

**COMPARISON OF THE PATTERNS OF SPREAD OF HUMAN
METAPNEUMOVIRUS (HMPV) AND RESPIRATORY SYNCYTIAL VIRUS
(RSV) IN AFRICA USING VIRUS SEQUENCE DATA**

John W. Oketch


**A thesis submitted in partial fulfilment of the requirements for the Degree of
Master of Science in Bioinformatics of Pwani University**

January, 2021

DECLARATION

This thesis is my original work and has not been presented in any other University or any other Award.

John W. Oketch

Signature.....  Date 26/01/2021.....

Supervisors' recommendation

We confirm that the work reported in this thesis was carried out by the candidate under our supervision.

James Nokes

Signature ...  . Date 26/012021.....

Dr. Everlyne Isoe

Signature Date 26/01/2021.....

DEDICATION

This thesis is dedicated to EANBIT/IDeAL for the fellowship. To my supervisors for their mentorship and to my family for their moral support and encouragement.

ACKNOWLEDGEMENT

All gratitude to God for favour and protection. I greatly appreciate my supervisors: Professor James Nokes, Dr. Everlyne Isoe and Dr. Charles Agoti for their detailed feedback and prompt responses throughout my study. Thanks to Dr. James Otieno for the collaborative analysis and for the insights as well. Special thanks to James Nokes for the great mentorship, not forgetting Charles Agoti.

I appreciate my lecturers for seeing me through the course. My classmates in EANBIT cohort1 for the friendship and encouragement, thank you all.

My family was committed to supporting me in numerous ways. Thank you all for giving me the necessary foundation to study and for the moral support.

I am grateful to EANBIT/IDeAL for the scholarship which provided me much needed support, mentorship and networking. Finally, I am grateful to my host institution KILIFI KEMRI Wellcome Trust that provided me a good research and learning environment.

ABSTRACT

Background: Human metapneumovirus (HMPV) and respiratory syncytial virus (RSV) are the leading causes of viral severe acute respiratory disease in childhood. They are related viruses from the *Pneumoviridae* family and show overlapping clinical, epidemiological and transmission features. Whether the two viruses also share similar patterns of geographic spread remains unknown; this may provide insight on common modalities of control.

Materials and Methods: Using 232 HMPV and 842 RSV attachment (G) glycoprotein gene sequences obtained from Gambia, Zambia, Mali, South Africa, and Kenya, between August 2011 and January 2014, we conducted a comparative phylogenetic and phylogeographic analyses to explore the spatial-temporal patterns of HMPV and RSV across Africa using Bayesian discrete phylogeography.

Results: HMPV and RSV epidemics are characterised by co-circulation of multiple genetic variants. Similar genotype dominance patterns were observed between neighbouring countries. Phylogeographic analyses indicate sequences largely cluster by geographical region i.e., West Africa (Mali, Gambia), East Africa (Kenya) and Southern Africa (Zambia, South Africa), with strong regional links between neighbouring locations. African sequences were well-mixed with global sequences. Sequences from different African subregions fell into separate clusters interspersed with sequences from other countries globally.

Conclusion: HMPV and RSV share similar patterns of geographic spread across Africa, characterised by co-circulation of multiple genetic variants within epidemics, geographic clustering of sequences and strong regional links between neighbouring locations. Geographical clustering of sequences suggests independent introduction of HMPV and RSV variants in Africa from the global pool, and further local diversification. The

genotype dominance patterns observed further supports strong epidemiological linkage between neighbouring countries. Globally, African strains are not different from globally circulating strains.

TABLE OF CONTENTS

DECLARATION	I
DEDICATION	II
ACKNOWLEDGEMENT.....	III
ABSTRACT.....	IV
LIST OF TABLES	IX
LIST OF FIGURES	X
ABBREVIATIONS.....	XV
INTRODUCTION	1
<i>Background information</i>	<i>1</i>
<i>Statement of the Problem</i>	<i>3</i>
<i>Main objectives</i>	<i>4</i>
Specific objectives:.....	4
<i>Research Questions.....</i>	<i>4</i>
<i>Justification.....</i>	<i>4</i>
<i>Limitations and scope of the Study.....</i>	<i>5</i>
LITERATURE REVIEW	6
<i>Epidemiology.....</i>	<i>6</i>
<i>Transmission</i>	<i>8</i>
<i>Genetic structure and molecular epidemiology</i>	<i>8</i>

<i>Phylogeographic Analyses</i>	12
<i>Phylogeographic methods</i>	16
MATERIALS AND METHODS	19
<i>Study population and study sites</i>	19
<i>Laboratory methods</i>	22
Respiratory pathogen screening	22
HMPV and RSV G genes sequencing	22
<i>Data analysis</i>	24
Phylogenetic and phylogeographic analyses	24
Statistical analysis	26
RESULTS	27
<i>HMPV subtyping and subgroup temporal patterns</i>	27
<i>HMPV genetic diversity and inter-country transmission network</i>	27
<i>HMPV spatial origins and dispersal patterns</i>	30
<i>RSV subtyping and subgroup temporal patterns</i>	40
<i>RSV intra and inter-country genetic diversity</i>	40
<i>RSV spatial patterns and Origins</i>	41
DISCUSSION	45
CONCLUSIONS AND RECOMMENDATIONS	51
REFERENCES	53
APPENDICES	70

VIII

Appendix A 70

Appendix B 71

Appendix C 86

Funding 86

LIST OF TABLES

Table 1:Virus positive by site and Number sequenced.....	20
Table 2: WHO Clinical Criteria for Severe and Very Severe Pneumonia.....	21
Table 3: HMPV and RSV genotype detection patterns	30
Table 4: Inferred locations of tMRCA of African sequences.....	52

LIST OF FIGURES

Figure 1: Genomic organization of human metapneumovirus (HMPV) and respiratory syncytial virus (RSV), (Rima et al., 2017). HMPV is classified in genus Metapneumovirus, RSV belongs to genus Orthopneumovirus, both in the Pneumoviridae family. Each box represents a gene encoding a different mRNA and is drawn to scale in 3' to 5' orientation both within the virus genome and between the two genomes. The figure shows the differences in gene order between HMPV and RSV. The reading frames of gene M2 and gene L overlapping in RSV. RSV also encodes two extra genes i.e. NS1 and NS2 (Rima et al., 2017). 8

Figure 2: Temporal-scaled phylogeographic DENV-1 tree (Nunes et al., 2014). Each branch is coloured according to the most probable location as inferred using a discrete phylogeographic diffusion model. Geographic locations considered are shown in the left. Phylogenetic posterior probabilities percentages are shown next to relevant nodes along with the location-state posterior support. The number of sequences falling in Brazilian monophyletic lineages (highlighted in grey) is shown in brackets. For each lineage, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals (BCI) are shown in a black box 13

Figure 3: Phylogeographic reconstruction and spatial history of the trunk lineage (Lemey et al., 2014). Phylogeographic reconstruction and spatial history of the trunk lineage (Lemey et al., 2014). Maximum clade credibility (MCC) tree colored according to the time spent in the air communities as inferred by the GLM diffusion model. The tree represents one of the three different sub-sampled data sets discretized according to the 14 air communities. Branches are colored according the Markov reward estimates for each location. The uncertainty of these estimates is represented by superimposing an additional gray color proportional to the Shannon entropy of the Markov reward values. The trunk

lineage in the tree is represented by the thick upper branch path from the root to the nodes that represent the ancestors of samples that are exclusively from December 2006. The total time spent in each location (in years) along the trunk between 2002 and 2006 is plotted on the left of the tree. The trunk reward proportion for each location through time between 2002 and 2006 is summarized at the top of the tree. Both the total trunk time and the trunk reward proportions through time are averaged over the three sub-sampled data sets. In the trunk proportion through time plot, the number of Southeast Asian and Chinese samples are represented by a white full and dashed line respectively (secondary Y-axis). 14

Figure 4: The locations of PERCH study sites are shown by the coloured location-pointers. A single site was enrolled in each country i.e. Kilifi; Kenya, Lusaka; Zambia, Bamako; Mali, Soweto; South Africa and Basse; The Gambia. 19

Figure 5: HMPV subgroup prevalence and temporal patterns derived from G gene sequence data collected from Kenya, Mali, Gambia, South Africa and Zambia between August 2011 and January 2014. 32

Figure 6: RSV subgroup prevalence and temporal patterns derived from G gene sequence data collected from Kenya, Mali, Gambia, South Africa and Zambia between August 2011 and January 2014. 33

Figure 7: ML phylogenetic tree of HMPV subgroup B1 G gene sequences collected from Kenya, Mali, Gambia, South Africa and Zambia between August 2011 and January 2014. Tip shapes are coloured by country of sampling. Taxon labels are coloured in blue and green to differentiate the major phylogenetic clusters. Bootstrap values for each clade are indicated next to the nodes. Panel b, PopART minimum spanning network of the genetic distances of the viruses between and within countries. Clusters in red margin indicate

higher sequence similarity between the sequences and potential inter-country transmission links. 34

Figure 8:Temporal scaled maximum clade credibility (MCC) tree constructed using B1 G gene sequences obtained from Africa and GenBank collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix GB. Posterior probabilities are shown next to nodes. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade..... 35

Figure 9:Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV A2b G gene sequences obtained from Africa and GenBank collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Posterior probabilities are shown next to nodes. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix GB. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade..... 36

Figure 10: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV B2 G gene sequences obtained from Africa and GenBank collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Posterior probabilities are shown next to nodes. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix GB. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade..... 37

Figure 11: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV B1 (panel a) and B2 (panel b) G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. The tips were coloured according to the country of sampling. Posterior probabilities support for each node are indicated next to the nodes..... 38

Figure 12: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV A2c (panel a), A2b (panel c) and A2a (panel c) G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. The tips were coloured according to the country of sampling. Posterior probabilities support for each node are indicated next to the nodes. 39

Figure 13: Temporal scaled maximum clade credibility (MCC) trees constructed using RSV BA (panel a), RSV and RSV GA2 (panel b) G gene sequences obtained from Kenya,

Mali, Gambia, South Africa and Zambia. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. The tips were coloured according to the country of sampling Posterior probabilities support for each node are indicated next to the nodes.	42
Figure 14:Temporal scaled maximum clade credibility (MCC) trees constructed using RSV ON1 G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. The tips were coloured according to the country of sampling.....	43
Figure 15: The inferred continental (Africa) migration pathways for RSV BA (panel a), GA2 (panel b) and ON1 (panel c) viruses constructed under symmetric diffusion model. The lines indicate connection between different countries. The colour gradient from red to blue of the lines indicate the relative strength of connection between countries according to the Bayes Factor test. Only statistically supported migration pathways (Bayes Factor >3) are shown. Posterior probabilities >95% shows very strong supported rates.....	44

ABBREVIATIONS

HMPV	Human metapneumovirus
RSV	Respiratory syncytial virus
ALRTI	Acute lower respiratory tract infection
tMRCA	Time to the most recent common ancestor
PERCH	Pneumonia Etiology Research for Child Health
ESS	Effective sample size
KML	Keyhole markup language
BSSVS	Bayesian Stochastic Search Variable Selection

INTRODUCTION

Background information

Human metapneumovirus (HMPV) and respiratory syncytial virus (RSV) are both important respiratory pathogens that cause seasonal epidemics of acute respiratory tract infection which contribute significantly to childhood pneumonia. A recent multi-country study, PERCH (Pneumonia Etiology Research for Child Health), in Africa and Asia reported RSV as the main cause of pneumonia in children under five years of all the pathogens (O'Brien et al., 2019). RSV accounted for about 31% of the aetiological distribution. HMPV was also important and accounted for at least 5% of the aetiological distribution. This thesis presents a molecular-epidemiological analysis of samples collected by the PERCH study from the five African countries, i.e., Kenya, South Africa, Zambia, Mali and Gambia.

HMPV and RSV share similar clinical and epidemiological profiles (Moe et al., 2017; Schildgen et al., 2011). Clinical presentation of HMPV and RSV infections ranges from mild upper respiratory tract illness to severe lower respiratory tract disease (Schildgen et al., 2011). Infections with both pathogens can occur across all ages with severe disease in children under five years, immunocompromised and the elderly. HMPV and RSV have a seasonal distribution which overlap and tends to peak in winter and spring months in the southern and northern hemispheres. In the tropics, outbreaks are associated with rainy seasons or high relative humidity (Li et al., 2019).

The two viruses have a similar genetic structure and are all from the same family. HMPV has a genome size of about 13kb and encodes 8 genes i.e. 3`Nucleoprotein (N) - Phosphoprotein(P) - Matrix(M) - Fusion(F) - Matrix2(M2-1&M2-2) - Small hydrophobic protein(SH) -Glycoprotein(G) - Polymerase(L)5` (Kim et al., 2016). RSV has a genome size of about 15kb that encodes 10 genes i.e. 3`Non-structural 1(NS1) - Non-structural

2(NS2) - Nucleoprotein(N) - Phosphoprotein(P) - Matrix(M) - Small hydrophobic protein(SH) - Glycoprotein(G) - Fusion(F) - Matrix2(M2-1&M2-2) - Polymerase(L)5' (Agoti et al., 2015a). The differences in the genomic structures of HMPV and RSV is that HMPV has a different gene order and lacks non-structural proteins NS1 and NS2 (Hoogen et al., 2002). Of all the genes, the G gene is the most variable gene and is used to discriminate the genetic variants. There are two distinct HMPV antigenic groups, A and B (Hoogen et al., 2004). The two groups are further subdivided into four subgroups, A1, A2 (group A) and B1 and B2 (group B) based on phylogenetic analysis of F and G genes but do not show clear antigenic differences (Hoogen et al., 2004;Kim et al., 2016). Subgroup A2 is the most heterogeneous classified into clades A2a, A2b and A2c (Huck et al., 2006; Jagušić et al., 2017). Similarly, RSV is classified into two distinct groups (RSV A and B) based on antigenic and genetic variability (Johnson et al., 1987; Mufson et al., 1985). The two groups are further classified into multiple genotypes based on nucleotide differences within the RSV G gene.

HMPV and RSV epidemics are characterised by co-circulation of multiple subgroups/genotypes and the dominant subgroup/genotype may vary from year-to-year (Agoti et al., 2015b; Oketch et al., 2019). There is appearance and disappearance of some genotypes. Continued accumulation of genetic changes in circulating types occasionally gives rise to new genotypes. Due to G gene diversification, two new emergent clades of HMPV A2b with 111 or 180 nucleotide duplication within the G gene have been reported, first observed in 2017 and 2014 respectively (Saikusa et al., 2017a; Saikusa et al., 2017b). Similar evolutionary changes have also been reported for RSV with the emergence of two novel genotypes, RSV A ON1(2011) and RSV BA(1999) (Eshaghi et al., 2012; Trento et al., 2003). RSV A ON1 has 72 nucleotide duplications while RSV BA has a 60 nucleotide duplication, all within G gene (Eshaghi et al., 2012; Trento et al., 2003).

HMPV and RSV variants are reported to cluster temporally with limited geographical clustering implying widespread movement of HMPV and RSV variants (Agoti et al., 2015b; Oketch et al., 2019).

It is not clear whether HMPV and RSV share geographic spread patterns. This is as a result of scarce sequence data and strongly metachronous sampling in time and space, especially in continental Africa. As a result, the origins and interconnectedness of RSV and HMPV epidemic across Africa are not well understood. Viral sequence sampling on a time scale allows real-time tracking of viral strains. In addition, integration of sequence data with spatial-temporal data, allows reconstruction of transmission histories necessary for tracing of epidemiological linkages especially when there is limited case surveillance and tracing (Lemey et al., 2009). This study provides a unique set of data, well clinically phenotyped, collected at the same time and across five African countries; The Gambia, Kenya, Mali, South Africa and Zambia. Using this data set, we aimed to assess the virus relatedness across the continent, to trace origins and to infer dispersal patterns.

Statement of the Problem

RSV is the leading cause of pneumonia in children aged under five years and accounts for ~33 million cases of acute lower respiratory tract infection (ALRTI) annually (Shi et al., 2017). HMPV is also an important respiratory pathogen and accounts for ~10% of cases of ALRTI in children under five years. To date, there are no vaccines for HMPV and RSV. Understanding HMPV and RSV spatial-temporal patterns could help inform on possible sources of HMPV and RSV introductions into Africa. Also, by comparing the geographic patterns of HMPV and RSV across Africa, our study illuminates on the patterns of spread of seasonally recurring respiratory viruses necessary for public health interventions.

Main objectives

To undertake comparative phylogenetic and phylogeographic analyses of HMPV and RSV using virus sequence data to understand the geographic spread patterns of these two viruses.

Specific objectives:

- To compare and contrast the genetic diversity of HMPV and RSV using G gene sequence data collected from 5 African countries
- To compare and contrast the geographic spread patterns of the two viruses across the 5 African countries
- To assess whether the G gene is adequate to examine RSV and HMPV phylogeography in detail.

Research Questions

This study aimed to investigate the relatedness of the viruses across the five African countries to explore the degree of inter-country virus mixing or isolation and then compare the patterns of spread of HMPV and RSV. Do the two viruses differ in their phylogenetic profiles suggestive of different rates of transmission and is there evidence of different spatial patterns of circulation suggesting different pathways of geographical spread?

Justification

Integration of virus sequence data and epidemiological data in phylogenetic analyses has proven useful in the reconstruction of spatial-temporal histories of respiratory viruses and give insights into evolutionary dynamics underlying epidemics (Hadfield et al., 2018; Lemey et al., 2009). This has allowed real-time assessment of virus evolution and assessment of local, national and global spread patterns of respiratory pathogens such as

influenza virus (Bedford et al., 2010; Hadfield et al., 2018), necessary for the optimization of influenza virus vaccine design and delivery. However, such inferences are limited by the availability of data. In Africa where the greatest pneumonia burden occurs, sequence data remain sparse. As a result, the origins and epidemic interconnectedness are not well understood and little is known about the geographic spread patterns of respiratory viruses including HMPV and RSV. Using sequence data collected from 5 African countries we undertook a comparative phylogenetic analysis of spatial-temporal patterns of HMPV and RSV to explore their patterns of spread across Africa. The samples analysed provide a unique set; collected over the same time between August 2011 to January 2014 and well clinically phenotyped.

We further explored the relationship between viruses in Africa and globally to provide context within which African viruses are found and inform on potential sources of African continental introductions and role in worldwide circulation. Our analysis provides a precedent description of HMPV phylogeography across Africa and therefore represent a significant reference in our understanding of HMPV spatial-temporal transmission patterns. Besides, identifying similarities and differences in transmission characteristics of HMPV and RSV could validate the generalisation of inferences on pathogen spread.

Limitations and scope of the Study

The limitation of this study is that the data was restricted to five countries over a two-year period (all sites enrolled for 24 months) and the modest sample size, HMPV n=278 and RSV n=934. This study was also limited to hospital cases which may not reflect community transmission. Only children were recruited whereas these are also diseases of the elderly. Finally, samples were collected from a single site per country and therefore, fine grain detail of the transmission pathways were not possible.

LITERATURE REVIEW

Epidemiology

HMPV and RSV are single-stranded, negative-sense RNA viruses classified in the *Pneumoviridae* family (Rima et al., 2017). The two viruses share similar clinical and epidemiological characteristics (Moe et al., 2017; Schildgen et al., 2011). Clinical manifestations of HMPV and RSV infections range from mild upper respiratory tract illness to severe lower respiratory tract disease, or may be asymptomatic. Clinically, there are no specific characteristics that distinguish HMPV from RSV infection (Schildgen et al., 2011). Clinical characteristics commonly associated with both pathogens are bronchiolitis and pneumonia (Moe et al., 2017; Xepapadaki et al., 2004). Diagnoses of both pathogens rely mostly on molecular techniques (Miller et al., 2018; Kumar et al., 2014). Similar to RSV, HMPV infections can occur across all ages with severe disease in young children, immunocompromised and the elderly (Schildgen et al., 2011; Shi, 2019). Hospitalization due to RSV and HMPV is highest in infants less than 1 year of age but can occur throughout childhood (Panda et al., 2014; Shi et al., 2017). Many studies report that the peak age of hospitalization for HMPV tends to occur in older children (6- 12 months) compared to RSV peak age of hospitalization that occurs in early infancy (2-3 months) (Peiris et al., 2003). HMPV and RSV co-infections or co-infections with other respiratory viruses are possible, however, co-infections are not very common (Woensel et al., 2006). Reinfections with both pathogens are also possible and can occur across all ages (Glezen et al., 1986; Pelletier et al., 2002; Shi, 2019; Shafagati & Williams, 2018). This may be due to incomplete immunity that wanes overtime upon previous-exposures combined with ongoing antigenic variation in viral epitopes that may support immune escape (Henderson et al., 1979; Schildgen et al., 2011). As yet, there is unclear knowledge of HMPV and RSV immune responses in humans, and the evasion mechanisms of the

immune system used by the two pathogens (Soto et al., 2018; Tognarelli et al., 2019). The current studies suggest that both HMPV and RSV have low innate immune responses that translate to aberrant adaptive immunity (Soto et al., 2018; Tognarelli et al., 2019). One of the mechanisms in HMPV immune responses involves the promotion of an anergic state in T cells associated with low levels of cytokine secretions (Soto et al., 2018). These studies suggest that RSV-specific T cells fail to expand *in vivo* following reinfections. RSV infects dendritic cells altering their maturation, migration to lymph nodes, and their ability to activate virus-specific T cells consequently affecting the host response (BONT, 2002; Tognarelli et al., 2019).

Globally, RSV is the leading cause of LRTI in children with the highest hospitalization rates in children aged less than 6 months (Shi et al., 2017). Almost all children are affected by the age of 2 years. Global estimates for children under five years of age indicate RSV causes about 33 million cases of ALRTI, ~3.2 million of which lead to hospitalization and ~60,000 lead to hospital mortality (Shi et al., 2017). HMPV accounts for ~10% cases of ALRTI in children (Kahn, 2006). Although not as common as RSV, HMPV is a frequent respiratory pathogen and has a global distribution (Panda et al., 2014). Seroprevalence studies suggest that most children are affected by the age of five years (Schildgen et al., 2011). Both RSV and HMPV have a seasonal distribution of occurrence with outbreaks mainly in cold seasons in temperate climatic regions (Li et al., 2019). In the tropical climates, the association of peaks transmission months and weather patterns is less clearly defined (Chow et al., 2016; Li et al., 2019; Owor et al., 2016). Studies have reported that HMPV peak periods may overlap with RSV peak season or occur after the RSV season (Mizuta et al., 2013; Owor et al., 2016).

Transmission

Transmission of the two viruses primarily occurs through direct or close contact with contaminated secretions (e.g. large droplets) or materials and surfaces (i.e. fomites) (Hall et al., 1981; Kahn, 2006). Upon infection with RSV, the incubation period may range from 4 to 6 days (Lessler et al., 2009). The period of viral shedding is about a week (Okiro et al., 2010), however, the duration of shedding may vary with age and other factors such as history of exposure (King et al., 1993). Similar to RSV, HMPV incubation period varies from individual to individual and lasts about 5 days (Matsuzaki et al., 2013). The shedding period is thought to last between 1 to 2 weeks after acute illness (Panda et al., 2014).

Genetic structure and molecular epidemiology

The genomic structures of HMPV and RSV closely resemble each other except that HMPV has a different gene order and lacks two non-structural proteins (NS1 and NS2) that are present in RSV (Hoogen et al., 2002), Figure 1.

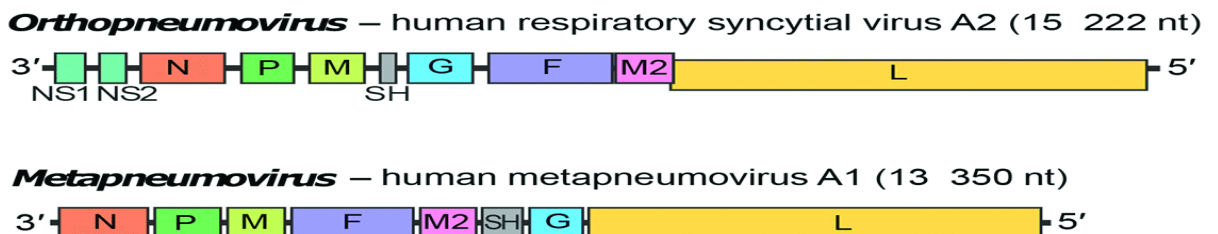


Figure 1: Genomic organization of human metapneumovirus (HMPV) and respiratory syncytial virus (RSV), (Rima et al., 2017). HMPV is classified in genus Metapneumovirus, RSV belongs to genus Orthopneumovirus, both in the Pneumoviridae family. Each box represents a gene encoding a different mRNA and is drawn to scale in 3' to 5' orientation both within the virus genome and between the two genomes. The figure shows the differences in gene order between HMPV and RSV. The reading frames

of gene M2 and gene L overlapping in RSV. RSV also encodes two extra genes i.e. NS1 and NS2 (Rima et al., 2017).

HMPV has a genome size of about 13Kb composed of 8 genes i.e. $3'N-P-M-F-M2-SH-G-L5'$ (Kim et al., 2016). RSV genome is about 15.2Kb and encodes 10 genes i.e. $3'NS1-NS2-N-P-M-SH-G-F-M2-L5'$ (Agoti et al., 2015a). Similar to RSV, the N, L and P protein form the viral replication complex (Piyaratna et al., 2011). The F, SH and G genes code for surface proteins i.e., the fusion glycoprotein (F protein), the small hydrophobic glycoprotein (SH protein) and the attachment glycoprotein (G protein), respectively. F and G proteins are the primary immunogenic proteins in both pathogens and targets for vaccine development (Xiao et al., 2019). Among the three surface proteins, F gene is the most highly conserved gene in both HMPV and RSV and even between these two viruses (33%-38%) (Hoogen et al., 2002). The G gene is the most variable and has been widely used to characterize virus evolution and in establishing the genetic variants (Bastien et al., 2004; Johnson et al., 1987). Both HMPV G and RSV G genes are characterised by high sequence polymorphism within and between groups, often occurring due to use of alternate start codon and sequence substitutions (Peret et al., 2004; Sullender et al., 1991). The evolution of these two proteins (HMPV G and RSV G) is hypothesised to occur by positive selection due to immune pressure (Peret et al., 2004; Woelk & Holmes, 2001). Compared to RSV G gene, the HMPV G gene is reported to have higher evolutionary rates (Padhi & Verghese, 2008). This is attributed to the fact that the HMPV G gene is not a major neutralizing or protective antigen as is the case for RSV and paramyxoviruses and therefore not limited by structural or functional constraints (Biacchesi et al., 2005; Skiadopoulos et al., 2006).

HMPV is classified into two groups, A and B, based on antigenic and nucleotide differences in the nucleoprotein (N), fusion (F) and attachment (G) glycoprotein genes

(Hoogen et al., 2004; Kim et al., 2016). Based on phylogenetic analysis of the F and G genes, the two groups are further classified into subgroups A1 and A2 (group A) and B1 and B2 (group B) (Kim et al., 2016). HMPV A2 is the most heterogeneous with multiple circulating clades (A2a, A2b and A2c) (Huck et al., 2006; Jagušić et al., 2017). Additionally, two novel clades within subgroup A2 with 111 (HMPV_A2b111nt-dup) or 180 (HMPV A2b180nt-dup) nucleotide duplication within the G gene have been reported (Saikusa et al., 2017a; Saikusa et al., 2017b). HMPV A2b180nt-dup was first detected in 2014, in Spain (Piñana et al., 2017) and Japan (Saikusa et al., 2017b), and became major epidemic strains. Similarly, upon first detection of HMPV_A2b111nt-dup in 2017, in Japan (Saikusa et al., 2017a), it became a predominant strain replacing the classical HMPV A2b (Saikusa et al., 2019). Clinically, there is no difference in disease severity between the subgroups (Shafagati & Williams, 2018). HMPV subgroups can co-circulate with a shift in predominant subgroup. Subgroup prevalence's can also vary by year and location (Panda et al., 2014). Geographically, HMPV subgroups are known to circulate widely and cluster temporally except subgroup A1 that has been identified in few countries (Oketch et al., 2019; Reiche et al., 2014). Molecular epidemiology studies reveal that HMPV epidemic seasons are characterised by multiple circulating subgroups, and appearance and disappearance of the subgroups (Oketch et al., 2019; Pitoiset et al., 2010; Reiche et al., 2014). Virus persistence during local-community HMPV epidemics has been reported to often result from multiple entry of new viruses from the global pool into the local community (Oketch et al., 2019).

Similar to HMPV, RSV is classified into two distinct groups, RSV A and RSV B, based on antigenic and genetic variability (Johnson et al., 1987; Mufson et al., 1985). The two groups are further classified into genotypes based on phylogenetic divergence in the G gene. Identical to HMPV subgroup assignment (Hoogen et al., 2004), RSV genotype

assignment is based on pragmatic reasoning to discriminate the genetic variants within the RSV groups. The genotypes designate clusters of viruses within the groups with viruses from a single genotype having less genetic distance between each other than to viruses of any other genotype (Agoti et al., 2015b). The genotypes have been characterised further into (a) imported variants which show greater genetic difference than expected from in situ diversification and (b) locally persisting variants that are sustained in situ (Agoti et al., 2015b; Otieno et al., 2016). RSV genotypes are reported to cluster temporally with limited geographical clustering suggesting widespread movement of RSV variants (Agoti et al., 2015b). Owing to duplication in RSV G gene, two new emergent genotypes have been reported, RSV A ON1 and RSV BA (Eshaghi et al., 2012; Alfonsina Trento et al., 2003). The BA genotype was first detected in Buenos Aires, Argentina in 1999 (Alfonsina Trento et al., 2003). The genotype had a 60-nucleotide duplication within the C terminal of G gene. This genotype (BA) subsequently spread rapidly throughout the world and has become a predominant form of RSV B (A. Trento et al., 2010). Similarly, RSV A ON1 genotype has circulated worldwide since its first description in Ontario Canada, in 2001 (Agoti et al., 2014; Pierangeli et al., 2014). The genotype has a 72-nucleotide duplication, also within the C terminal of G gene. Several studies have described long-term molecular epidemiology of RSV in different geographic regions including Kenya (Agoti et al., 2015b) and United Kingdom (Cane et al., 1994). These studies reveal RSV epidemics characterised by co-circulation of multiple genotypes/variants, predominance of different genotypes and, appearance and disappearance of some genotypes. There is limited persistence of RSV genotypes/variants across multiple epidemics surviving the interepidemic troughs. Genotypic dominance can also vary based on year and location (Sullender et al., 2000). The genotype/subgroup replacement patterns observed for both HMPV and RSV is attributed to herd-immunity

against circulating strains and possible infection with the heterologous strains. Virus persistence within the communities is characterized by continual introduction of new variants from the global pool alongside establishment of local variants that may persist between epidemics (Agoti et al., 2015b; Giallonardo et al., 2018).

To date, many studies have characterised the molecular epidemiology of RSV. Besides, the environmental and critical social determinants of RSV transmission such as social contacts, air pollution, temperature, in-house smokers and relative humidity which might increase the risk of RSV infections have been well studied (Munywoki et al., 2014; Okiro et al., 2008; Pitzer et al., 2015). Conversely, little is known about the global dispersal patterns and epidemiological factors that influence the spread patterns. Such inferences could inform public health intervention measures. Similar to RSV, we lack an understanding of HMPV global phylodynamic and epidemiological processes.

Phylogeographic Analyses

Use of pathogen sequence data in phylogeographic analysis has proven useful in reconstructing the evolutionary history and to infer spatial patterns of the viral pathogens. For instance, to trace timings of introduction and spread of dengue (Figure 2) (Nunes et al., 2014) and Zika viruses in Brazil (Faye et al., 2014), and global spread of the Middle East respiratory syndrome (MERS) (Min et al., 2016). Phylogeographic analysis has also allowed testing of hypothesis and identifying potential predictors of viral spatial spread such as air travel, geographic proximity and population density (Lemey et al., 2014). For example, phylogeographic analysis has proven useful in unravelling the global source-sink dynamics of influenza virus, Figure 3, (Lemey et al., 2014). In addition, the integration of sequence data with epidemiological surveillance data together with computational methods into a statistical framework permit simultaneous inferences of virus population dynamics and spatial-temporal histories (Lemey et al., 2014).

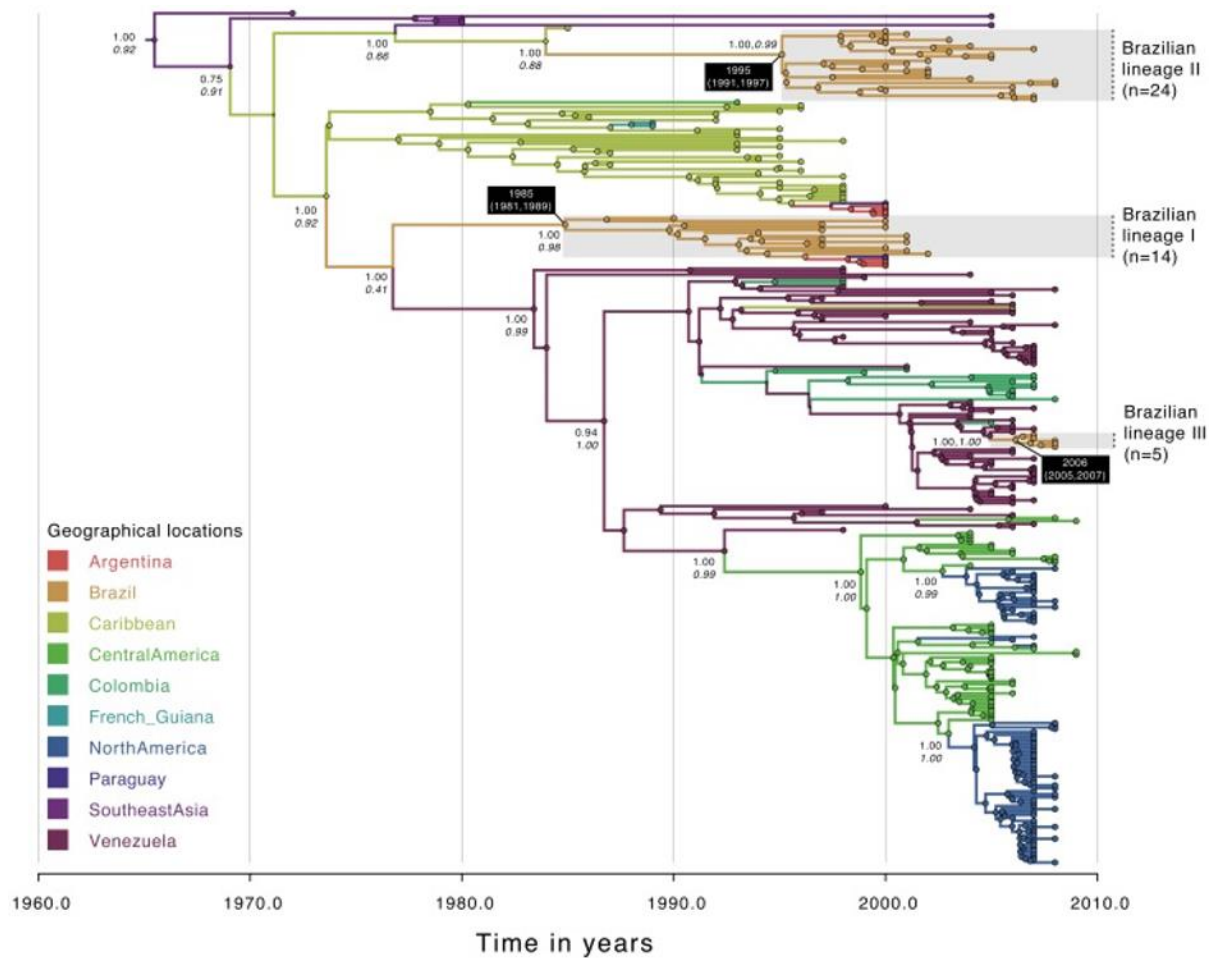


Figure 2: Temporal-scaled phylogeographic DENV-1 tree (Nunes et al., 2014). Each branch is coloured according to the most probable location as inferred using a discrete phylogeographic diffusion model. Geographic locations considered are shown in the left. Phylogenetic posterior probabilities percentages are shown next to relevant nodes along with the location-state posterior support. The number of sequences falling in Brazilian monophyletic lineages (highlighted in grey) is shown in brackets. For each lineage, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals (BCI) are shown in a black box

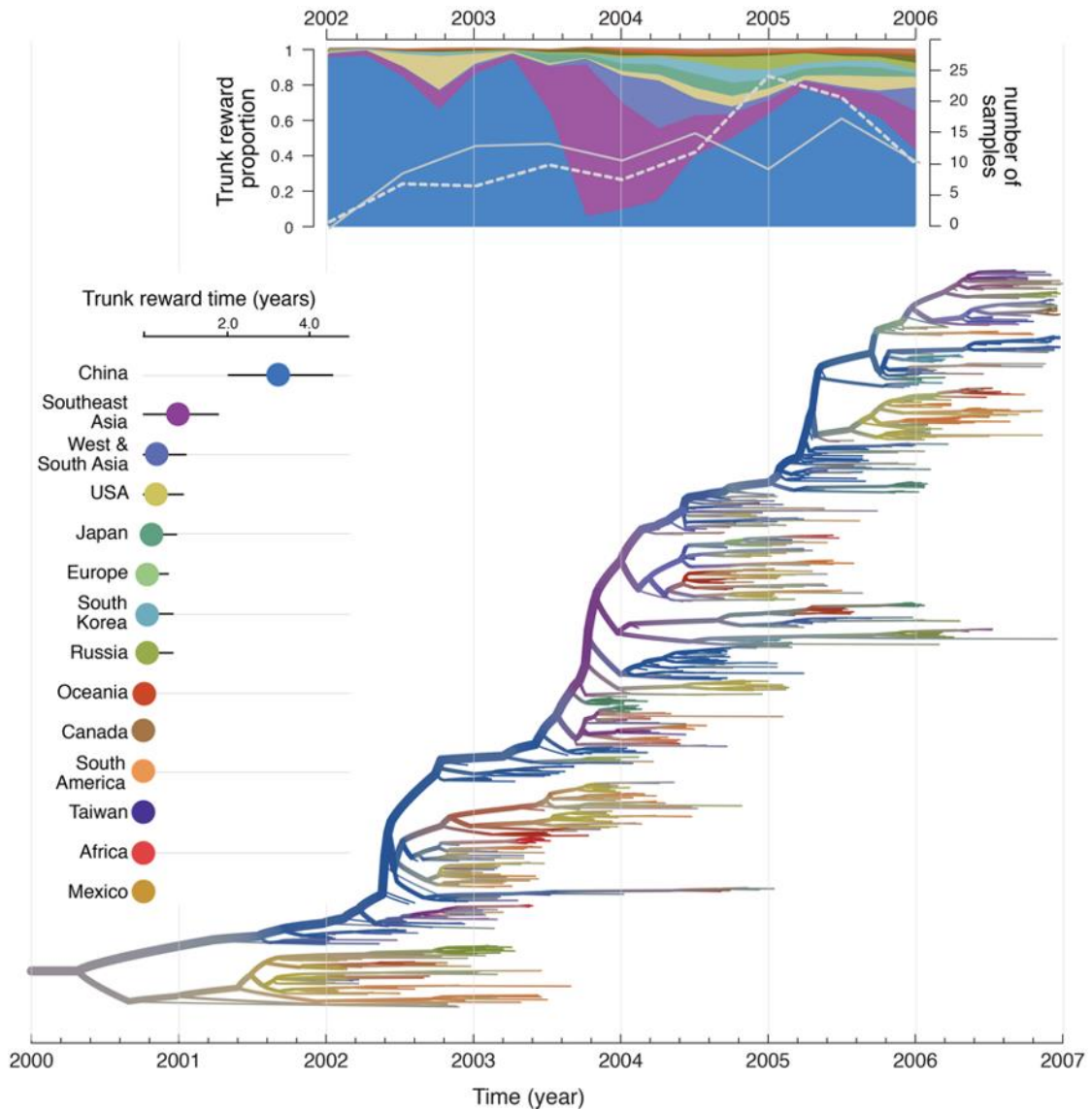


Figure 3: Phylogeographic reconstruction and spatial history of the trunk lineage (Lemey et al., 2014). Phylogeographic reconstruction and spatial history of the trunk lineage (Lemey et al., 2014). Maximum clade credibility (MCC) tree colored according to the time spent in the air communities as inferred by the GLM diffusion model. The tree represents one of the three different sub-sampled data sets discretized according to the 14 air communities. Branches are colored according to the Markov reward estimates for each location. The uncertainty of these estimates is represented by superimposing an additional gray color proportional to the Shannon entropy of the Markov reward values. The trunk lineage in the tree is represented by the thick upper branch path from the root to the nodes

that represent the ancestors of samples that are exclusively from December 2006. The total time spent in each location (in years) along the trunk between 2002 and 2006 is plotted on the left of the tree. The trunk reward proportion for each location through time between 2002 and 2006 is summarized at the top of the tree. Both the total trunk time and the trunk reward proportions through time are averaged over the three sub-sampled data sets. In the trunk proportion through time plot, the number of Southeast Asian and Chinese samples are represented by a white full and dashed line respectively (secondary Y-axis).

Unlike influenza virus, reconstruction of global phylodynamic processes of both HMPV and RSV is compromised by limited sequence data and strongly metachronous sampling in time and space. Phylogeographic studies of both pathogens have been limited to within country or regional surveillance and thus little is known about the global diffusion patterns of the two pathogens. These studies (Rojo et al., 2017a; Velez et al., 2013) reveal HMPV and RSV migration events occur both globally and locally. Whether this is a natural phenomenon for seasonally recurring respiratory viruses requires further investigation.

Previous studies have reported on the importance of full-genome sequences to identify and trace individual transmission chains of viruses such as influenzas (Baillie et al., 2012) and Ebola (Arias et al., 2016). Similar findings have been reported by RSV house-hold transmission studies (Agoti et al., 2019). Coupled with spatial information, whole-genome sequences provide increased resolution useful for detailed transmission studies, applicable in real-time tracking of infections (Houlihan et al., 2018). Whole-genome sequences also provide better estimates of time to the most recent common ancestors and rates of evolution (Agoti et al., 2015a; Otieno et al., 2018). Also, changes across the genome can be determined for enhanced diversity studies (Otieno et al., 2018). Although

HMPV G and RSV G ORFs show similar phylogenetic relationships as their full-genome sequences, the single ORFs may not discriminate genetic variants collected over a short epidemiological timescale within the same epidemic (Agoti et al., 2015; Kim et al., 2016). In this study, samples were collected across 2 years. Therefore, we hypothesized that the patterns of spread across different countries will require shorter sequence lengths to resolve than spread patterns within the country since across countries there is likely to be longer time and more transmission events to allow selection and hence diversification.

Phylogeographic methods

Different phylogeographic methods have been developed including maximum parsimony (MP) (Cunningham et al., 1998), Maximum likelihood (ML) (Pagel, 1999) and Bayesian methods (Lemey et al., 2009; Lemey et al., 2010). These methods aim to reconstruct the diffusion history between locations of the sampled viruses. Locations are treated as inherited traits of the viruses and used to estimate ancestral locations for ancestral viruses. MP involve reconstruction of phylogenetic tree omitting the spatial data and then conditions the phylogeographic inferences on the reconstructed phylogeny. Characters are mapped onto a single phylogeny in a heuristic approach. The model aims at minimising the number of historical state exchange between locations, necessary for the localities of the sampled sequences to be consistent with the genetic phylogeny. As a result, phylogeographic inferences based on MP can be misleading when the evolution rates are rapid and when the character exchange probabilities are unequal (Cunningham et al., 1998). Moreover, MP does not account for phylogenetic uncertainty and for uncertainty in the dispersal process (Cunningham et al., 1998). Compared to MP, ML and Bayesian methods employ an evolutionary model with the ability to glance over the entire state history over the phylogeny and conveniently draw statistical inferences (Lemey et al., 2009). ML and Bayesian methods offer the ability to reconstruct the dispersal process

throughout evolutionary history. The methods account for uncertainty in the estimation of character evolution (Lemey et al., 2009). A classic example of ML method is implemented in Nextstrain (Hadfield et al., 2018) for real-time phylogeographic inferences of endemic viral disease such as RSV subtype A, seasonal influenza, dengue, and emergent viral outbreaks such as Zika and Ebola. The model estimates the probabilities of all possible location states at every node on the tree determined by the distribution of the terminal traits (discrete traits such as country/region of isolation). The reconstructed ancestral states allow identification of probable transmission routes plus the inferred probability distribution of ancestral states at every node.

Similar to ML method, the Bayesian methods utilises the probabilistic model to estimate the relative posterior probabilities of each location state at every node along the phylogenetic tree. Bayesian methods permit the simultaneous reconstruction of spatial-temporal and demographic dynamics of the evolving pathogen (Lemey et al., 2009). Besides, Bayesian method employs a model that offers the ability to simultaneously reconstruct spatial-temporal history and test the contribution of potential predictors of spatial spread (Lemey et al., 2014). Two Bayesian methods exist, discrete (Lemey et al., 2009) and continuous (Lemey et al., 2010) phylogeography, all implemented in BEAST software package (Drummond et al., 2012). Discrete phylogeography inferences are performed using continuous-time Markov chain process. Sequence sampling locations are considered as discrete states. Markov chains are used to model dispersal history between discrete states and infers a posterior distribution of phylogenies whose internal nodes are associated with an estimated ancestral location. Because of the discretised dispersal process, the model does not explicitly model diffusion in a continuous space (Faria et al., 2011). The inferred location of ancestral viruses at any node along the phylogeny can only be drawn in the sampled locations (Faria et al., 2011). Nevertheless,

the method provides a useful framework for hypothesis testing to investigate the epidemiological linkage between sampled locations. In contrast, Continuous phylogeography relies on Brownian motion process (analogue to the Markov chain transition model) that employs a relaxed random walk diffusion model (Lemey et al., 2010). This model allows to reconstruct the phylogeographic history on a continuous landscape and generates a posterior distribution of phylogenies whose internal nodes are associated with geographic coordinates. In contrast to discrete phylogeography, the continuous model allows ancestral viruses to reside at any location in a continuous geographic landscape. As a result, the model can provide a more realistic spatial inference (Faria et al., 2011). However, the model may only apply to limited observation scales such as within country for locally circulating viruses whose spatial dispersal process is likely to adhere to a homogenous Brownian process (Faria et al., 2011). Because this model considers the relationship between dispersal and geographic distance, it may not be tenable to viruses whose dispersal patterns are likely to be influenced by host mobility patterns. Due to complementary approach of the continuous and discrete models, the two methods provide a combined strategy to examine spatial processes at both local (e.g. within country) and wider (e.g. between country) geographic scales, respectively (Baele et al., 2017; Faria et al., 2011). In this study, HMPV and RSV dispersal patterns between countries were reconstructed using Bayesian discrete method.

MATERIALS AND METHODS

Study population and study sites

This study analysed sequence data for the G gene (encoding the attachment protein) for samples positive for HMPV (n=278) and RSV (n=934) collected from 5 African countries (The Gambia, Zambia, Mali, South Africa and Kenya), Figure 4.



Figure 4: The locations of PERCH study sites are shown by the coloured location-pointers. A single site was enrolled in each country i.e. Kilifi; Kenya, Lusaka; Zambia, Bamako; Mali, Soweto; South Africa and Basse; The Gambia.

Table 1: Virus positive by site and Number sequenced

A)								B)						
HMPV								RSV						
Site	Enrollment date	Cases			Controls		Total sequenced	No. of samples	Cases			Controls		Total sequenced
		No. of samples	No. of cases	No. of sequenced	No. of controls	No. of sequenced			No. of sequenced cases	No. of sequenced	No. of controls	No. of sequenced		
Gambia	November 2011 - October 2013	46	37	32	9	9	41	117	113	97	4	2	99	
Kenya	August 2011 - November 2013	62	50	50	13	8	58	263	251	251	12	12	263	
Mali	January 2012 - January 2014	46	39	34	7	6	40	182	154	138	28	20	158	
South Africa	August 2011 - August 2013	77	55	44	22	14	58	260	232	208	28	22	230	
Zambia	October 2011 - October 2013	47	39	30	8	5	35	112	94	82	18	10	92	
Totals		278	220	190	59	42	232	934	844	776	90	66	842	

Abbreviations: HMPV, human metapneumovirus; RSV, respiratory syncytial virus. Total number of HMPV and RSV positive samples collected between August 2011 and January 2014 from 5 African countries. Panel A: Total number of HMPV sequences stratified by cases and controls, and total sequenced. Panel B: Total number of RSV sequences stratified by cases and controls, and total sequenced.

These samples were collected by the PERCH (Pneumonia Etiology Research for Child Health). All sites enrolled participants over 24 months, between August 2011 to January 2014, Table 1. The study design has been previously reported (Deloria-Knoll et al., 2012; Levine et al., 2012). The PERCH study recruited both cases and controls all of whom were selected to be living within defined catchment populations of known population size, and who presented to hospital for admission (cases) or to outpatient facilities for mild illness or vaccination (controls). A case was defined as a child aged between 28 weeks and 59 months from within the catchment area admitted to a participating hospital with WHO defined severe or very severe pneumonia, as previously described (Dale, 2006; O'Brien et al., 2019). Cases were eligible to enter into the PERCH study if they met the following inclusion criteria:

- Admitted to hospital
- Meets the WHO clinical criteria for severe or very severe pneumonia on admission

Table 2: WHO Clinical Criteria for Severe and Very Severe Pneumonia	
Classification	Cough or difficulty breathing plus any of the following signs or symptoms:
Severe Pneumonia	Lower chest wall indrawing
Very Severe Pneumonia*	Central cyanosis Unable to feed, or vomiting everything Convulsions, lethargy, or unconsciousness Head nodding

- Aged 1 -59 months
- Accompanied by written informed parental/guardian consent
- Lives in defined study catchment area (may be defined as all or part of a geopolitically defined area or distance zone from the hospital)

Controls were randomly selected from residents of the same catchment area as cases and frequency matched to cases by age group (1 to <6 months, 6 to <12 months, 12 to <24 months, and 24–59 months of age), as previously described (Deloria-Knoll et al., 2012; O’Brien et al., 2019). Controls were enrolled regardless of the respiratory symptoms. A written informed consent was obtained from the parent or a guardian. A nasopharyngeal

(NP) flocked swab or combination of nasopharyngeal swab and oro-pharyngeal(OP) swab was collected from each participant into viral transport medium for respiratory pathogen screening (Driscoll et al., 2017).

Laboratory methods

Respiratory pathogen screening

Standardize protocols as described below were implemented across all the PERCH study sites. Total nucleic acid extraction was performed on NP/OP specimens using the NucliSENS easyMAG platform (bioMérieux, Marcy l'Etoile, France) and virus screening done using the Fast-track Diagnostics Respiratory Pathogens 33 multiplex Realtime PCR kit (Fast-track Diagnostics, Sliema, Malta) (Driscoll et al., 2017). Quantitative PCR was carried out at each site (The Gambia, Zambia, Mali, South Africa and Kenya) using Applied Biosystems 7500 (ABI-7500) platform (Applied Biosystems, Foster City, California). Cycling conditions were performed at 50°C for 15 minutes, 95°C for 10 minutes, and 40 cycles of 95°C for 8 seconds followed by 60°C for 34 seconds. Samples with a rRT-PCR cycle threshold (Ct) value <40 were considered positive for the respective targets. PERCH study specimens were archived at the study reference laboratory (Canterbury Health Laboratories, Christchurch, New Zealand) (Driscoll et al., 2017).

HMPV and RSV G genes sequencing

Sequencing of was performed at one of PERCH's study site laboratory (KEMRI Wellcome Trust) in Kilifi, Kenya. All HMPV and RSV positive samples from The Gambia, Mali, South Africa and Zambia were obtained from the PRECH study biorepository reference laboratory (Canterbury Health Laboratories, Christchurch, New Zealand). Sequencing was done using capillary sequencing as described below and

sequence data deposited to the KEMRI-Wellcome Trust Research Programme data server or to GenBank.

Briefly, RNA extraction from HMPV and RSV positive samples was done using QIAmp Viral RNA Minikit (Qiagen, Germany) manual extraction method, following the manufacturer's instructions. HMPV PCR primers amplified full G gene, approximately 700bp (Oketch et al., 2019). HMPV subgroup specific primers were used in a one-step reverse transcription (RT) PCR (Qiagen). RT-PCR and sequencing primers are shown in Appendix A, Table A1. Thermocycling conditions were set at: 50 °C for 30 min, 95 °C for 15 min, 38cycles of 94°C for 1min, 53°C for 1min, 72 °C for 1 min, and a final extension of 10 min at 72 °C (Oketch et al., 2019). For RSV, a two-step PCR protocol was employed. The first-round amplification was performed using Qiagen one-step RT-PCR kit with outer forward (AG20) and outer reverse (F164) primers. Thermocycling conditions were set at: 50°C for 30 minutes, 95°C for 15 minutes, and then 40 cycles of 94°C for 30 seconds, 54°C for 30 seconds, and 72°C for 1 minute, followed by a final extension of 10 minutes at 72°C (Agoti et al., 2012). The primers for both first-round and second-round PCR and sequencing are listed in Appendix A, Table A1. Two micro- litres of the One-Step RT-PCR products were amplified in the second-round nested PCR using Qiagen TaqMan PCR kit mastermix with the inner primers BG10 and F1(Appendix A, Table A1). Thermocycling conditions were set at: 95°C for 2 minutes, followed by 30 cycles of 95°C for 45 seconds, 54°C for 45 seconds, and 72°C for 1 minute. A final extension at 72°C for 5 minutes was allowed. PCR products were purified using Qiagen PCR product purification kit. The amplified fragments were sequenced in both forward and reverse strands using Big dye terminator v1.3 chemistry run on ABI 3130xl instrument. For sequencing, in addition to the two nest primers BG10 and F1, two additional RSV subgroup-specific were included to increase the coverage (Appendix A,

Table A1). Primers G523F and G523R for RSV group A, and primers 533F and 533R for RSV group B. The raw sequence data (sequencing chromatograms) were deposited to the KEMRI-Wellcome Trust Research Programme data server except for Kilifi (Kenya) site which had been previously assembled and deposited to Genbank. For this study, the raw HMPV G and RSV G genes sequence sequenced fragments from the Gambia, Mali, South Africa and Zambia were edited and assembled using Sequencher v5.4.6 (Gene Codes Corporation). The newly assembled sequences were deposited to GenBank. For the subsequent analysis, sequences collected from Kilifi (Kenya) were retrieved from GenBank and collated with the newly assembled data from the rest of the study sites. The GenBank accession numbers of the sequences collected from all the five sites are listed in Additional file 1.

Data analysis

Phylogenetic and phylogeographic analyses

The collated nucleotide sequences were aligned using MAFFT v7.407 (Katoh & Standley, 2013) and manually curated in AliView v1.26 (Larsson, 2014). Best fitting nucleotide substitution and site heterogeneity models for the alignments were determined using ModelFinder (Kalyaanamoorthy et al., 2017) in IQTree v1.6.11 (Nguyen et al., 2015). Phylogenetic trees were then constructed using Maximum Likelihood (ML) approach implemented in IQTree v1.6.11 to classify the genetic diversity of HMPV and RSV. The branch support was evaluated by bootstrapping. HMPV and RSV subgroups/genotypes were confirmed if sequences clustered with reference sequences of HMPV (Miranda et al., 2008; Huck et al., 2006; Jagušić et al., 2017) and RSV (Eshaghi et al., 2012; Peret et al., 1998; Venter et al., 2001) genotypes.

To infer the genetic evolution, temporal phylogenetic signal in the sequence data i.e. a root-to-tip regression of genetic distance against year of sampling was tested using

TempEst software v1.5.3 (Rambaut et al., 2016). Best demographic models were estimated using the established methods in BEAST i.e. path sampling (Lartillot & Philippe, 2006) and stepping-stone sampling (Xie et al., 2011), under an uncorrelated lognormal relaxed molecular clock. The Markov Chain Monte Carlo (MCMC) chains were run to convergence. The MCMC convergence (effective sample size [ESS] > 200) for all parameters were evaluated in TRACER v1.7.1 (Rambaut et al., 2018). The best demographic models were selected for subsequent analysis.

To explore spatial patterns in continental Africa, sequences were assigned to geographical traits, that is; Western Africa (Mali, Gambia), Eastern Africa (Kenya) and Southern Africa (South Africa and Zambia) as discrete traits. To attain high spatial resolution, country locations were also assigned to sequences as discrete traits. Viral migration events between the discrete locations were inferred using Bayesian symmetric discrete phylogeography implemented in BEAST v1.10.4 software under an uncorrelated lognormal relaxed molecular clock and the best selected demographic model. The symmetric diffusion model, infers ancestral reconstruction using the standard continuous-time Markov chain (CTMC), in which the transition rates between locations are reversible (Lemey et al., 2009). We choose to explore the patterns in the context of other global G gene sequences within the detected subgroups to inform on the viral introduction into Africa. Contemporaneous sequences were retrieved from GenBank, only sequences with collection date and overlapping with the sequenced fragments were included in the analysis. Due to the scarcity of HMPV global G gene sequence data, all the data collected between 2000 and 2018 was included. A total of 714 sequences were retrieved from 20 different countries. The accession numbers for the retrieved sequences are listed in additional file 1. For RSV, sequences collected a year before and after our study were analysed to place our data into immediate context. A total of 1810 sequences from 28

different countries were retrieved, additional file 1. The collated sequences were aligned using MAFFT v7.407. Country and continent of sampling were assigned to sequences as discrete traits. The global phylogeographic analysis was carried out under the same symmetric diffusion model. The MCMC convergence was evaluated in TRACER v1.7.1. The BEAST trees were summarised using Tree annotator v2.6.0 (Drummond et al., 2012), and the summarised maximum clade credibility tree (MCC) visualized in FigTree v1.4.4. (<http://tree.bio.ed.ac.uk/software/figtree/>). To summarise BEAST reconstructed spatial diffusion histories, Spatial Phylogenetic Reconstruction of Evolutionary Dynamics using Data Driven Documents (Spread3) software was used (Bielejec et al., 2016). Spread3 v0.9.7.1 was used to generate html files and the reconstructed viral migration events visualized in a web browser.

Statistical analysis

Significant diffusion rates between discrete locations were determined using the Bayes factor (Lartillot & Philippe, 2006). Rate matrix log file for location states generated under the phylogeographic analysis using the Bayesian Stochastic Search Variable Selection (BSSVS) procedure was used as input to identify frequently invoked rates between locations. Significant migration pathways were summarised based on Bayes factor (BF) as follows: $BF \geq 1000$ indicates very strong support, $10 \leq BF \leq 1000$ strong support, and $3 \leq BF \leq 10$ indicates supported.

RESULTS

HMPV subtyping and subgroup temporal patterns

A total of 232 HMPV G gene sequences were obtained, 44% (102/232) were group A and 56% (130/231) were group B (Table1). All HMPV A sequences belonged to subgroup A2 and further clustered into clades A2a (18/102), A2b (35/102) and A2c (48/102), Appendix B - Figure S1a. Subgroup A1 viruses were not detected. Among group B, 82% (107/130) were subgroup B1 and 18% (23/130) were subgroup B2 (Appendix B - Figure S1a). Multiple subgroups co-circulated in each country (The Gambia, Kenya, Mali, South Africa and Zambia), Figure 5. Notably, subgroup A2a viruses were only detected in South Africa and Zambia. HMPV subgroup temporal patterns in Mali mirrored those in The Gambia (Figure 5).

HMPV genetic diversity and inter-country transmission network

From the G gene phylogenies, all subgroup B1 viruses clustered into two major phylogenetic clusters supported with strong bootstrap values >95%, clusters b1C1 (coloured blue) and b1C2 (coloured green), Figure 7. Within each cluster, sequences from the same Africa subregion i.e. West Africa (Mali and Gambia), East Africa (Kenya) and Southern Africa (South Africa and Zambia) largely clustered together into monophyletic clades. We used PopART software (<http://popart.otago.ac.nz/screenshots.shtml>) to construct minimum spanning network (MSN) of genetic distances to investigate potential inter-country transmission patterns. MSN shows pairwise distances between sequences without regard to an evolutionary model. Clusters were named and coloured to reflect their placement on the G gene ML phylogeny (Figure 7). Similar to B1 ML phylogeny (Figure 7) sequences largely clustered by geographical region, into major clusters and sub-clusters as depicted by ML phylogeny (Figure 7). MSN genetic clusters (highlighted in red margins) containing sequences that were deemed similar and from multiple

countries, suggesting possible transmission linkage, were identified. The patterns of clustering of sequences from this cluster (b1C2) were assessed on the global G gene time-resolved phylogeny. On the global time-resolved phylogeny, the two clusters (b1C1, b1C2) were placed into separate clades interspersed with global sequences (Figure 8). This suggests at least two distinct variants of B1 viruses circulated in these countries. Sequences in cluster b1C2 that were deemed similar equally clustered by Africa subregion into monophyletic clades interspersed with global sequences (Figure 8). Similar analysis was done for the rest of the detected HMPV genetic groups (A2b, A2c, A2a and B2).

Consistent with B1, ML phylogenies of B2 and A2b sequences revealed geographical clustering of sequences and at least two distinct variants of each subgroup circulated in these regions (Appendix B - Figure S2a and S2d). For A2c viruses, sequences formed single phylogenetic clusters (Appendix B - Figure S2c). Sequences from the same geographical region clustered into minor monophyletic clades. A2a sequences formed single phylogenetic clusters and were only detected in South Africa and Zambia (Appendix B - Figure S2d). The minimum spanning networks (MSN) for each subgroup (B2, A2b, A2c and A2a) are also shown in Appendix B - Figure S2 next to the subgroup specific ML phylogenies. Similar to ML phylogenies, sequences largely clustered by geographical region, into major clusters and sub-clusters as depicted by ML phylogenies. Lesser genetic distances were observed within A2a and A2c viruses concordant with high sequence similarity and lesser phylogenetic resolution observed for A2a and A2c viruses. Pairwise mean sequence identities were estimated at: 98% for A2a, 99% for A2c, 93% for A2b, 94% for B2, 95% for B1. For each subgroup, MSN of genetic clusters (highlighted in red margins) containing sequences that were deemed similar and from multiple countries, suggesting possible transmission linkage, were identified (Appendix

B - Figure S2). The patterns of clustering of sequences from these clusters i.e., b2C1, A2bC1 and, all A2a and A2c sequences (Appendix B - Figure S2) were assessed on the global G gene time-resolved phylogenies.

From the B2 and A2b globally time-resolved phylogenies, the major genetic clusters observed in B2 and A2b ML phylogenies fell into separate clades interspersed with global sequences (Figures 9 and 10). This suggests at least two distinct strains of each subgroup circulated in these African countries. The clusters were named to reflect their placement in ML phylogenies. A2c sequences fell into a single clade interspersed with global sequences (Appendix B - Figure S3). For A2a, sequences were placed on a single monophyletic cluster indicating a single introduction (Appendix B - Figure S4). Notably A2a sequences were only detected in Zambia and South Africa and indicates a unique introduction of A2a viruses in these locations (Appendix B - Figure S4). The clustering patterns of genetic clusters (b2C1, A2bC1 and all A2c sequences) that were regarded as possible transmission linkages were assessed in the global context, (Appendix B - Figure S2). Similarly, sequences mostly clustered by geographical region into monophyletic clades interspersed with global sequences (Figures 9, 10 and Appendix B - Figure S3).

We further assessed within-country sequence diversity. Only HMPV subgroup B1 viruses were detected in high frequencies in all the five countries and were analysed (Table 3). From the country-specific phylogenies, sequences formed two major phylogenetic clusters supported by strong bootstrap values >95 (Appendix B - Figure S5), representing the two genetic clusters that were observed on the G gene ML phylogeny of the collated B1 sequenced data, Figure 7. Clusters were coloured to reflect their placement on the ML phylogenies of collated B1 sequenced data Figure 7). We also analysed within-country clustering patterns of sequences for sites (Kenya, Mali and South Africa) who's within-country sampling location information was available. Sequences from different locations

were interspersed within the phylogenetic clusters (Appendix B - Figure S5). Sequences from cases and controls were mixed within the clades (Appendix B - Figure S5).

Table 3: HMPV and RSV genotype detection patterns

HMPV subgroup detection						
A)						
Country	A2a	A2b	A2c	B1	B2	Total
Kenya	0	12	9	21	16	58
Gambia	0	0	12	27	2	41
Mali	0	1	7	32	0	40
South Africa	6	16	17	15	4	58
Zambia	12	6	4	12	1	35
Total	18	35	49	107	23	232

RSV subgroup detection				
B)				
Country	RSVA_ON1	RSVA_GA2	RSVB_BA	Total
Kenya	114	42	107	263
Gambia	2	8	89	99
Mali	5	47	106	158
South Africa	13	188	29	230
Zambia	29	61	2	92
Total	163	346	333	842

HMPV spatial origins and dispersal patterns

Our phylogeographic analysis revealed clustering of sequences by geographic locations (Figures 11 and 12). The inferred viral migration pathways indicated very strong supported links between Mali and Gambia (BF >1000, posterior probability > 95%), Appendix B - Figure S6. To inform on the viral introductions into continental Africa, we analysed the patterns in the global context within the detected subgroups (A2a, A2b, A2c, B1 and B2). The temporal distribution of the combined Africa and global sequence data is shown in Appendix B - Figure S7. From the global time-resolved phylogenies, sequences from Africa fell into separate clades interspersed with global sequences (Figures 8, 9, 10 and Appendix B - Figure S3). The geographical clustering of Africa sequences was evident (Figures 8, 9,10 and Appendix B - Figure S3). Sequences from

South Africa and Zambia closely clustered together. Similarly, sequences from Gambia and Mali clustered more closely among themselves, indicating an epidemiological linkage between neighbouring locations and independent introductions of HMPV variants in Africa. The African clades were named to reflect their placement on the ML phylogenies. The most probable ancestral location at the branches leading to each African clade is indicated next to the nodes alongside tMRCA for each clade (Figures 8, 9, 10 and Appendix B - Figure S3). Only ancestral locations with posterior probability support of >70% were indicated. On the B1 (Figure 8) and A2c (Appendix B - Figure S3) phylogenies, although sequences from Africa were interspersed with global sequences, they mostly clustered together. Of note, most of the B1 (178/228) and A2c (165/232) sequences were from Africa and Asia, making it difficult to assess viral introductions from unsampled locations.

The inferred viral movements were visualized on the global map (Appendix B - Figures S8 and S9). Our analysis indicated wide spread movement of HMPV variants. Significant migration links were summarised based on the Bayes factor (BF). Very strong (BF >1000, posterior probability > 95%) and strongly supported (BF >10, posterior probability > 95%) migration pathways were indicated between the Gambia and Mali, Canada and Kenya, Malaysia and Mali, South Africa and Zambia, Mali and Cameroon, Kenya and Zambia, Kenya and Malaysia, Peru and South Africa (Appendix B, Figures S8 and S9). Other strongly supported links were also observed between other regions globally. The interactive visualization files of the reconstructed viral migration patterns are listed in additional file 2. Overall, North America, Australia and South East Asia acted as a major links and could have been central in the dissemination of infection to other regions worldwide (Appendix B - Figures S8 and S9).

A) HMPV

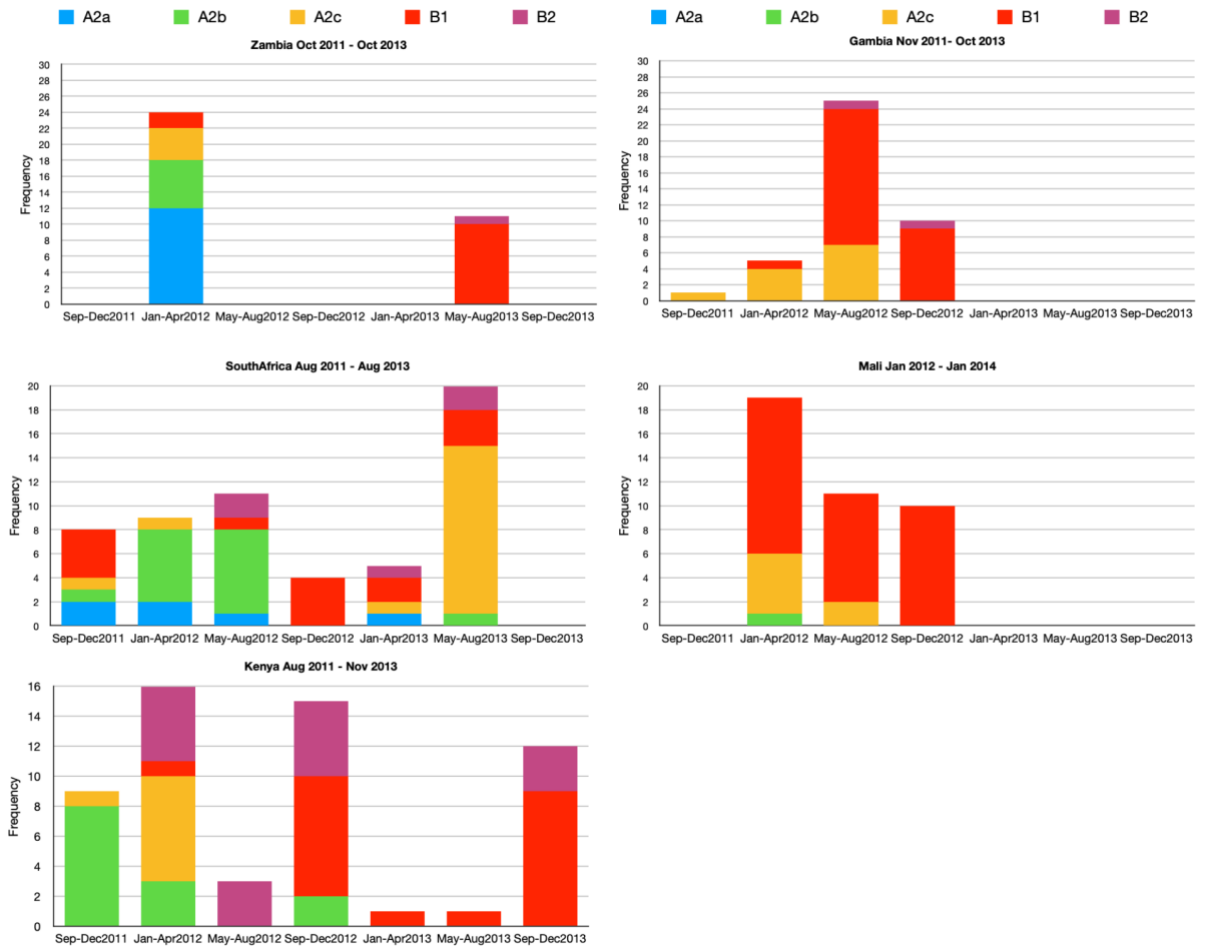


Figure 5: HMPV subgroup prevalence and temporal patterns derived from G gene sequence data collected from Kenya, Mali, Gambia, South Africa and Zambia between August 2011 and January 2014.

B) RSV

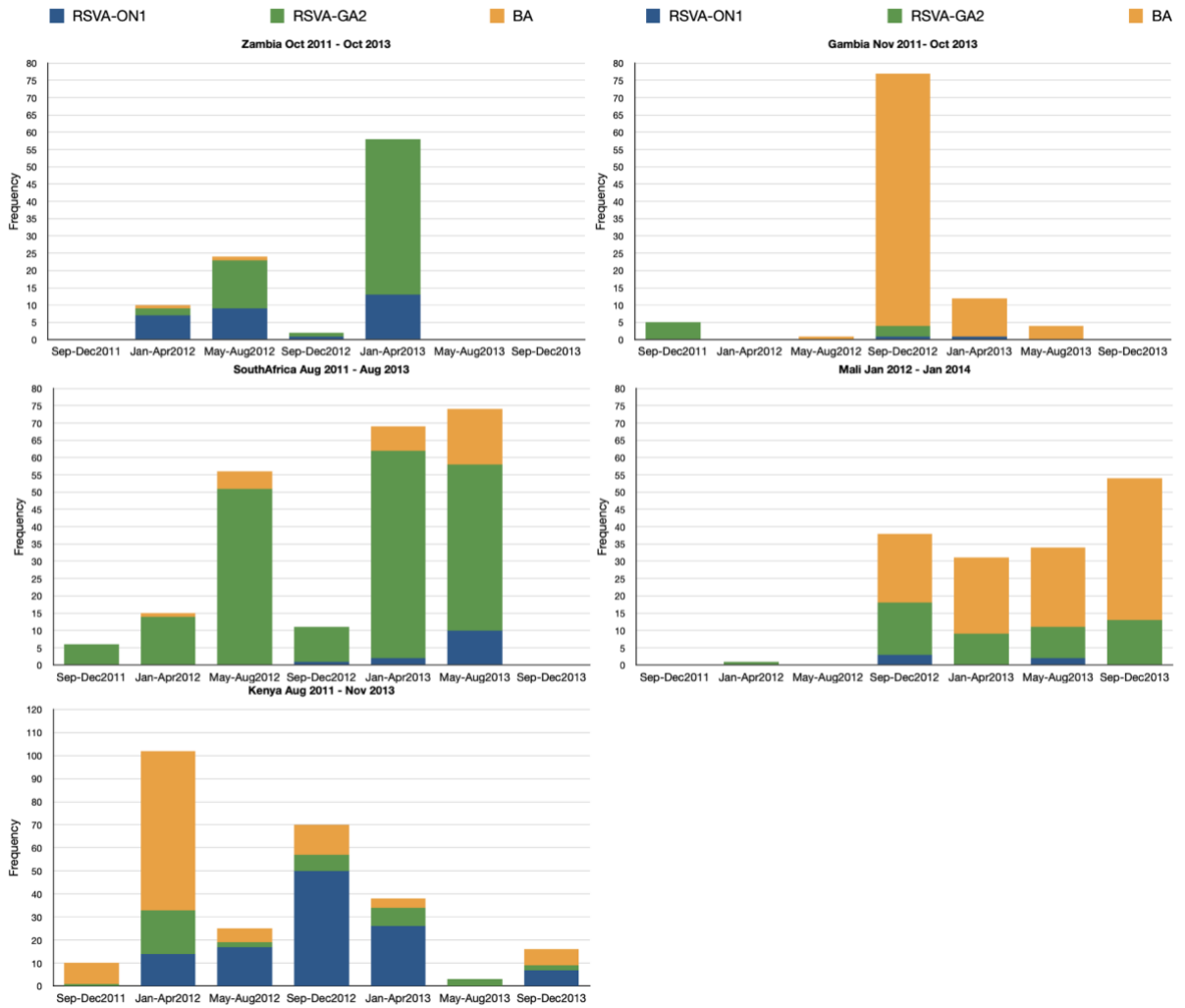


Figure 6:RSV subgroup prevalence and temporal patterns derived from G gene sequence data collected from Kenya, Mali, Gambia, South Africa and Zambia between August 2011 and January 2014.

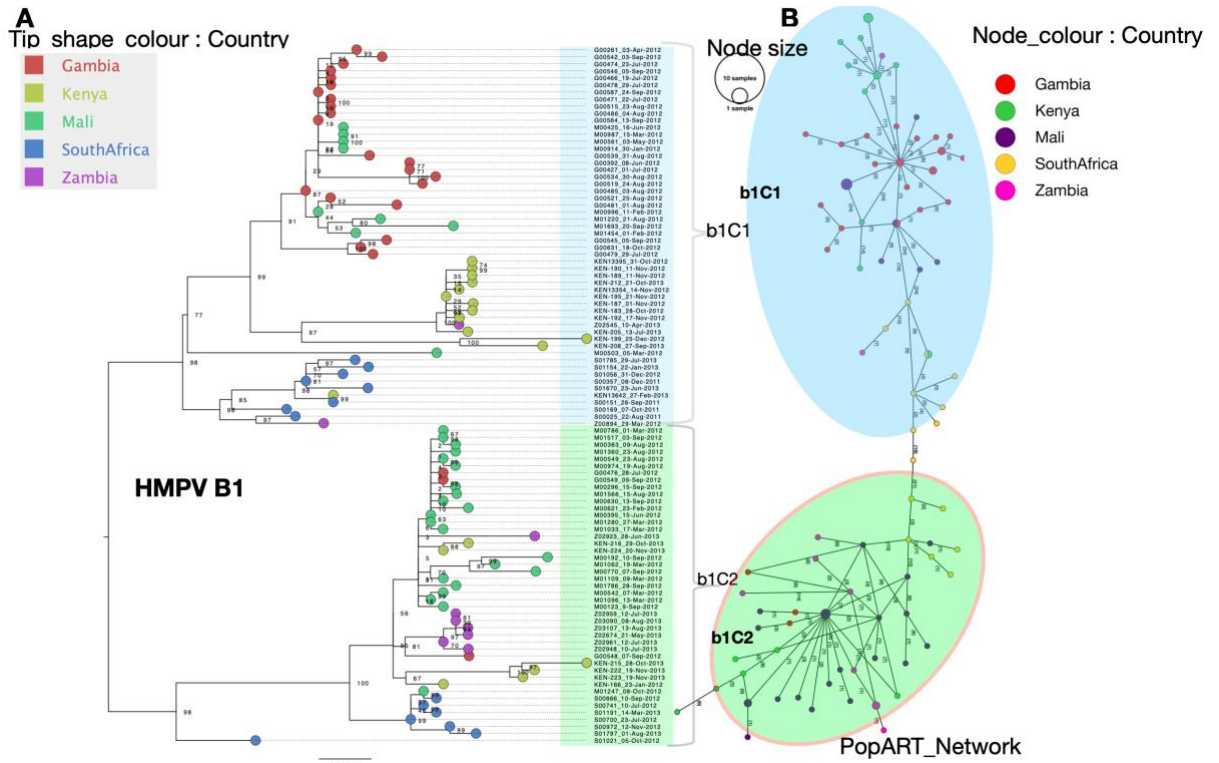


Figure 7: ML phylogenetic tree of HMPV subgroup B1 G gene sequences collected from Kenya, Mali, Gambia, South Africa and Zambia between August 2011 and January 2014. Tip shapes are coloured by country of sampling. Taxon labels are coloured in blue and green to differentiate the major phylogenetic clusters. Bootstrap values for each clade are indicated next to the nodes. Panel b, PopART minimum spanning network of the genetic distances of the viruses between and within countries. Clusters in red margin indicate higher sequence similarity between the sequences and potential inter-country transmission links.

HMPV B1

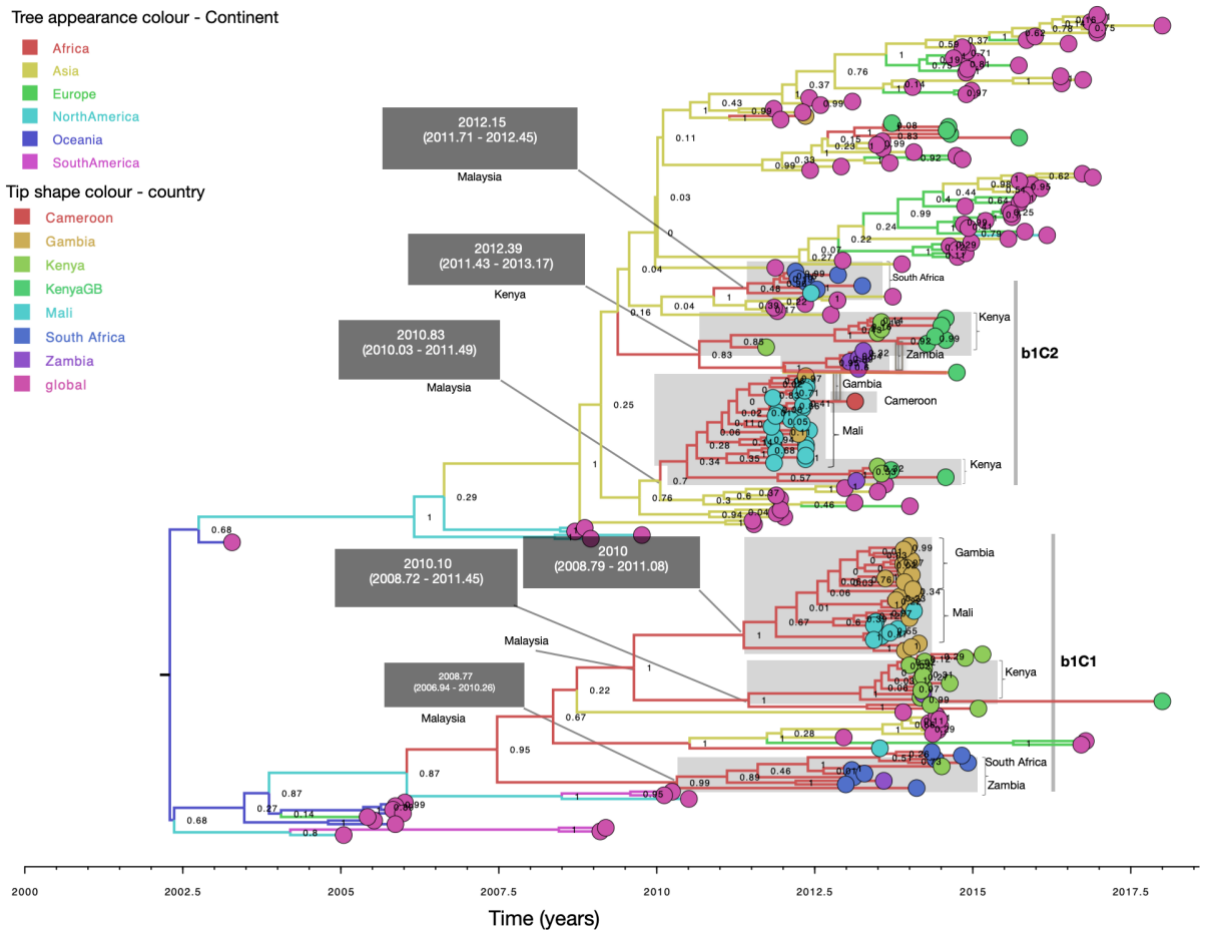


Figure 8: Temporal scaled maximum clade credibility (MCC) tree constructed using B1 G gene sequences obtained from Africa and GenBank collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix GB. Posterior probabilities are shown next to nodes. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade.

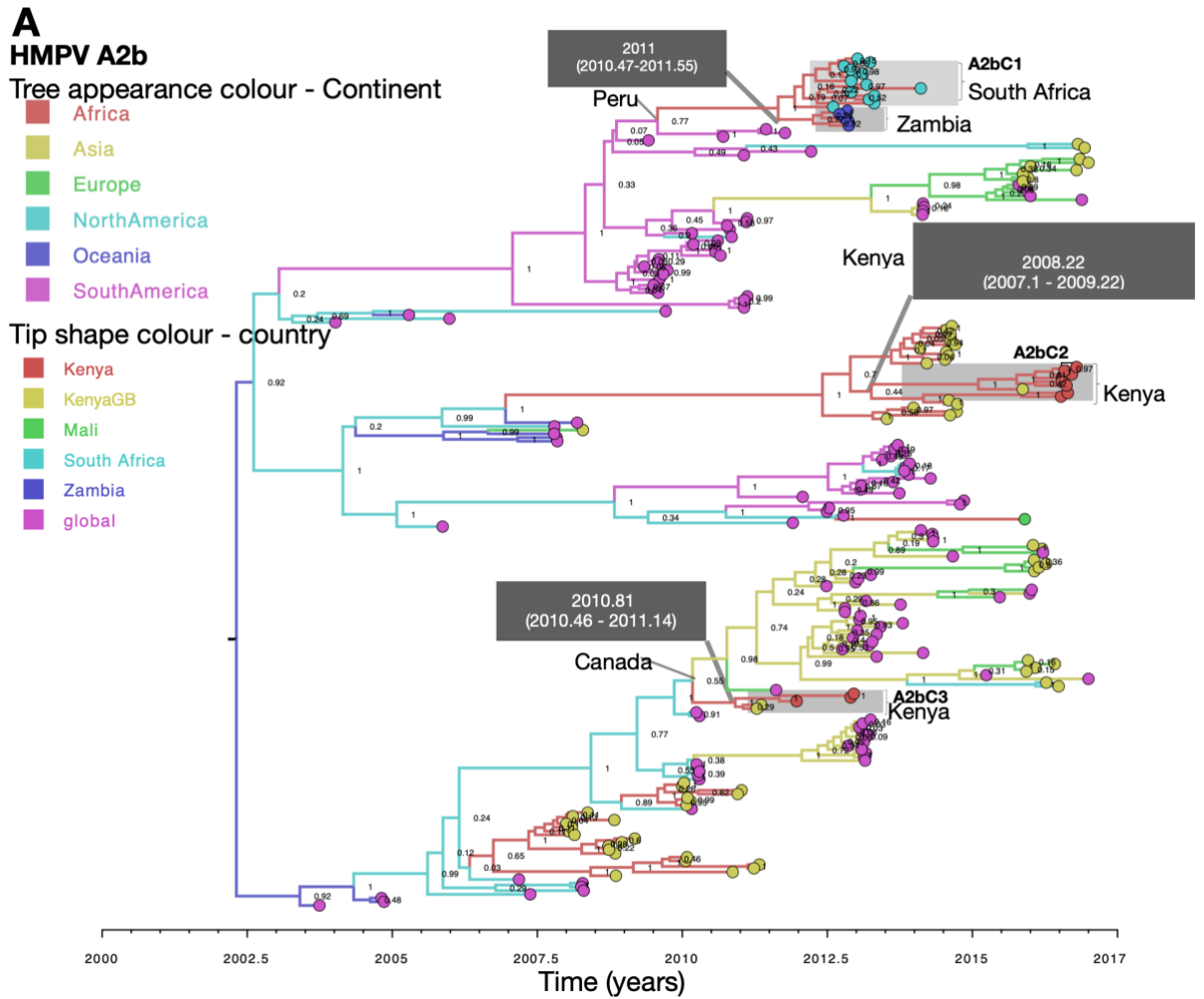


Figure 9: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV A2b G gene sequences obtained from Africa and GenBank collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Posterior probabilities are shown next to nodes. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix GB. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade.

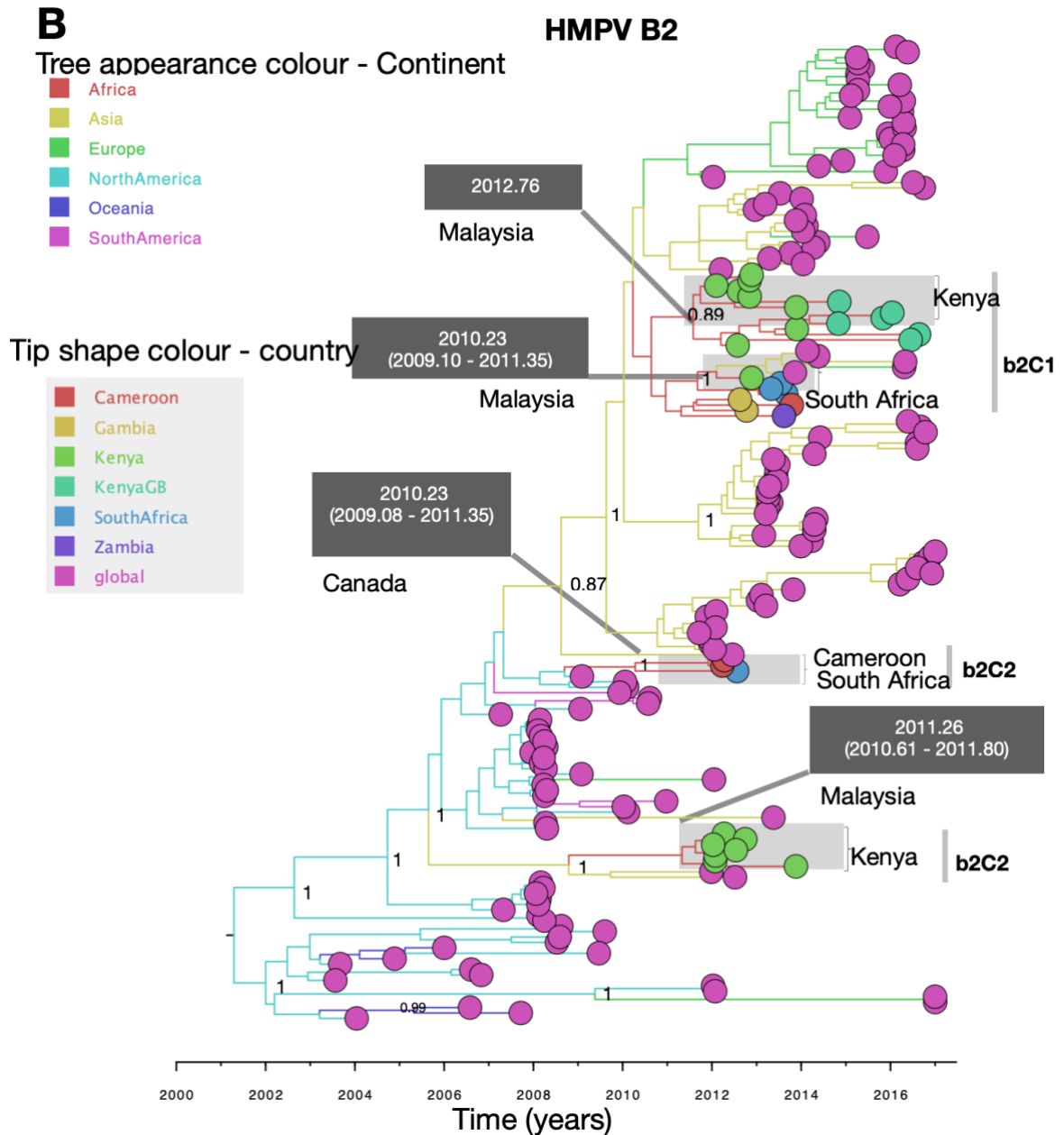


Figure 10: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV B2 G gene sequences obtained from Africa and GenBank collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Posterior probabilities are shown next to nodes. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix GB. Clades containing African sequences falling in

monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade.

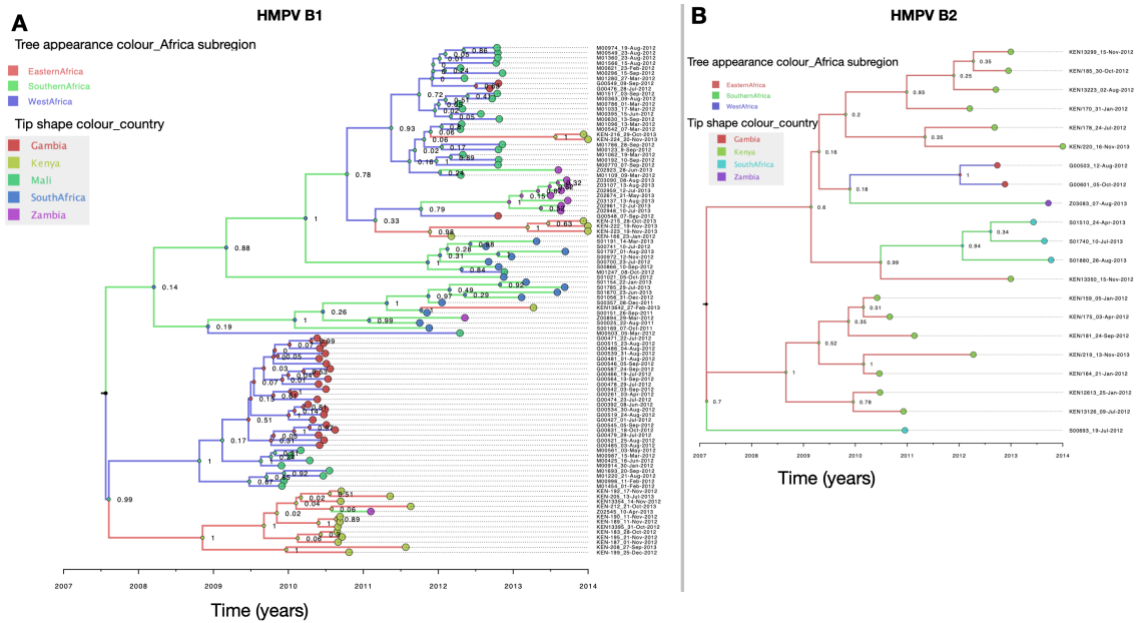


Figure 11: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV B1 (panel a) and B2 (panel b) G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. The tips were coloured according to the country of sampling. Posterior probabilities support for each node are indicated next to the nodes.

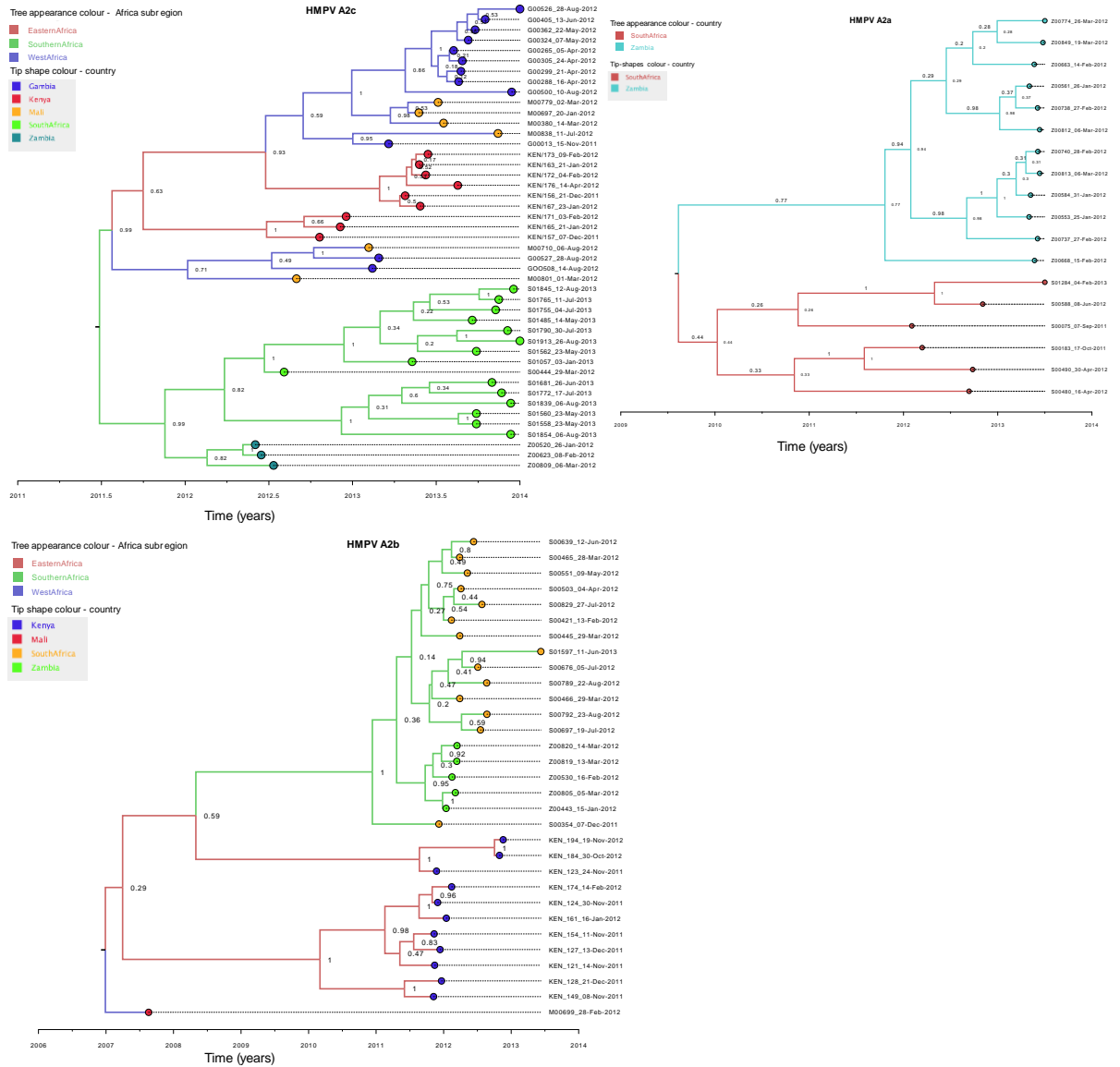


Figure 12: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV A2c (panel a), A2b (panel b) and A2a (panel c) G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. The tips were coloured according to the country of sampling. Posterior probabilities support for each node are indicated next to the nodes.

RSV subtyping and subgroup temporal patterns

From the G gene phylogeny, all RSV sequences fell into two major clusters- RSV A (512/846) and RSV B (334/846), Appendix B - Figures S1b. All RSV B sequences were genotype BA. Among RSV A, 32% (164/512) were genotype ON1 and 68% (348/512) were genotype GA2, Appendix B - Figures S10. Similar to HMPV, multiple RSV genetic groups co-circulated within epidemics (Figure 6). Similar genotype dominance patterns were observed between Mali and Gambia, South Africa and Zambia, and were all different from Kenya (Figure 6).

RSV intra and inter-country genetic diversity

To assess within country genetic diversity, RSV BA genotype sequences were selected (Table 3). From the country specific phylogenies, sequences from the different locations were interspersed within the phylogenetic clusters (Appendix B - Figures S11). Results shown for sites (Mali and South Africa) who's within-country sampling location information was available (Appendix B - Figures S11). Conversely, on the continental scale, sequences from the same Africa sub-regions (West Africa, Southern Africa and East Africa) largely clustered together into major and minor monophyletic clades (Appendix B - Figures S12 and S13a). We further investigated, potential inter-country transmission links using PopART. Similarly, sequences from the same geographic region closely clustered among themselves into minor clusters seen in ML phylogenies (Appendix B - Figures S13). Generally lesser genetic distances were observed between the sequences across all the RSV genotypes (Appendix B - Figures S13). Suggesting wide spread movement of similar variants. The mean sequence identities were estimated at: 99% for ON1, 98% for GA2 and 97% for genotype BA. Clusters (highlighted in red margins) containing sequences from multiple countries that were deemed similar suggesting potential intercountry transmission were identified (Appendix B - Figures

S13). Clustering patterns of these sequences were also assessed in global context on time resolved phylogenies.

RSV spatial patterns and Origins

RSV phylogeographic analysis revealed markedly similar patterns of spatial spread to those of HMPV. On the continental scale, geographical clustering was evident (Figure 13 and 14). The inferred migration pathways indicated very strongly supported links between neighbouring countries (BF >1000, posterior probability > 95%), Figure 15. These results suggest high sequence relatedness between neighbouring countries and possible transmission links. To be consistent in our analysis we also sought to explore the RSV spatial patterns globally within the detected genotypes (ON1, GA2 and BA) to inform on the viral introductions. However, our RSV BEAST runs failed to converge and needed more time to run due to larger RSV data sets. The effective sample size [ESS] for some of the estimated parameters were < 200 and therefore needed more time to run to attain MCMC convergence. The temporal distribution of the collated Africa and global contemporaneous sequence data is shown in Appendix B, Figure S14.

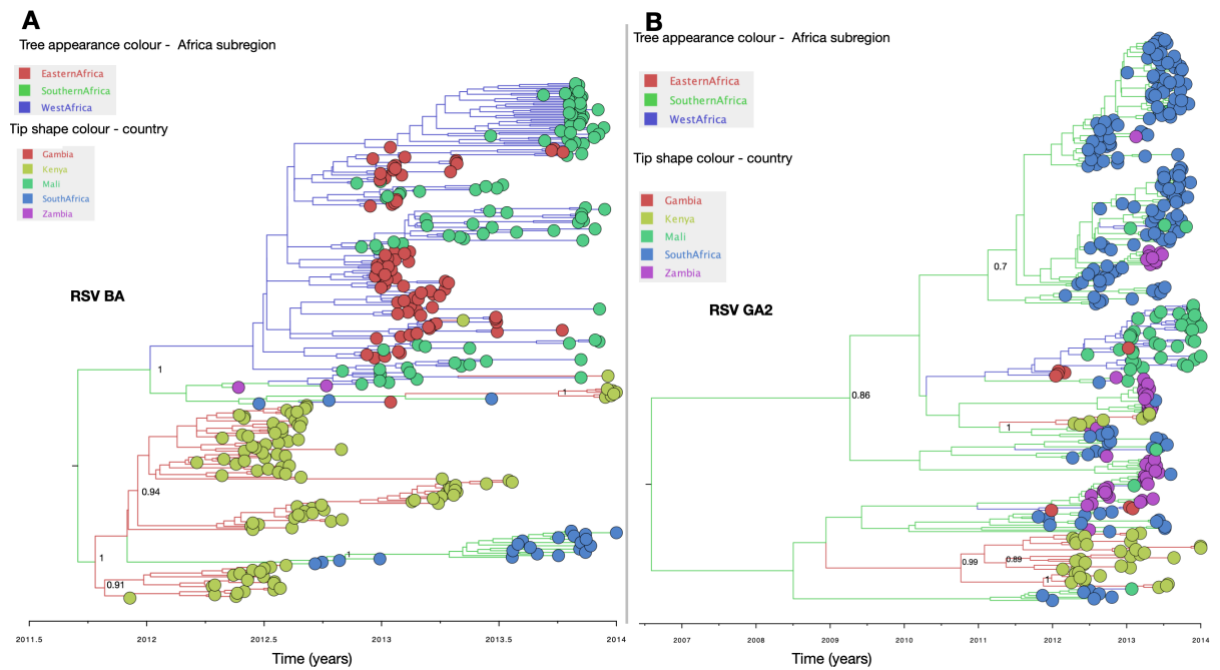


Figure 13: Temporal scaled maximum clade credibility (MCC) trees constructed using RSV BA (panel a), RSV and RSV GA2 (panel b) G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. Branches are coloured according to the most probable location as inferred using symmetric discrete phylogeographic diffusion model. The tips were coloured according to the country of sampling. Posterior probabilities support for each node are indicated next to the nodes.

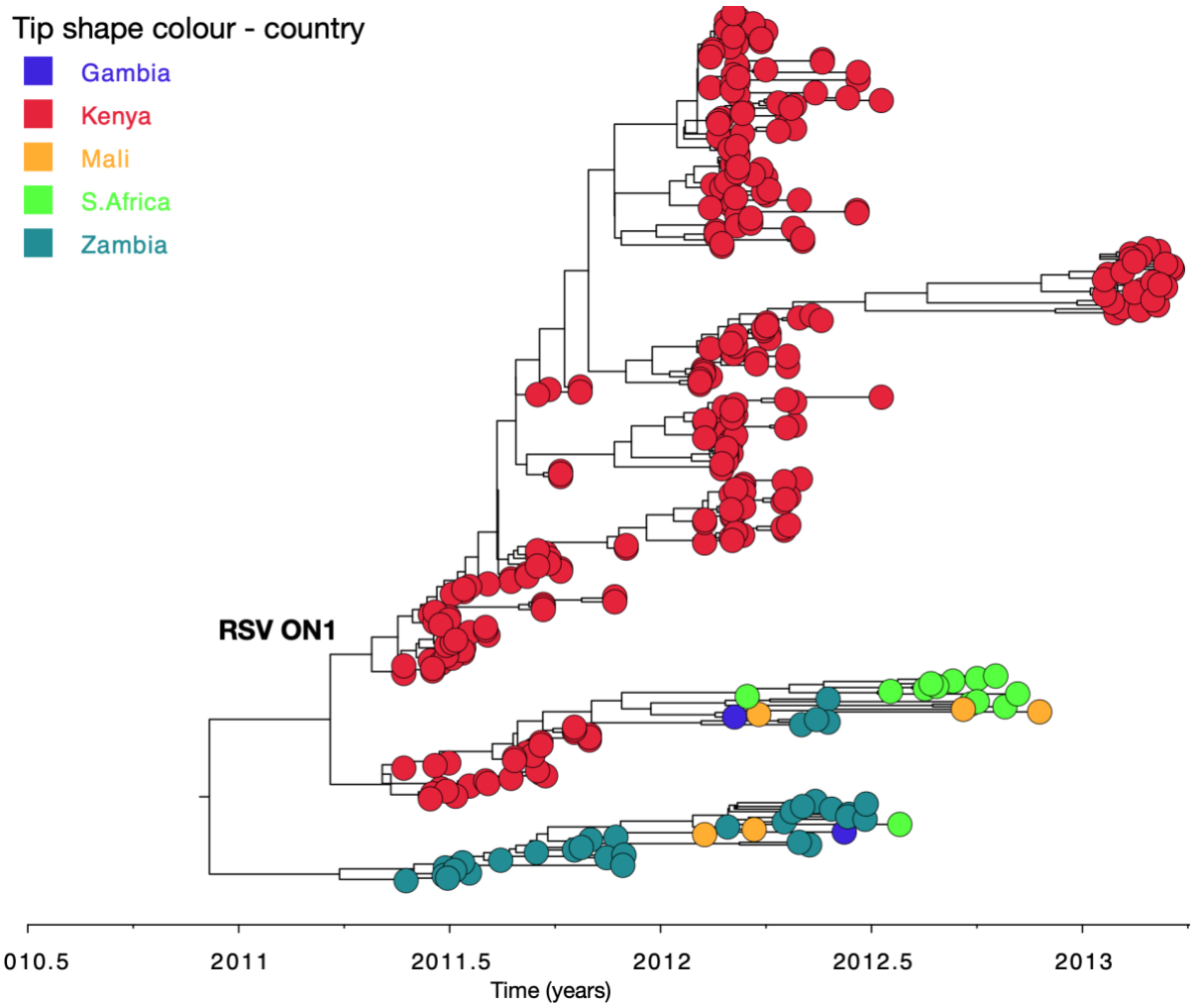


Figure 14: Temporal scaled maximum clade credibility (MCC) trees constructed using RSV ON1 G gene sequences obtained from Kenya, Mali, Gambia, South Africa and Zambia. The tips were coloured according to the country of sampling.

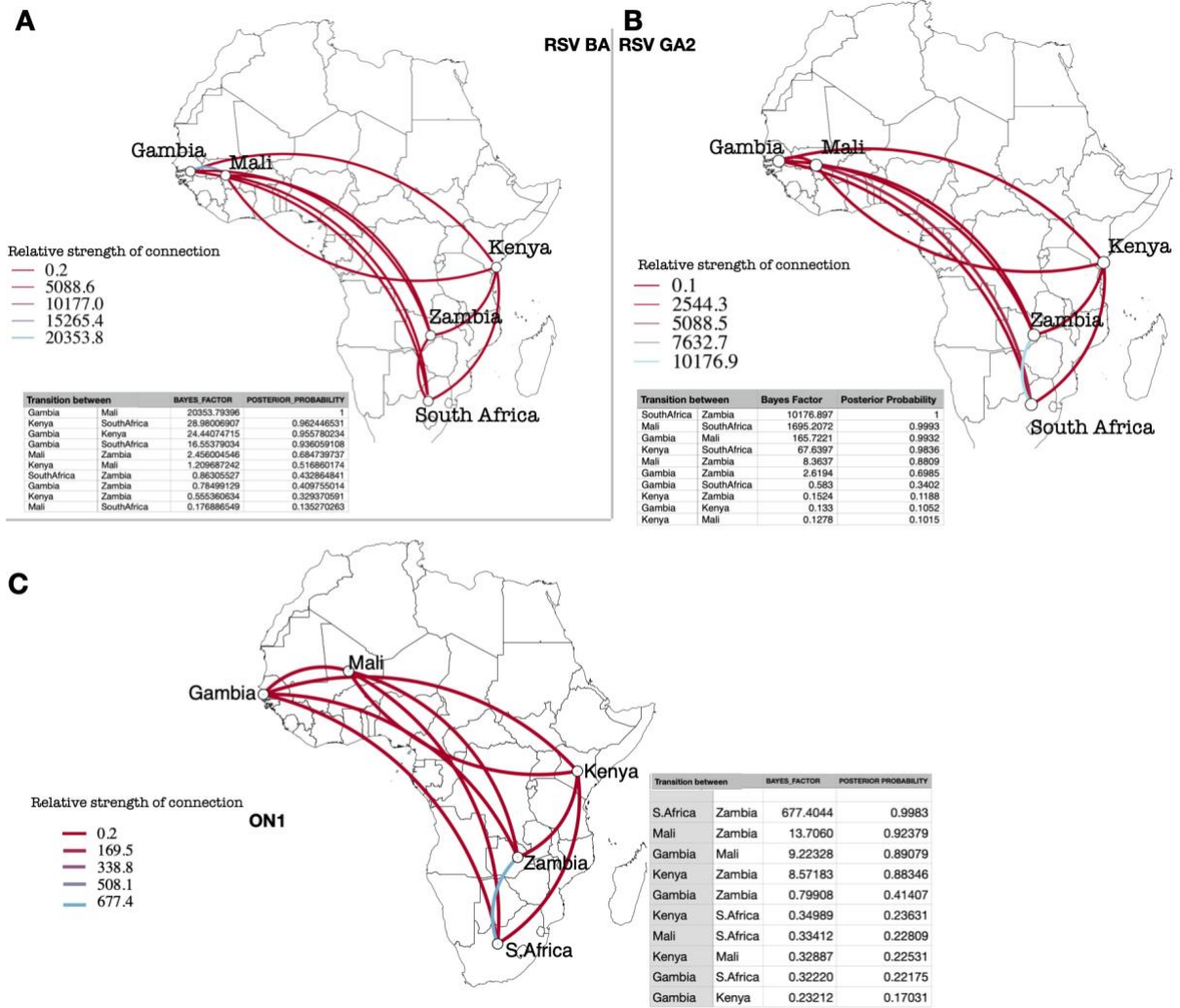


Figure 15: The inferred continental (Africa) migration pathways for RSV BA (panel a), GA2 (panel b) and ON1 (panel c) viruses constructed under symmetric diffusion model. The lines indicate connection between different countries. The colour gradient from red to blue of the lines indicate the relative strength of connection between countries according to the Bayes Factor test. Only statistically supported migration pathways (Bayes Factor >3) are shown. Posterior probabilities >95% shows very strong supported rates

DISCUSSION

Our comparative phylogeographic analysis revealed markedly similar patterns of geographic spread of HMPV and RSV in Africa. Geographical clustering of sequences by Africa subregion was evident and suggests high sequence relatedness between neighbouring countries and separate introduction of HMPV and RSV variants into continental Africa. Of note, our sequence data was collected at the same time across all the five sites (Kenya, Mali, Gambia, South Africa and Zambia) over 24 months period. The sites are well distributed across different Africa subregions (East, West and South) and therefore allowed the exploration of the patterns of spread across the continent. The geographical clustering of sequences suggests predominant local transmission and points to common source of introduction between neighbouring African countries. At the country levels, sequences from different locations were interspersed within the phylogenetic clusters. This indicate that following introduction in the country, there is rapid movement of HMPV and RSV variants between locations in a relatively short time and further local diversification. However, we cannot ignore the fact that only single site and hospital cases were sampled in each country and therefore we may not have characterised all the locally circulating strains.

HMPV and RSV epidemics were characterised by co-circulation of multiple genetic groups within epidemics. Our analysis revealed similar RSV genotype dominance patterns between neighbouring African countries. South Africa and Zambia which are in the Southern Africa subregion recoded similar RSV genotype dominance patterns. Similarly, Mali and Gambia which are immediate neighbours in the West Africa subregion, their genotype dominance patterns mirrored each other and were all different from Kenya (East Africa). Consistent with RSV results, Mali and Gambia HMPV genotype dominance patterns mirrored each other. South Africa and Zambia HMPV

epidemics were characterised by a unique circulation of HMPV A2a viruses and were not detected in the rest of the study sites. Similarities in genotype dominance patterns between neighbouring countries further supports the epidemiological linkage between neighbouring African countries and the independent introduction of multiple HMPV and RSV variants into Africa subregions from the global pool.

We sought to explore the spatial patterns of HMPV and RSV globally. On the global scale, geographical clustering of HMPV sequences by Africa subregion was still evident and fell into different clusters interspersed with global sequences. This further supports common source of introduction between neighbouring African countries, and on the global scale, suggests wide spread movement of similar HMP variants. Our RSV BEAST analysis failed to converge and needed more time to run to attain MCMC convergence i.e. effective sample size [ESS] > 200 for all estimated parameters (Rambaut et al., 2018). For this reason, we did not explore the RSV spatial patterns globally and inform on the potential sources of RSV introductions into Africa. To make further comparison, analysis of RSV global patterns will be necessary. In addition, this will also help validate our inferences on the potential sources of introductions of HMPV and RSV variants into Africa.

In this study, we hypothesized that the partial length (G gene) would be sufficient to resolve patterns of spread across different countries than spread patterns within the country. Indeed, using the G gene data we were able to discriminate the sequences, characterise the genetic diversity and assess potential inter-country transmission patterns. Genetic clusters containing sequences from multiple countries that were regarded similar will require whole genome sequencing to further discriminate the sequences and assess the possible transmission links between the African countries. The high sequence relatedness in this context, could also suggest a common source of introduction into

Africa and or possible inter-country transmission or global rapid movement of similar variants. However, based on the global HMPV G gene phylogenies, these sequences largely clustered by Africa subregion interspersed by other global sequences. This further points to the independent introduction of HMPV variants into Africa regions and rapid spread of similar variants globally. On the other hand, within country transmission dynamics will require whole genome sequencing especially for sequences collected over short epidemiological time-scale (Agoti et al., 2015a).

Previous studies of HMPV (Velez et al., 2013) and RSV (Rojo et al., 2017b) done in Argentina, reveal dispersal patterns of the two viruses occur both locally and globally. Similar findings have been reported for influenza viruses in Asia (Zar et al., 2019) and USA (Lemey et al., 2014). Respiratory viruses are transmitted by respiratory route and can easily transmit to distant locations harboured in hosts. Air travel has been shown to be the dominant determinant of influenza H3N2 and H1N1 viruses on the global scale (Lemey et al., 2014; Su et al., 2015). However, on smaller geographic scales, geographic distance which may represent other forms of mobility can act as a major predictor of spatial spread (Lemey et al., 2014). Our results corroborate these reports, the spatial diffusion pathways revealed strong connections between countries in the same African subregions and weak connections between other African countries in a far distant location. Globally, the clustering patterns of African HMPV sequences observed in this study suggested weak connection between Africa subregions but well connected with other countries globally. Overall, the patterns of spread of HMPV and RSV observed in this study may reflect underlying host mobility patterns. In particular, Africa experiences separate introduction of HMPV and RSV variants from the global pool influenced by human migration patterns. Following introduction, there is an establishment of local epidemic in Africa subregions that are in close contact due to more interactions, linked to

predominant migration between neighbouring countries (Flahaux & De Haas, 2016), due to environmental and socioeconomic factors such as distribution of ethnic groups, colonial and regional trade ties (Flahaux & De Haas, 2016). Recent reports on the role of long-distance truck drivers from neighbouring countries on the spread of COVID-19 in Uganda underscores these links between neighbouring countries (Bajunirwe et al., 2020). Also, HMPV and RSV epidemics are known to occur seasonally and, in the tropics, peak epidemics tend to coincide with high relative humidity and rainy seasons (Chow et al., 2016; Li et al., 2019; Owor et al., 2016). It is possible that the environmental factors could influence establishment of a local epidemic. In addition, studies reveal HMPV and RSV virus persistence within communities is often characterized by continual introduction of new variants from the global pool (Agoti et al., 2015b; Oketch et al., 2019). Therefore, these importations can be influenced by human migration patterns. We acknowledge that due to biased sampling, we did not assess possible introductions from unsampled locations.

The inferred global migration pathways indicated worldwide circulation of HMPV variants. Very strong and strongly supported links were identified between Africa and other countries globally (Table 4). These links reveal high sequence relatedness between these regions and points to possible sources of introduction of HMPV variants into these African subregions. However, we cannot rule out introductions from unsampled locations. Due to disproportionate sampling our ancestral location estimates could be biased and it was difficult to pinpoint the source sinks for HMPV epidemics. This is because the discrete trait analysis is inherently biased by the sampling intensities of locations (Wilson et al., 2015). Therefore, overrepresentations of sampled locations are often associated with high rates out of, or into, these specific locations (Baele et al., 2017). Our study suggests North America and Asia act as major links in the dispersal patterns of

HMPV globally. This suggests the two locations play a central role in the spread of HMPV variants globally. On the MCC phylogenies, the basal branches were occupied by sequences sampled outside Africa. This shows ancestral strains may reside in other regions outside Africa or from unsampled locations. To pin-point the source-sinks, more representative sampling will be required globally. Comparable results have been reported by global phylogeographic studies of influenza A H3N2 and H1N1 viruses (Lemey et al., 2014; Su et al., 2015). USA, China and Southeast Asia were reported as major source sinks with significant but minor contribution from other temperate regions and tropics, associated to high net migration outflow from these regions. Analysis of RSV global spatial patterns will further help validate our inferences on the potential source sinks and possible sources of introductions of HMPV and RSV into Africa. In addition, representative sampling and assessing the contribution of human mobility and other potential predictors on spatial spread will be important to clarify the complex migration dynamics of these two viruses.

Although our analysis was based on modest sample size (HMPV $n=232$ and RSV $n=842$ sequences), this did not hinder our ability to assess sequence relatedness and to infer spatial-temporal spread of HMPV and RSV in Africa. Overall, 83% (232/278) of HMPV and 90% (842/934) of RSV positive samples identified were successfully sequenced and analysed. Our sequencing ability improved the study power to characterise the different HMPV and RSV strains. It is possible that we may have missed to characterise some of the strains from the samples not sequenced. Although the non-sequenced samples had relatively higher mean Ct. values than those sequenced (Appendix B - Figure S15), the difference was not statistically significant for both HMPV (0.0756) and RSV (0.3147). Therefore, it is unlikely there was any bias in sequencing. The failure to sequence some of the samples could be as a result of RNA degradation as

samples were sequenced retrospectively. Another strength of this study is that sequences were collected at the time over a two-year period and therefore allowed exploration of the spatial patterns to assess possible epidemiological linkages between the five sites (Kenya, Mali, Gambia, South Africa, and Zambia). Conversely, we did not assess possible epidemiological links from unsampled locations in Africa. Future studies across different countries in different Africa subregions (East, West, South, Central and North) will be important for tracing transmission patterns of HMPV and RSV in Africa. Also, representative sampling in Africa and other regions globally will be necessary to assess the role of Africa in worldwide circulation and establishing the source sinks for HMPV and RSV epidemics.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, our study provides the first contemporaneous HMPV and RSV sequences across 5 African countries. In addition, this study provides an initial report on HMPV spatial patterns in Africa, acting as a major reference for future work. We also compared the geographic patterns of HMPV and RSV across Africa to give insights on respiratory pathogen spread that can help validate inferences on the spread of seasonally recurring respiratory viruses. The continental spatial-temporal analysis indicates similar patterns of spread of HMPV and RSV across Africa. Multiple strains can co-circulate and distinct strains can circulate in different Africa subregions at the same time. To explore the patterns further, representative sampling from different locations within and between African countries will be necessary. Also, representative sampling globally will help pinpoint source sink for HMPV and RSV epidemics, and inform on HMPV and RSV sources of introductions into Africa. The strong regional links observed, suggests that regional, tailored public health intervention measures should be considered.

Table 4: Inferred locations of tMRCA of African sequences

		Origin				
A)		HMPV subgroup				
<i>Region</i>	A2a	A2b	A2c	B1	B2	
West Africa	Not detected	*Canada (North America)	#Spain (Europe) Malaysia (Asia)	**Malaysia (Asia)	*Malaysia (Asia)	
East Africa	Not detected	**Canada (North America)	# Spain (Europe) Malaysia (Asia)	**Malaysia (Asia)	*Malaysia, Nepal (Asia)	
Southern Africa	*Peru (Southern America)	*Peru (South America)	#Spain (Europe) Malaysia (Asia)	*Malaysia (Asia)	#Canada (North America) *Malaysia (Asia)	
B)		RSV subgroup				
<i>Region</i>	ON1	GA2	BA			
West Africa	Not determined	Not determined	Not determined			
East Africa	Not determined	Not determined	Not determined			
Southern Africa	Not determined	Not determined	Not determined			

The support for the tMRCA leading to African clades was indicated as follows:

***Very strong support: $BF \geq 1000$, Posterior probability $\geq 95\%$

**Strong support: $100 \leq BF \leq 1000$

*Supported: $3 \leq BF \leq 100$

Not supported but shared the tMRCA with African clades

REFERENCES

- Agoti, C. N., , Otieno, J. R., Munywoki, P. K., Mwihuri, A. G., Cane, P. A., Nokes, D. J., ... Cotten, M. (2015). Local Evolutionary Patterns of Human Respiratory Syncytial Virus Derived from Whole-Genome Sequencing. *Journal of Virology*, 89(7), 3444–3454. <https://doi.org/10.1128/jvi.03391-14>
- Agoti, C. N., Otieno, J. R., Gitahi, C. W., Cane, P. A., & James Nokes, D. (2014). Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya. *Emerging Infectious Diseases*. <https://doi.org/10.3201/eid2006.131438>
- Agoti, C. N., Otieno, J. R., Ngama, M., Mwihuri, A. G., Medley, G. F., Cane, P. A., & Nokes, D. J. (2015). Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence. *Journal of Virology*, 89(22), 11630–11642. <https://doi.org/10.1128/jvi.01972-15>
- Agoti, C. N., Phan, M. V. T., Munywoki, P. K., Githinji, G., Medley, G. F., Cane, P. A., ... Nokes, D. J. (2019). Genomic analysis of respiratory syncytial virus infections in households and utility in inferring who infects the infant. *Scientific Reports*, 9(1), 1–14. <https://doi.org/10.1038/s41598-019-46509-w>
- Arias, A., Watson, S. J., Asogun, D., Tobin, E. A., Lu, J., Phan, M. V. T., ... Cotten, M. (2016). Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evolution*, 2(1), vew016. <https://doi.org/10.1093/ve/vew016>
- Van den Hoogen, B. G., Herfst, S., Sprong, L., Cane, P. A., Forleo-Neto, E., De Swart, R. L., ... & Fouchier, R. A. (2004). Antigenic and genetic variability of human metapneumoviruses. *Emerging infectious diseases*, 10(4), 658.

- Baele, G., Suchard, M. A., Rambaut, A., & Lemey, P. (2017). Emerging concepts of data integration in pathogen phylodynamics. *Systematic biology*, 66(1), e47-e65
- Baillie, G. J., Galiano, M., Agapow, P.-M., Myers, R., Chiam, R., Gall, A., ... Zambon, M. (2012). Evolutionary Dynamics of Local Pandemic H1N1/2009 Influenza Virus Lineages Revealed by Whole-Genome Analysis. *Journal of Virology*, 86(1), 11–18. <https://doi.org/10.1128/jvi.05347-11>
- Bajunirwe, F., Izudi, J., & Asiimwe, S. (2020). Long-distance truck drivers and the increasing risk of COVID-19 spread in Uganda. *International Journal of Infectious Diseases*. <https://doi.org/10.1016/j.ijid.2020.06.085>
- Bastien, N., Liu, L., Ward, D., Taylor, T., & Li, Y. (2004). Genetic variability of the G glycoprotein gene of human metapneumovirus. *Journal of Clinical Microbiology*, 42(8), 3532–3537. <https://doi.org/10.1128/JCM.42.8.3532-3537.2004>
- Bedford, T., Cobey, S., Beerli, P., & Pascual, M. (2010). Global migration dynamics underlie evolution and persistence of human influenza a (H3N2). *PLoS Pathogens*, 6(5). <https://doi.org/10.1371/journal.ppat.1000918>
- Biacchesi, S., Pham, Q. N., Skiadopoulos, M. H., Murphy, B. R., Collins, P. L., & Buchholz, U. J. (2005). Infection of Nonhuman Primates with Recombinant Human Metapneumovirus Lacking the SH, G, or M2-2 Protein Categorizes Each as a Nonessential Accessory Protein and Identifies Vaccine Candidates. *Journal of Virology*. <https://doi.org/10.1128/jvi.79.19.12608-12613.2005>
- Bielejec, F., Baele, G., Vrancken, B., Suchard, M. A., Rambaut, A., & Lemey, P. (2016). Spread3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msw082>

- Bont, L., Versteegh, J., Swelsen, W. T., Heijnen, C. J., Kavelaars, A., Brus, F., ... & Kimpen, J. L. (2002). Natural reinfection with respiratory syncytial virus does not boost virus-specific T-cell immunity. *Pediatric research*, 52(3), 363-367.
- Cane, P. A., Matthews, D. A., & Pringle, C. R. (1994). Analysis of respiratory syncytial virus strain variation in successive epidemics in one city. *Journal of Clinical Microbiology*.
- Chow, W. Z., Chan, Y. F., Oong, X. Y., Ng, L. J., Nor'E, S. S., Ng, K. T., ... & Tee, K. K. (2016). Genetic diversity, seasonality and transmission network of human metapneumovirus: identification of a unique sub-lineage of the fusion and attachment genes. *Scientific reports*, 6, 27730.
- Cunningham, C. W., Omland, K. E., & Oakley, T. H. (1998). Reconstructing ancestral character states: A critical reappraisal. *Trends in Ecology and Evolution*.
[https://doi.org/10.1016/S0169-5347\(98\)01382-2](https://doi.org/10.1016/S0169-5347(98)01382-2)
- Dale, H. (2006). WHO Pocket Book of Hospital Care for Children - Guidelines for the Management of Common Illnesses with Limited Resources WHO Pocket Book of Hospital Care for Children - Guidelines for the Management of Common Illnesses with Limited Resources. *Nursing Standard*.
<https://doi.org/10.7748/ns2006.07.20.44.36.b492>
- de Graaf, M., Osterhaus, A. D. M. E., Fouchier, R. A. M., & Holmes, E. C. (2008). Evolutionary dynamics of human and avian metapneumoviruses. *Journal of General Virology*, 89(12), 2933–2942. <https://doi.org/10.1099/vir.0.2008/006957-0>
- De Maio, N., Wu, C. H., O'Reilly, K. M., & Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics*, 11(8), 1–22. <https://doi.org/10.1371/journal.pgen.1005421>

- Deloria-Knoll, M., Feikin, D. R., Scott, J. A. G., O'Brien, K. L., DeLuca, A. N., Driscoll, A. J., ... & Pneumonia Methods Working Group. (2012). Identification and selection of cases and controls in the Pneumonia Etiology Research for Child Health project. *Clinical infectious diseases*, 54(suppl_2), S117-S123.
- Di Giallonardo, F., Kok, J., Fernandez, M., Carter, I., Geoghegan, J. L., Dwyer, D. E., ... Eden, J. S. (2018). Evolution of human respiratory syncytial virus (RSV) over multiple seasons in New South Wales, Australia. *Viruses*, 10(9), 1–13. <https://doi.org/10.3390/v10090476>
- Driscoll, A. J., Karron, R. A., Morpeth, S. C., Bhat, N., Levine, O. S., Baggett, H. C., ... Murdoch, D. R. (2017). Standardization of laboratory methods for the PERCH study. *Clinical Infectious Diseases*, 64(Suppl 3), S245–S252. <https://doi.org/10.1093/cid/cix081>
- Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, 29(8), 1969-1973.
- Eshaghi, A., Duvvuri, V. R., Lai, R., Nadarajah, J. T., Li, A., Patel, S. N., ... & Gubbay, J. B. (2012). Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. *PloS one*, 7(3), e32807.
- Faria, N. R., Suchard, M. A., Rambaut, A., & Lemey, P. (2011). Toward a quantitative understanding of viral phylogeography. *Current opinion in virology*, 1(5), 423-429.
- Faye, O., Freire, C. C. M., Iamarino, A., Faye, O., de Oliveira, J. V. C., Diallo, M., ... Sall, A. A. (2014). Molecular Evolution of Zika Virus during Its Emergence in the 20th Century. *PLoS Neglected Tropical Diseases*, 8(1), 36.

<https://doi.org/10.1371/journal.pntd.0002636>

- Flahaux, M. L., & De Haas, H. (2016). African migration: trends, patterns, drivers. *Comparative Migration Studies*. <https://doi.org/10.1186/s40878-015-0015-6>
- Glezen, W. P., Taber, L. H., Frank, A. L., & Kasel, J. A. (1986). Risk of Primary Infection and Reinfection With Respiratory Syncytial Virus. *American Journal of Diseases of Children*. <https://doi.org/10.1001/archpedi.1986.02140200053026>
- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., ... Neher, R. A. (2018). NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. <https://doi.org/10.1093/bioinformatics/bty407>
- Hall, C. B., Douglas, R. G., Schnabel, K. C., & Geiman, J. M. (1981). Infectivity of respiratory syncytial virus by various routes of inoculation. *Infection and Immunity*, *33*(3), 779–783.
- Henderson, F. W., Collier, A. M., Clyde Jr, W. A., & Denny, F. W. (1979). Respiratory-syncytial-virus infections, reinfections and immunity: a prospective, longitudinal study in young children. *New England Journal of Medicine*, *300*(10), 530-534
- Houlihan, C. F., Frampton, D., Bridget Ferns, R., Raffle, J., Grant, P., Reidy, M., ... Nastouli, E. (2018). Use of whole-genome sequencing in the investigation of a nosocomial influenza virus outbreak. *Journal of Infectious Diseases*, *218*(9), 1485–1489. <https://doi.org/10.1093/infdis/jiy335>
- Huck, B., Scharf, G., Neumann-Haefelin, D., Puppe, W., Weigl, J., & Falcone, V. (2006). Novel human metapneumovirus sublineage. *Emerging Infectious Diseases*, *12*(1), 147–150. <https://doi.org/10.3201/eid1201.050772>
- Jagušić, M., Slović, A., Ljubin-Sternak, S., Mlinarić-Galinović, G., & Forčić, D. (2017).

- Genetic diversity of human metapneumovirus in hospitalized children with acute respiratory infections in Croatia. *Journal of Medical Virology*, 89(11), 1885–1893. <https://doi.org/10.1002/jmv.24884>
- Johnson, P. R., Spriggs, M. K., Olmsted, R. A., & Collins, P. L. (1987). 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 of the United States of America*, 84(16), 5625–5629. <https://doi.org/10.1073/pnas.84.16.5625>
- Kahn, J. S. (2006). Epidemiology of human metapneumovirus. *Clinical Microbiology Reviews*, 19(3), 546–557. <https://doi.org/10.1128/CMR.00014-06>
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., & Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6), 587–589.
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kim, J. I., Park, S., Lee, I., Park, K. S., Kwak, E. J., Moon, K. M., ... & Song, K. J. (2016). Genome-wide analysis of human metapneumovirus evolution. *PloS one*, 11(4), e0152962.
- King, J. C., Burke, A. R., Clemens, J. D., Nair, P., Farley, J. J., Vink, P. E., ... & Johnson, J. P. (1993). Respiratory syncytial virus illnesses in human immunodeficiency virus- and noninfected children. *The Pediatric infectious disease journal*, 12(9), 733–738.
- Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large

- datasets. *Bioinformatics*, 30(22), 3276-3278.
- Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*. <https://doi.org/10.1080/10635150500433722>
- Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., ... & Suchard, M. A. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PloS pathog*, 10(2), e1003932.
- Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9). <https://doi.org/10.1371/journal.pcbi.1000520>
- Lemey, P., Rambaut, A., Welch, J. J., & Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution*, 27(8), 1877-1885.
- Lessler, J., Reich, N. G., Brookmeyer, R., Perl, T. M., Nelson, K. E., & Cummings, D. A. (2009). Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet Infectious Diseases*. [https://doi.org/10.1016/S1473-3099\(09\)70069-6](https://doi.org/10.1016/S1473-3099(09)70069-6)
- Levine, O. S., O'Brien, K. L., Deloria-Knoll, M., Murdoch, D. R., Feikin, D. R., DeLuca, A. N., ... & Scott, J. A. (2012). The Pneumonia Etiology Research for Child Health Project: a 21st century childhood pneumonia etiology study. *Clinical infectious diseases*, 54(suppl_2), S93-S101.
- Li, Y., Reeves, R. M., Wang, X., Bassat, Q., Brooks, W. A., Cohen, C., ... Zar, H. J. (2019). Global patterns in monthly activity of influenza virus, respiratory syncytial

- virus, parainfluenza virus, and metapneumovirus: a systematic analysis. *The Lancet Global Health*, 7(8), e1031–e1045. [https://doi.org/10.1016/S2214-109X\(19\)30264-5](https://doi.org/10.1016/S2214-109X(19)30264-5)
- Matsuzaki, Y., Itagaki, T., Ikeda, T., Aoki, Y., Abiko, C., & Mizuta, K. (2013). Human metapneumovirus infection among family members. *Epidemiology & Infection*, 141(4), 827-832.
- Miller, J. M., Binnicker, M. J., Campbell, S., Carroll, K. C., Chapin, K. C., Gilligan, P. H., ... Yao, J. D. (2018). A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiology. *Clinical Infectious Diseases*, 67(6), 813–816. <https://doi.org/10.1093/cid/ciy584>
- Min, J., Cella, E., Ciccozzi, M., Pelosi, A., Salemi, M., & Prosperi, M. (2016). The global spread of Middle East respiratory syndrome: an analysis fusing traditional epidemiological tracing and molecular phylodynamics. *Global Health Research and Policy*, 1(1), 1–14. <https://doi.org/10.1186/s41256-016-0014-7>
- Mizuta, K., Abiko, C., Aoki, Y., Ikeda, T., Matsuzaki, Y., Itagaki, T., ... & Ahiko, T. (2013). Seasonal patterns of respiratory syncytial virus, influenza A virus, human metapneumovirus, and parainfluenza virus type 3 infections on the basis of virus isolation data between 2004 and 2011 in Yamagata, Japan. *Japanese journal of infectious diseases*, 66(2), 140-145.
- Moe, N., Krokstad, S., Stenseng, I. H., Christensen, A., Skanke, L. H., Risnes, K. R., ... & Døllner, H. (2017). Comparing human metapneumovirus and respiratory syncytial virus: viral co-detections, genotypes and risk factors for severe disease. *PLoS One*, 12(1), e0170200.

- Mufson, M. A., Orvell, C., Rafnar, B., & Norrby, E. (1985). Two distinct subtypes of human respiratory syncytial virus. *Journal of General Virology*.
<https://doi.org/10.1099/0022-1317-66-10-2111>
- Munywoki, P. K., Koech, D. C., Agoti, C. N., Lewa, C., Cane, P. A., Medley, G. F., & Nokes, D. J. (2014). The source of respiratory syncytial virus infection in infants: a household cohort study in rural Kenya. *The Journal of infectious diseases*, 209(11), 1685-1692.
- Nguyen, Tung, Lam, Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274.
<https://doi.org/10.1093/molbev/msu300>
- Nunes, M. R. T., Palacios, G., Faria, N. R., Sousa, E. C., Pantoja, J. A., Rodrigues, S. G., ... Lipkin, W. I. (2014). Air Travel Is Associated with Intracontinental Spread of Dengue Virus Serotypes 1-3 in Brazil. *PLoS Neglected Tropical Diseases*, 8(4).
<https://doi.org/10.1371/journal.pntd.0002769>
- O'Brien, K. L., Baggett, H. C., Brooks, W. A., Feikin, D. R., Hammitt, L. L., Higdon, M. M., ... Zaman, S. M. A. (2019). Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *The Lancet*, 394(10200), 757–779.
[https://doi.org/10.1016/s0140-6736\(19\)30721-4](https://doi.org/10.1016/s0140-6736(19)30721-4)
- Oketch, J. W., Kamau, E., Otieno, G. P., Otieno, J. R., Agoti, C. N., & Nokes, D. J. (2019). Human metapneumovirus prevalence and patterns of subgroup persistence identified through surveillance of pediatric pneumonia hospital admissions in coastal Kenya, 2007–2016. *BMC Infectious Diseases*, 19(1), 1–13.

<https://doi.org/10.1186/s12879-019-4381-9>

Okiro, E. A., Ngama, M., Bett, A., Cane, P. A., Medley, G. F., & James Nokes, D. (2008).

Factors associated with increased risk of progression to respiratory syncytial virus-associated pneumonia in young Kenyan children. *Tropical Medicine and International Health*. <https://doi.org/10.1111/j.1365-3156.2008.02092.x>

Okiro, E. A., White, L. J., Ngama, M., Cane, P. A., Medley, G. F., & Nokes, D. J. (2010).

Duration of shedding of respiratory syncytial virus in a community study of Kenyan children. *BMC Infectious Diseases*, *10*. <https://doi.org/10.1186/1471-2334-10-15>

Otieno, J R, Kamau, E. M., Oketch, J. W., Ngoi, J. M., Gichuki, A. M., Binter, Š., ...

Nokes, D. J. (2018). Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. *Virus Evolution*, *4*(2), 1–13. <https://doi.org/10.1093/ve/vey027>

Otieno, James R., Agoti, C. N., Gitahi, C. W., Bett, A., Ngama, M., Medley, G. F., ...

Nokes, D. J. (2016). Molecular Evolutionary Dynamics of Respiratory Syncytial Virus Group A in Recurrent Epidemics in Coastal Kenya. *Journal of Virology*, *90*(10), 4990–5002. <https://doi.org/10.1128/jvi.03105-15>

Owor, B. E., Masankwa, G. N., Mwangi, L. C., Njeru, R. W., Agoti, C. N., & Nokes, D.

J. (2016). Human metapneumovirus epidemiological and evolutionary patterns in Coastal Kenya, 2007-11. *BMC Infectious Diseases*, *16*(1), 301. <https://doi.org/10.1186/s12879-016-1605-0>

Padhi, A., & Verghese, B. (2008). Positive natural selection in the evolution of human

metapneumovirus attachment glycoprotein. *Virus Research*, *131*(2), 121–131. <https://doi.org/10.1016/j.virusres.2007.08.014>

- Pagel, M. (1999). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*.
<https://doi.org/10.1080/106351599260184>
- Panda, S., Mohakud, N. K., Pena, L., & Kumar, S. (2014). Human metapneumovirus: Review of an important respiratory pathogen. *International Journal of Infectious Diseases*, 25, 45–52. <https://doi.org/10.1016/j.ijid.2014.03.1394>
- Peiris, J. S. M., Tang, W. H., Chan, K. H., Khong, P. L., Guan, Y., Lau, Y. L., & Chiu, S. S. (2003). Children with respiratory disease associated with metapneumovirus in Hong Kong. *Emerging Infectious Diseases*, 9(6), 628–633.
<https://doi.org/10.3201/eid0906.030009>
- Pelletier, G., Déry, P., Abed, Y., & Boivin, G. (2002). Respiratory tract reinfections by the new human Metapneumovirus in an immunocompromised child. *Emerging Infectious Diseases*, 8(9), 976–978. <https://doi.org/10.3201/eid0809.020238>
- Peret, T. C. T., Abed, Y., Anderson, L. J., Erdman, D. D., & Boivin, G. (2004). Sequence polymorphism of the predicted human metapneumovirus G glycoprotein. *Journal of General Virology*, 85(3), 679–686. <https://doi.org/10.1099/vir.0.19504-0>
- Peret, T. C. T., Hall, C. B., Schnabel, K. C., Golub, J. A., & Anderson, L. J. (1998). Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *Journal of General Virology*.
<https://doi.org/10.1099/0022-1317-79-9-2221>
- Pierangeli, A., Trotta, D., Scagnolari, C., Ferreri, M. L., Nicolai, A., Midulla, F., ... Bagnarelli, P. (2014). Rapid spread of the novel respiratory syncytial virus a on1 genotype, central Italy, 2011 to 2013. *Eurosurveillance*.
<https://doi.org/10.2807/1560-7917.ES2014.19.26.20843>

- Piñana, M., Vila, J., Gimferrer, L., Valls, M., Andrés, C., Codina, M. G., ... Antón, A. (2017). Novel human metapneumovirus with a 180-nucleotide duplication in the G gene. *Future Microbiology*, *12*(7), 565–571. <https://doi.org/10.2217/fmb-2016-0211>
- Pitoiset, C., Darniot, M., Huet, F., Aho, S. L., Pothier, P., & Manoha, C. (2010). Human metapneumovirus genotypes and severity of disease in young children (n = 100) during a 7-year study in Dijon Hospital, France. *Journal of Medical Virology*. <https://doi.org/10.1002/jmv.21884>
- Pitzer, V. E., Viboud, C., Alonso, W. J., Wilcox, T., Metcalf, C. J., Steiner, C. A., ... Grenfell, B. T. (2015). Environmental Drivers of the Spatiotemporal Dynamics of Respiratory Syncytial Virus in the United States. *PLoS Pathogens*, *11*(1), 1–14. <https://doi.org/10.1371/journal.ppat.1004591>
- Piyaratna, R., Tollefson, S. J., & Williams, J. V. (2011). Genomic analysis of four human metapneumovirus prototypes. *Virus Research*, *160*(1–2), 200–205. <https://doi.org/10.1016/j.virusres.2011.06.014>
- Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*. <https://doi.org/10.1093/sysbio/syy032>
- Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*. <https://doi.org/10.1093/ve/vew007>
- Reiche, J., Jacobsen, S., Neubauer, K., Hafemann, S., Nitsche, A., Milde, J., ... Schweiger, B. (2014). Human metapneumovirus: Insights from a ten-year molecular and epidemiological analysis in Germany. *PLoS ONE*, *9*(2). <https://doi.org/10.1371/journal.pone.0088342>

- Rima, B., Collins, P., Easton, A., Fouchier, R., Kurath, G., Lamb, R. A., ... Wang, L. (2017). ICTV virus taxonomy profile: Pneumoviridae. *Journal of General Virology*, 98(12), 2912–2913. <https://doi.org/10.1099/jgv.0.000959>
- Rojo, G. L., Goya, S., Orellana, M., Sancilio, A., Rodriguez Perez, A., Montali, C., ... Viegas, M. (2017a). Unravelling respiratory syncytial virus outbreaks in Buenos Aires, Argentina: Molecular basis of the spatio-temporal transmission. *Virology*, 508(February), 118–126. <https://doi.org/10.1016/j.virol.2017.04.030>
- Rojo, G. L., Goya, S., Orellana, M., Sancilio, A., Rodriguez Perez, A., Montali, C., ... Viegas, M. (2017b). Unravelling respiratory syncytial virus outbreaks in Buenos Aires, Argentina: Molecular basis of the spatio-temporal transmission. *Virology*. <https://doi.org/10.1016/j.virol.2017.04.030>
- Saikusa*, M., Kawakami, C., Nao, N., Takeda, M., Usuku, S., Sasao, T., ... Toyozawa, T. (2017). 180-nucleotide duplication in the G Gene of Human metapneumovirus A2b subgroup strains circulating in Yokohama City, Japan, since 2014. *Frontiers in Microbiology*, 8(MAR), 1–11. <https://doi.org/10.3389/fmicb.2017.00402>
- Saikusa, M., Nao, N., Kawakami, C., Usuku, S., Sasao, T., Toyozawa, T., ... Okubo, I. (2017). A novel 111-nucleotide duplication in the G gene of human metapneumovirus. *Microbiology and Immunology*. <https://doi.org/10.1111/1348-0421.12543>
- Saikusa, M., Nao, N., Kawakami, C., Usuku, S., Tanaka, N., Tahara, M., ... Okubo, I. (2019). Predominant detection of the subgroup a2b human metapneumovirus strain with a 111-nucleotide duplication in the g gene in Yokohama city, Japan in 2018. *Japanese Journal of Infectious Diseases*, 72(5), 350–352. <https://doi.org/10.7883/yoken.JJID.2019.124>

- Schildgen, V., van den Hoogen, B., Fouchier, R., Tripp, R. A., Alvarez, R., Manoha, C., ... Schildgen, O. (2011). Human metapneumovirus: Lessons learned over the first decade. *Clinical Microbiology Reviews*, 24(4), 734–754. <https://doi.org/10.1128/CMR.00015-11>
- Shafagati, N., & Williams, J. (2018). Human metapneumovirus - what we know now. *F1000Research*, 7, 135. <https://doi.org/10.12688/f1000research.12625.1>
- Shi, T. (2018). The etiological role of common respiratory viruses in acute respiratory infections in older adults: A systematic review and meta-analysis, PA4504. <https://doi.org/10.1183/13993003.congress-2018.pa4504>
- Shi, T., McAllister, D. A., O'Brien, K. L., Simoes, E. A. F., Madhi, S. A., Gessner, B. D., ... Nair, H. (2017). 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*, 390(10098), 946–958. [https://doi.org/10.1016/S0140-6736\(17\)30938-8](https://doi.org/10.1016/S0140-6736(17)30938-8)
- Skiadopoulos, M. H., Biacchesi, S., Buchholz, U. J., Amaro-Carambot, E., Surman, S. R., Collins, P. L., & Murphy, B. R. (2006). Individual contributions of the human metapneumovirus F, G, and SH surface glycoproteins to the induction of neutralizing antibodies and protective immunity. *Virology*, 345(2), 492–501. <https://doi.org/10.1016/j.virol.2005.10.016>
- Soto, J. A., Gálvez, N. M. S., Benavente, F. M., Pizarro-Ortega, M. S., Lay, M. K., Riedel, C., ... Kalergis, A. M. (2018). Human metapneumovirus: Mechanisms and molecular targets used by the virus to avoid the immune system. *Frontiers in Immunology*, 9(OCT), 1–11. <https://doi.org/10.3389/fimmu.2018.02466>
- Su, Y. C. F., Bahl, J., Joseph, U., Butt, K. M., Peck, H. A., Koay, E. S. C., ... Smith, G.

- J. D. (2015). Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nature Communications*, 6. <https://doi.org/10.1038/ncomms8952>
- Sullender, W. M., Mufson, M. A., Anderson, L. J., & Wertz, G. W. (1991). Genetic diversity of the attachment protein of subgroup B respiratory syncytial viruses. *Journal of Virology*, 65(10), 5425–5434.
- Sullender, Wayne M. (2000). Respiratory syncytial virus genetic and antigenic diversity. *Clinical Microbiology Reviews*, 13(1), 1–15. <https://doi.org/10.1128/CMR.13.1.1-15.2000>
- Tognarelli, E. I., Bueno, S. M., & González, P. A. (2019). Immune-modulation by the human respiratory syncytial virus: Focus on dendritic cells. *Frontiers in Immunology*, 10(MAR), 1–13. <https://doi.org/10.3389/fimmu.2019.00810>
- Trento, A., Casas, I., Calderon, A., Garcia-Garcia, M. L., Calvo, C., Perez-Brena, P., & Melero, J. A. (2010). Ten Years of Global Evolution of the Human Respiratory Syncytial Virus BA Genotype with a 60-Nucleotide Duplication in the G Protein Gene. *Journal of Virology*, 84(15), 7500–7512. <https://doi.org/10.1128/jvi.00345-10>
- Trento, Alfonsina, Galiano, M., Videla, C., Carballal, G., García-Barreno, B., Melero, J. A., & Palomo, C. (2003). Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *Journal of General Virology*. <https://doi.org/10.1099/vir.0.19357-0>
- Van Den Hoogen, B. G., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2002a). Analysis of the genomic sequence of a human metapneumovirus. *Virology*, 295(1), 119–132. <https://doi.org/10.1006/viro.2001.1355>

- Van Den Hoogen, B. G., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2002b). Analysis of the genomic sequence of a human metapneumovirus. *Virology*. <https://doi.org/10.1006/viro.2001.1355>
- van den Hoogen, B. G., Herfst, S., Sprong, L., Cane, P. A., Forleo-Neto, E., de Swart, R. L., ... Fouchier, R. A. M. (2004). Antigenic and genetic variability of human metapneumoviruses. *Emerging Infectious Diseases*, *10*(4), 658–666. <https://doi.org/10.3201/eid1004.030393>
- Van Woensel, J. B. M., Bos, A. P., Lutter, R., Rossen, J. W. A., & Schuurman, R. (2006). Absence of human metapneumovirus co-infection in cases of severe respiratory syncytial virus infection. *Pediatric Pulmonology*, *41*(9), 872–874. <https://doi.org/10.1002/ppul.20459>
- Velez Rueda, A. J., Mistchenko, A. S., & Viegas, M. (2013). Phylogenetic and Phylodynamic Analyses of Human Metapneumovirus in Buenos Aires (Argentina) for a Three-Year Period (2009-2011). *PLoS ONE*, *8*(4). <https://doi.org/10.1371/journal.pone.0063070>
- Venter, M., Madhi, S. A., Tiemessen, C. T., & Schoub, B. D. (2001). 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*. <https://doi.org/10.1099/0022-1317-82-9-2117>
- Williams, J., & Shafagati, N. (2018). Human metapneumovirus - what we know now. *F1000Research*, *7*(0), 1–11. <https://doi.org/10.12688/f1000research.12625.1>
- Woelk, C. H., & Holmes, E. C. (2001). Variable immune-driven natural selection in the attachment (G) glycoprotein of respiratory syncytial virus (RSV). *Journal of Molecular Evolution*. <https://doi.org/10.1007/s002390010147>

- Xepapadaki, P., Psarras, S., Bossios, A., Tsolia, M., Gourgiotis, D., Liapi-Adamidou, G., ... Papadopoulos, N. G. (2004). Human metapneumovirus as a causative agent of acute bronchiolitis in infants. *Journal of Clinical Virology*. <https://doi.org/10.1016/j.jcv.2003.12.012>
- Xiao, X., Tang, A., Cox, K. S., Wen, Z., Callahan, C., Sullivan, N. L., ... Zhang, L. (2019). Characterization of potent RSV neutralizing antibodies isolated from human memory B cells and identification of diverse RSV/hMPV cross-neutralizing epitopes. *MAbs*. <https://doi.org/10.1080/19420862.2019.1654304>
- Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M. H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic biology*, 60(2), 150-160.
- Zar Htwe, K. T., Dapat, C., Shobugawa, Y., Odagiri, T., Hibino, A., Kondo, H., ... Saito, R. (2019). Phylogeographic analysis of human influenza A and B viruses in Myanmar, 2010–2015. *PLoS ONE*, 14(1), 2010–2015. <https://doi.org/10.1371/journal.pone.0210550>

APPENDICES

Appendix A

Table A1: HMPV G and RSV G genes PCR and sequencing primers

	Primer name		Sequence	Gene	Organism	Polarity	Subgroup
1	AG20 F	1 st round PCR	GGGGCAAATGCAAACATGTCC	G	RSV	+	AB
2	F164 R	1 st round PCR	GTTATGACACTGGTATAC CAACC	G	RSV	-	AB
3	BG10 F	2 nd round PCR	GCAATGATAATCTCAACCTC	G	RSV	+	AB
4	F1 R	2 nd round PCR	CAACTCCATTGTTATTTGCC	G	RSV	-	AB
5	G523 F	Sequencing	ATATG CAGCAACAATCCAAC	G	RSV	+	A
6	G523 R	Sequencing	GTTG GATTGTTGCTGCATAT	G	RSV	-	A
7	G533 F	Sequencing	TGTAGTATATGTGGCAACAA	G	RSV	+	B
8	G533 R	Sequencing	TGTTGCCACATATACTACA	G	RSV	-	B
9	13F	PCR + sequencing	GTRGAGAACATTCGAGCAATAGACA	G	HMPV	+	A
10	264F	Sequencing	TCCAAACTCACAGCATCCAAC	G	HMPV	+	A
11	1163R	PCR + sequencing	AGGGAGATAGACATTAACAGTGGA	G	HMPV	-	A
12	2F	PCR + sequencing	TGGAAGTAAGAGTGGAGAACATTC	G	HMPV	+	B
13	222F	Sequencing	YAARAAGACCCCAATGACCTC	G	HMPV	+	B
15	718R	Sequencing	ACTACTGGATGAGATACCTGTGT	G	HMPV	-	B
16	1098R	PCR + sequencing	TGACTGCATTTCTAAGCCTTACAT	G	HMPV	-	B

Appendix B

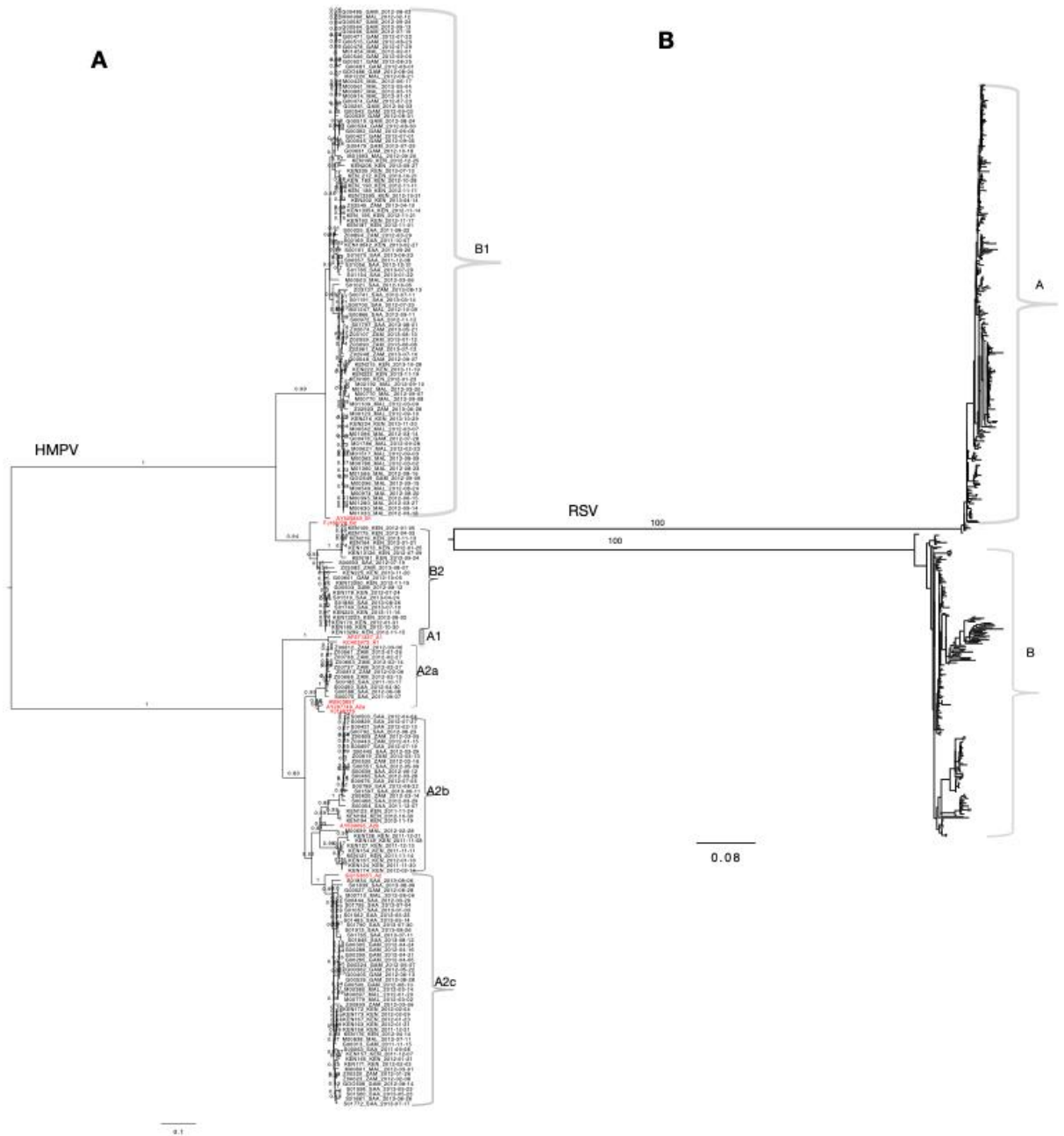


Figure S1: ML phylogenies of G gene sequences collected from Kenya, Mali, Gambia, South Africa and Zambia. Sequences were subtyped using subgroup reference sequences retrieved from GenBank. Panel a: HMPV G gene sequences constructed using 231G gene sequences. Reference sequences are coloured in red. The numbers next to branches indicate the bootstrap values. Subgroups were confirmed if sequences clustered with the

reference sequences within a major branch with > 70% bootstrap support. Panel b: RSV G ML phylogeny constructed using 627 unique gene sequences.

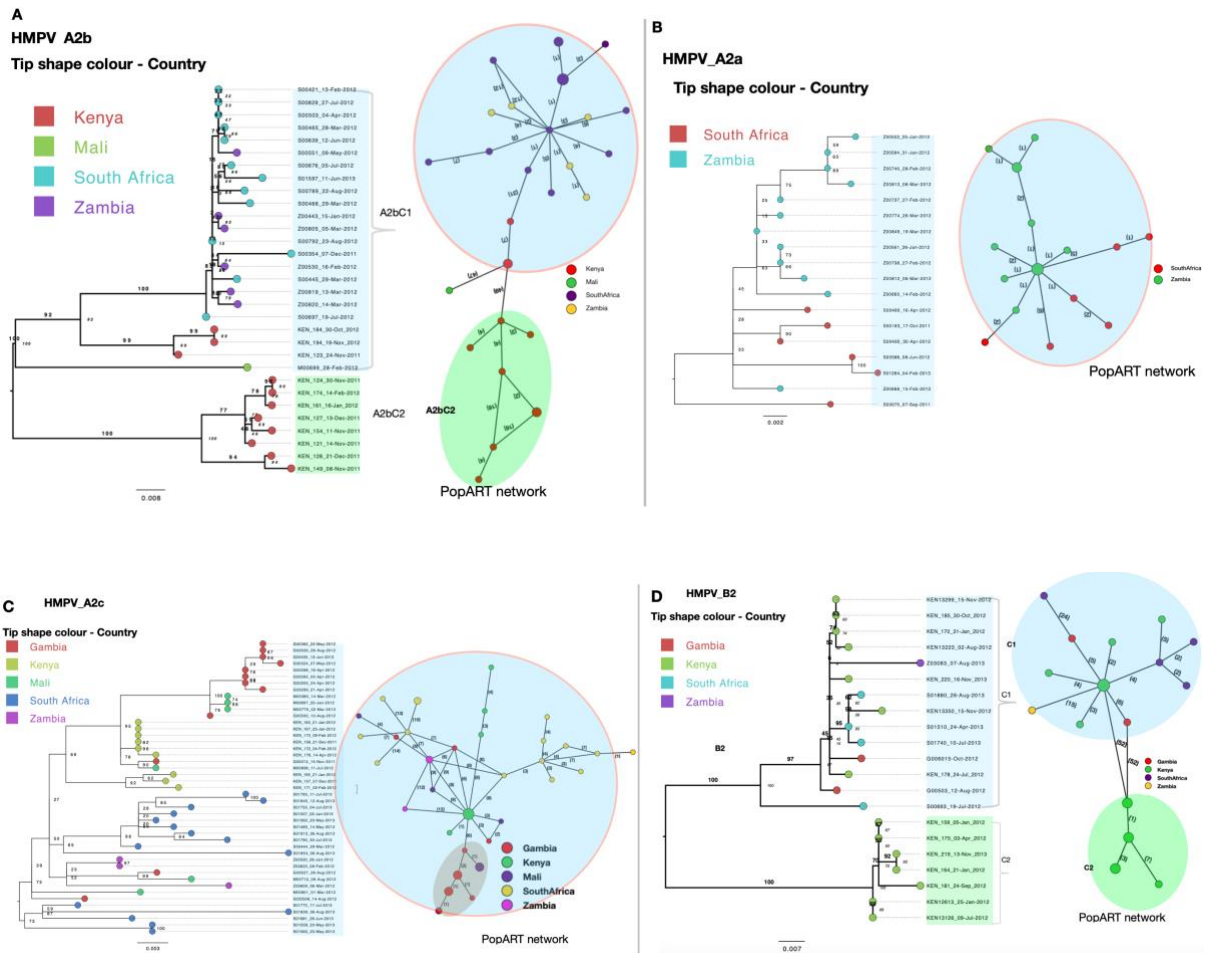


Figure S2: ML phylogenetic tree of HMPV clade A2b (panel a), clade A2a (panel b), clade A2c (panel d) and subgroup B2 (panel d) of G gene sequences collected from five African countries. Tip shapes are coloured by country of sampling. Taxon labels are coloured in blue and green to differentiate the major phylogenetic clusters. Bootstrap values for each clade are indicated next to the nodes. Network next to the phylogenies shows PopART minimum spanning network of the genetic distances of the viruses between and within countries. Clusters in red margin indicate higher sequence similarity between the sequences and potential inter-country transmission links.

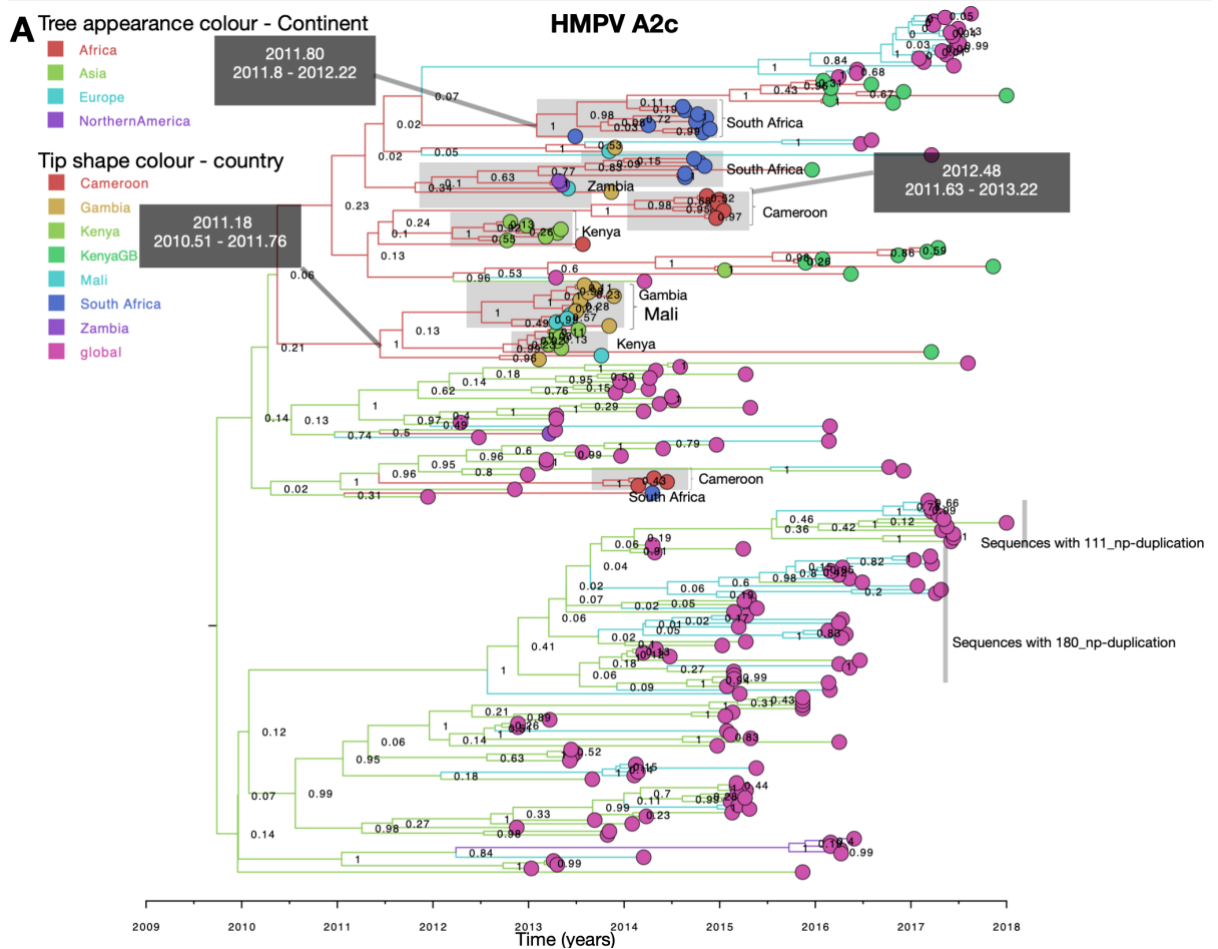


Figure S3: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV A2c (panel a) G gene sequences obtained from Africa and GenBank, collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using a discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Sequences from Kenya, Mali, Gambia, South Africa and Zambia collected beyond the study period are indicated with a suffix “GB”. Posterior probabilities are shown next to nodes. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade.

B Tree appearance colour - continent

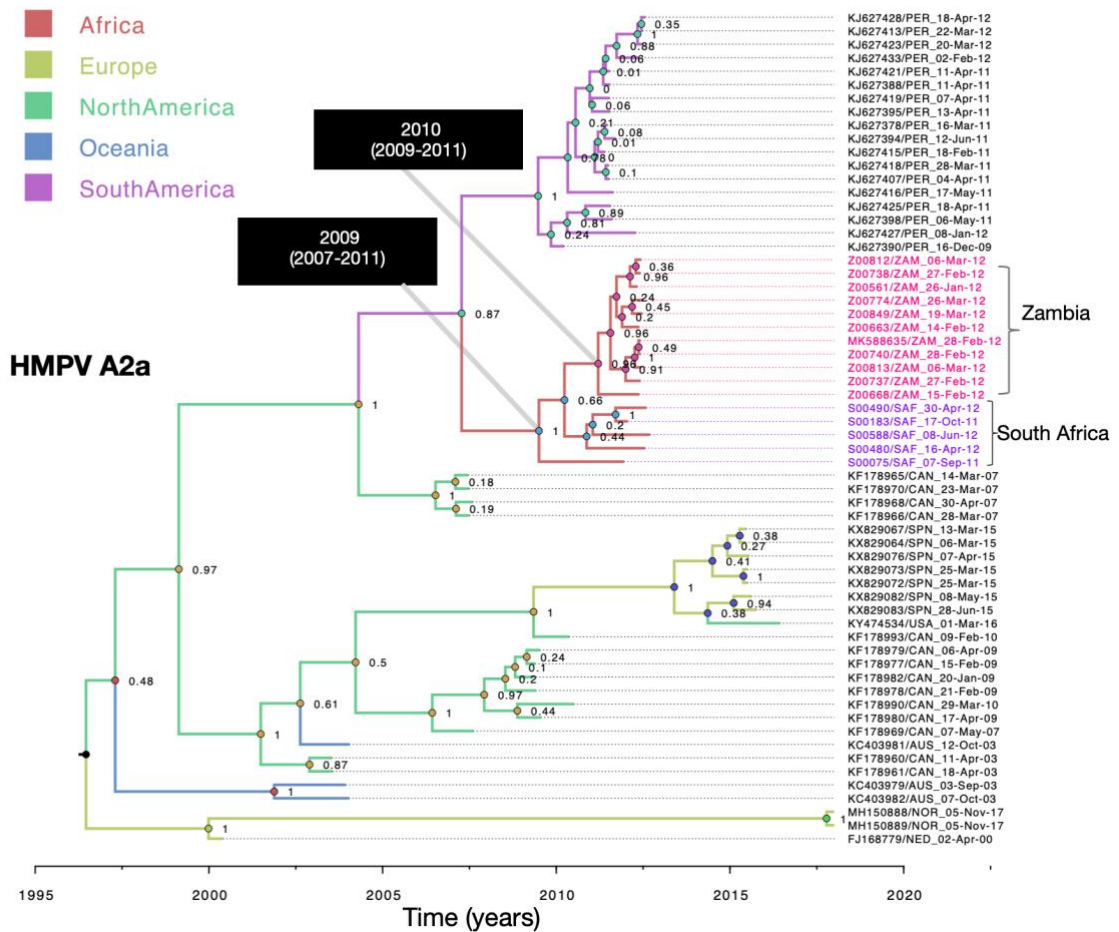


Figure S4: Temporal scaled maximum clade credibility (MCC) trees constructed using HMPV A2a G gene sequences obtained from Africa and GenBank, collected between 2000 to 2018. Branches are coloured according to the most probable location as inferred using a discrete phylogeographic diffusion model. Geographic locations considered are shown in the figure key. Posterior probabilities are shown next to nodes. Clades containing African sequences falling in monophyletic clades are highlighted in grey boxes. For each clade, the mean estimated time of the most recent common ancestor (tMRCA) and respective 95% Bayesian credible intervals are shown in a black box alongside the most probable location leading to each clade.

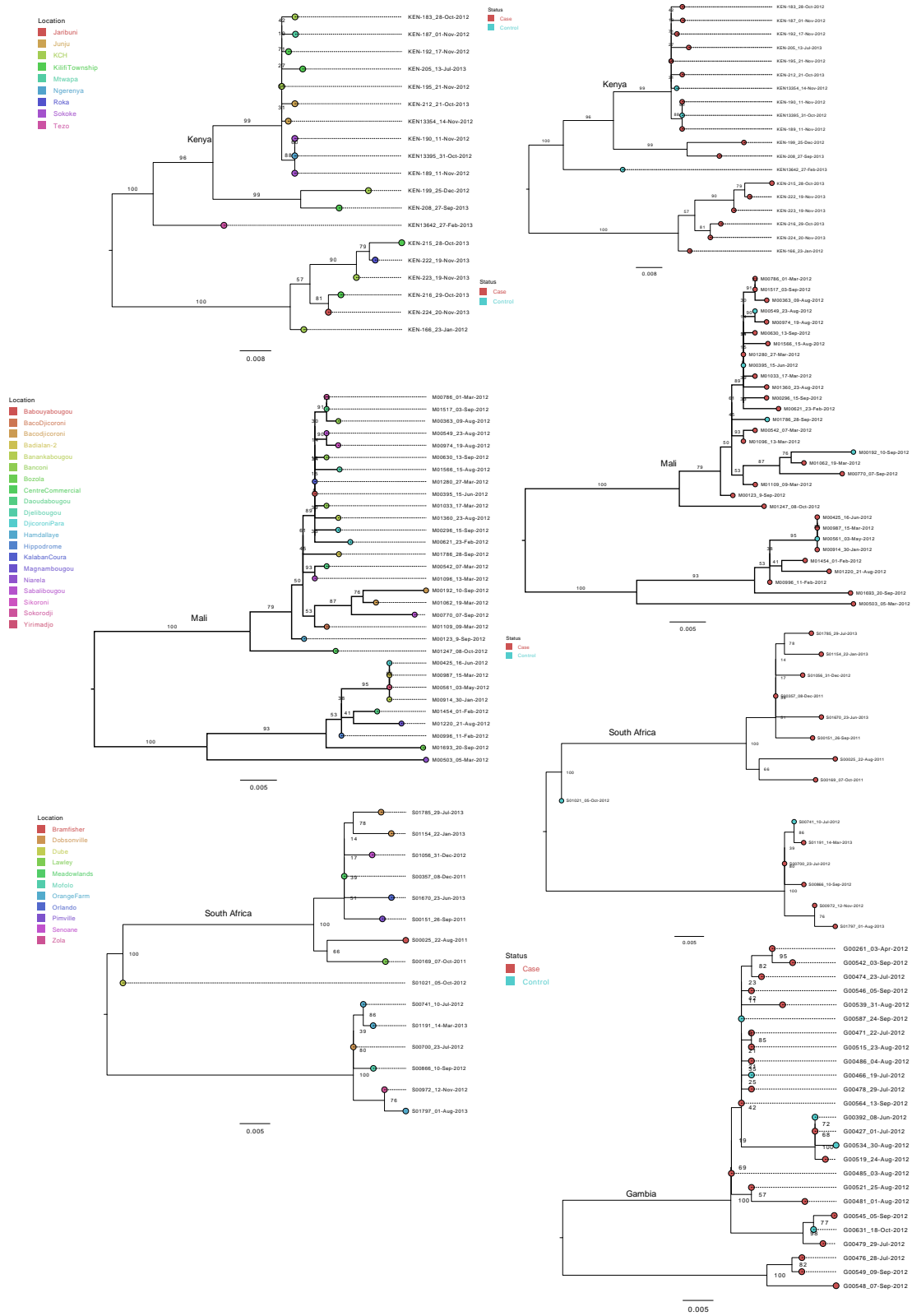


Figure S5: ML phylogenies of subgroup B1 sequences showing within country sequence diversity for Kenya, Gambia, Mali and South Africa sequences. Clustering patterns were

determined by within country sampling location and or case/control status. For Gambia, only case/control clustering patterns were determined.

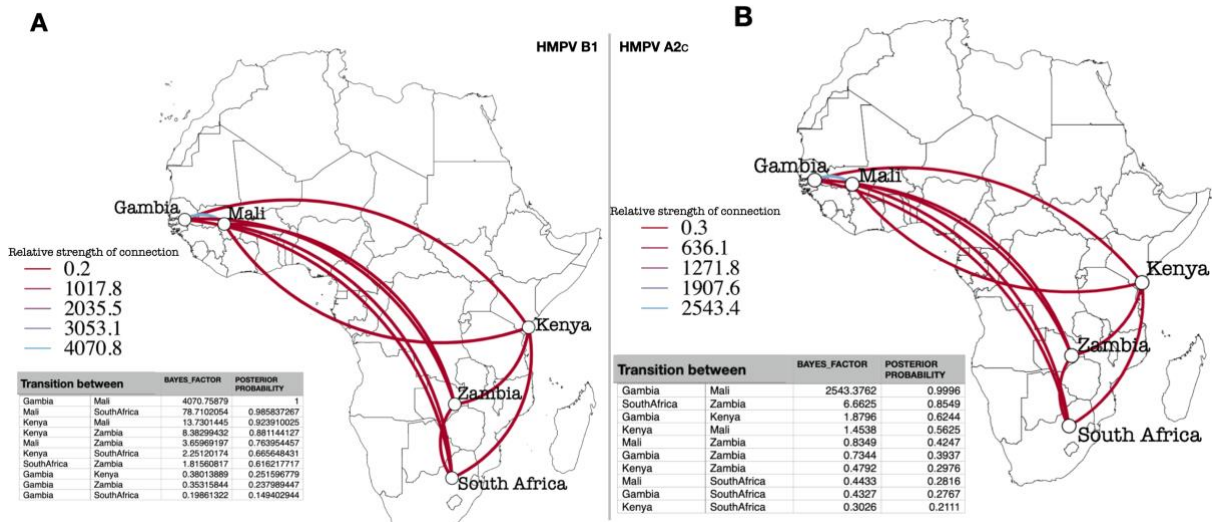


Figure S6: The inferred continental (Africa) migration pathways for HMPV B1 (panel a) and A2c (panel b) viruses constructed under symmetric diffusion model. The lines indicate connection between different countries. The colour gradient from red to blue of the lines indicate the relative strength of connection between countries according to the Bayes Factor test. Only statistically supported discrete diffusion rates (Bayes Factor >3) are shown. Posterior probabilities >95% shows very strong supported rates.

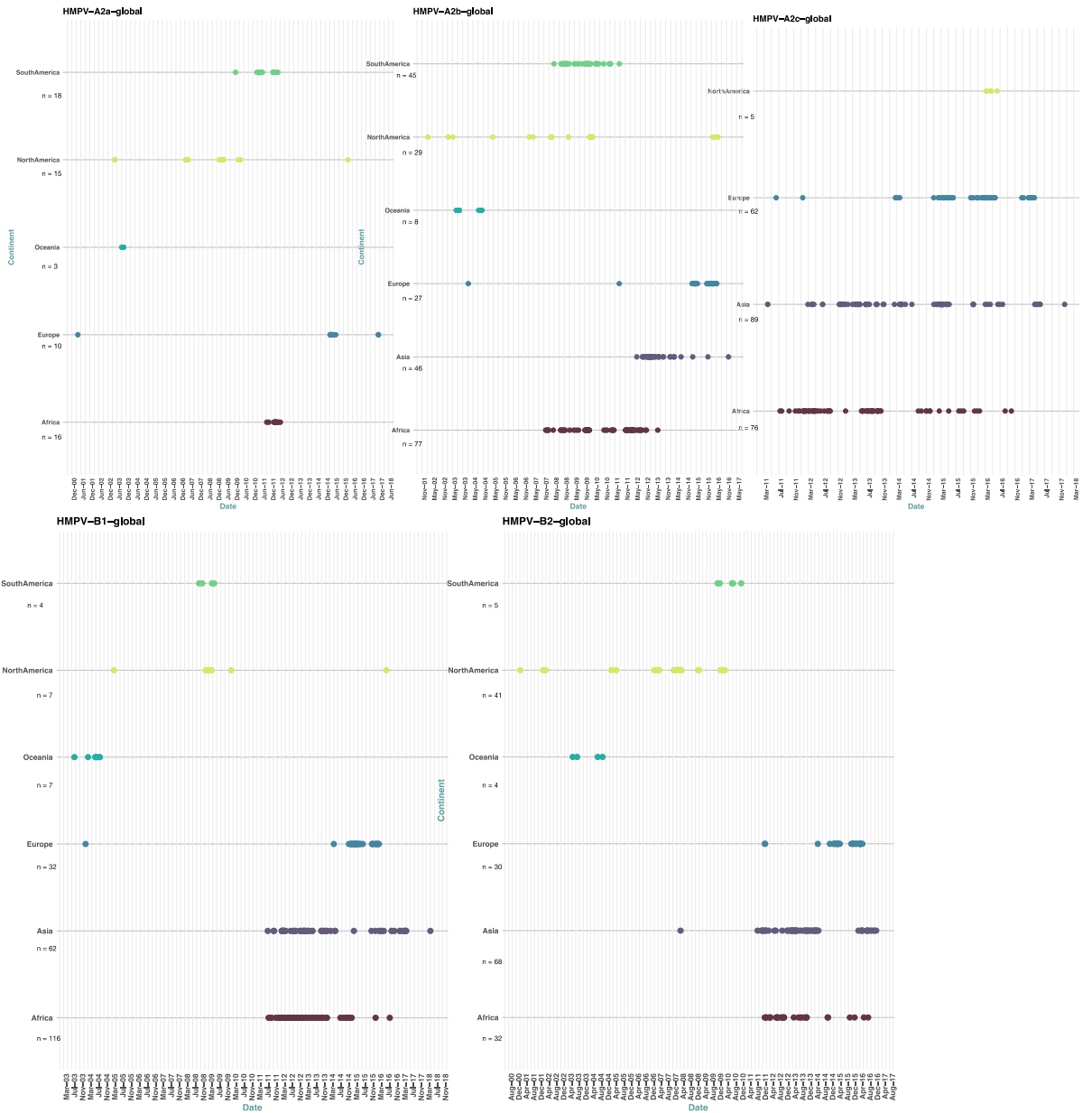


Figure S7: Temporal distribution of HMPV sequence data by subgroup obtained from Africa and GenBank collected between 2000 to 2018.

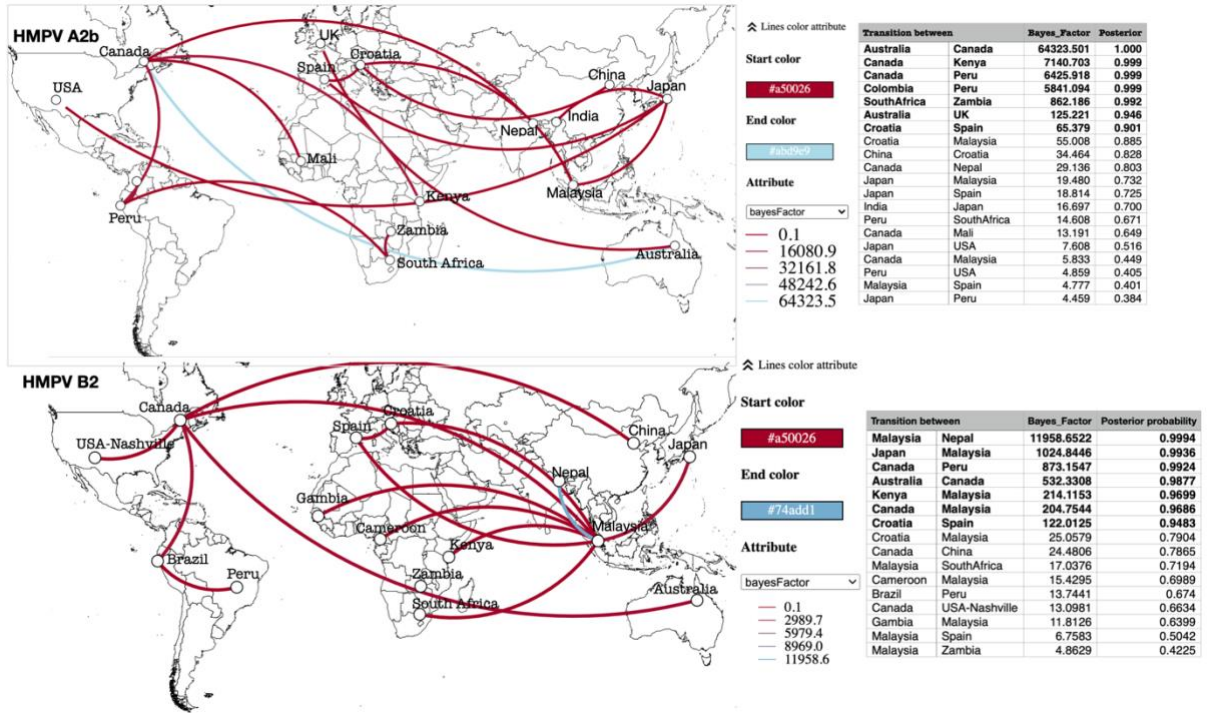


Figure S8: Global spatial diffusion pathways of HMPV A2b (panel a) and B2 (panel b) viruses constructed under symmetric diffusion model. Only statistically supported discrete diffusion rates (Bayes Factor >3) are shown; strongly supported rates with posterior probabilities >95% are highlighted in bold. The lines indicate connection between different countries. The colour gradient from red to blue of the lines indicate the relative strength of connection between countries according to the Bayes Factor test.

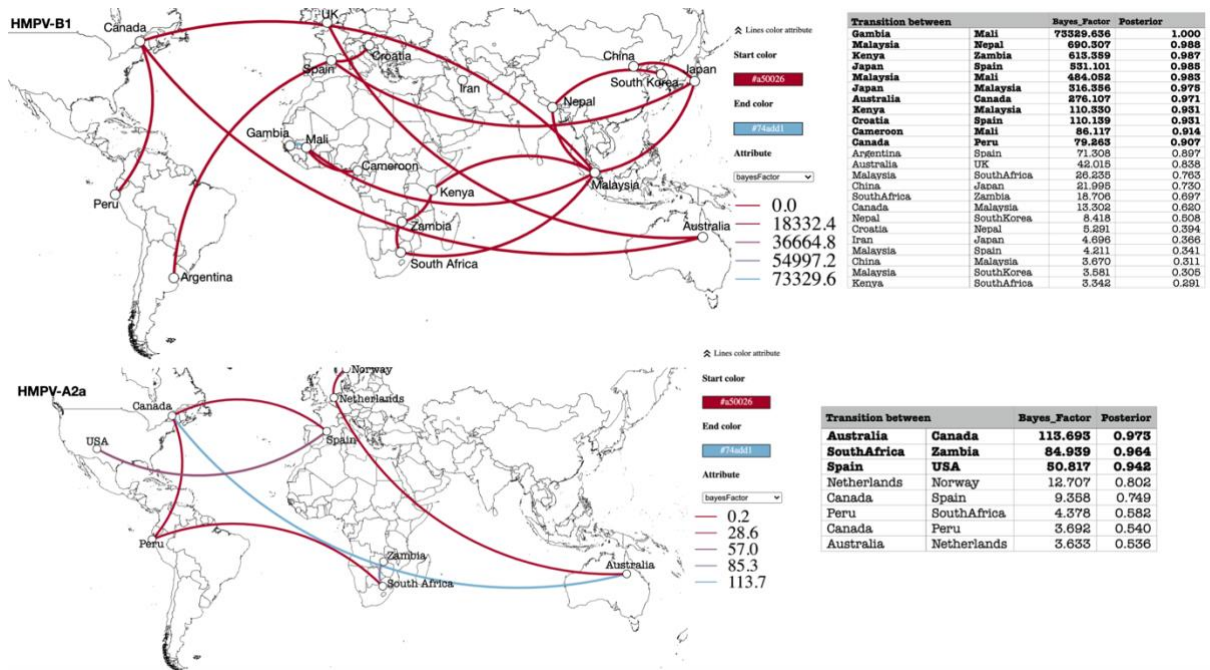


Figure S9: Global spatial diffusion pathways of HMPV B1 (panel a) and A2a (panel b) viruses constructed under symmetric diffusion model. Only statistically supported discrete diffusion rates (Bayes Factor >3) are shown; strongly supported rates with posterior probabilities >95% are highlighted in bold. The lines indicate connection between different countries. The colour gradient from red to blue of the lines indicate the relative strength of connection between countries according to the Bayes Factor test.

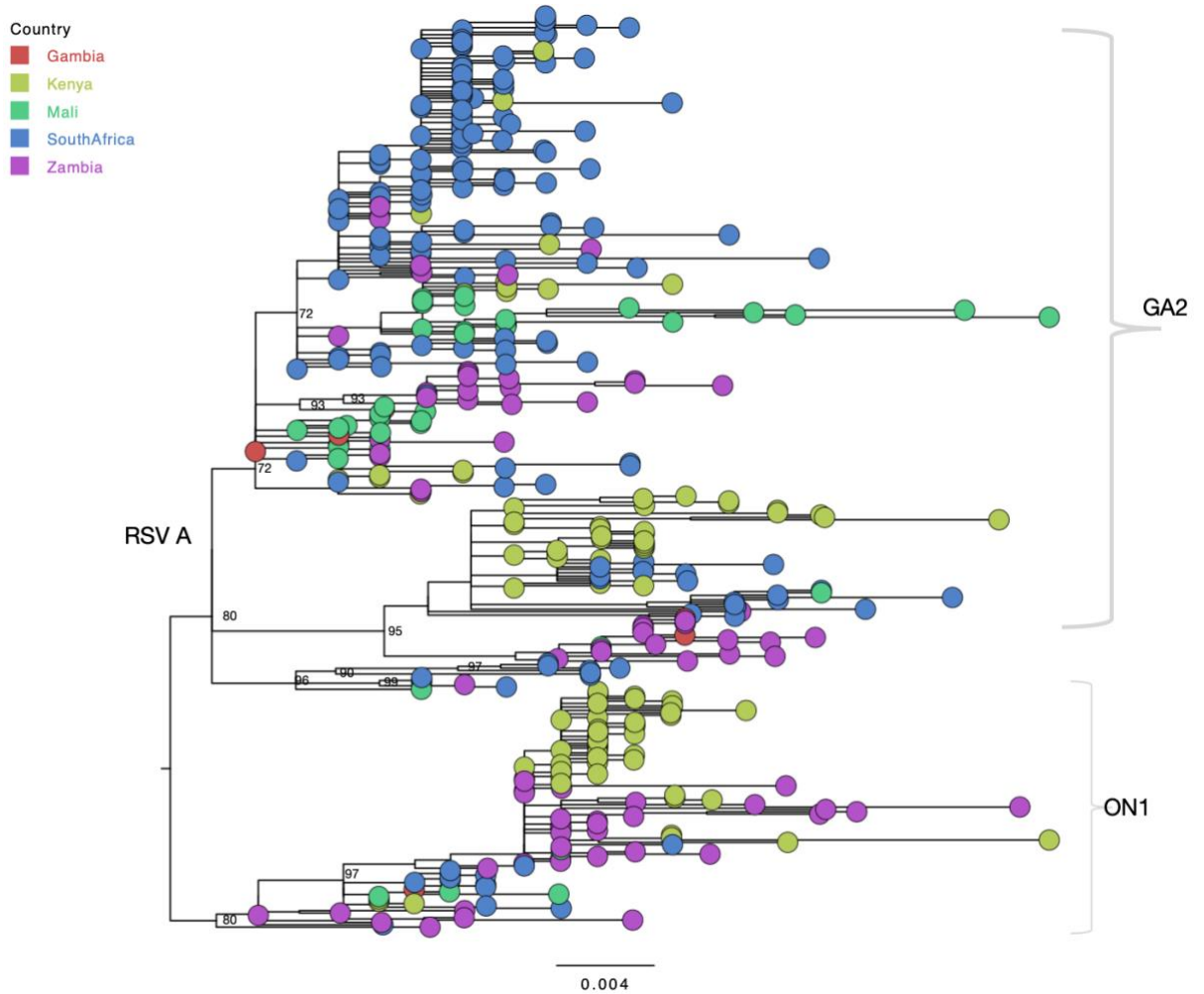


Figure S10: Phylogenetic relatedness of RSV A G gene sequences collected from Kenya, Mali, Gambia, South Africa and Zambia. The numbers next to branches indicate the bootstrap values. The tree tip-shapes are coloured by country of sampling. The genotype assignment is shown to the right-hand side of the major clades. Subgroups were confirmed if sequences clustered with the reference sequences within a major branch with > 70% bootstrap support.

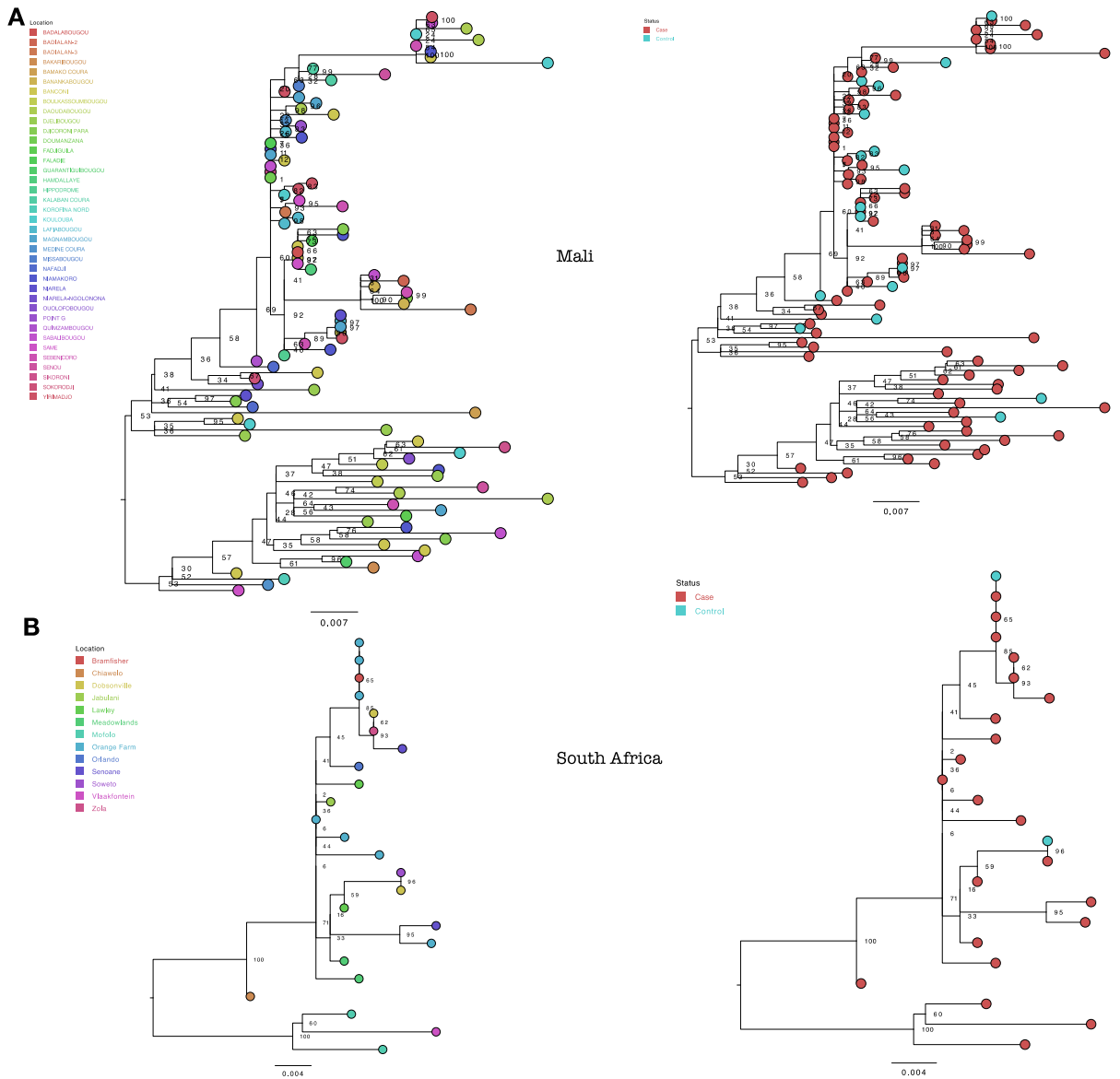


Figure S11: ML phylogenies of RSV genotype BA sequences showing within country sequence diversity for Mali and South Africa sequences. Clustering patterns were determined by within country sampling location and or case/control status. The tree tip-shapes are coloured by within-country sampling locations or by cases/control status.

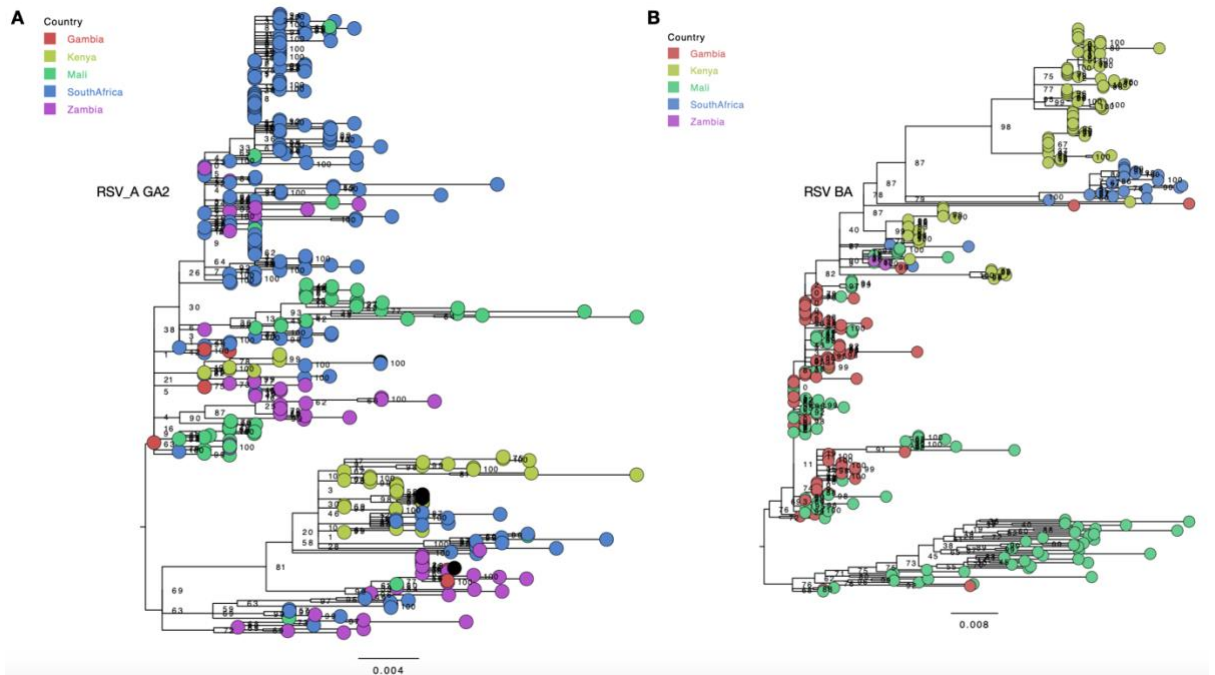


Figure S12: Phylogenetic relatedness of RSV G gene sequences collected from Kenya, Mali, Gambia, South Africa and Zambia. Tip shapes are coloured by country of sampling. The numbers next to branches indicate the bootstrap values. Panel a, ML phylogeny of RSV genotype GA2 sequences. Panel b, ML phylogeny of RSV genotype BA sequences.

RSV A ON1

Tip shapes colour_country

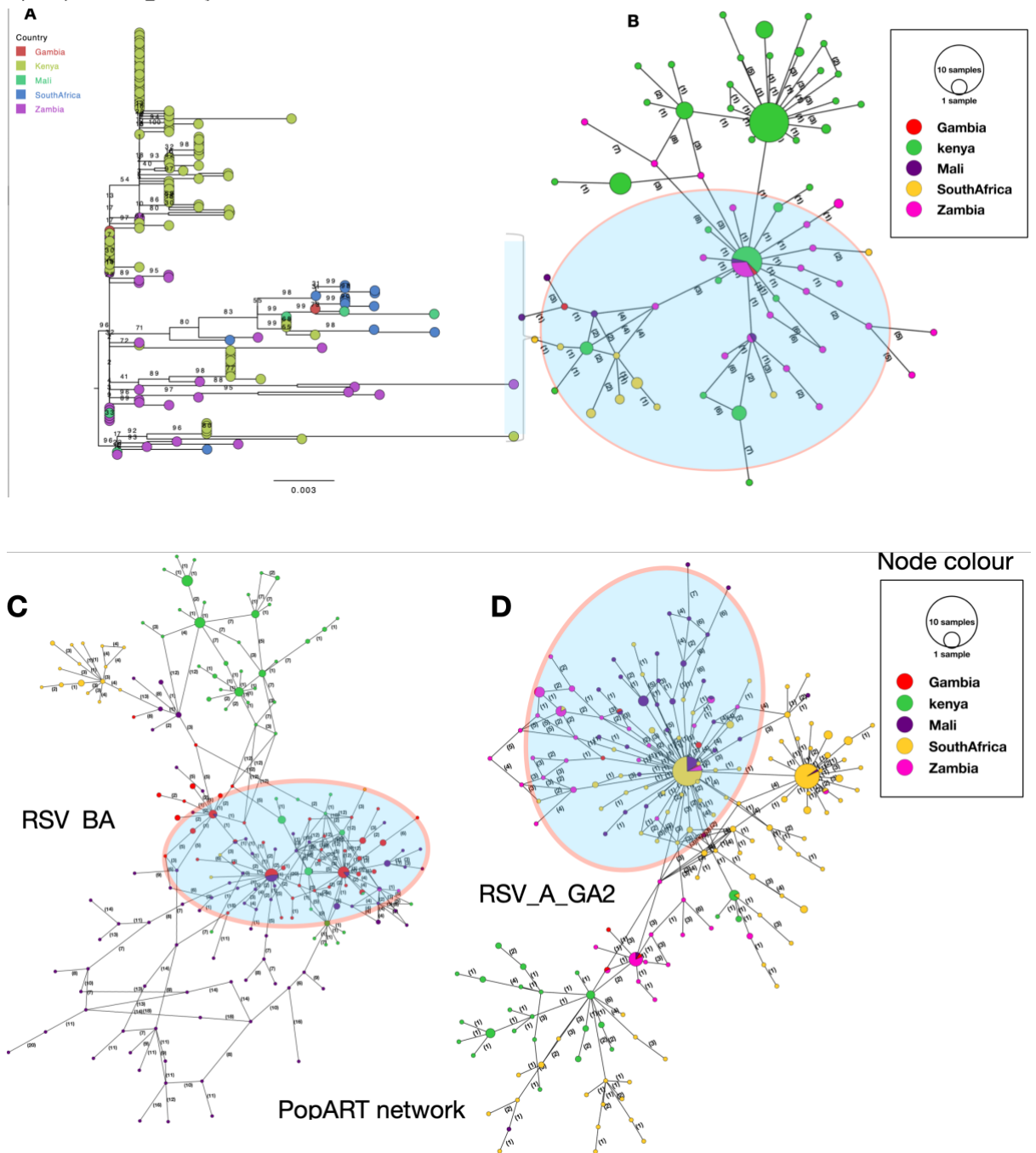


Figure S13: Genetic relatedness of RSV G gene sequences collected from Kenya, Mali, Gambia, South Africa and Zambia. Panel a, ML phylogeny of RSV ON1 sequences. Tip shapes are coloured by country of sampling. The numbers next to branches indicate the bootstrap values. Panel b, c, and d, shows minimum spanning network of genetic

distances for RSV genotype ON1, BA, and GA2 respectively, constructed using PopART.

Clusters in red margin indicate potential inter-country transmission links.

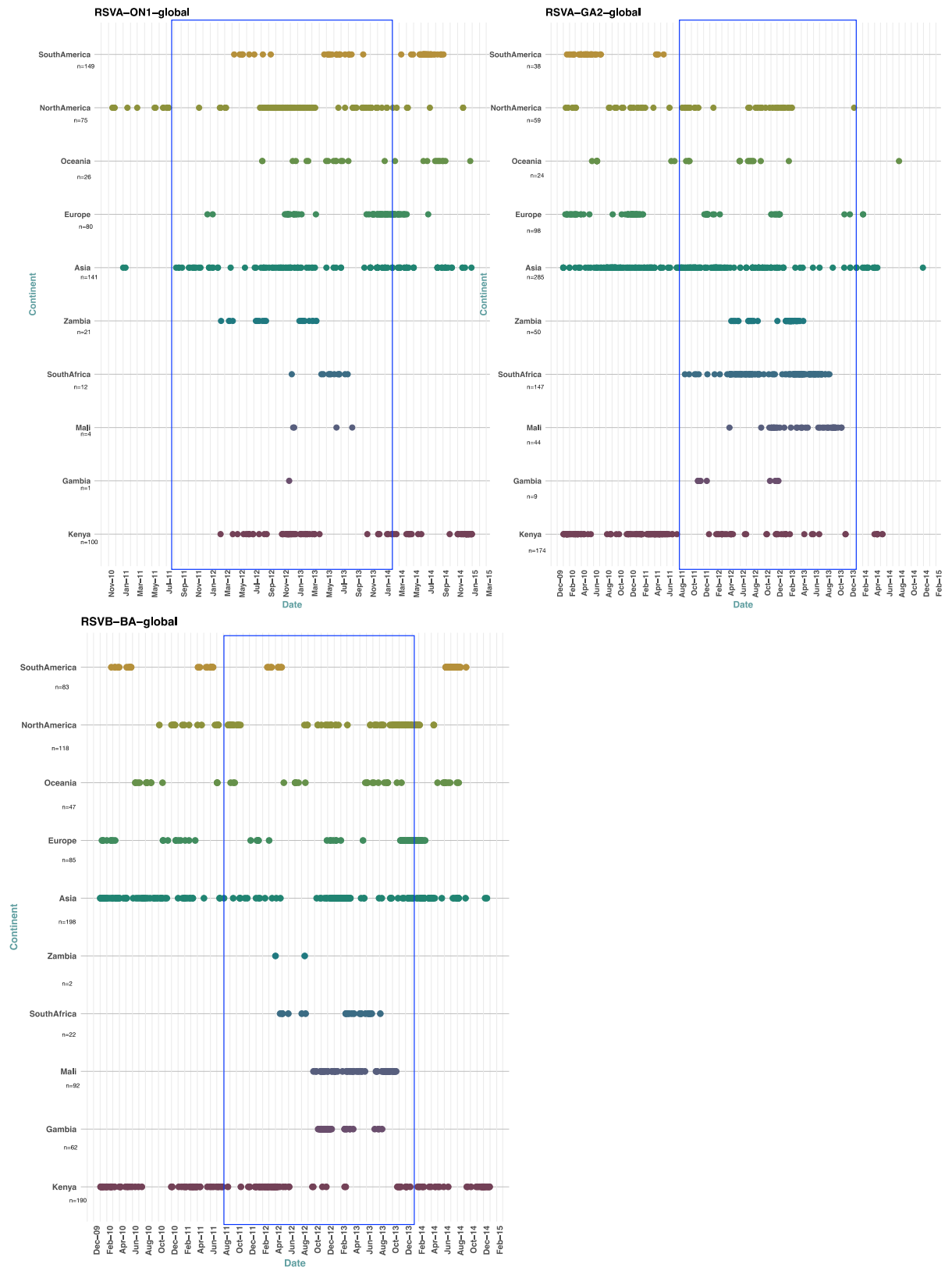


Figure S14: Temporal distribution of RSV sequence data by subgroup obtained from Africa and GenBank collected between 2010 to 2015. The blue margins indicate sampling period contemporaneous with our study sampling period (August 2011 to January 2014).

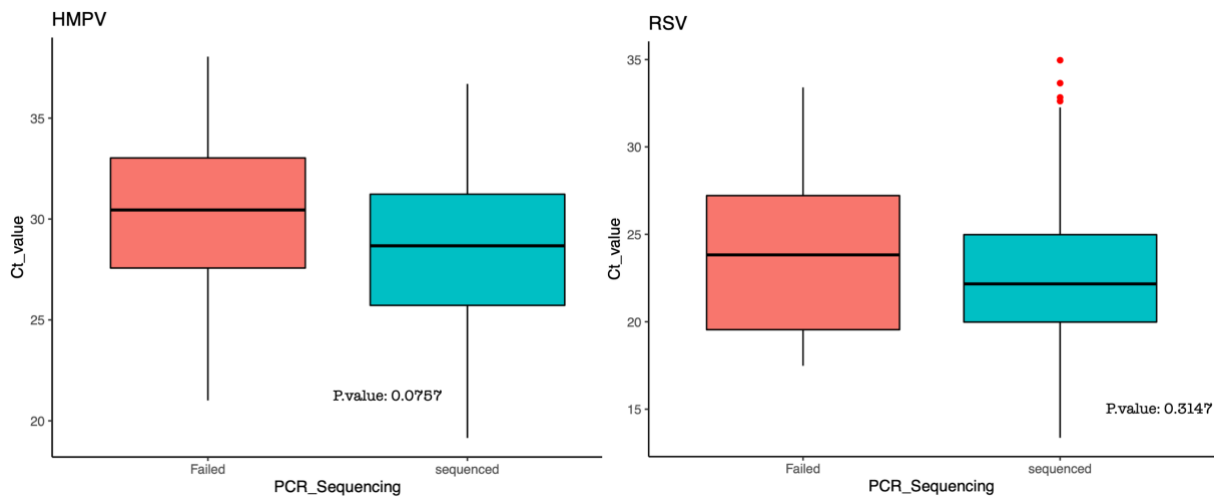


Figure S15: Distribution of Ct. values for samples sequenced and those that failed sequencing for HMPV and RSV. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The horizontal line in the middle of the box represents the median Ct. value. The p.values indicate the mean differences in Ct. values between samples sequenced and those that failed sequencing for each virus.

Appendix C

Funding

This work was supported by the Fogarty International Center of the National Institutes of Health under Award Number U2RTW010677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Additional support was provided through the DELTAS Africa Initiative [DEL-15-003]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107769/Z/10/Z] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government