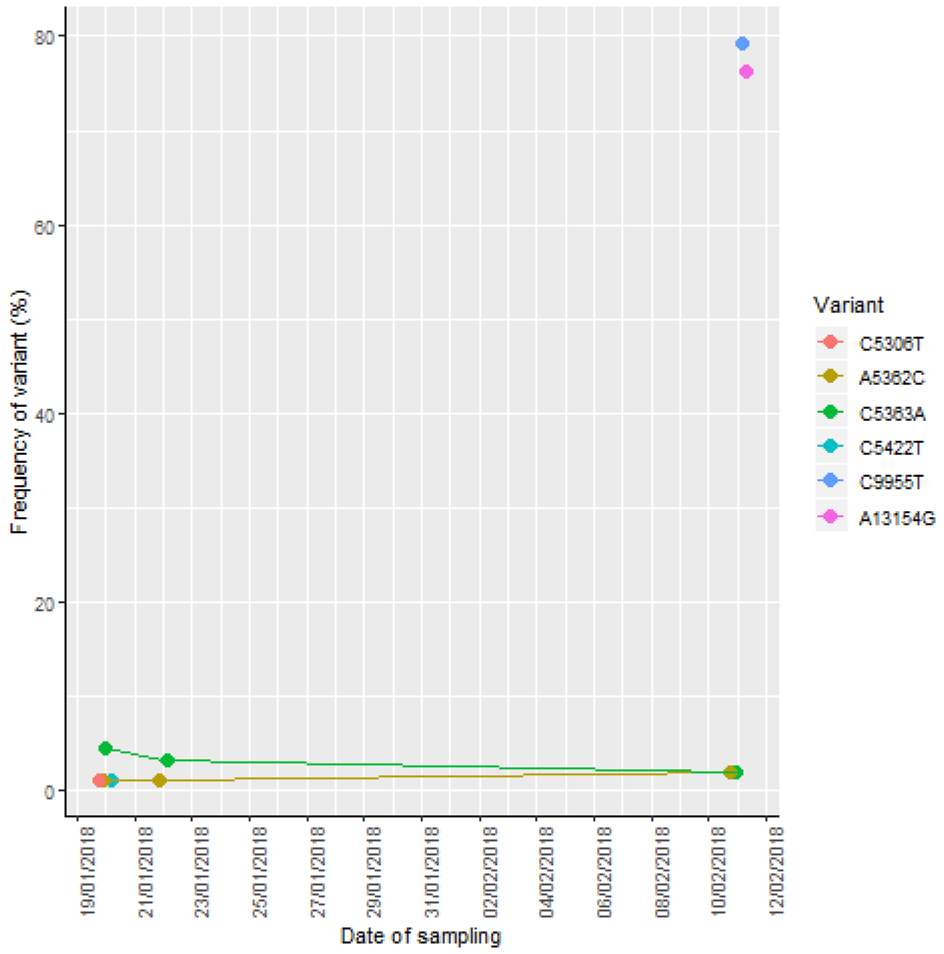


Figure 7.14: Frequency of each variants over time in infant on prophylactic palivizumab shows little fluctuations over time.

7.4. Discussion

In this chapter of the thesis, it was shown that during acute RSV infection, minority variants are present. Variants can be transmitted, but new variants can also arise. On average 9.1 minority variants were detected in each sample. During RSV infection, it is assumed that infection results from multiple virions becoming seeded into the respiratory tract, one or more of which achieves infection. This might be a mix of quasi-species rather than a set of identical viruses.

Several non-synonymous variants were detected in G of which two increased in frequency, *i.e.* R258H and H176N, which might indicate that these variants carry some advantage, while others were not tolerated and disappeared quickly. It has indeed been shown in earlier studies that position 258 is under a positive selection pressure and accumulates more non-synonymous variation than other positions in G (224, 226). However, similar studies with the same inoculum have not seen this positive selection pressure at position 258 of the G gene (171). Certain differences carried by quasi-species can be cost-neutral and therefore persist or they might contain an advantageous trait for transmissibility, replication or virulence. Transmissibility advantages during bottleneck events and development of *de* *novo* variants during acute RSV infection have not been studied in detail before. This study showed that, with inoculation, not all established variations in the inoculum survive the bottleneck of infection, but that most variations do and these variants might even persist during the infection.

Besides transmissibility advantages, minority variants might be better at evading the immune system or might not be recognised as an invader as quickly as virions without that variant. Another potential advantage might lie in the fact that certain variations could increase replication speed or improve infection of neighbouring cells. Other studies on minority variants in RSV are limited, but some have shown the presence of minority variants during infection (35). A study by Agoti *et al.* showed that duplications, deletions non-synonymous and nonsense substitutions are all found as minority variants during natural infection. These alterations to the G gene include larger deletions and premature stop codons confirming that the G gene is prone to errors that are likely to affect viral function (238). Although both synonymous and non-synonymous variants have been detected here, these more extreme variants have not been seen in this study.

Earlier studies on persistent infection in immune suppressed patients have shown that intra-host variation is present, even without a strong immune response, although a bone marrow transplant in this patient did strongly increase intra-host diversity, suggesting that variation in the RSV genome is immune driven. This patient also received palivizumab which might have affected intra-host diversity somewhat, but not as much as the bone marrow transplant (172). This important study highlights the ease with which the RSV genome adapts to its environment by allowing variation to accumulate in its genome.

Detailed studies on the inoculum used here were conducted in earlier years. They found a minority variant, namely N176H substitution in the L gene, with a prevalence of 18%. This was the only minority variant present above 5% in this study (171). In this research, the opposite was noticed, with a 16% prevalence of minority variant H176N. Two variants found in this study became dominant during experimental infection of healthy adults, namely in the M2 gene at position 193 and in the L gene at position 2072 (171). Other minority variants included substitutions at both CD4 and CD8 epitopes and one variant in G known to be in a B cell epitope.

The G14102T variant did not survive the bottleneck, but more research has to be done to be able to investigate the reason for this. The amino acid change D1880Y in the polymerase gene (L) might affect the function of the protein which could explain why no clinical samples show evidence of this variant or virulence might be affected. It should be noted that other studies with the same inoculum have seen this variant in healthy, infected adults (171). It might also be coincidental although this seems unlikely since not one clinical sample from any of the 10 volunteers showed signs of this variant.

Literature shows that L forms a distinct complex with P (155), which might be affected when L is mutated.

Furthermore, *de novo* variations developed during acute infection. Most of these disappeared quickly. Transmission of virions during this stage could lead to the spreading of that variant into the population if it survives the bottleneck of transmission. Further *in vitro* experiments could shed light on the virulence of this type of variant without the presence of immune pressure. Non-synonymous variants tend to disappear within 24 hours, while synonymous variants seem to experience less pressure to be removed from the viral population. This could be due to the development of defective particles which has been shown to happen *in vitro* before (239).

The F glycoprotein rarely develops variants during acute infection and if variants do arise, they are synonymous variants and therefore do not to influence protein function or structure. This might be good news for new F-targeting vaccines being tested knowing the RSV F protein does not alter during acute infection. The other major glycoprotein, the G protein, tolerates more amino acid variation than any other protein. This is to be expected according to the literature and these variations are located in either one of the two known hypervariable regions of G. Both M2 proteins seem to accumulate synonymous variations more easily than any other protein, but did not acquire any non-synonymous variants. L and P seem to tolerate the least amount of variation out of all proteins.

Certain variants can occur within the same codon and affect the amino acid sequence differently when both variants are present compared to one variant. If they are present in similar frequencies, this might suggest they are present in the same quasi-species. Variants that lie close together can be seen in the same reads, but when variants lie further apart in the genome, it is impossible to determine whether they are part of the same quasi-species with short read technologies. Newer technologies can sequence long reads and even complete viral genomes. This could be an interesting research question to find out which variants are always expressed together and what effect that might have on the function or structure of the protein.

This study showed the variability of the genome of RSV even during acute infection in healthy adults, but it also raises more questions. Some variants seem to be advantageous to the virus while others are not. Why is that? Further studies with reverse engineered viruses could show differences in infectivity or replication *in vitro* or *in vivo.* There are no standard models for transmissibility studies of RSV, but setting up new models could help investigate the bottleneck events during transmission. Some experiments have been done with ferrets before as they can sneeze and infect others by droplet transmission (240).

Experiments based on samples from naturally infected infants would be the best starting point to determine the dynamics and effects of variants as these viruses have not been grown *in vitro* without the presence of the immune system or in mice, which are not natural carriers of RSV. These conditions could eliminate or introduce other variants that would never occur in humans and are therefore not the most useful to investigate, but they could show which changes in which regions have certain effects on the function of the protein or its structure.

Unfortunately, it was not possible to retrieve a good sequence of the complete F gene in these samples. It has been shown before that certain amino acid substitutions can influence palivizumab sensitivity. All these variants are found in the antigenic site II of the F gene, which contains the palivizumab-binding site. Specific variants that have been linked to decreased sensitivity are substitutions are position 816, 827 and 828 which cause (partial) resistance by substitutions in the protein at position 268 or 272 (232, 233). Substitutions are position 276 are also linked to palivizumab resistance (231). Another site which has been suggested as having an influence on palivizumab sensitivity is at position 526, where the substitution of methionine by isoleucine is thought to induce palivizumab resistance (232). However, a surveillance study by Devincenzo *et al.* in 2004 has not detected these substitutions occurring naturally (234). Further attempts with different primer sites might have helped to determine the sequence of F in theses samples and could have shown the presence of palivizumab resistance variants or the lack thereof.

It has been shown in the experiments on samples from a naturally infected infant though that it is difficult to cover the entire genomic sequence on the first sequencing round. This might be due to the fact that the exact strain is not known and primers should be adapted to the sequence to be able to retrieve the entire sequence. Therefore, developing new primers and re-sequencing these samples could provide much more information.

One of the (dis)advantages of protein sequence changes could be an impaired protein function due to structural changes. Not all protein structures of RSV have been solved yet, so it is difficult to predict function changes. Further X-ray crystallography experiments could lead to solving the structures of all proteins, however, some RSV proteins are notoriously difficult to keep in the same structural formation. Once the structures have been solved, this would become a lot easier to predict the change in function with a single amino acid change compared to the wild-type sequence.

With this type of study, it is important to keep the limitations of methods in mind. PCR errors are difficult to distinguish from real variants and sequencing is not perfect, so sequencing bias has to be accounted for. Several of those errors can be removed via bioinformatics quality checks, but these methods are not fool proof. Furthermore, it is fairly sure not all variants have been detected as only a part of the collected sample was used for sequencing and sampling itself might not be enough to obtain all quasi-species at once. If variation is located in the primer binding-sites, this might also prevent amplification and therefore would reduce the ability to detect variants in these sites or introduce variants that are not actually present in the sample. This study was therefore not focussed on obtaining a complete variant overview, but on detecting the presence of variability during acute infection over time. This is in line with an earlier paper by Devincenzo *et al*. (35).

By combining more techniques and expanding our current protein structure knowledge, these dynamics can be explored in more detail to improve our understanding of the recurrence of RSV infections. Samples from this study should be re-sequenced and more samples will need to be tested to confirm the dynamics seen in this study and studies of naturally infected children over time could show even more accurate information on infection dynamics. Constantly changing the genome in strategic places might prevent the host immune system from developing lasting immune responses against the virus. The effect of variants can always be studied *in vitro* and/or *in vivo* by testing antibody reactivity, neutralisation assays, developing recombinant viruses, protein conformation studies, using infection models such as mice or transmission models such as ferrets.

8. General discussion

8.1. Comprehensive bioinformatics analysis

The first publication detailing the division of RSV strains into groups was in 1966 by Coates *et al.* as they showed a difference in antibody reactivity (71). The current genotyping system is based on the genome and was described in 1998 (82), although genetic diversity had been shown before that point too by restriction mapping, digestions with RNase and electrophoresis, and sequencing (78, 79, 81, 84, 202). Over time, more genotypes were discovered and more genotype names were appointed. However, it was not a standardised process to decide when a new strain was part of a new genotype or how to name that genotype. In the first papers, partial sequences were used for genotyping and this was kept as the standard, although this was not always the same part of the genome. The partial sequence of G that was used in the first papers is too short to determine genotypes currently. This part of the genome does not carry enough information to base confident phylogenetic analysis on. This raises the question whether we should stick to this partial G gene to determine genotypes.

Besides this issue, genotypes have been appointed based on reference strains of choice from other genotypes. This way, new genotypes might have been appointed by one research group, while another research group would not have seen separate clustering in their phylogenetic tree, solely because they used different reference strains. This raises a second question: should we keep all previously appointed genotypes as they are? Should we establish a standard method for genotyping RSV strains and re-analyse all known genotypes so far with this new standard? Perhaps a new system based on current and future RSV strains would be more useful than the current system. It could prove to be difficult to find all necessary information from previous strains that are no longer around to include them in the new system.

Which part of the genome is necessary to determine genotypes if partial G sequences are not enough? This thesis discusses the option of using full length genomes, and, although this would not be a cost-effective measure, this would provide the most information. However, full length genomes might not be necessary to determine genotypes. The variability of each gene was therefore determined based on all available full-length genomes in the NCBI database. The background variability of both L and F genes are noteworthy, however, G carries the most variability by far. The HVR2 is the most variable region as was established in earlier studies. It is also this part of G that carries the 72- or 60-nucleotide duplication in ON1 and BA strains for RSV A and B respectively. This indicates that G would be the best gene to base genotyping efforts on.

To determine how much of the genome was necessary to determine the correct genotypes, phylogenetic trees of full genomes were compared to trees based on full G genes and partial G genes.

This showed that full G genes showed a similar structure as full genome trees, albeit with lower bootstrap values and branching support. However, when comparing these trees to the trees based on partial G, more specifically HVR2, this showed another tree structure with very low support for most branches. This strengthens the hypothesis that this region is not informative enough to confidently determine genotypes using phylogenetic analysis.

However, F would be a viable candidate to base genotyping on as well. It is a larger gene than G, but contains less variation than G. There are also fewer sequences available in online databases as G has always been targeted for genotyping purposes. An advantage of basing genotypes on F would be the fact that most current vaccines are targeting F. If responses of vaccines are different based on the differences in the F protein, it would be crucial to have a system to divide strains into groups based on F protein variability. Then again, it will not be an easy shift to introduce into the RSV community. Every laboratory would have to re-evaluate their genotyping methods and adjust them to target F and all genotypes determined so far might become useless.

Based on the previous analyses, a reference dataset was set up containing G gene sequences from most known genotypes or alternatives that were most similar if full G was not available. A standard reference dataset like this would be a good tool to standardise genotyping efforts around the world. This dataset would have to be determined and intensively tested by a panel of experts and would have to be revised and tested periodically or after big impact discoveries.

In this thesis, the reference dataset was used to test the capabilities to determine genotypes from clinical samples based on G genes only. Sequences of 116 clinical samples were determined of which 56 were RSV A strains and 60 were RSV B strains. Maximum likelihood trees were built based on RSV A strains only, RSV B strains only or a combined dataset of both RSV A and B strains. Phylogenetic trees were better supported when RSV A and B strains were inspected separately. This analysis showed that clinical strains from seasons 2014-2015 to 2017-2018 clustered mostly with a cluster containing GA2, TN1, TN2, ON1 and ON2 strains. Only 3 RSV A strains did not carry the 72-nucleotide duplication in the G gene. Two of those belonged to the GA5 genotype. All RSV B strains clustered in the BA cluster and contained the 60-nucleotide duplication in the G gene.

The level variation in RSV is striking and what genotyping is based on. However, it is still unclear how much variation is present during an acute infection and when new mutations arise. This was studied in further chapters of this thesis.

8.2. Viral cell culture experiments

First, cell culture experiments were optimised. The time of harvest affects virus yields and the best point in time is when the cytopathic effect is greatest, while the number of dead cells is at its lowest. Temporal optimisation experiments showed that the best time point that was tested in these experiments was 55 hours post infection.

Once experiments were optimised, clinical-like samples were tested to be grown in culture. A known virus strain was used to infect cell cultures in the form of nasal lavage samples in different concentrations. To be able to cause visible cytopathic effects and detectable viral titres, a C_T value of 30 or lower was necessary.

These methods could be used to grow clinical samples containing RSV strains with distinctive genome characteristics. Cell culture experiments could then be used to investigate the replication rate and the cytopathic effects of that specific strain as well as the timing of these processes. However, clinical samples were not tested in cull culture yet.

8.3. Set-up of deep-sequencing methods

The best standard operating procedures should be established before new experiments are started. Each step should be optimised to have the best protocol that can be used throughout all experiments. When optimising protocols, the practical possibilities should not be forgotten. The theoretically best protocol is not necessarily feasible regarding financial or time aspects or for large-scale experiments. In this thesis, the aim was to increase cDNA yield as much as possible to identify minority variants in clinical samples. The tests conducted were aimed at enzymes and sample volumes.

The enzymes that were compared were PfuUltra II Fusion HS DNA polymerase and PlatinumTM Taq DNA polymerase High Fidelity of which the first resulted in more and brighter bands on agarose gels, which is equivalent to higher amounts of (c)DNA being present.

Further research investigated the effect of sample volumes used in each step of the process. Different volumes of clinical sample were used for total nucleic acid extraction. Between 25µl and 100µl of sample, the latter resulted in the most feasible protocol. The tested volumes used for conversion of viral RNA to cDNA were 15µl and 42µl and again the latter turned out to be the best option. Finally, the volume of cDNA used for amplification of fragments was tested. Both 3µl and 10µl returned good yields, but 3µl of cDNA was more feasible as 16 amplification reactions were to be performed.

A second primer set was tested as well, however, this produced lower viral yields, so the choice fell on the first set of primers. Further optimisation could be done by combining the strategy of the second primer set, where multiple primers at similar positions were used in the same amplification reaction in combination with the primers from the first set. This could increase yields in regions which are prone to genomic variation.

The optimised protocol was successful and full genomes of RSV could be determined by deep-sequencing clinical samples. Viral loads were high enough to detect minority variants in these samples as well.

8.4. Testing of clinical samples from community and hospital patients

Newly established methods were used to deep-sequence RSV positive samples from patients with severe disease and mild disease. The purpose of deep-sequencing the samples was to determine the presence of minority variants and measure the genetic variance in these samples to compare samples from patients with mild and severe disease. Disease severity might be caused by high genetic variance in the viral genome or severe disease might be a favourable condition for the virus to thrive and to drive increased viral population and therefore increased variation.

In this thesis, a cohort of hospitalised patients was compared to a cohort of community patients. To reduce the risk of bias, samples were spatiotemporally matched and the viral load between the two inspected cohorts was studied, which showed no difference.

The samples from both cohorts were deep-sequenced and the dataset, containing all clinical samples and reference sequences as determined earlier in this work, was quality-checked for phylogenetic analysis. Completeness scores were calculated for three datasets, *i.e.* one with only RSV A sequences, one with only RSV B sequences and a combined dataset for RSV A and B. The last one could be useful when samples are sequenced of which the subtype is unknown. Likelihood mapping analysis showed that the combined dataset was by far the worst dataset. Indicative parsimony trees showed that all sequences belonged to the RSV A subtype. This was confirmed after a maximum likelihood tree was built. The genotypes were shown to be GA2 or ON1 genotypes. This is in line with the current genotypes, which are GA2/ON1 and GA5 for RSV A and BA genotypes for RSV B.

The assembly of unknown RSV strains proved to be difficult. First, *de novo* assembly was performed, which only returned partial sequences. The largest contig was used to BLAST against the online NCBI database. The most similar strain was then used as a reference strain for a second assembly round. This returned more complete sequences. A standardised reference sequence could have been a useful tool at this point of the analysis. A strain with ambiguous nucleotide annotation at varying positions could be used to assemble *de novo* sequences. Nevertheless, this is not available and would be flawed as well as unknown or new variations might not be mapped and therefore lost.

Further in-depth research into the variations found within each clinical sample showed that the number of variants was highest in the L gene, but when normalised for gene length G carried the most variants per nucleotide. G also contained the most non-synonymous variations out of all genes and this was to be expected.

When comparing the amount of variation between community samples and hospital samples, the Shannon Entropy showed no significant difference between the two cohorts and was therefore not correlated with disease severity in this study. However, numbers were low in these tests and type II statistical errors are possible in this analysis. Further research might reveal more subtle differences than the ones that could be detected with this small-scale study.

Overall, there seems to be no clear association between RSV variants or strains and disease severity. However, most clinical samples do seem to carry minority variants.

8.5. Deep-sequencing of RSV over time and minority variant dynamics

It was shown in the previous chapters that variation within a sample is present. It was unclear whether these were transmitted during infection or whether these developed during acute infection. In the last part of this thesis, healthy volunteers were inoculated with a known RSV strain and daily samples were taken. These samples were deep-sequenced to investigate the bottleneck of transmission and detect the origination of new variants.

The inoculum was sequenced to determine which minority variants might be present in the starting material. Several variants were detected of which one non-synonymous variant did not survive the bottleneck of transmission via inoculation. Two non-synonymous variants increased in frequency over time. These might have some advantage over the more prevalent variant at that position. Other non-synonymous variants did not increase or decrease in frequency over time. There did not seem to be any significant change over time in synonymous variants. Certain differences carried by quasi-species can be cost-neutral and therefore persist or even be advantageous for transmissibility, replication or virulence.

New variations also developed during acute infection, although most of these disappeared quickly. These variants might be disadvantageous or the quasispecies might not have replicated enough to persist due to other factors. Synonymous variants seemed to experience less pressure to be removed from the population than non-synonymous variants. Non-synonymous variants can cause the development of defective particles, which is not the case for synonymous variants and might be an explanation for these dynamics.

It was shown that the F protein rarely develops non-synonymous variants during acute infection. This is excellent news as several vaccines are being tested targeting the F protein. The M2 gene which codes for both M2 proteins, seemed to accumulate variants quite easily, although these were all synonymous variants. The L and P proteins were least tolerable to *de novo* variant development in this study.

Further experiments on samples from a naturally infected patient receiving palivizumab treatment showed similar dynamics being present. Certain variants arose during infection and others were already present and persisted over time. Further experiments of naturally infected patients being sampled over time could provide a lot more insight into *de novo* development of variants during RSV infection and the effects on disease severity.

Protein conformation analysis should be performed to identify the effect of the variants found in this study on protein folding and perhaps function. Our current knowledge on protein structures should be expanded to inspect the effect of these dynamics in more detail and improve our understanding of disease severity and related biomechanisms.

8.6. Future work

8.6.1. Future research in bioinformatics analysis

The comprehensive study on sequences collected from the NCBI database was limited to the G gene in this thesis. Further inspection of the F gene could elucidate if the F gene could be used as the basis for genotyping. Large, encompassing maximum likelihood trees could be built containing strains from all over the world and from a large timeframe.

This research focussed on trees based on G gene nucleotides which could be compared to amino acid trees of G. All synonymous variations would be taken out of the equation which would reduce the variability. The trees of full genomes, full G and HVR2 could be compared to each other in a similar fashion as shown before to investigate if these trees are more useful for genotyping endeavours.

Similar research could be carried out focussing on the F gene. Trees based on full genome and full F gene could be compared for both nucleotide and amino acid trees to find out if F would be a viable candidate for genotyping. This might be useful for vaccine-related research as well. G- and F-based trees could be compared to determine whether they have similar outcomes. This research could be applied to all genes, although G and F seem to be the most informative ones.

The analysis described above will be carried out in the coming months and will be submitted for publication afterwards.

When it comes to studying disease severity, the disruption of cell homeostasis can be detected by single-cell sequencing. Different types of cells can be divided into infected and non-infected groups. Single-cell sequencing can determine the up- or downregulation of all of their genes compared to each other. It might even be possible to train artificial intelligence software to find the differences between infected and non-infected cells, mild disease and severe disease to detect genes that are dysregulated in such circumstances. This could provide advanced insights into new therapeutic strategies.

The expansion of the groups that are included in research could increase the understanding of differences in RSV disease. Babies, young children, healthy or immunocompromised adults and elderly people all have different immune systems, which could affect disease severity. Where healthy adults can carry RSV asymptomatically, people at the extremes of ages are often severely affected, but it is still unclear why that is. This extra information could be included in research conducted using artificial intelligence to gain further insights.

8.6.2. Future research in experimental analysis

8.6.2.1. Host genome effect

It is still unclear if viral strain variation contributes to disease severity. The influence of the host genome could theoretically be studied by identical twin infection studies, although this would be extremely difficult to organise and would be a work of decades. This type of study would not take the epigenome into account either, which could also play a role in disease severity. However, if there are differences found between identical twins infected with the same virus at the same time in the same environment, it would suggest that epigenomic differences and infection history are the drivers behind disease severity.

8.6.2.2. Viral genome effect

When specific variants are encountered, it would be enlightening to be able to study the effect of that variant with *in vitro* or *in vivo* experiments. Reverse engineering a virus containing only those variants needed for the investigation could produce enough virus to investigate the specific effect on cytokine production in cell culture or in mice. Further transcriptional changes in the cell could be detected by single-cell analysis of samples collected during these experiments.

Further research should be conducted to elucidate the conformational analysis of all proteins with and without mutations, which could help to fine-tune theoretical predictions of protein conformation changes. Artificial intelligence could also play a role in this part of RSV research. Once the software has learned which conformations are physically possible, sequences could be fed to the software that would run an algorithm and return the conformational effect resulting from mutations in the viral genome.

8.6.3. Future directions

Overall, artificial intelligence could teach us a lot if we combined all the data we can collect. The combination of disease severity, complete virus genome, complete host genome, host microbiome, host infection history and even single-cell data could be the key to discovering the cause of severe disease caused by RSV. It would be an enormous project that could only be accomplished by several international research institutions working together, but it might be worth to collect the wealth of information such an endeavour would produce. It would certainly be an innovative project bringing the future of research to the current day and age.

8.7. Conclusions

a. Genotyping RSV is currently based on an inadequate part of the genome and not standardized. Working groups have been established to address this problem. It might even be better to use a different part of the genome all together. Further investigation might shed further light on that notion.

b. Full-genome deep-sequencing and minority variant identification are possible with the methods established in this thesis.

c. RSV infected patients are likely to carry minority variants, regardless of disease severity. Further research could elucidate if the amount of minority variants and their frequency is different between both cohorts.

d. *De novo* minority variants frequently emerge during acute RSV infection in healthy adults.

Bibliography

1. Glezen WP, Taber LH, Frank AL, Kasel JA. Risk of primary infection and reinfection with respiratory syncytial virus. American journal of diseases of children. 1986;140(6):543-6.

2. Taylor S, Taylor RJ, Lustig RL, Schuck-Paim C, Haguinet F, Webb DJ, et al. Modelling estimates of the burden of respiratory syncytial virus infection in children in the UK. BMJ Open. 2016;6(6):e009337.

3. Mazur NI, Higgins D, Nunes MC, Melero JA, Langedijk AC, Horsley N, et al. The respiratory syncytial virus vaccine landscape: lessons from the graveyard and promising candidates. The Lancet Infectious diseases. 2018;18(10):e295-e311.

4. Shi T, McAllister DA, O'Brien KL, Simoes EA, Madhi SA, Gessner BD, 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. The Lancet. 2017;390(10098):946-58.

5. Stein RT, Bont LJ, Zar H, Polack FP, Park C, Claxton A, et al. Respiratory syncytial virus hospitalization and mortality: systematic review and meta-analysis. Pediatric pulmonology. 2017;52(4):556-69.

6. Nair H, Nokes DJ, Gessner BD, Dherani M, Madhi SA, Singleton RJ, et al. Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. The Lancet. 2010;375(9725):1545-55.

7. Nair H, Brooks WA, Katz M, Roca A, Berkley JA, Madhi SA, et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. The Lancet. 2011;378(9807):1917-30.

8. Fleming DM, Taylor RJ, Lustig RL, Schuck-Paim C, Haguinet F, Webb DJ, et al. Modelling estimates of the burden of Respiratory Syncytial virus infection in adults and the elderly in the United Kingdom. BMC Infect Dis. 2015;15:443.

9. Wilkinson TM, Donaldson GC, Johnston SL, Openshaw PJ, Wedzicha JA. Respiratory syncytial virus, airway inflammation, and FEV1 decline in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2006;173(8):871-6.

10. Falsey AR, Hennessey PA, Formica MA, Cox C, Walsh EE. Respiratory syncytial virus infection in elderly and high-risk adults. New England Journal of Medicine. 2005;352(17):1749-59.

11. Collins PL. Respiratory syncytial virus and metapneumovirus. Fields virology. 2007:1601-46.

12. Jha A, Jarvis H, Fraser C, J.M. Openshaw P. Respiratory syncytial virus. 2016:84-109.

13. Openshaw PJ, Tregoning JS. Immune responses and disease enhancement during respiratory syncytial virus infection. Clin Microbiol Rev. 2005;18(3):541-55.

14. Culley FJ, Pollott J, Openshaw PJM. Age at First Viral Infection Determines the Pattern of T Cell–mediated Disease during Reinfection in Adulthood. The Journal of Experimental Medicine. 2002;196(10):1381-6.

15. Zhivaki D, Lemoine S, Lim A, Morva A, Vidalain P-O, Schandene L, et al. Respiratory syncytial virus infects regulatory B cells in human neonates via chemokine receptor CX3CR1 and promotes lung disease severity. Immunity. 2017;46(2):301-14.

16. Hall CB, Walsh EE, Long CE, Schnabel KC. Immunity to and frequency of reinfection with respiratory syncytial virus. Journal of Infectious Diseases. 1991;163(4):693-8.

17. Ascough S, Paterson S, Chiu C. Induction and subversion of human protective immunity: contrasting influenza and respiratory syncytial virus. Frontiers in immunology. 2018;9:323.

18. Acosta PL, Caballero MT, Polack FP. Brief history and characterization of enhanced respiratory syncytial virus disease. Clin Vaccine Immunol. 2015.

19. Iversen MB, Reinert LS, Thomsen MK, Bagdonaite I, Nandakumar R, Cheshenko N, et al. An innate antiviral pathway acting before interferons at epithelial surfaces. Nature immunology. 2016;17(2):150.

20. Spann KM, Tran KC, Chi B, Rabin RL, Collins PL. Suppression of the Induction of Alpha, Beta, and Gamma Interferons by the NS1 and NS2 Proteins of Human Respiratory Syncytial Virus in Human Epithelial Cells and Macrophages. Journal of Virology. 2004;78(8):4363-9.

21. Collins PL. Human Respiratory Syncytial Virus. In: Mahy BW, van Regenmortel MH, editors. Encyclopedia of Virology. Third Edition ed: Elsevier Ltd.; 2008. p. 542-50.

22. Swedan S, Musiyenko A, Barik S. Respiratory syncytial virus nonstructural proteins decrease levels of multiple members of the cellular interferon pathways. Journal of virology. 2009;83(19):9682-93.

23. Legg JP, Hussain IR, Warner JA, Johnston SL, Warner JO. Type 1 and type 2 cytokine imbalance in acute respiratory syncytial virus bronchiolitis. Am J Respir Crit Care Med. 2003;168(6):633-9.

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

25. Gonzalez PA, Prado CE, Leiva ED, Carreno LJ, Bueno SM, Riedel CA, et al. Respiratory syncytial virus impairs T cell activation by preventing synapse assembly with dendritic cells. Proc Natl Acad Sci U S A. 2008;105(39):14999-5004.

26. Habibi MS, Jozwik A, Makris S, Dunning J, Paras A, DeVincenzo JP, 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(9):1040-9.

27. SIMOES EA, CARBONELL-ESTRANY X. Impact of severe disease caused by respiratory syncytial virus in children living in developed countries. The Pediatric infectious disease journal. 2003;22(2):S13-S20.

28. Homaira N, Mallitt KA, Oei JL, Hilder L, Bajuk B, Lui K, et al. Risk factors associated with RSV hospitalisation in the first 2 years of life, among different subgroups of children in NSW: a whole-of-population-based cohort study. BMJ Open. 2016;6(6):e011398.

29. Bont L, Checchia PA, Fauroux B, Figueras-Aloy J, Manzoni P, Paes B, et al. Defining the Epidemiology and Burden of Severe Respiratory Syncytial Virus Infection Among Infants and Children in Western Countries. Infect Dis Ther. 2016.

30. Zeng R, Li C, Li N, Wei L, Cui Y. The role of cytokines and chemokines in severe respiratory syncytial virus infection and subsequent asthma. Cytokine. 2011;53(1):1-7.

31. Hull J, Rowlands K, Lockhart E, Moore C, Sharland M, Kwiatkowski D. Variants of the chemokine receptor CCR5 are associated with severe bronchiolitis caused by respiratory syncytial virus. Journal of Infectious Diseases. 2003;188(6):904-7.

32. Puthothu B, Forster J, Heinzmann A, Krueger M. TLR-4 and CD14 polymorphisms in respiratory syncytial virus associated disease. Disease markers. 2006;22(5, 6):303-8.

33. Pasanen A, Karjalainen MK, Bont L, Piippo-Savolainen E, Ruotsalainen M, Goksör E, et al. Genome-wide association study of polymorphisms predisposing to bronchiolitis. Scientific reports. 2017;7:41653.

34. Teng S, Wang L, Srivastava AK, Schwartz CE, Alexov E. Structural assessment of the effects of amino acid substitutions on protein stability and protein-protein interaction. International journal of computational biology and drug design. 2010;3(4):334.

35. DeVincenzo JP, Wilkinson T, Vaishnaw A, Cehelsky J, Meyers R, Nochur S, et al. Viral load drives disease in humans experimentally infected with respiratory syncytial virus. Am J Respir Crit Care Med. 2010;182(10):1305-14.

36. International Committee on Taxonomy of Viruses 2016 [Available from: http://www.ictvonline.org/virusTaxonomy.asp.

37. Morris J, Blount R, Savage R. Recovery of Cytopathogenic Agent from Chimpanzees with Goryza. Experimental Biology and Medicine. 1956;92(3):544-9.

38. Chatterjee S, Luthra P, Esaulova E, Agapov E, Yen BC, Borek DM, et al. Structural basis for human respiratory syncytial virus NS1-mediated modulation of host responses. Nature microbiology. 2017;2(9):17101.

39. Swedan S, Andrews J, Majumdar T, Musiyenko A, Barik S. Multiple functional domains and complexes of the two nonstructural proteins of human respiratory syncytial virus contribute to interferon suppression and cellular location. Journal of virology. 2011;85(19):10090-100.

40. Ling Z, Tran KC, Teng MN. Human respiratory syncytial virus nonstructural protein NS2 antagonizes the activation of beta interferon transcription by interacting with RIG-I. J Virol. 2009;83(8):3734-42.

41. Xie J, Long X, Gao L, Chen S, Zhao K, Li W, et al. Respiratory syncytial virus nonstructural protein 1 blocks glucocorticoid receptor nuclear translocation by targeting ipo13 and may account for glucocorticoid insensitivity. The Journal of infectious diseases. 2017;217(1):35-46.

42. Bakre A, Wu W, Hiscox J, Spann K, Teng MN, Tripp RA. Human respiratory syncytial virus non-structural protein NS1 modifies miR-24 expression via transforming growth factor-β. The Journal of general virology. 2015;96(Pt 11):3179.

43. Zhang Y, Yang L, Wang H, Zhang G, Sun X. Respiratory syncytial virus non-structural protein 1 facilitates virus replication through miR-29a-mediated inhibition of interferon-alpha receptor. Biochem Biophys Res Commun. 2016.

44. Atherton LJ, Jorquera PA, Bakre AA, Tripp RA. Determining immune and miRNA biomarkers related to respiratory syncytial virus (RSV) vaccine types. Frontiers in immunology. 2019;10:2323.

45. Liesman RM, Buchholz UJ, Luongo CL, Yang L, Proia AD, DeVincenzo JP, et al. RSV-encoded NS2 promotes epithelial cell shedding and distal airway obstruction. The Journal of clinical investigation. 2014;124(5).

46. Whelan JN, Tran KC, Van Rossum DB, Teng MN. Identification of respiratory syncytial virus nonstructural protein 2 residues essential for exploitation of the host ubiquitin system and inhibition of innate immune responses. Journal of virology. 2016;90(14):6453-63.

47. Hendricks DA, McIntosh K, Patterson JL. Further characterization of the soluble form of the G glycoprotein of respiratory syncytial virus. Journal of virology. 1988;62(7):2228-33.

48. Roberts SR, Lichtenstein D, Ball LA, Wertz G. The membrane-associated and secreted forms of the respiratory syncytial virus attachment glycoprotein G are synthesized from alternative initiation codons. Journal of Virology. 1994;68(7):4538-46.

49. Satake M, Coligan JE, Elango N, Norrby E, Venkatesan S. Respiratory syncytial virus envelope glycoprotein (G) has a novel structure. Nucleic Acids Research. 1985;13(21):7795-812.

50. Wertz GW, Collins PL, Huang Y, Gruber C, Levine S, Ball LA. Nucleotide sequence of the G protein gene of human respiratory syncytial virus reveals an unusual type of viral membrane protein. Proceedings of the National Academy of Sciences. 1985;82(12):4075-9.

51. Lambert DM. Role of oligosaccharides in the structure and function of respiratory syncytial virus glycoproteins. Virology. 1988;164(2):458-66.

52. Johnson PR, Spriggs MK, Olmsted RA, Collins PL. The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins. Proceedings of the National Academy of Sciences. 1987;84(16):5625-9.

53. Melero JA, Mas V, McLellan JS. Structural, antigenic and immunogenic features of respiratory syncytial virus glycoproteins relevant for vaccine development. Vaccine. 2017;35(3):461-8.

54. Norrby E, Mufson MA, Alexander H, Houghten RA, Lerner RA. Site-directed serology with synthetic peptides representing the large glycoprotein G of respiratory syncytial virus. Proceedings of the National Academy of Sciences. 1987;84(18):6572-6.

55. Cane PA. Analysis of linear epitopes recognised by the primary human antibody response to a variable region of the attachment (G) protein of respiratory syncytial virus. Journal of medical virology. 1997;51(4):297-304.

56. GARCÍA-BEATO R, MARTÍNEZ I, FRANCÍ C, REAL FX, GARCÍA-BARRENO B, MELERO JA. Host cell effect upon glycosylation and antigenicity of human respiratory syncytial virus G glycoprotein. Virology. 1996;221(2):301-9.

57. Trento A, Abrego L, Rodriguez-Fernandez R, Gonzalez-Sanchez MI, Gonzalez-Martinez F, Delfraro A, et al. Conservation of G-Protein Epitopes in Respiratory Syncytial Virus (Group A) Despite

Broad Genetic Diversity: Is Antibody Selection Involved in Virus Evolution? J Virol. 2015;89(15):7776-85.

58. Hause AM, Henke DM, Avadhanula V, Shaw CA, Tapia LI, Piedra PA. Sequence variability of the respiratory syncytial virus (RSV) fusion gene among contemporary and historical genotypes of RSV/A and RSV/B. PloS one. 2017;12(4):e0175792.

59. Mas V, Nair H, Campbell H, Melero JA, Williams TC. Antigenic and sequence variability of the human respiratory syncytial virus F glycoprotein compared to related viruses in a comprehensive dataset. Vaccine. 2018;36(45):6660-73.

60. González-Reyes L, Ruiz-Argüello MB, García-Barreno B, Calder L, López JA, Albar JP, et al. Cleavage of the human respiratory syncytial virus fusion protein at two distinct sites is required for activation of membrane fusion. Proceedings of the National Academy of Sciences. 2001;98(17):9859-64.

61. McLellan JS, Yang Y, Graham BS, Kwong PD. Structure of respiratory syncytial virus fusion glycoprotein in the postfusion conformation reveals preservation of neutralizing epitopes. Journal of virology. 2011;85(15):7788-96.

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

63. Melero JA, Mas V, McLellan JS. Structural, antigenic and immunogenic features of respiratory syncytial virus glycoproteins relevant for vaccine development. Vaccine. 2016.

64. Karron RA, Buonagurio DA, Georgiu AF, Whitehead SS, Adamus JE, Clements-Mann ML, et al. Respiratory syncytial virus (RSV) SH and G proteins are not essential for viral replication in vitro: clinical evaluation and molecular characterization of a cold-passaged, attenuated RSV subgroup B mutant. Proceedings of the National Academy of Sciences. 1997;94(25):13961-6.

65. Li D, Jans DA, Bardin PG, Meanger J, Mills J, Ghildyal R. Association of respiratory syncytial virus M protein with viral nucleocapsids is mediated by the M2-1 protein. Journal of virology. 2008;82(17):8863-70.

66. Richard C-A, Rincheval V, Lassoued S, Fix J, Cardone C, Esneau C, et al. RSV hijacks cellular protein phosphatase 1 to regulate M2-1 phosphorylation and viral transcription. PLoS pathogens. 2018;14(3):e1006920.

67. Braun MR, Deflubé LR, Noton SL, Mawhorter ME, Tremaglio CZ, Fearns R. RNA elongation by respiratory syncytial virus polymerase is calibrated by conserved region V. PLoS pathogens. 2017;13(12):e1006803.

68. Bermingham A, Collins PL. The M2–2 protein of human respiratory syncytial virus is a regulatory factor involved in the balance between RNA replication and transcription. Proceedings of the National Academy of Sciences. 1999;96(20):11259-64.

69. Tan L, Lemey P, Houspie L, Viveen MC, Jansen NJ, van Loon AM, et al. Genetic variability among complete human respiratory syncytial virus subgroup A genomes: bridging molecular evolutionary dynamics and epidemiology. PLoS One. 2012;7(12):e51439.

70. Otieno JR, Agoti CN, Gitahi CW, Bett A, Ngama M, Medley GF, et al. Molecular evolutionary dynamics of respiratory syncytial virus group A in recurrent epidemics in coastal Kenya. Journal of virology. 2016:JVI. 03105-15.

71. Coates H, Alling D, Chanock R. An antigenic analysis of respiratory syncytial virus isolates by a plaque reduction neutralization test. American Journal of Epidemiology. 1966;83(2):299-313.

72. Anderson LJ, Hierholzer JC, Tsou C, Hendry RM, Fernie BF, Stone Y, et al. Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies. Journal of Infectious Diseases. 1985;151(4):626-33.

73. Mufson MA, Örvell C, Rafnar B, Norrby E. Two distinct subtypes of human respiratory syncytial virus. Journal of general virology. 1985;66(10):2111-24.

74. Gimenez H, Hardman N, Keir H, Cash P. Antigenic variation between human respiratory syncytial virus isolates. Journal of General Virology. 1986;67(5):863-70.

75. Morgan L, Routledge E, Willcocks M, Samson A, Scott R, Toms G. Strain variation of respiratory syncytial virus. Journal of general virology. 1987;68(11):2781-8.

76. Örvell C, Norrby E, Mufson MA. Preparation and characterization of monoclonal antibodies directed against five structural components of human respiratory syncytial virus subgroup B. Journal of general virology. 1987;68(12):3125-35.

77. Garcia-Barreno B, Palomo C, Penas C, Delgado T, Perez-Brena P, Melero J. Marked differences in the antigenic structure of human respiratory syncytial virus F and G glycoproteins. Journal of Virology. 1989;63(2):925-32.

78. Cane P, Pringle C. Respiratory syncytial virus heterogeneity during an epidemic: analysis by limited nucleotide sequencing (SH gene) and restriction mapping (N gene). Journal of general virology. 1991;72(2):349-57.

79. García O, Martin M, Dopazo J, Arbiza J, Frabasile S, Russi 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. Journal of virology. 1994;68(9):5448-59.

80. Sullender WM, Sun L, Anderson L. Analysis of respiratory syncytial virus genetic variability with amplified cDNAs. Journal of clinical microbiology. 1993;31(5):1224-31.

81. Cane PA, Matthews DA, Pringle CR. Analysis of respiratory syncytial virus strain variation in successive epidemics in one city. Journal of clinical microbiology. 1994;32(1):1-4.

82. Peret T, Hall CB, Schnabel KC, Golub JA, Anderson LJ. Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. Journal of General Virology. 1998;79(9):2221-9.

83. Trento A, Galiano M, Videla C, Carballal G, Garcia-Barreno B, Melero JA, 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(Pt 11):3115-20.

84. Cane PA, Pringle CR. Evolution of subgroup A respiratory syncytial virus: evidence for progressive accumulation of amino acid changes in the attachment protein. Journal of virology. 1995;69(5):2918-25.

85. Woelk CH, Holmes EC. Variable immune-driven natural selection in the attachment (G) glycoprotein of respiratory syncytial virus (RSV). Journal of molecular evolution. 2001;52(2):182-92.

86. Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, 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(3):e32807.

87. Trento A, Casas I, Calderon A, Garcia-Garcia ML, Calvo C, Perez-Brena P, 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(15):7500-12.

88. Venter M, Madhi SA, Tiemessen CT, 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. Journal of General Virology. 2001;82(9):2117-24.

89. Malasao R, Furuse Y, Okamoto M, Dapat C, Saito M, Saito-Obata M, et al. Complete Genome Sequences of 13 Human Respiratory Syncytial Virus Subgroup A Strains of Genotypes NA1 and ON1 Isolated in the Philippines. Genome Announc. 2018;6(10):e00151-18.

90. Ivancic-Jelecki J, Slovic A, Ljubin-Sternak S, Galinović GM, Forcic D. Variability analysis and inter-genotype comparison of human respiratory syncytial virus small hydrophobic gene. Virology journal. 2018;15(1):109.

91. Peret TC, Hall CB, Hammond GW, Piedra PA, Storch GA, Sullender WM, et al. Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America. Journal of infectious diseases. 2000;181(6):1891-6.

92. Pretorius MA, van Niekerk S, Tempia S, Moyes J, Cohen C, Madhi SA, et al. Replacement and positive evolution of subtype A and B respiratory syncytial virus G-protein genotypes from 1997-2012 in South Africa. J Infect Dis. 2013;208 Suppl 3:S227-37.

93. Shobugawa Y, Saito R, Sano Y, Zaraket H, Suzuki Y, Kumaki A, et al. Emerging genotypes of human respiratory syncytial virus subgroup A among patients in Japan. J Clin Microbiol. 2009;47(8):2475-82.

94. Cui G, Zhu R, Qian Y, Deng J, Zhao L, Sun Y, et al. Genetic variation in attachment glycoprotein genes of human respiratory syncytial virus subgroups a and B in children in recent five consecutive years. PLoS One. 2013;8(9):e75020.

95. Hirano E, Kobayashi M, Tsukagoshi H, Yoshida LM, Kuroda M, Noda M, 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.

96. Schobel SA, Stucker KM, Moore ML, Anderson LJ, Larkin EK, Shankar J, 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.

97. Trento A, Viegas M, Galiano M, Videla C, Carballal G, Mistchenko AS, 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(2):975-84.

98. Dapat IC, Shobugawa Y, Sano Y, Saito R, Sasaki A, Suzuki Y, et al. New genotypes within respiratory syncytial virus group B genotype BA in Niigata, Japan. J Clin Microbiol. 2010;48(9):3423-7.

99. Arnott A, Vong S, Mardy S, Chu S, Naughtin M, Sovann L, et al. A study of the genetic variability of human respiratory syncytial virus (HRSV) in Cambodia reveals the existence of a new HRSV group B genotype. J Clin Microbiol. 2011;49(10):3504-13.

100. Blanc A, Delfraro A, Frabasile S, Arbiza J. Genotypes of respiratory syncytial virus group B identified in Uruguay. Arch Virol. 2005;150(3):603-9.

101. Sullender W, Mufson M, Anderson L, Wertz G. Genetic diversity of the attachment protein of subgroup B respiratory syncytial viruses. Journal of virology. 1991;65(10):5425-34.

102. Malasao R, Okamoto M, Chaimongkol N, Imamura T, Tohma K, Dapat I, et al. Molecular Characterization of Human Respiratory Syncytial Virus in the Philippines, 2012-2013. PLoS One. 2015;10(11):e0142192.

103. Kim Y-J, Kim D-W, Lee W-J, Yun M-R, Lee HY, Lee HS, et al. Rapid replacement of human respiratory syncytial virus A with the ON1 genotype having 72 nucleotide duplication in G gene. Infection, Genetics and Evolution. 2014;26:103-12.

104. Tabatabai J, Prifert C, Pfeil J, Grulich-Henn J, Schnitzler P. Novel respiratory syncytial virus (RSV) genotype ON1 predominates in Germany during winter season 2012–13. PLoS One. 2014;9(10):e109191.

105. Agoti CN, Otieno JR, Gitahi CW, Cane PA, Nokes DJ. Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya. Emerg Infect Dis. 2014;20(6):950-9.

106. Tsukagoshi H, Yokoi H, Kobayashi M, Kushibuchi I, Okamoto-Nakagawa R, Yoshida A, et al. Genetic analysis of attachment glycoprotein (G) gene in new genotype ON 1 of human respiratory syncytial virus detected in Japan. Microbiology and immunology. 2013;57(9):655-9.

107. Auksornkitti V, Kamprasert N, Thongkomplew S, Suwannakarn K, Theamboonlers A, Samransamruajkij R, et al. Molecular characterization of human respiratory syncytial virus, 2010-2011: identification of genotype ON1 and a new subgroup B genotype in Thailand. Arch Virol. 2014;159(3):499-507.

108. Comas-García A, Noyola DE, Cadena-Mota S, Rico-Hernández M, Bernal-Silva S. Respiratory syncytial virus-A ON1 genotype emergence in central Mexico in 2009 and evidence of multiple duplication events. The Journal of infectious diseases. 2018;217(7):1089-98.

109. Furuse Y. Multiple or single duplication events leading to the emergence of a novel genotype of respiratory syncytial virus. The Journal of infectious diseases. 2018;217(12):2008-10.

110. Otieno JR, Kamau EM, Agoti CN, Lewa C, Otieno G, Bett A, et al. Spread and evolution of respiratory syncytial virus A genotype ON1, Coastal Kenya, 2010–2015. Emerging infectious diseases. 2017;23(2):264.

111. Agoti CN, Mayieka LM, Otieno JR, Ahmed JA, Fields BS, Waiboci LW, et al. Examining strain diversity and phylogeography in relation to an unusual epidemic pattern of respiratory syncytial virus (RSV) in a long-term refugee camp in Kenya. BMC Infect Dis. 2014;14:178.

112. Agoti CN, Otieno JR, Ngama M, Mwihuri AG, Medley GF, Cane PA, et al. Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence. J Virol. 2015;89(22):11630-42.

113. Agoti CN, Munywoki PK, Phan MV, Otieno JR, Kamau E, Bett A, et al. Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis. Virus evolution. 2017;3(1).

114. Khor CS, Sam IC, Hooi PS, Chan YF. Displacement of predominant respiratory syncytial virus genotypes in Malaysia between 1989 and 2011. Infect Genet Evol. 2013;14:357-60.

115. Tan L, Coenjaerts FE, Houspie L, Viveen MC, van Bleek GM, Wiertz EJ, et al. The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. Journal of virology. 2013;87(14):8213-26.

116. Haider MSH, Khan WH, Deeba F, Ali S, Ahmed A, Naqvi IH, et al. BA9 lineage of respiratory syncytial virus from across the globe and its evolutionary dynamics. PloS one. 2018;13(4):e0193525.

117. Otieno JR, Kamau EM, Oketch JW, Ngoi JM, Gichuki AM, Binter Š, et al. Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. Virus evolution. 2018;4(2):vey027.

118. Geoghegan JL, Holmes EC. The phylogenomics of evolving virus virulence. Nature Reviews Genetics. 2018;19(12):756-69.

119. Jackson D, Hossain MJ, Hickman D, Perez DR, Lamb RA. A new influenza virus virulence determinant: the NS1 protein four C-terminal residues modulate pathogenicity. Proceedings of the National Academy of Sciences. 2008;105(11):4381-6.

120. Cotter CR, Jin H, Chen Z. A single amino acid in the stalk region of the H1N1pdm influenza virus HA protein affects viral fusion, stability and infectivity. PLoS pathogens. 2014;10(1):e1003831.

121. Urbanowicz RA, McClure CP, Sakuntabhai A, Sall AA, Kobinger G, Müller MA, et al. Human adaptation of Ebola virus during the West African outbreak. Cell. 2016;167(4):1079-87. e5.

122. Diehl WE, Lin AE, Grubaugh ND, Carvalho LM, Kim K, Kyawe PP, et al. Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. Cell. 2016;167(4):1088-98. e6.

123. Melero JA, Moore ML. Influence of respiratory syncytial virus strain differences on pathogenesis and immunity. Current topics in microbiology and immunology. 2013;372:59-82.

124. Walsh EE, McConnochie KM, Long CE, Hall CB. Severity of respiratory syncytial virus infection is related to virus strain. Journal of Infectious Diseases. 1997;175(4):814-20.

125. Hall C, Walsh E, Schnabel K, Long C, McConnochie K, Hildreth S, et al. Occurrence of groups A and B of RSV over 15 years: associated epidemiologic and clinical characteristics in hospitalized and ambulatory children. J infect dis. 1990;162(6):1283-90.

126. McConnochie KM, Hall CB, Walsh EE, Roghmann KJ. Variation in severity of respiratory syncytial virus infections with subtype. The Journal of pediatrics. 1990;117(1):52-62.

127. Gilca R, De Serres G, Tremblay M, Vachon M-L, Leblanc E, Bergeron MG, et al. Distribution and clinical impact of human respiratory syncytial virus genotypes in hospitalized children over 2 winter seasons. Journal of Infectious Diseases. 2006;193(1):54-8.

128. Vandini S, Biagi C, Lanari M. Respiratory syncytial virus: the influence of serotype and genotype variability on clinical course of infection. International journal of molecular sciences. 2017;18(8):1717.

129. Devincenzo JP. Natural infection of infants with respiratory syncytial virus subgroups A and B: a study of frequency, disease severity, and viral load. Pediatr Res. 2004;56(6):914-7.

130. Midulla F, Nenna R, Scagnolari C, Petrarca L, Frassanito A, Viscido A, et al. How respiratory syncytial virus genotypes influence the clinical course in infants hospitalized for bronchiolitis. The Journal of infectious diseases. 2018;219(4):526-34.

131. Kwilas S, Liesman RM, Zhang L, Walsh E, Pickles RJ, Peeples ME. Respiratory syncytial virus grown in Vero cells contains a truncated attachment protein that alters its infectivity and dependence on glycosaminoglycans. J Virol. 2009;83(20):10710-8.

132. Hotard AL, Laikhter E, Brooks K, Hartert TV, Moore ML. Functional Analysis of the 60-Nucleotide Duplication in the Respiratory Syncytial Virus Buenos Aires Strain Attachment Glycoprotein. J Virol. 2015;89(16):8258-66.

133. Meng J, Hotard AL, Currier MG, Lee S, Stobart CC, Moore ML. Respiratory Syncytial Virus Attachment Glycoprotein Contribution to Infection Depends on the Specific Fusion Protein. J Virol. 2015;90(1):245-53.

134. Lawlor HA, Schickli JH, Tang RS. A single amino acid in the F2 subunit of respiratory syncytial virus fusion protein alters growth and fusogenicity. J Gen Virol. 2013;94(Pt 12):2627-35.

135. Hotard AL, Lee S, Currier MG, Crowe JE, Jr., Sakamoto K, Newcomb DC, et al. Identification of residues in the human respiratory syncytial virus fusion protein that modulate fusion activity and pathogenesis. J Virol. 2015;89(1):512-22.

136. Corry J, Johnson SM, Cornwell J, Peeples ME. Preventing Cleavage of the Respiratory Syncytial Virus Attachment Protein in Vero Cells Rescues the Infectivity of Progeny Virus for Primary Human Airway Cultures. J Virol. 2015;90(3):1311-20.

137. Whitehead S, Hill M, Firestone C, Claire MS, Elkins W, Murphy B, et al. Replacement of the F and G proteins of respiratory syncytial virus (RSV) subgroup A with those of subgroup B generates chimeric live attenuated RSV subgroup B vaccine candidates. Journal of virology. 1999;73(12):9773-80.

138. Connors M, Crowe Jr JE, Firestone C-Y, Murphy BR, Collins PL. A cold-passaged, attenuated strain of human respiratory syncytial virus contains mutations in the F and L genes. Virology. 1995;208(2):478-84.

139. Whitehead SS, Juhasz K, Firestone C-Y, Collins PL, Murphy BR. Recombinant respiratory syncytial virus (RSV) bearing a set of mutations from cold-passaged RSV is attenuated in chimpanzees. Journal of virology. 1998;72(5):4467-71.

140. Teng MN, Whitehead SS, Bermingham A, Claire MS, Elkins WR, Murphy BR, et al. Recombinant respiratory syncytial virus that does not express the NS1 or M2-2 protein is highly attenuated and immunogenic in chimpanzees. Journal of virology. 2000;74(19):9317-21.

141. Whitehead SS, Bukreyev A, Teng MN, Firestone C-Y, Claire MS, Elkins WR, et al. Recombinant respiratory syncytial virus bearing a deletion of either the NS2 or SH gene is attenuated in chimpanzees. Journal of virology. 1999;73(4):3438-42.

142. Lukacs NW, Moore ML, Rudd BD, Berlin AA, Collins RD, Olson SJ, et al. Differential immune responses and pulmonary pathophysiology are induced by two different strains of respiratory syncytial virus. The American journal of pathology. 2006;169(3):977-86.

143. Moore ML, Chi MH, Luongo C, Lukacs NW, Polosukhin VV, Huckabee MM, et al. A chimeric A2 strain of respiratory syncytial virus (RSV) with the fusion protein of RSV strain line 19 exhibits enhanced viral load, mucus, and airway dysfunction. J Virol. 2009;83(9):4185-94.

144. Stokes KL, Chi MH, Sakamoto K, Newcomb DC, Currier MG, Huckabee MM, et al. Differential pathogenesis of respiratory syncytial virus clinical isolates in BALB/c mice. Journal of virology. 2011;85(12):5782-93.

145. Stokes KL, Currier MG, Sakamoto K, Lee S, Collins PL, Plemper RK, et al. The respiratory syncytial virus fusion protein and neutrophils mediate the airway mucin response to pathogenic respiratory syncytial virus infection. J Virol. 2013;87(18):10070-82.

146. Beaird O, Freifeld A, Ison M, Lawrence S, Theodoropoulos N, Clark N, et al. Current practices for treatment of respiratory syncytial virus and other non-influenza respiratory viruses in high-risk patient populations: a survey of institutions in the Midwestern Respiratory Virus Collaborative. Transplant Infectious Disease. 2016;18(2):210-5.

147. Aljabr W, Touzelet O, Pollakis G, Wu W, Munday DC, Hughes M, et al. Investigating the influence of ribavirin on human respiratory syncytial virus RNA synthesis by using a high-resolution transcriptome sequencing approach. Journal of virology. 2016;90(10):4876-88.

148. Mehedi M, McCarty T, Martin SE, Le Nouën C, Buehler E, Chen Y-C, et al. Actin-related protein 2 (ARP2) and virus-induced filopodia facilitate human respiratory syncytial virus spread. PLoS pathogens. 2016;12(12):e1006062.

149. Group I-RS. Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. Pediatrics. 1998;102(3):531-7.

150. Jaberolansar N, Toth I, Young PR, Skwarczynski M. Recent advances in the development of subunit-based RSV vaccines. Expert Rev Vaccines. 2016;15(1):53-68.

151. Guvenel AK, Chiu C, Openshaw PJ. Current concepts and progress in RSV vaccine development. Expert Rev Vaccines. 2014;13(3):333-44.

152. Jorquera PA, Anderson L, Tripp RA. Understanding respiratory syncytial virus (RSV) vaccine development and aspects of disease pathogenesis. Expert Rev Vaccines. 2016;15(2):173-87.

153. KIM HW, CANCHOLA JG, BRANDT CD, PYLES G, CHANOCK RM, JENSEN K, et al. Respiratory syncytial virus disease in infants despite prior administration of antigenic inactivated vaccine. American journal of epidemiology. 1969;89(4):422-34.

154. Tian D, Battles MB, Moin SM, Chen M, Modjarrad K, Kumar A, et al. Structural basis of respiratory syncytial virus subtype-dependent neutralization by an antibody targeting the fusion glycoprotein. Nature communications. 2017;8(1):1877.

155. Gilman MS, Liu C, Fung A, Behera I, Jordan P, Rigaux P, et al. Structure of the Respiratory Syncytial Virus Polymerase Complex. Cell. 2019;179(1):193-204. e14.

156. Vaughan K, Ponomarenko J, Peters B, Sette A. Analysis of human RSV immunity at the molecular level: learning from the past and present. PloS one. 2015;10(5):e0127108.

157. ClinicalTrials.gov. A Study to Evaluate Different Dose Levels of GlaxoSmithKline (GSK) Biologicals' Investigational Respiratory Syncytial Virus (RSV) Vaccine (GSK3888550A), Based on the Vaccine Safety and the Antibodies (Body Defences) Produced Following Vaccine Administration, When Given to Healthy Non-pregnant Women: National Library of Medicine

(US). 2019 [Available from: https://clinicaltrials.gov/ct2/show/record/NCT03674177.

158. Ascough S, Vlachantoni I, Kalyan M, Haijema B-J, Wallin-Weber S, Dijkstra-Tiekstra M, et al. Local and Systemic Immunity against Respiratory Syncytial Virus Induced by a Novel Intranasal Vaccine. A Randomized, Double-Blind, Placebo-controlled Clinical Trial. American Journal of Respiratory and Critical Care Medicine. 2019;200(4):481.

159. Sun Y, Jain D, Koziol-White CJ, Genoyer E, Gilbert M, Tapia K, et al. Immunostimulatory Defective Viral Genomes from Respiratory Syncytial Virus Promote a Strong Innate Antiviral Response during Infection in Mice and Humans. PLoS Pathog. 2015;11(9):e1005122.

160. PATH. RSV Vaccine and mAb snapshot 2018-2019 [Available from: https://vaccineresources.org/files/RSV-snapshot-2019_08_28_High%20Resolution_PDF.pdf.

161. Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences. 1977;74(2):560-4.

162. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences. 1977;74(12):5463-7.

163. Tracy T, Mulcahy L. A simple method for direct automated sequencing of PCR fragments. Biotechniques. 1991;11(1):68-75.

164. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.

165. Dunham I, Hunt A, Collins J, Bruskiewich R, Beare D, Clamp M, et al. The DNA sequence of human chromosome 22. Nature. 1999;402(6761):489-95.

166. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931.

167. Consortium GP. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061.

168. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Frontiers in microbiology. 2012;3:329.

169. Vandenhende M-A, Bellecave P, Recordon-Pinson P, Reigadas S, Bidet Y, Bruyand M, et al. Prevalence and evolution of low frequency HIV drug resistance mutations detected by ultra deep sequencing in patients experiencing first line antiretroviral therapy failure. PloS one. 2014;9(1):e86771.

170. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nature Reviews Microbiology. 2017;15(3):183.

171. Lau JW, Kim Y-I, Murphy R, Newman R, Yang X, Zody M, 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.

172. Grad Y, Newman R, Zody M, Yang X, Murphy R, Qu J, et al. Within-host whole genome deep sequencing and diversity analysis of human RSV infection reveals dynamics of genomic diversity in the absence and presence of immune pressure. Journal of virology. 2014:JVI. 00038-14.

173. Check Hayden E. Genome sequencing: the third generation. Nature Publishing Group; 2009.

174. Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ, et al. Direct RNA sequencing of the coding complete influenza A virus genome. Scientific reports. 2018;8.

175. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.

176. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31(20):3350-2.

177. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26(5):589-95.

178. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

179. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.

180. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10(1):421.

181. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792-7.

182. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics. 2004;5(1):113.

183. Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002;30(14):3059-66.

184. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7(1).

185. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 1988;73(1):237-44.

186. Higgins DG. CLUSTAL V: multiple alignment of DNA and protein sequences. Computer analysis of sequence data: Springer; 1994. p. 307-18.

187. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994;22(22):4673-80.

188. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic acids research. 1997;25(24):4876-82.

189. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic acids research. 2003;31(13):3497-500.

190. Strimmer K, Von Haeseler A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proceedings of the National Academy of Sciences. 1997;94(13):6815-9.

191. Jukes TH, Cantor CR. Evolution of protein molecules. Mammalian protein metabolism. 1969;3(21):132.

192. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution. 1980;16(2):111-20.

193. Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods. 2017;14(6):587.

194. Moore GW, Goodman M, Barnabas J. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. Journal of theoretical biology. 1973;38(3):423-57.

195. Robinson DF. Comparison of labeled trees with valency three. Journal of Combinatorial Theory, Series B. 1971;11(2):105-19.

196. Money D, Whelan S. Characterizing the phylogenetic tree-search problem. Systematic biology. 2012;61(2):228.

197. Bryant D. The splits in the neighborhood of a tree. Annals of Combinatorics. 2004;8(1):1-11.

198. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Molecular biology and evolution. 2013;30(5):1188-95.

199. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology. 2010;59(3):307-21.

200. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

201. Wu NC, Wilson IA. A perspective on the structural and functional constraints for immune evasion: insights from influenza virus. Journal of molecular biology. 2017;429(17):2694-709.

202. Cristina J, Arbiza J, Albo C, Garci B, Garci O, Portela A. Evolution of the G and P genes of human respiratory syncytial virus (subgroup A) studied by the RNase A mismatch cleavage method. Virology. 1991;184(1):210-8.

203. Baek YH, Choi EH, Song M-S, Pascua PNQ, Kwon H-i, Park S-J, et al. Prevalence and genetic characterization of respiratory syncytial virus (RSV) in hospitalized children in Korea. Archives of virology. 2012;157(6):1039-50.

204. Ren L, Xiao Q, Zhou L, Xia Q, Liu E. Molecular characterization of human respiratory syncytial virus subtype B: a novel genotype of subtype B circulating in China. Journal of medical virology. 2015;87(1):1-9.

205. Gimferrer L, Andrés C, Campins M, Codina MG, Rodrigo JA, Melendo S, et al. Circulation of a novel human respiratory syncytial virus Group B genotype during the 2014–2015 season in Catalonia (Spain). Clinical Microbiology and Infection. 2016;22(1):97. e5-. e8.

206. Gaymard A, Bouscambert-Duchamp M, Pichon M, Frobert E, Vallee J, Lina B, et al. Genetic characterization of respiratory syncytial virus highlights a new BA genotype and emergence of the ON1 genotype in Lyon, France, between 2010 and 2014. Journal of Clinical Virology. 2018;102:12-8.

207. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2014;32(1):268-74.

208. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular evolution. 1994;39(3):306-14.

209. Organization WH. Expanded Programme on Immunization (EPI): Standardization of the nomenclature for describing the genetic characteristics of wild-type measles viruses. Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire. 1998;73(35):265-9.

210. Organization WH. Genetic diversity of wildtype measles viruses and the global measles nucleotide surveillance database (MeaNS). Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire. 2015;90(30):373-80.

211. Muwonge A, Nanyunja M, Rota PA, Bwogi J, Lowe L, Liffick SL, et al. New measles genotype, Uganda. Emerging infectious diseases. 2005;11(10):1522.

212. Organization WH. Update of the nomenclature for describing the genetic characteristics of wild-type measles viruses: new genotypes and reference strains: Background. Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire. 2003;78(27):229-32.

213. Organization WH. New genotype of measles virus and update on global distribution of measles genotypes. Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire. 2005;80(40):347-51.

214. Oberste MS, Maher K, Kilpatrick DR, Pallansch MA. Molecular evolution of the human enteroviruses: correlation of serotype with VP1 sequence and application to picornavirus classification. Journal of virology. 1999;73(3):1941-8.

215. Lau SK, Yip CC, Tsoi H-w, Lee RA, So L-y, Lau Y-l, et al. Clinical features and complete genome characterization of a distinct human rhinovirus (HRV) genetic cluster, probably representing a previously undetected HRV species, HRV-C, associated with acute respiratory illness in children. Journal of clinical microbiology. 2007;45(11):3655-64.

216. Blomqvist S, Savolainen C, Råman L, Roivainen M, Hovi T. Human rhinovirus 87 and enterovirus 68 represent a unique serotype with rhinovirus and enterovirus features. Journal of clinical microbiology. 2002;40(11):4218-23.

217. Savolainen C, Blomqvist S, Mulders MN, Hovi T. Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70. Journal of General Virology. 2002;83(2):333-40.

218. Tapparel C, Junier T, Gerlach D, Cordey S, Van Belle S, Perrin L, et al. New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. BMC genomics. 2007;8(1):224.

219. Ren L, Yang D, Ren X, Li M, Mu X, Wang Q, et al. Genotyping of human rhinovirus in adult patients with acute respiratory infections identified predominant infections of genotype A21. Scientific reports. 2017;7(1):1-9.

220. Yoshihara K, Le MN, Nagasawa K, Tsukagoshi H, Nguyen HA, Toizumi M, et al. Molecular evolution of respiratory syncytial virus subgroup A genotype NA1 and ON1 attachment glycoprotein (G) gene in central Vietnam. Infection, Genetics and Evolution. 2016;45:437-46.

221. Bose ME, He J, Shrivastava S, Nelson MI, Bera J, Halpin RA, et al. Sequencing and analysis of globally obtained human respiratory syncytial virus A and B genomes. Plos one. 2015;10(3):e0120098.

222. Agoti CN, Otieno JR, Munywoki PK, Mwihuri AG, Cane PA, Nokes DJ, et al. Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. J Virol. 2015;89(7):3444-54.

223. Melero JA, Pringle CR, Cane PA. Antigenic structure, evolution and immunobiology of human respiratory syncytial virus attachment (G) protein. 1997.

224. Zlateva KT, Lemey P, Vandamme A-M, Van Ranst M. Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup A: positively selected sites in the attachment G glycoprotein. Journal of virology. 2004;78(9):4675-83.

225. Gaunt ER, Jansen RR, Poovorawan Y, Templeton KE, Toms GL, Simmonds P. Molecular epidemiology and evolution of human respiratory syncytial virus and human metapneumovirus. PloS one. 2011;6(3).

226. Zlateva KT, Lemey P, Moes E, Vandamme AM, Van Ranst M. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. J Virol. 2005;79(14):9157-67.

227. Garcia-Barreno B, Delgado T, Akerlind-Stopner B, Norrby E, Melero JA. Location of the epitope recognized by monoclonal antibody 63G on the primary structure of human respiratory syncytial virus G glycoprotein and the ability of synthetic peptides containing this epitope to induce neutralizing antibodies. Journal of General Virology. 1992;73(10):2625-30.

228. Lu B, Liu H, Tabor DE, Tovchigrechko A, Qi Y, Ruzin A, et al. Emergence of new antigenic epitopes in the glycoproteins of human respiratory syncytial virus collected from a US surveillance study, 2015–17. Scientific reports. 2019;9(1):1-10.

229. McCarthy M, Villafana T, Stillman E, Esser MT. Respiratory syncytial virus protein structure, function and implications for subunit vaccine development. Future Virology. 2014;9(8):753-67.

230. Zhu Q, Lu B, McTamney P, Palaszynski S, Diallo S, Ren K, et al. Prevalence and significance of substitutions in the fusion protein of respiratory syncytial virus resulting in neutralization escape from antibody MEDI8897. The Journal of infectious diseases. 2018;218(4):572-80.

231. Adams O, Bonzel L, Kovacevic A, Mayatepek E, Hoehn T, Vogel M. Palivizumab-resistant human respiratory syncytial virus infection in infancy. Clinical infectious diseases. 2010;51(2):185-8.

232. Zhao X, Chen F-P, Sullender WM. Respiratory syncytial virus escape mutant derived in vitro resists palivizumab prophylaxis in cotton rats. Virology. 2004;318(2):608-12.

233. Zhao X, Sullender WM. In vivo selection of respiratory syncytial viruses resistant to palivizumab. Journal of virology. 2005;79(7):3962-8.

234. DeVincenzo JP, Hall CB, Kimberlin DW, Sánchez PJ, Rodriguez WJ, Jantausch BA, et al. Surveillance of clinical isolates of respiratory syncytial virus for palivizumab (Synagis)–resistant mutants. Journal of Infectious Diseases. 2004;190(5):975-8.

235. Habibi MS. Correlates of protection and disease in experimental human respiratory syncytial virus infection. 2014.

236. Mac S, Sumner A, Duchesne-Belanger S, Stirling R, Tunis M, Sander B. Cost-effectiveness of Palivizumab for Respiratory Syncytial Virus: A Systematic Review. Pediatrics. 2019;143(5):e20184064.

237. PERFORM. Personalised Management of Febrile Illness 2019 [https://www.perform2020.org/]. Available from: https://www.perform2020.org/.

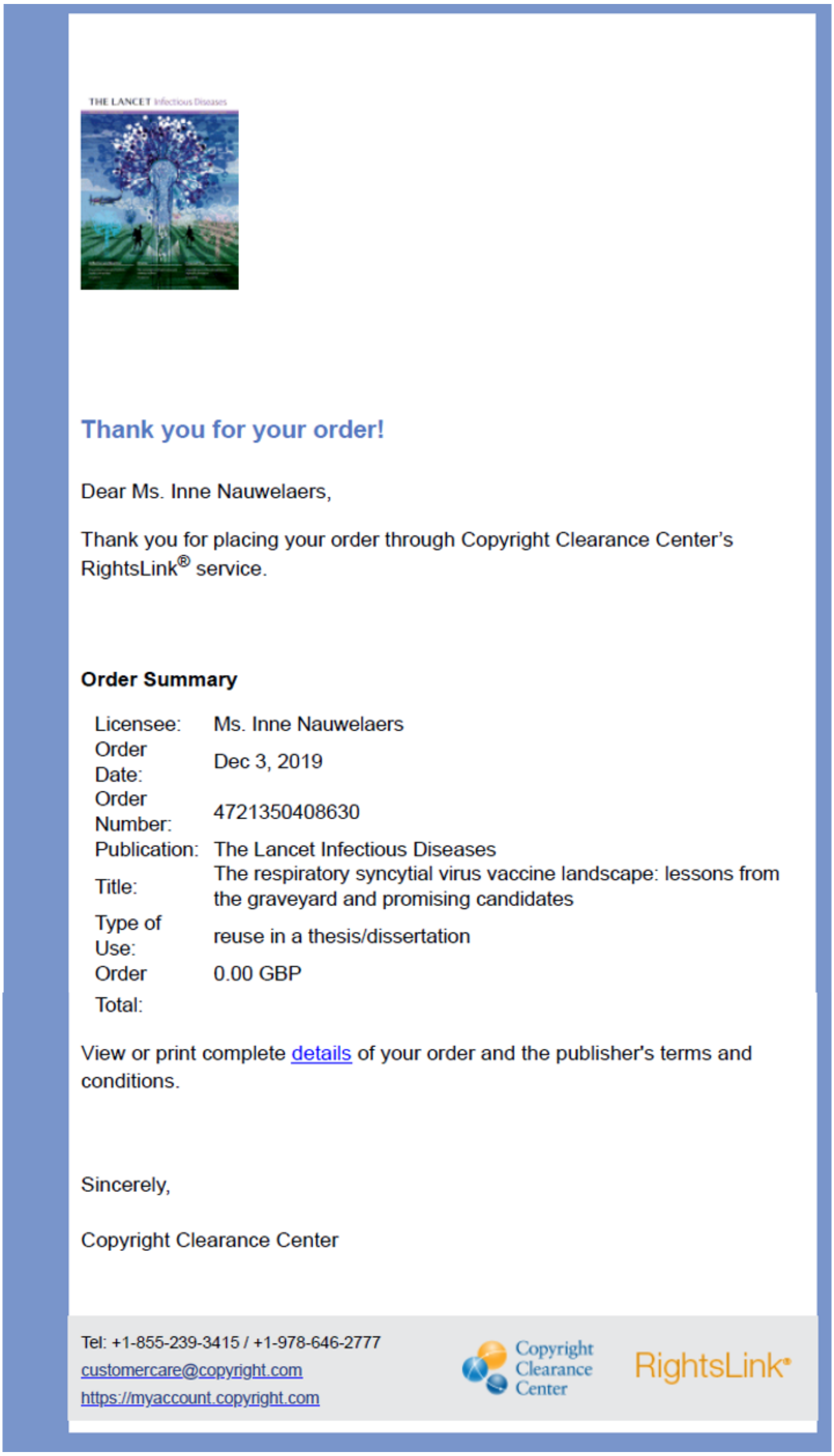
238. Agoti CN, Mbisa JL, Bett A, Medley GF, Nokes DJ, Cane PA. Intrapatient variation of the respiratory syncytial virus attachment protein gene. Journal of virology. 2010;84(19):10425-8.

239. Sun Y, López CB. Preparation of respiratory syncytial virus with high or low content of defective viral particles and their purification from viral stocks. Bio-protocol. 2016;6(10).

240. Chan KF, Carolan LA, Druce J, Chappell K, Watterson D, Young P, et al. Pathogenesis, humoral immune responses, and transmission between cohoused animals in a ferret model of human respiratory syncytial virus infection. Journal of virology. 2018;92(4):e01322-17.

Annexes

Permissions to use figures in this thesis not produced by the author can be found below.



Thank you for your order!

Dear Ms. Inne Nauwelaers,

Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

Order Summary

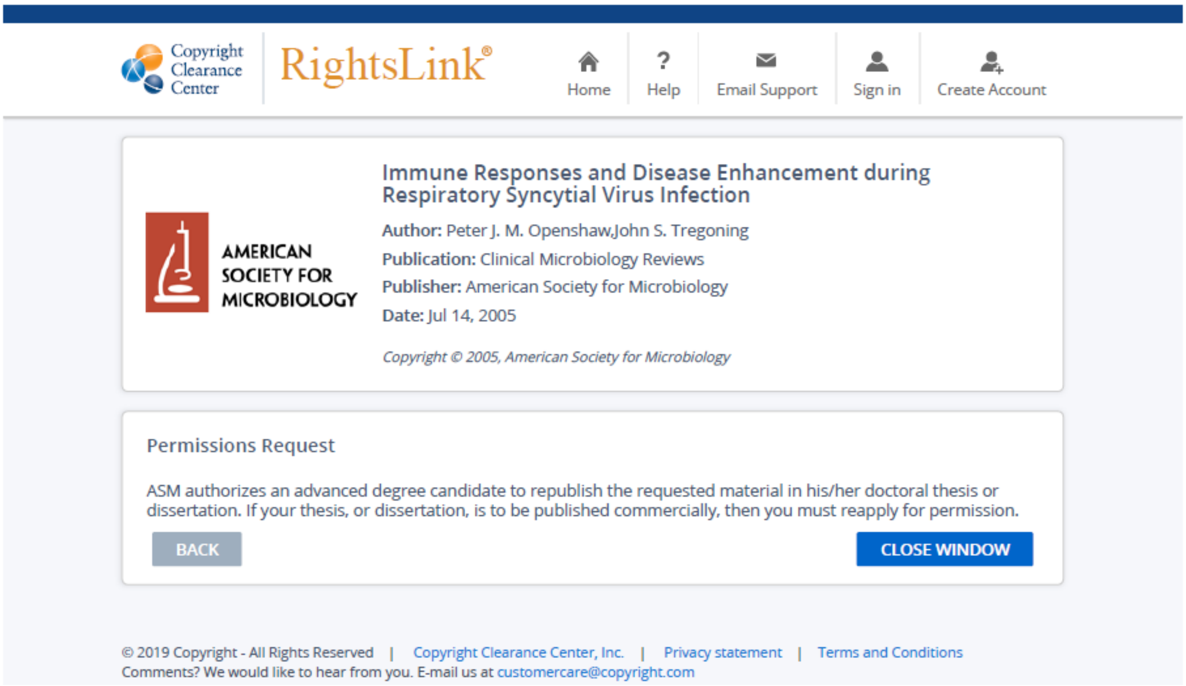
Licensee:	Ms. Inne Nauwelaers
Order Date:	Dec 3, 2019
Order Number:	4721350408630
Publication:	The Lancet Infectious Diseases
Title:	The respiratory syncytial virus vaccine landscape: lessons from the graveyard and promising candidates
Type of Use:	reuse in a thesis/dissertation
Order Total:	0.00 GBP

View or print complete details of your order and the publisher's terms and conditions.

Sincerely,

Copyright Clearance Center

Tel: +1-855-239-3415 / +1-978-646-2777
customercare@copyright.com
https://myaccount.copyright.com



Immune Responses and Disease Enhancement during Respiratory Syncytial Virus Infection

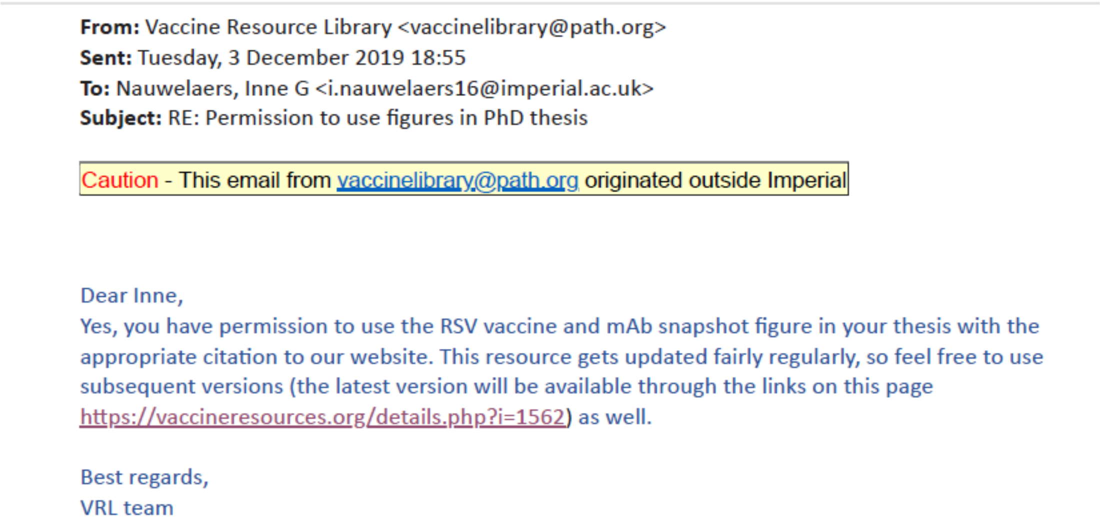
Author: Peter J. M. Openshaw,John S. Tregoning
Publication: Clinical Microbiology Reviews
Publisher: American Society for Microbiology
Date: Jul 14, 2005

Copyright © 2005, American Society for Microbiology

Permissions Request

ASM authorizes an advanced degree candidate to republish the requested material in his/her doctoral thesis or dissertation. If your thesis, or dissertation, is to be published commercially, then you must reapply for permission.

© 2019 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

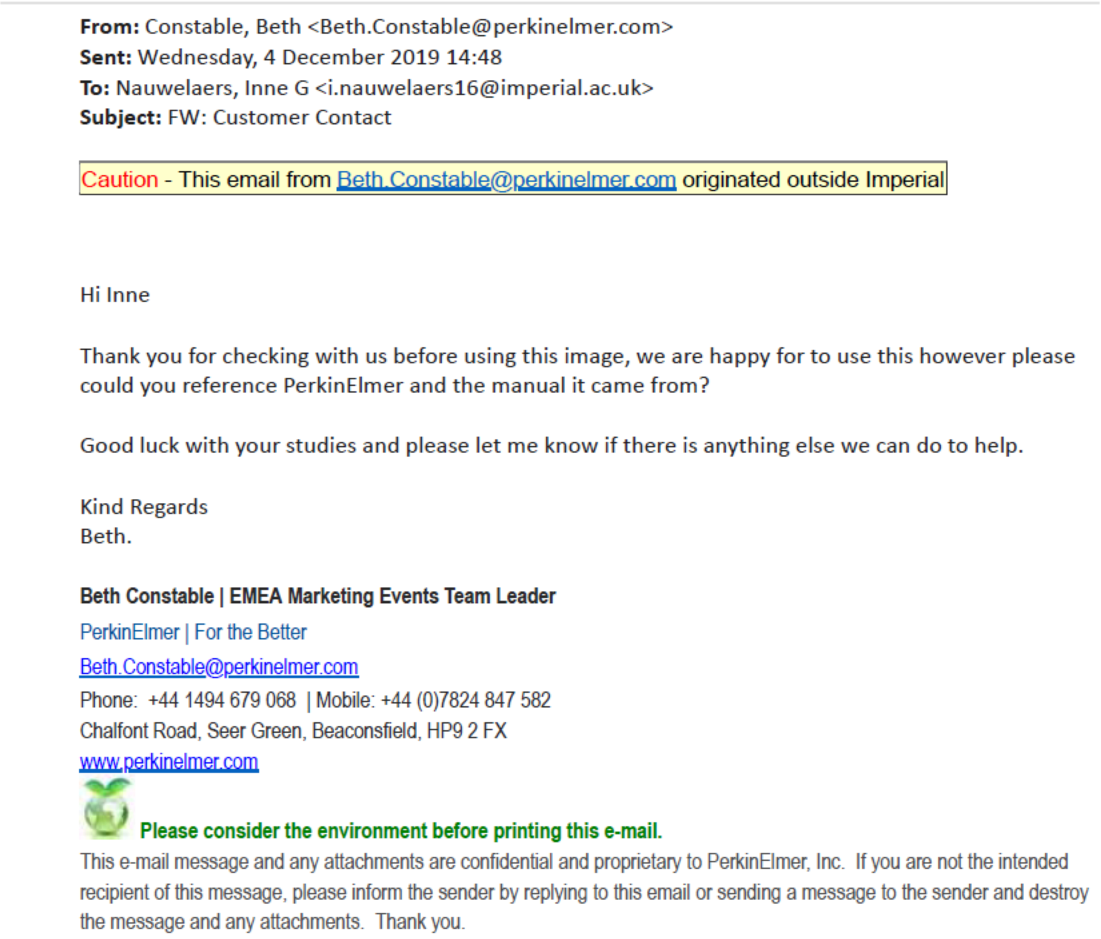


From: Vaccine Resource Library <vaccinelibrary@path.org>
Sent: Tuesday, 3 December 2019 18:55
To: Nauwelaers, Inne G <i.nauwelaers16@imperial.ac.uk>
Subject: RE: Permission to use figures in PhD thesis

Caution - This email from vaccinelibrary@path.org originated outside Imperial

Dear Inne,
Yes, you have permission to use the RSV vaccine and mAb snapshot figure in your thesis with the appropriate citation to our website. This resource gets updated fairly regularly, so feel free to use subsequent versions (the latest version will be available through the links on this page https://vaccineresources.org/details.php?i=1562) as well.

Best regards,
VRL team



From: Constable, Beth <Beth.Constable@perkinelmer.com>
Sent: Wednesday, 4 December 2019 14:48
To: Nauwelaers, Inne G <i.nauwelaers16@imperial.ac.uk>
Subject: FW: Customer Contact

Caution - This email from Beth.Constable@perkinelmer.com originated outside Imperial

Hi Inne

Thank you for checking with us before using this image, we are happy for to use this however please could you reference PerkinElmer and the manual it came from?

Good luck with your studies and please let me know if there is anything else we can do to help.

Kind Regards
Beth.

Beth Constable | EMEA Marketing Events Team Leader
PerkinElmer | For the Better
Beth.Constable@perkinelmer.com
Phone: +44 1494 679 068 | Mobile: +44 (0)7824 847 582
Chalfont Road, Seer Green, Beaconsfield, HP9 2 FX
www.perkinelmer.com

Please consider the environment before printing this e-mail.

This e-mail message and any attachments are confidential and proprietary to PerkinElmer, Inc. If you are not the intended recipient of this message, please inform the sender by replying to this email or sending a message to the sender and destroy the message and any attachments. Thank you.