The Texas Medical Center Library
DigitalCommons@TMC

UT School of Public Health Dissertations (Open Access)

School of Public Health

Fall 12-2018

RELATING RSV GENETIC DIVERSITY TO GLOBAL TRANSMISSION DYNAMICS

Rebecca J. Fisk
UTHealth School of Public Health, becca.fisk@gmail.com

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen

Part of the Community Psychology Commons, Health Psychology Commons, and the Public Health Commons

Footer Logo

RELATING RSV GENETIC DIVERSITY TO GLOBAL TRANSMISSION DYNAMICS

AND DISEASE SEVERITY

by

REBECCA J FISK, BS

APPROVED:

_____

CRAIG L HANIS, PHD

_____

JUSTIN BAHL, PHD

_____

_____

RELATING RESPIRATORY SYNCYTIAL VIRUS GENETIC DIVERSITY TO GLOBAL

TRANSMISSION DYNAMICS AND DISEASE SEVERITY

by

REBECCA J FISK
BS, College of William & Mary, 2015

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF PUBLIC HEALTH

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
December, 2018

ACKNOWLEDGEMENTS

RELATING RESPIRATORY SYNCYTIAL VIRUS GENETIC DIVERSITY TO GLOBAL

TRANSMISSION DYNAMICS AND DISEASE SEVERITY

Rebecca J Fisk, BS, MPH
The University of Texas
School of Public Health, 2018

CE/Thesis/Dissertation Chair: Craig L Hanis, PhD

Many studies of Respiratory Syncytial Virus (RSV) have relied on analyses of the Major Surface Glycoprotein G gene (G gene). Global transmission patterns have not been well studied due to lack of systematic global surveillance efforts. This study used phylogenetic analysis of full genome data, categorized by geo-region, to determine the sources of RSV A and B infection in Chile and Houston, Texas. Additionally, disease severity studies have generally focused on outcomes associated with a single genotype. In this study we developed a statistical phylogenetic approach to explore the relationship between tree topology and disease severity. Disease severity data included if the patient was given oxygen, if they were hospitalized, and if they were admitted to an intensive care unit.

Global data was downloaded from GenBank, separated into RSV A and RSV B, aligned, and manually optimized. The United States and Canada region was overrepresented in the publicly available data, so subsampling was conducted to reduce selection bias. Starting trees were generated from the subsampled datasets using RAxML. Geographic traits and trait state transition rates were jointly estimated in a Bayesian statistical framework using

BEAST. The global transmission network was estimated using the Bayesian stochastic search variable selection and a constant population with a HKY genetic substitution model. Trait associations were calculated using BaTS.

For RSV A, the time to most recent common ancestor (tMRCA) was 1963.40 (95% BCI: 1946.15, 1969.60). For RSV B, the tMRCA was 1963.80 (95% BCI: 1959.50, 1967.33). Europe and Central Asia was a key source of RSV A and B transmissions for both Chile and Houston. In addition, the Middle East and North Africa and Latin America and the Caribbean were sources of RSV transmission into Houston. For the RSV A clinical data, there were significant associations between disease severity and tree topology when analyzing all three traits together (AI 3.13 p<0.01, PS 22.39 p<0.01) and for oxygen (AI 0.98 p<0.01, PS 9.32 p<0.01) and hospitalization independently (AI 1.92 p<0.01, PS 11.78 p<0.01). No significant association was found between tree topology and ICU admission. No significant associations were found in the RSV B clinical data, which may be due to the small sample size and homogeneous outcomes in this group.

Improved surveillance systems are needed to gain a better understanding of global transmission patterns to complement studies done of local transmission patterns, as global introductions play an important role in local outbreaks. Identifying genetic mutations that lead to more severe outcomes may help researchers target vaccine development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

BACKGROUND

Literature Review

Respiratory Syncytial Virus (RSV) was first isolated in 1956 in Chimpanzees by Dr. J.A. Morris and colleagues (Orga, 2004). It was isolated from humans for the first time later that year from two infants, one with pneumonia and one with croup (Chanock, et al., 1962). One of the first RSV epidemiological studies in 1962 noted that "RS virus appears to become disseminated extensively in the pediatric population every year (Chanock, et al., 1962)." Today we have more information on the epidemiology of the disease and its ubiquitous nature; it is estimated that almost every child has been infected by the age of two (Bont, et al., 2016). In the era of genetic sequencing, efforts to study transmission and severity through phylogenetics originally focused on partial sequencing of the G gene, which codes for the attachment glycoprotein (Schobel, et al., 2016), but the focus has recently shifted to collecting and using full genome data in analyses to better understand transmission dynamics not readily apparent in traditional epidemiologic and G gene data.

RSV Genome and Genotypes

RSV is an enveloped, negative-sense, single-stranded RNA virus, with a non-segmented genome that is about 15,000 nucleotides long (Borchers, et al., 2013). Along with the mumps virus and parainfluenza, RSV is a member of the *Paramyxoviridae* family and of the genus *Pneumovirus*. The virus has two major antigenic types, A and B, which are classified by differences in reaction to monoclonal antibodies, although there is significant genetic variability within each group (Duvvuri, et al., 2015; Orga, 2004).

1

The RSV genome has 10 genes that code for 11 proteins (Schobel, et al., 2016). Genes G, F, and SH code respectively for attachment glycoprotein, fusion glycoprotein, and small hydrophobic protein, which play roles in viral attachment and entry to host cells (McLellan, Ray, & Peeples, 2013; Tan, et al., 2013). The remaining genes code for nonstructural proteins (NS1 and NS2), nucleocapsid protein (N), phosphoprotein (P), matrix protein (M), transcription regulators (M2-1 and M2-2), and large polymerase (L) (Schobel, et al., 2016; Tan, et al., 2013).

The G gene produces the key surface glycoprotein in viral binding to host cells and is often called the attachment protein (McLellan, Ray, & Peeples, 2013). This gene has traditionally been the focus of studies on the evolutionary history of RSV because there is a hypervariable region at the C-terminus that contains most of the genetic variation in the genome (Schobel, et al., 2016). Before whole genome sequencing for RSV was widespread, it was easier to base analyses on this section of the gene since it was thought to be the location demonstrating most evolutionary signals (Schobel, et al., 2016). It has also been suggested that the lack of cross-immunity between types and genotypes is due to variation in the G gene (Tan, et al., 2013). Genotypes of RSV are currently classified by this hypervariable region in the G gene (Hibino, et al., 2018). There has been disagreement on the classification and naming of RSV genotypes, but there are two that are easily identified due to insertions in the G gene: RSV A genotype ON1 and RSV B genotype BA. RSV A genotype ON1 has a 72-nucleotide duplication that is not present in other common A genotypes, such as NA1 and GA2 (Schobel, et al., 2016). RSV B genotype BA has a similar duplication in the same region that is 60-nucleotides long, this duplication is also missing in

other genotypes of B (Schobel, et al., 2016). ON1 and BA have been the dominant genotypes in recent outbreaks (Cui, et al., 2013; Tabatabai, et al., 2014).

The F gene, which is sometimes analyzed in conjunction with the G gene, codes for another surface glycoprotein that is key to viral entry to host cells (Schobel, et al., 2016). This protein has a pre-fusion form and a post-fusion form, and critical antigenic sites in the protein have been used in vaccine design (Hause, et al., 2017). Compared to G gene, the F gene is well-conserved across all genotypes of both RSV A and B (McLellan, Ray, and Peeples, 2013). A study conducted by Hause and colleagues found more genetic variability than previously thought, especially in the pre-fusion antigenic sites, and more variations in RSV B sequences than RSV A sequences (2017). Another gene that plays a role in viral entry and is often not included in epidemiologic analyses is the SH gene. SH codes for the small hydrophobic protein that changes the permeability of the host membrane and aids in viral entry to host cells, although it is not necessary for entry (Tan, et al., 2013).

Epidemiology

In temperate climates, RSV circulation shows seasonal patterns similar to seasonal influenza, with peaks in the winter months and fewer infections in the summer. In the Northern Hemisphere, the season normally starts in October or November and ends in March or April, with an epidemic peak occurring between December and February (Bont, et al., 2016). Like influenza, RSV can spread through aerosolized droplets generated by coughs or sneezes, and the virus is able to live on hard surfaces for several hours (CDC, 2017). According to the US Centers for Disease Control and Prevention (CDC), children are most often exposed to the virus at school or day-care and then cause household transmission

(CDC, 2017). Historical data are largely focused on the United States, although more recent studies on the epidemiology of RSV have been conducted in a large number of countries around the world. Increasing the number of countries and regions represented in the publicly available dataset makes it possible to study the larger patterns of disease transmission.

RSV infections usually cause mild, cold-like signs and symptoms, but high-risk populations can develop more severe complications, including bronchiolitis, pneumonia, and/or death. RSV "has a propensity for causing bronchiolitis," and often produces a form of the disease that is longer in duration and more severe than bronchiolitis from other causes (Pickles & DeVincenzo, 2017). High-risk populations listed by the CDC include infants under 6 months of age, infants born prematurely, especially if the infant has chronic lung or heart disease, children with reduced immune function, and adults over the age of 65, especially those with weakened immune system or chronic heart or lung disease (CDC, 2017). Although the risk of severe complication requiring hospitalization is higher in infants with chronic diseases and premature birth at the individual level, the majority of admissions occur in children who are not high-risk (Bont, et al., 2016). Over 70% of pediatric hospital admissions for RSV have no underlying medical conditions, so although their individual risk is lower, previously healthy children make up the majority of cases (Bont, et al., 2016).

A meta-analysis of all available epidemiologic studies on RSV to determine its global impacts, conducted by Shi and colleagues, found that 22% of all severe cases of acute lower respiratory infections were due to RSV in children under the age of 5, resulting in around 3.2 million hospitalizations and 59,600 deaths in hospitals (2017). Almost half of the hospitalizations and deaths, 1.4 million and 27,300 respectively, occurred in children under

six months of age (Shi, et al., 2017). The original Global Burden of Disease Study estimated that 6.7% of all deaths in children 1 month to 1 year old were due to RSV, as were 1.6% of all deaths in children ages 1 to 4 (Lozano, et al., 2012). The rate of hospitalization and case fatality rates within hospitals varied by the country's income level as defined by the World Bank. Lower hospitalization rates were found in low income countries and rates increased as the economic situation of the country improved (Shi, et al., 2017). Two reasons for this discrepancy is poor access to care in resource constrained locations, and poor care-seeking behavior, although Shi and colleagues do not state whether this is due to a family's inability to pay or mistrust of the medical system (2017). Case fatality rates trended in the opposite direction; death rates were higher in low income countries and the rates decreased as income status increased (Shi, et al., 2017).

To date, the Shi study has been one of the most comprehensive reviews of the burden of RSV globally, although it is likely that the true burden of the disease has been underestimated because the study only counted cases that were admitted to the hospital. In places where access to medical care is limited, patients may not be able to go the hospital and those cases would go unreported and unaccounted for in this burden analysis. The underestimation is compounded by gaps in the data, as there are many areas of the world where data on RSV have not been collected (Shi, et al., 2017). To address this problem, the World Health Organization launched the Global RSV Surveillance pilot, which "aims to test the feasibility of leveraging the Global Influenza Surveillance and Response System platform for RSV surveillance without adversely affecting the well-established surveillance of influenza (WHO, 2017)." More data on the distribution of RSV and its molecular

epidemiology will aid efforts to better understand the transmission patterns and determinants of disease in order to better control its spread, and eventually to design effective RSV vaccines (WHO, 2017).

Molecular Epidemiology and Global Transmission

Many studies have focused on local circulation patterns of RSV genotypes across seasons or within a single season using local surveillance data (Hibino, et al., 2018; de-Paris, et al., 2014; Esposito, et al., 2015; Tran, et al., 2013; Panayiotou, et al., 2014). The dominant type can alternate between A and B across seasons, and the dominant genotypes can also change across seasons. For example, RSV A genotype NA1 was the dominant type in Japan before the introduction of ON1, which then became the dominant genotype (Hibino, et al., 2018). Multiple genotypes from both RSV A and B types can circulate concurrently within a single season, increasing the complexity of circulation patterns. (Schobel, et al., 2016). Many studies conducted on the distribution of RSV focus on identifying the dominant and co-circulating genotypes in one location and provide limited analysis of the geographic transmission (Hibino, et al., 2018; de-Paris, et al., 2014; Esposito, et al., 2015; Tran, et al., 2013). Very few studies connect this local data to publicly available global data to place local outbreaks in the context of the larger phylogenic tree, which would allow investigators to make inferences about the source of the disease in their country or locality and about global transmission patterns.

Global transmission patterns have not been well studied due to lack of systematic global surveillance efforts and the resulting large gaps in the information available to investigators (Duvvuri, et al., 2015; Zou, et al., 2016). One of the few studies to take the

6

global view, focused only on the distribution of the ON1 genotypes, as defined by the G gene (Duvvuri, et al., 2015). They estimated that ON1 emerged in 2007 or 2008 in Ontario, Canada, and by the time the study was conducted it had been detected on every continent (Duvvuri, et al., 2015). The phylogenic tree produced in the study did not appear to show a strong association between clade and geography, although this could be due to the limited amount of data available (Duvvuri, et al., 2015). Duvvuri and colleagues suggested that irregular clusters, such as one including Canada, United States, Thailand, and Italy, could be due to patterns in travel (2015). One limitation of this study is that they used data only on the second hypervariable region of the G gene, instead of using a whole genome approach to infer global patterns (Duvvuri, et al., 2015).

Another study analyzing sequences from Guangdong, China, in conjunction with sequences from publicly available strains indicated the dominant strains GA2 and ON1 originated in the Americas before spreading to other regions (Zou, et al., 2016). The investigators noted that this result could be due to bias in the dataset, since most early samples are from the United States and much of the available data is also from this country (Zou, et al., 2016). In the phylogenetic tree generated by the study, they found that many clades contained viral strains from multiple geographic areas, but several clades represented only Guangdong (Zou, et al., 2016). This indicates that there were introductions of RSV from other regions, but seasonal epidemics may also have been seeded by locally persistent strains of the virus (Zou, et al., 2016). However, analyses are biased towards identifying local persistence because of a lack of information from other regions (Zou, et al., 2016).

The Zou study had two main limitations: their analysis relied on just G gene sequences, similar to many other molecular epidemiologic studies; and their geographic analyses relied on four large regions, which limits the utility of the findings (Zou, et al., 2016). Reliance on the G gene does not present the whole evolutionary history of the virus, but RSV whole genome analyses are rare. One study conducted by Bose and colleagues found a similar pattern to Zou (2016). Their analysis found several genotypes were circulating simultaneous across the globe, but that some rarer genotypes were only found within one area; which is probably due to inconsistent surveillance around the world as these rare genotypes were found in areas with more data (Bose, et al., 2015). Unlike the Zou study, Bose and colleagues did not note a strong relationship between geographic location and clades on the phylogenic tree, which indicates that RSV is not evolving independently and separately in one locale and there is interaction between regions (Bose, et al., 2015).

As more whole genome data has become available and more advanced phylogeographic methods are developed, global transmission patterns of RSV can be better explored. Understanding both local and global dynamics using a phylogeographic approach will reveal the relative importance of international introductions in these seasonal epidemics and inform surveillance programs and control measures in the future.

Patterns in Severity of Disease

Although much attention has been paid to the severity of RSV infection in relation to host factors, such as immunosuppression or chronic health conditions, many studies indicate that severity is related to more than just host factors. Some viral genome characteristics can also be important determinants of viral pathogenicity, resulting in different levels of disease

8

severity. However, viral genetic characteristics and RSV pathogenicity have not been well connected. Most studies focus on the relationship between disease severity and RSV genotype with inconsistent results. One such study conducted over three epidemic seasons in Japan found that 35.6% of those infected with ON1 were hospitalized, indicating more severe disease, and this proportion was greater than all other measured genotype hospitalization rates (Hibino, et al., 2018). The odds ratio of hospitalization for those infected with ON1 to those infected with NA1 was 6.92:1, and there was no significant difference in the odds of hospitalization between NA1, BA9, and BA10 infections (Hibino, et al., 2018). Yoshihara and colleagues found similar results in Vietnam, where the risk of lower respiratory tract infections was 2.26 times higher in those with ON1 compared to those with NA1 (2016). This is opposed by results from a study in Northern Italy, in which NA1 was more likely to cause upper respiratory tract infections and require hospitalizations than ON1 (Esposito, et al., 2015). In a population where ON1 was not detected, Luchsinger and colleagues found that infection with NA1 was more likely to cause hospitalization and severe disease than two RSV B genotypes (2014). In Cyprus, researchers found that RSV A genotype GA2 and RSV B genotype BA both caused more severe disease than ON1 (Panayiotou, et al., 2014). They also found that a larger proportion of those infected with BA required oxygen, suggesting BA causes severe outcomes more frequently than other genotypes, in contrast to common results indicating that RSV A is more likely to cause severe disease (Panayiotou, et al., 2014). Espinosa and colleagues also looked for a relationship between genotype and disease severity in Chile but did not find any significant interaction between viral type or genotype and disease severity (2017).

Since multiple studies have yielded mixed results, the links between genetic characteristics of RSV and disease severity are equivocal. Many studies have focused on relating genotype to severity, but these studies have been inconclusive due to limited sample size and the large variations within genotypes. The reliance on just the G gene in genotyping may also play a role in the muddled findings, since other genes inducing host immune responses are ignored. These other genes, especially the F gene, may have important genetic differences that lead to differences in disease severity, but these characteristics have not yet been identified because a genomic approach has rarely been applied in these analyses.

A few studies on differences of disease severity among RSV strains have been conducted in vitro, which reported different levels of cytokine activation across strains, but not across genotypes. Levitz and colleagues (2012) studied the effects of different strains collected from patients to produce varying inflammatory cytokine (IL-6) responses. An increase in IL-6 production is part of the body's immune response to RSV infection. Their analysis focused on RSV A, genotypes GA2, GA3, GA4, and RSV B and found that the level of IL-6, an inflammatory cytokine, varied greatly across strains within the same genotype, between RSVA and B, and by genotypes at the same level (Levitz, et al., 2012). A similar study by Thompson and colleagues (2015) observed disease severity in patients and compared severity with *in vitro* cytokine induction for four cytokines (IL-1α, IL-6, IL-8 and RANTES). There was no significant difference between strains in IL-6 levels, which differs from the results of the Levitz study, but instead there were significant differences between strains in the production of two other inflammatory cytokines, IL-8 and RANTES (Levitz, 2012; Thompson, et al., 2015). There were no significant trends in activation of cytokines

across RSV types and genotypes, in contrast to many studies on disease severity which focus on genotyping discussed above (Thompson, et al., 2015; Hibino, et al., 2018; Esposito, et al., 2015; Panayiotou, et al., 2014). These data support the use of full genome data when investigating the connection between disease severity and the strain of RSV, since few trends were seen across genotypes, which indicates that the genetic predictor of disease severity could be located in another part of the genome.

Knowledge Gaps

In summary, we have identified two major gaps in the literature on RSV transmission and disease severity: 1) RSV whole genome data has not been used for inferring global transmission patterns, and current studies rely solely on the G gene, so incomplete lineage sorting could be a confounding variable in relationships proposed by most of these studies; 2) RSV disease severity has been mostly studied at the genotype level, but strain-specific genetic characteristics have not been explored. These two gaps are critical to understanding the epidemiology and burden of RSV. A better understanding of RSV transmission patterns will help inform prevention strategies, and genetic characteristics related to disease severity could provide valuable information for antiviral drug and vaccine design.

Public Health Significance

RSV is important to public health because of its ubiquitous nature and the potential for severe complications in high-risk populations, including children under the age of 5 and adults over the age of 65. Assessments of the impact of RSV have focused on young children, so the burden of disease in the elderly often has been overlooked. It is estimated

that close to 60,000 hospitalizations each year in the United States could be attributed to RSV in children 5 years old and younger, and half a million emergency room visits every year (Breese Hall, et al., 2009). The number of hospitalizations of those over the age of 64 was estimated to be 180,000 per year in the United States, close to three times the number in children (Falsey, et al., 2005). A review of insurance claims data showed that the expense of hospital stays related to RSV was significantly higher than the cost for non-RSV controls across all age categories except for patients ages 5 to 17 (Amand, et al., 2018). The highest discrepancy in cost occurred in those ages 75 to 84, where an average of $17,211 was spent on non-RSV controls and an average of $40,405 was spent on RSV cases (Amand, et al., 2018). These estimates do not include outpatient costs or lost productivity, so there is a large economic impact in addition to the burden on healthcare utilization (Amand, et al., 2018; Falsey, et al., 2005).

It is estimated that almost everyone has been infected by the age of two, but this history of infection does not induce lifelong immunity (Brochers, et al., 2013). Since no vaccine currently exists, although there are several in various clinical stages of development, it is especially important to understand global and local transmission dynamics to inform infection control measures. In this study, we propose to explore the RSV strains from Chile and Houston, Texas, in the context of global transmission patterns inferred from publicly available full genome data to identify potential sources of infection during seasonal outbreaks. Chile and Houston were selected due to collaboration agreements between investigators at Baylor College of Medicine, University of Chile College of Medicine, and

Dr. Bahl at the University of Texas Health Sciences Center. These data will facilitate further exploration of transmission paths between geo-regions and inform prevention strategies.

Viral genetic information paired with patient outcome information can elucidate the relation between the genetic characteristics of RSV and disease severity to target vaccine development on strains with the highest contributions to the burden of disease. Although host factors, viral dose, and region of the lung infected all play important roles in the severity of disease, it is equally important to understand strain specific responses and severity since previous *in vitro* and epidemiological studies have investigated the link between RSV genotype and disease severity. The strains collected from Chile and Houston include information on patient outcomes (if the patient needed oxygen and admission to a hospital or intensive care unit), so clusters or specific genetic characteristics that are associated with high disease severity can be identified.

Specific Aims

Previously, much attention has been paid to local patterns of RSV infection and disease severity by genotype because local transmission was not well understood. Many of these studies indicate that dominant genotypes that were not previously present in their area were introduced from an outside source. As the understanding of local dynamics has improved and the amount of global data have increased, it has become possible to investigate transmission dynamics on a larger scale. Studies on these global patterns indicate that there is a mixture of local persistence and global introduction of RSV based on the G gene only, but no studies so far have tried to apply a source-sink model with RSV full genome data to infer

the global transmission pattern. Studies of disease severity have also focused on identifying the genotypes that caused the largest number of most severe cases, which has led to inconsistent conclusions due to the diversity within each genotype, and no studies have been conducted to investigate the relationships between strain-specific genetic characteristics and disease severity via the phylogenetic approach. With publicly available full-genome data from around the world, the samples from Chile and Houston, Texas, and detailed data of disease severity, we want to answer the questions: 1) What is the RSV global transmission pattern? 2) What are the sources of RSV samples collected in Chile and Houston, Texas between 1987 and 2014? 3) Are certain viral genetic characteristics associated with hospitalization and ICU admission in Chile and Houston? This information can help inform surveillance strategies, similar to what has been done with influenza, and help target vaccine development to protect against genetic characteristics associated with more severe outcomes.

Specific Aims:

1. To understand where the isolates from Chile and Houston are located in the global tree and identify global transmission patterns using phylodynamic source-sink modeling.

2. To explore the correlation between viral genetic characteristics and severe outcomes with clustering analyses.

METHODS

Study Population & Data Set Compilation

All publicly available RSV sequences with over 5,000 nucleotides were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). The accession numbers for each sequence were downloaded manually. Geographic and collection date information was abstracted from the textual and tabular contents of full-text publications by employing GeoBoost, which is a Java program to extract and normalize the location of infected host of viruses from the metadata files and related published records (Tahsin, et al., 2017). The associated metadata, including the date of collection and geographic data was matched with sequence data from GenBank via its unique accession number.

The G, F, and SH gene data with detailed information on disease severity from Chile and Houston was shared by the collaborating investigators in Houston and Chile. Data from Houston were collected from children with lower respiratory infections between 1987 and 2005. These samples have been used previously to study the genetic variability of the G, F, and SH genes (Tapia, et al., 2014). The data from Santiago de Chile were collected as part of a study investigating the effect of sequence variation of the G, F, and SH genes on immune response and disease severity (Tapia, et al., 2012). Older sequences from Chile are from 1990 to 2010 and were sampled from the biorepository at the University of Chile. Strains from 2010 to 2014 were taken from a cohort of RSV infected infants at Hospital Roberto del Rio during these outbreaks. Tapia and colleagues also conducted a chart review to abstract information on the patient's age, sex, and disease severity. Disease severity was evaluated

with three dichotomous variables: given oxygen therapy, hospitalization, and admission to the intensive care unit (ICU).

The full dataset from publicly available data and from collaborators was then separated into RSV A and RSV B for alignment and manual optimization. The data were aligned using MUSCLE v3.8.31 (Edgar, 2004). The aligned data were reviewed and cleaned using BioEdit (Hall, 2005). Manual optimization of the alignment was then conducted to correct artificial gaps inserted into the data by strains with non-base insertions (N instead of A, T, C, or G). Strains without geographic information or the year of collection were removed. If no day information was available then the strain was assigned to the 15th of that month; if day and month information was missing then the strain was assigned to the first of July of that year. Duplicate strains were removed in RAxML and the oldest strain was retained (Stamatakis, 2014). TempEst (http://tree.bio.ed.ac.uk/software/tempest/) was used to identify outliers and investigate the temporal structure or signals of molecular clock in the dataset (Rambaut, et al., 2016).

After outliers were removed from the dataset, the temporal and geographic distributions of RSV A and RSV B were analyzed in Tableau (https://www.tableau.com/) and it was determined that subsampling was necessary. Across several years, some countries were overrepresented in the data, which were then selected for subsampling. We defined overrepresentation as having more than 25 strains for RSV A and more than 20 strains for RSV B from one country in any given year. The selection of strains to be included in the subsample was conducted by randomly selecting accession numbers from the overrepresented groups to get the sample sizes above. Two subsampled datasets were

generated for both RSV A and B for the purpose of sensitivity analysis and quality control. Strains that were not selected were excluded from further analysis to reduce sampling bias in the dataset.

Maximum Likelihood Phylogenetic Analysis

A maximum likelihood (ML) tree to preliminarily explore the phylogeny in the full dataset was generated in RAxML using the General Time Reversible (GTR) nucleotide substitution model with GAMMA distribution of rate heterogeneity among sites. The two subsampled datasets were also run through RAxML using the same substitution model. The resulting ML trees estimated the time to most recent common ancestor (tMRCAs) and the substitution rates were compared to each other and to the full dataset, in order to check the quality of the subsampling strategy. To evaluate the reliability of tree topologies, bootstrapping analyses were conducted with the extended majority rule consensus tree criterion (that is, autoMRE option in RAxML) which automatically determines the sufficient bootstrap replicates for getting stable support values. Trees rooted with mid-point root were visualized and colored in FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). Bootstrapping values were shown on the tree nodes.

Bayesian Phylogenetic Analysis

The phylogeny of the subsampled datasets was reconstructed using Bayesian Markov chain Monte Carlo (MCMC) methods in the Bayesian Evolutionary Analysis by Sampling Trees (BEAST) v1.8.4 with the ML trees as starting trees (Drummond, et al., 2012). The tree topology, the evolutionary rates and tMRCAs were co-estimated. HKY nucleotide substitution model with GAMMA distribution of rate heterogeneity among sites was used to

17

depict the pattern of nucleotide changes (Hasegawa, Kishino, & Yano, 1985). The uncorrelated log-normal relaxed clock model was used to calibrate time information on tree branches (Drummond, et al., 2006). Due to computing constraints, a constant coalescent prior was used for the global datasets (Drummond, et al., 2002). To generate a set of empirical trees, two independent MCMC analyses with chain length of 150 million were run until reaching good convergence, with sampling every 15,000[th] generation and resulting in 10,000 states or trees output in log and tree files, respectively. The empirical trees for RSV A and RSV B were then used for the global transmission analysis. Tracer v1.7 was then used to visualize and assess how well these runs have converged; the convergence of combined chains from two runs was determined to be acceptable when the estimated effective sample size (ESS) > 200 after removing the burn-in (Rambaut, et al., 2018). With well-converged runs, log and tree files from these runs were combined in BEAST LogCombiner. Then BEAST TreeAnnotator was used to annotate one maximum clade credibility (MCC) tree from the combined tree file. Uncertainty from Bayesian analysis was indicated by 95% highest posterior density (95% HPD) also called the 95% Bayesian Credible Interval. The estimated tMRCA and evolutionary rate were reported with mean value and 95% HPD, respectively.

Discrete phylogeographical analysis

To make inferences about global transmission dynamics and to guarantee sufficient samples in each geographic discrete trait, collection location (country) of each sequence was coded into six regions defined by the World Bank: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa

(MEN), South Asia (SA), Sub-Saharan Africa (AF) (World Bank Group, 2018). High income

countries, which are not categorized by the World Bank, were added to the nearest region. In

the results, Western Europe was categorized as a part of ECA, and Australia was in EAP.

The United States and Canada (UC) region was created as an additional region. Houston and

Chile were categorized as additional regions to assess transmission into and out of these

locations, resulting in a total of nine regions as the discrete geographic traits. With the same

substitution model, molecular clock and coalescent models from the previous section

("Bayesian Phylogenetic Analyses"), the Bayesian stochastic search variable selection

(BSSVS) approach was applied to the sets of empirical trees generated above to find a

parsimonious set of rates explaining the geographic diffusions in the phylogeny. Three runs

with a chain length of 100 million were conducted. The process for checking convergence

and combining the runs was similar to the procedure outlined above.

  The BSSVS approach enabled us to construct a Bayes factor (BF) test to identify

significant diffusion processes between discrete geographic traits (Lemey, et al., 2009). The

criteria for significance of the BFs were in accordance with Kass and Raffery's definition:

BF≥3 is considered significant (1995). We further categorized the level of support into five

groups: 3≤BF<6 as weak supported, 6≤BF<10 as substantial support, 10≤BF<30 as strong

support, 30≤BF<100 as very strong support, and BF≥100 as decisive support (Bahl, et al.,

2013). Transmission pattern and BF results were visualized using SpreaD3 (Bielejec, et al.,

2016).

Furthermore, to quantify how many times the transmissions occurred between geo-regions, Markov jump counts for each node of the ancestral reconstruction were assessed based on the global tree (Minin & Suchard, 2008). A non-reversible model was used to determine the directionality of the transmission, so that an asymmetric matrix was constructed to assess source-sink dynamics among the sampled geo-regions. Heat maps to represent the estimated jump counts between two locations were used to show the connections of geo-regions.

Bayesian Tip-Association Significance Testing Analysis

To explore the correlation between viral genetic characteristics and disease severity, the subset of data from Houston and Chile during 2010-2014 with disease severity information was used to conduct clustering analysis. G, F and SH genes were linked and analyzed as one set in the BEAST phylogenetic analysis. The uncorrelated log-normal relaxed clock model and HKY nucleotide substitution model with GAMMA distribution were used for the analysis of the local data (Hasegawa, Kishino, & Yano, 1985; Drummond, et al, 2006). For the local data, the GMRF Bayesian Skyride with time-aware smoothing process was used as the coalescent prior to accommodate changes in the effective population size (Minim, Bloomquist, & Suchard, 2008; Drummond, et al., 2002). The chain length was set to 100 million with sampling every 10,000th generation, and three runs were conducted. LogCombiner and BEAST TreeAnnotator were used to combine the log files and create one MCC tree.

The MCC tree generated in BEAST for the data from Chile and Houston was analyzed further in Bayesian Tip-association Significance testing (BaTS) to determine if

there is a relationship between disease severity and tree topology (Parker, Rambaut, & Pybus, 2008). Disease severity traits, analyzed as dichotomous variables, included if oxygen therapy was needed, if they were admitted to the hospital, and if they had to be admitted to an intensive care unit (ICU). These traits were analyzed together, and also separately to identify the association between single disease severity trait and genetic characteristics. A single configuration was run in the BaTS program, where an Association Index (AI), Finch's parsimony score (PS), and a monophyletic clade (MC) size statistic were calculated to determine the strength of association. The AI tests for the consistency of a trait across internal nodes of the tree, and a lower number represents a stronger relationship between tree topology and the trait of interest. The PS takes into account the number of trait changes beyond each ancestral node, where the score is reported as an integer and lower numbers indicate a stronger relationship between the trait and tree structure. The MC statistic assumes that if a clade is monophyletic for the trait of interest, then the MC value is high and the relationship between tree topology is strong. Clusters that are significantly correlated with high disease severity were reported with the corresponding p-value.

Ethics Statement

All data used in this study were de-identified before we received the datasets from Chile and Houston. No further medical or demographic information was collected from these patients, and we did not have access to their medical records. The original data collection in Chile and Houston was carried out with IRB approval. Publicly available data from GenBank does not include protected health information or personal identifiers of patients from which

21

the samples were isolated. The protocol for this secondary data analysis was submitted for IRB approval within the University of Texas Health system and determined to be exempt.

RESULTS

Global Transmission Dynamics

Global Distribution of RSV Samples

Originally, 1,186 RSV A sequences were downloaded from GenBank and 143 samples came from collaborators in Chile and Houston. For RSV B, 486 sequences were downloaded from GenBank and 107 came from study collaborators. After removing duplicates, lab strains, and outliers, there were 1,123 RSV A sequences and 568 RSV B sequences (see Figures A1a and A1b for global distribution of these samples). After subsampling, there were 739 RSV A sequences and 474 RSV B sequences included in the analyses. Of 739 RSV A sequences, 54 were from Chile and 67 were from Houston. Of the 474 RSV B sequences, 44 taxa were from Chile and 37 were from Houston. By geo-region, the vast majority of RSV A sequences were from the UC region (n=284), as were the majority of RSV B sequences (n=178). EAP (n=115) was the second most common region in the RSV A dataset, and ECA (n=74) was the second most common region among the RSV B samples. SA was the least common region; only two RSV A sequences came from SA, and no RSV B sequences came from this region (see Figures 1a and 1b for the distribution of sequences by geo-region).

Figure 1. Distribution of Subsampled Sequences by Geo-Region

a) Distribution of RSV A Subsample
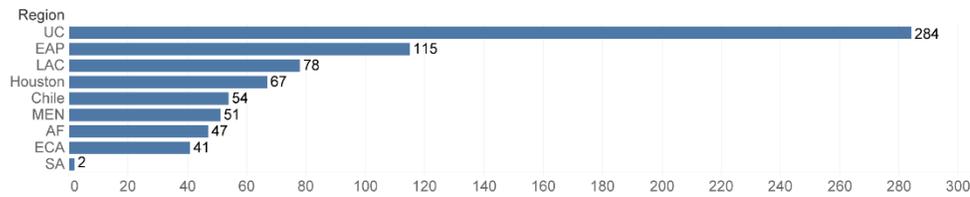
b) Distribution of RSV B Subsample
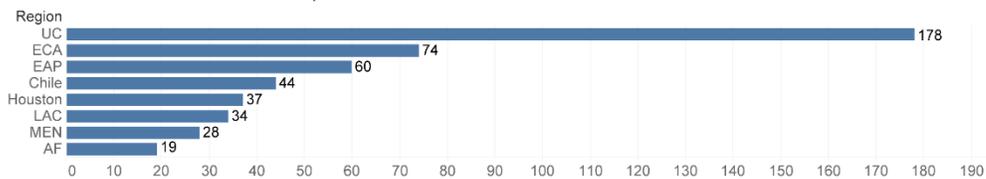
The number of RSV A (a) and B (b) sequences from each of the adapted World Bank regions included in the subsampled dataset for further analyses.

Bayesian Phylogenetic Analysis

The MCC trees for RSV A and B were generated and coded by geo-region to make inferences about global transmission dynamics. The MCC tree for RSV A (Figure 2) has two large clusters that have been co-evolving and co-circulating. One cluster is largely made up of sequences from the UC region, with some small clusters from Houston, ECA, and LAC. The other large cluster was less dominated by the UC region and contained clusters from EAP, LAC, and AF. There were also smaller clusters from the MEN, Houston, and Chile. The tree shows two major lineages with bushy leaves in the more recent time period, which indicates intense sampling efforts in more recent times. Multiple lineages co-circulate over time suggesting a higher standing genetic diversity than observed in other RNA viruses causing respiratory disease, such as Influenza A virus. The longer branches on the older part

23

of the tree indicate that there was missing information, probably due to a lack of RSV surveillance, which increases the uncertainty of the tMRCA estimation. The RSV A tMRCA for the sample of publicly available data and the clinical data from Chile and Houston was estimated to be 1963.40 (95% BCI: 1946.15, 1969.60), with a mean substitution rate of $7.43\text{x}10^{-4}$ substitutions/site/year. This is a relatively wide Bayesian credible interval, indicating the uncertainty of the estimation is increased by missing data at early stages of RSV surveillance.

Figure 2. Bayesian Maximum Clade Credibility Tree for Global RSV A
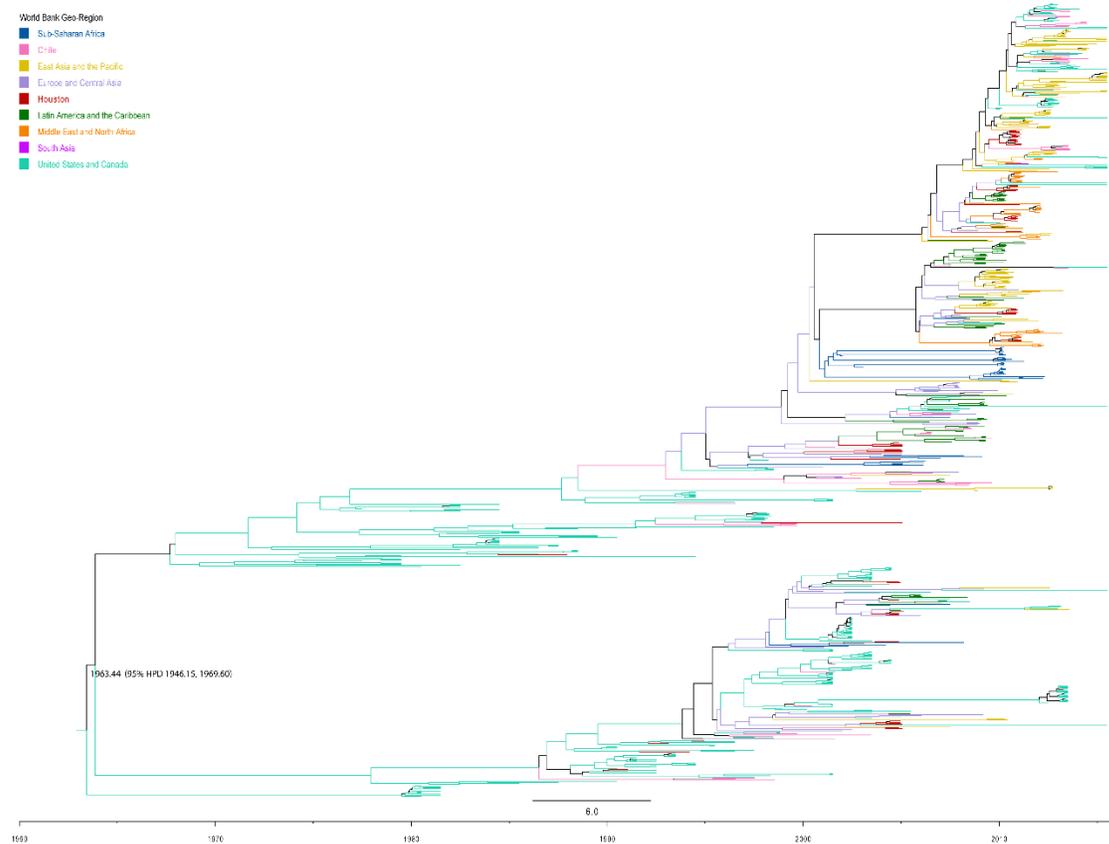
The MCC tree for RSV A color coded by geo- region. High income countries were added to the closest region, and the United States and Canada was added as another region.

The RSV B MCC tree (Figure 3) has a pronounced ladder shaped phylogeny with bushy leaves. Similar to the RSV A tree, there has been intensive surveillance in more recent times. In contrast, all co-circulating RSV B lineages have a more recent common ancestor from 2000.90 (95% BCI: 2000.88, 2001.93). The RSV B tree also has some long branches, which indicates that data is missing from the globally available dataset. The older clusters are mostly from the UC. The more recent clusters are mainly composed of strains from ECA, with smaller clusters from Chile, AF, EAP, and the UC. The RSV B tMRCA was 1963.80 (95% BCI: 1959.50, 1967.33) with a mean substitution rate of 8.34×10^{-4} substitutions/site/year. The Bayesian credible interval is relatively smaller, so this estimate has less uncertainty.

Figure 3. Bayesian Maximum Clade Credibility Tree for Global RSV B
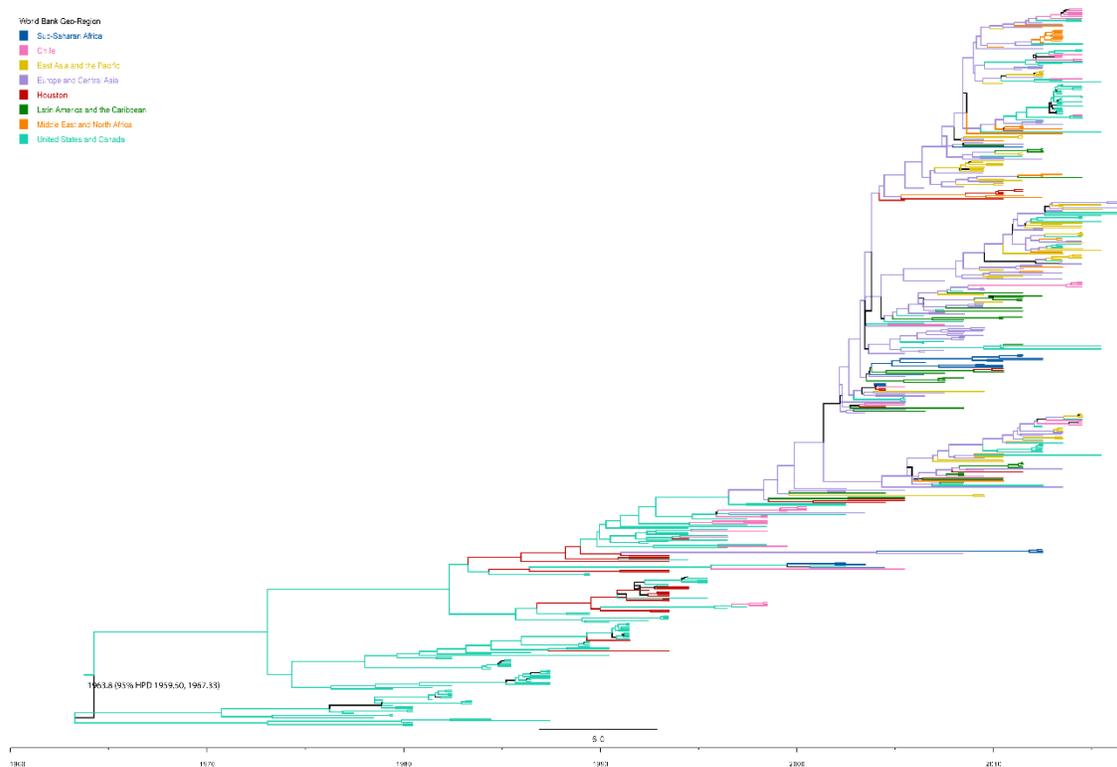
The MCC tree for RSV B color coded by geo-region. High income countries were added to the closest region, and the United States and Canada was added as another region.

RSV Global Transmission Patterns

The transmission rates and Bayes factors were generated to identify potentially important transmission routes on the global level. The key transmission routes for RSV A (Figure 4 and Table B1) were from ECA to LAC, EAP, AF, and EAP to MEN. These routes had the highest transmission rates (2.666, 2.086, 1.343, and 1.678 transitions per year, respectively) and had decisive statistical support (BF>100). ECA to the UC also had one of the highest transmission rates (2.523) but was not supported (BF<3).

Figure 4. Map of Global Transmission Rates for RSV A
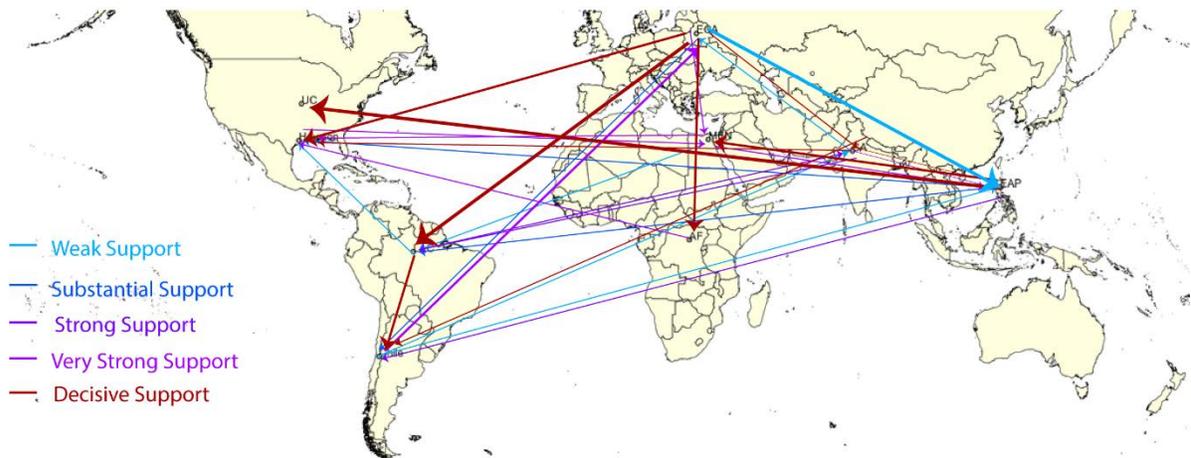
Map of global transmission patterns for RSV A. Only transmissions with BF≥3 are shown. The color of the line indicates the level of support, and the weight of the line indicates the rate of transmission with thicker lines indicating a higher rate. Arrowheads indicate the direction of transmission. UC is the United States and Canada region, LAC is Latin America and the Caribbean, ECA is Europe and Central Asia, MEN is the Middle East and North Africa, AF is Sub-Saharan Africa, SA is South Asia, and EAP is East Asia and the Pacific.

Transmission from LAC into Houston was weakly supported (3≤BF<6) and the rate was low (0.584 transitions per year). There was substantial support (6≤BF<10) for Houston

as a sink for EAP at a rate of 0.515 transitions per year. MEN and AF were strongly supported (10≤BF<30) sources of virus for Houston (0.841 and 0.425 transitions per year, respectively). Transmission routes from SA and ECA to Houston had decisive support (BF≥100), although the rate from SA (0.289 transitions per year) was much lower than the one from ECA (1.438 transitions per year). Houston was a very strongly supported (30≤BF<100) source of infection to MEN (0.787 transitions per year).

Transmission from Chile to SA and from Chile to EAP is weakly supported (3≤BF<6), and rates for these transmissions were low (0.192 and 0.767 transitions per year, respectively). ECA was a substantially supported source of infection for Chile, at a rate of 0.666 transitions per year, and EAP into Chile was strongly supported at a rate of 0.711 transitions per year. There was decisive support for SA and LAC as sources of infection for Chile. The rate of transmission from LAC into Chile was 1.500 transitions per year, the highest rate of all transmissions involving Chile, and the rate from SA was very low (0.269 transitions per year). Chile was a very strongly supported source of infection for ECA, with a rate of 1.476 transitions per year.

In RSV B (Figure 5 and table B2), transmission routes from ECA to UC and to EAP had the highest rates (2.854 and 2.815 transitions per year, respectively), and these routes had decisive support (BF>100). Routes from ECA to AF, Chile, LAC, and MEN all have decisive support as well, although the rates are lower (0.500, 1.270, 1.620, and 1.398 transitions per year, respectively). Transmission from MEN has strong support to Houston. Transmissions from LAC and ECA to Houston are very strongly supported. There is decisive support for transmission from the UC to Houston at a rate of 0.628 transitions per year and to

Chile at a rate of 1.27 transitions per year, and from Houston to the UC at a rate of 1.58

transitions per year.

Figure 5. Map of Global Transmission Rates for RSV B
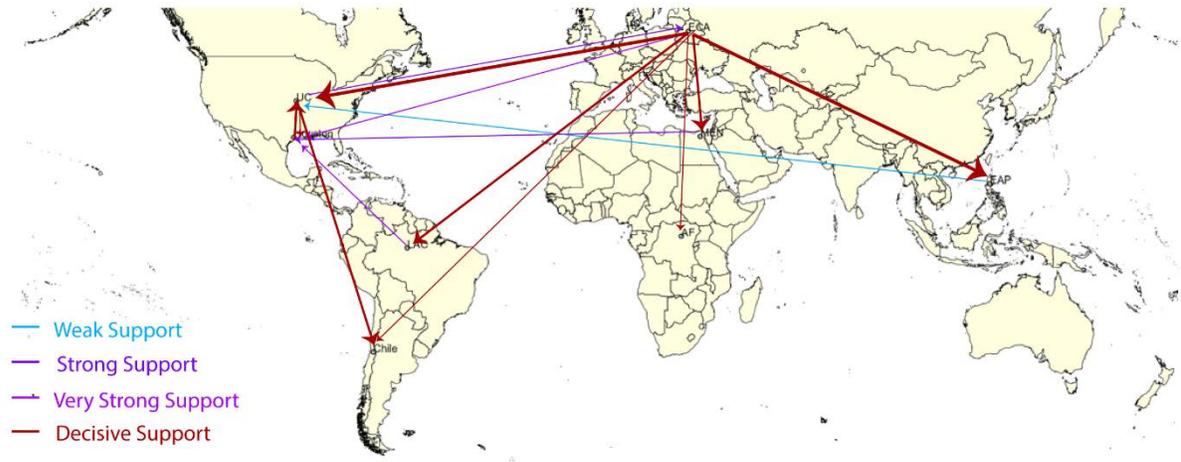
Map of global transmission patterns for RSV B. Only transmissions with BF≥3 are shown. The color of the line indicates the level of support, and the weight of the line indicates the rate of transmission with thicker lines indicating a higher rate. Arrowheads indicate the direction of transmission. UC is the United States and Canada region, LAC is Latin America and the Caribbean, ECA is Europe and Central Asia, MEN is the Middle East and North Africa, AF is Sub-Saharan Africa, and EAP is East Asia and the Pacific.

Markov Jump Counts

The complete history of the Markov jump counts was reconstructed to quantify jump

events in a source-sink model. For global RSV A (Figure 6), the average number of

transmissions between regions ("jump counts") were highest from ECA to LAC (19.90),

from ECA to the UC (18.62), and from the UC to Chile (17.94). AF, Houston, MEN, SA

were the least common sources of transmissions. ECA was the most common source but not

a major sink for any region. The most common source of transmission to Chile was the UC

region. Chile was the most common source of transmission for ECA at an average of 5.93

jumps. ECA and the UC were the most common sources of transmission for Houston, at 9.96 and 7.76 jumps, respectively. Houston was not a common source for any of the geo-regions.

Figure 6. Markov Jump Count Heat Map for RSV A

RSV A Markov Jump Counts

Source	AF	Chile	EAP	ECA	Sink Houston	LAC	MEN	SA	UC
AF		0.00	0.01	0.04	1.01*	0.01	0.01	0.00	0.02
Chile	0.01		1.23*	5.92*	0.23	0.53	0.02	0.00*	1.30
EAP	0.00	2.22*		0.04	1.12*	1.01*	9.96*	1.99*	15.84
ECA	9.22*	2.23*	15.23		9.96*	19.90	5.69*	0.00*	18.62
Houston	0.01	0.03	0.45	0.10		0.06	2.12*	0.00	0.56
LAC	0.07	6.97*	0.37	0.13	0.77*		0.09	0.00*	4.04
MEN	0.00	0.06	2.25*	0.15	2.29*	0.65*		0.03	0.16
SA	0.00	0.00*	0.00*	0.00*	0.00*	0.01*	0.02		0.02
UC	0.01	17.94	2.39	1.89	7.76	4.13	0.01	0.00	

Jump Counts: 0.00 — 19.90

* Indicates transmission route has BF>3

Darker colors indicate a higher average number of jumps from the source location (origin of transmission) to the sink location (destination of the transmission). A red star by the average number of jumps indicates that the transmission route has a BF≥3 meaning that it is supported.

For the global RSV B (Figure 7), we observed that ECA was the most common source of transmission, with the average number of jumps being highest to the UC (32.63), to EAP (32.39), and to LAC (18.16). AF, Chile, EAP, LAC, and MEN were rarely the source of transmissions. The UC was the most common sink. The UC and ECA were the most common sources of transmission for Chile, with average jump counts of 16.49 and 6.39, respectively. Chile was not a major source of transmission for any of the geo-regions. The main sources of transmission into Houston were the same as those for Chile. Houston was a source of transmission for the UC (8.95), although it was not the major source for UC.

Figure 7. Markov Jump Count Heat Map for RSV B

RSV B Markov Jump Counts

Source	AF	Chile	EAP	ECA	Sink Houston	LAC	MEN	UC
AF		0.03	0.00	0.11	0.14	0.01	0.00	0.32
Chile	0.31		0.01	0.46	0.04	0.03	0.00	0.36
EAP	0.03	0.02		0.05	0.02	0.15	0.01	1.11*
ECA	4.83*	6.39*	32.39*		6.29*	18.16*	15.64*	32.63*
Houston	0.29	0.01	0.00	0.91		0.23	0.04	8.95*
LAC	0.03	0.01	0.03	0.02	1.89*		0.02	0.02
MEN	0.01	0.01	0.01	0.01	2.13*	0.07		0.03
UC	0.56	16.49	0.16	4.08*	7.97*	0.67	0.00	

Jump Count: 0.00 — 32.63

* Indicates transmission route has BF>3

Darker colors indicate a higher average number of jumps from the source location (origin of transmission) to the sink location (destination of the transmission). A red star by the average number of jumps indicates that the transmission route has a BF≥3 meaning that it is supported.

Clinical Severity Analyses

Clinical Data Description

Clinical data from Chile and Houston was used to explore associations between viral genetic characteristics and disease severity. After cleaning the data, for RSV A, 60 taxa with clinical data from Chile and Houston were used. Twenty-three of the sequences were from Chile and 37 were from Houston. For RSV B severity, 41 taxa were from Chile and Houston. Thirty-one of the sequences were from Chile and 10 were from Houston. Detailed distribution and disease severity traits are in Table 1.

Table 1. Number of Cases by Disease Severity Traits for RSV A and RSV B

Disease Severity Traits	Number of RSV A Cases	Number of RSV B Cases
Oxygen, Hospitalized, Admitted to ICU	7	4
Oxygen, Hospitalized, Not Admitted to ICU	24	30
No Oxygen, Hospitalized, Admitted to ICU	2	0
No Oxygen, Hospitalized, Not Admitted to ICU	11	4
No Oxygen, Not Hospitalized, Not Admitted to ICU	16	3

Bayesian Phylogenetic Analysis of Clinical Data

In order to analyze the association between topologies and clinical traits in BaTS to identify clustering patterns, MCC trees were generated using BEAST. The trees for the RSV A and RSV B (Figures 8 and 9) clinical data do not exhibit the same ladder shape as the global trees, but both have bushy leaves, indicating intensive sampling for outbreaks at each location. Both trees have long branches, which indicates that missing data exist. The clades

30

of the RSV A tree are closely tied to the geographic origin, with two clades from Houston and one clade from Chile. The geographic origins are more interspersed in the RSV B tree, but there are very few sequences from Houston, where the unrepresented samples may explain the mixing. The RSV A clade from Chile is dominated by the "oxygen, hospitalized, not admitted to the ICU" severity traits. While the clades from Houston have a mixture of severity traits, with the "no oxygen, not hospitalized, not admitted to the ICU" trait making up the majority of cases. The estimated tMRCA for the clinical RSV A tree was 2007.78 (95% BCI: 2006.54, 2009.05) with a mean substitution rate 1.493×10^{-3} substitutions/site/year. The Bayesian credible interval is relatively narrow, so the tMRCA estimate is more certain.

Figure 8. Clinical Maximum Clade Credibility Trees for RSV A
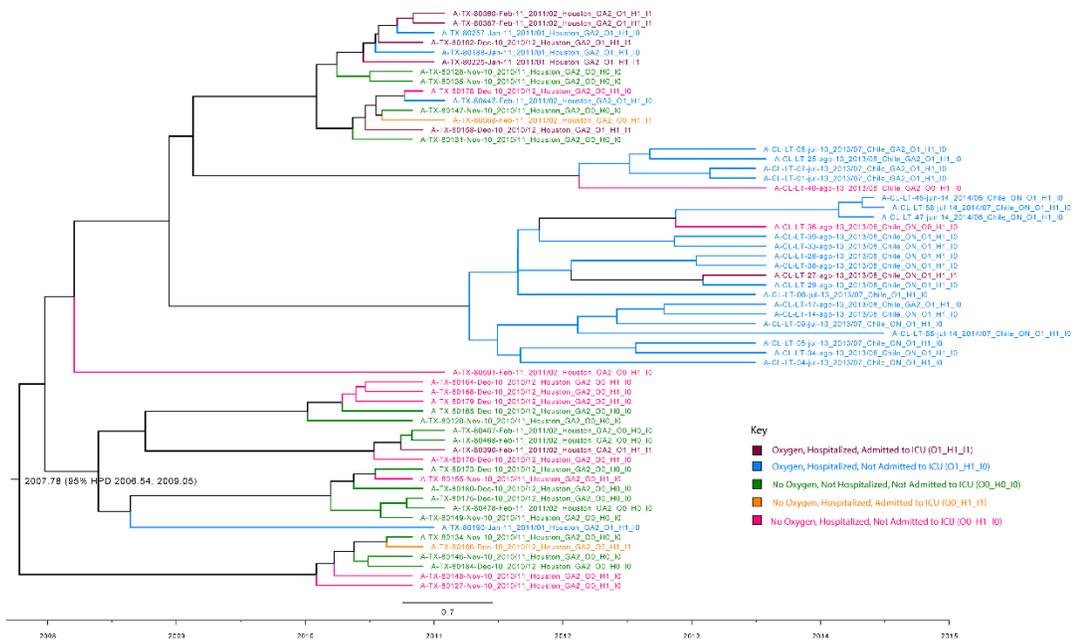
The MCC tree for RSV A from Chile and Houston color coded by severity traits: whether the patient was given oxygen therapy, whether they were hospitalized, and whether they were admitted to the intensive care unit.

Figure 9. Clinical Maximum Clade Credibility Trees for RSV B
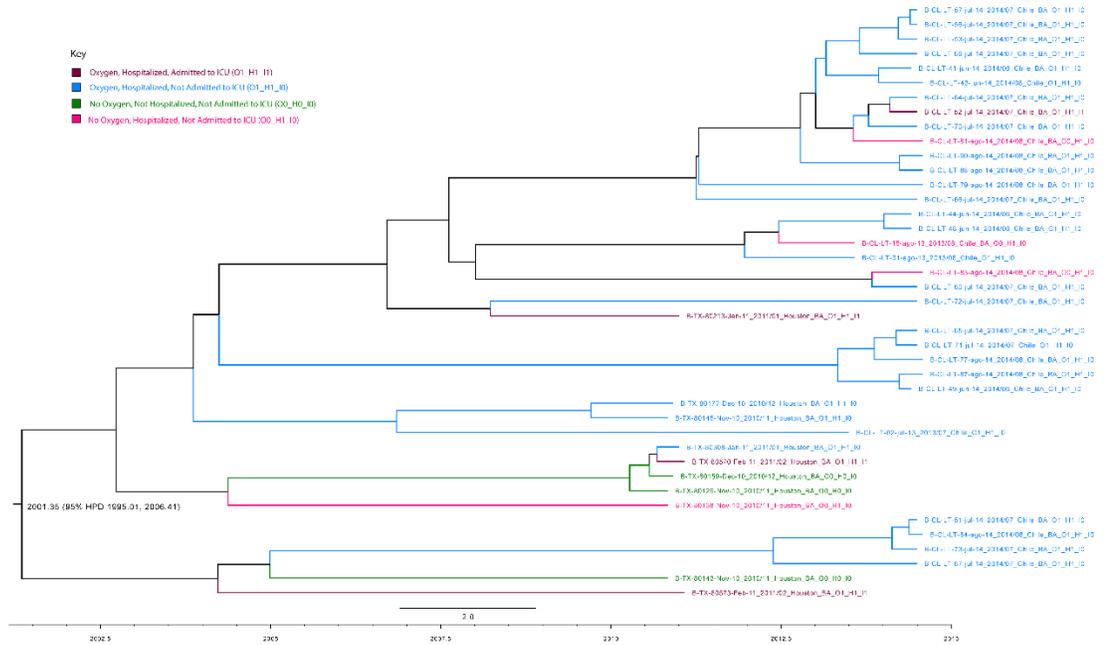
The MCC tree for RSV B from Chile and Houston color coded by severity traits: whether the patient was given oxygen therapy, whether they were hospitalized, and whether they were admitted to the intensive care unit.

For the RSV B clinical data the tMRCA was estimated to be 2001.35 (95% BCI: 1995.01, 2006.41) with a mean substitution rate of 9.562×10^{-4} substitutions/site/year. Unlike the RSV A Bayesian credible interval, the RSV B interval is quite wide so there is a large degree of uncertainty in the tMRCA estimate, which is probably due to missing data. The vast majority of sequences in the RSV B had the "oxygen, hospitalized, not admitted to the ICU" trait and they tended to be within the same cluster. The other severity traits were rare and were interspersed throughout the clades of the tree.

Clustering Analysis of Clinical Severity

Using BaTS, an analysis of the association between disease severity traits and viral genetic characteristics was conducted. For the RSV A samples, the BaTS analysis yielded a

32

significant association between the clinical traits and the phylogeny of the tree when all three

traits were combined (needed oxygen, hospitalized, and admitted to the ICU). The cluster

analysis yielded significant results when oxygen therapy and hospitalization were analyzed

separately. There were no significant associations between clinical traits and the tree

phylogeny for ICU admittance in RSV A. The RSV A AI and PS statistics, which assess all

traits, were significant (Table 2). The significant AI statistic indicates that the traits are

consistent across the internal nodes of the tree, and the significant PS indicates that the

number of trait changes through the ancestral nodes is small. However, not all of the MC

statistics, where each variant of the trait is assessed individually, were significant. The MC

statistic was significant for the "oxygen, hospitalized, not admitted to the ICU" trait and the

"no oxygen, not hospitalized, not admitted to the ICU", indicating that these traits have larger

monophyletic clades and therefore are associated with the phylogeny of the tree but the other

traits are not.

Table 2. Clustering of Disease Severity Traits for RSV A

RSV A Combined Clinical Traits		
Statistic	Observed Mean	Significance
AI	3.127	<0.01*
PS	22.392	<0.01*
MC (Oxygen, Hospitalized, No ICU)	5.593	0.02*
MC (Oxygen, Hospitalized, ICU)	1.601	1
MC (No Oxygen, Hospitalized, No ICU)	1.679	0.13
MC (No Oxygen, Not Hospitalized, No ICU)	3.444	0.01*
MC (Oxygen, Hospitalized, ICU)	1	1
Oxygen Therapy		
AI	0.983	0.00*
PS	9.318	0.00*
MC (Oxygen)	6.259	0.02*
MC (No Oxygen)	7.464	0.03*

Hospitalized		
AI	1.917	0.01*
PS	11.784	0.01*
MC (Hospitalized)	19.210	0.01*
MC (Not Hospitalized)	3.444	0.03*
Admitted to ICU		
AI	1.567	0.32
PS	8.116	0.13
MC (No ICU)	8.985	0.47
MC (ICU)	1.651	0.05
*indicates the null hypothesis can be rejected at a significance level of 0.05.		

For RSV B, no significant associations between clinical traits and phylogeny were found when analyzing all three traits together and each independently (Table 3).

Table 3. Clustering of Disease Severity Traits for RSV B

RSV B Combined Clinical Traits		
Statistic	Observed Mean	Significance
AI	1.613	0.06
PS	10.722	1
MC (Oxygen, Hospitalized, No ICU)	5.795	0.46
MC (No Oxygen, Hospitalized, No ICU)	1	1
MC (Oxygen, Hospitalized, ICU)	1	1
MC (No Oxygen, Not Hospitalized, No ICU)	1.154	1
Oxygen Therapy		
AI	1.038	0.16
PS	6.735	1
MC (Oxygen)	6.473	0.58
MC (No Oxygen)	1.154	1
Hospitalized		
AI	1.038	0.18
PS	6.735	1
MC (Hospitalized)	6.473	0.57
MC (Not Hospitalized)	1.154	1
Admitted to ICU		
AI	0.819	0.49
PS	4	1

MC (No ICU)	6.327	0.98
MC (ICU)	1	1
*indicates the null hypothesis can be rejected at a significance level of 0.05		

DISCUSSION

Although RSV usually presents with mild, flu-like symptoms, it can cause severe disease outcomes for infants, especially premature infants, and adults over the age of 65. In the absence of an FDA-approved vaccine that is widely available, outbreak prevention and control relies on a firm understanding of transmission patterns and dynamics. Since vaccine candidates and antiviral treatment are being developed, it is important to identify viral genetic characteristics that can predict disease severity and be better targets. The aim of this study was to use RSV full-genome sequence data to gain a better understanding of the global transmission patterns, the transmission of RSV into and out of Houston and Chile and to preliminarily test if there are viral characteristics within the F, G, and SH genes associated with higher severity of disease. To our knowledge, this is the first study to use full genome data to model and identify RSV global transmission patterns. While other studies have implied that global transmission dynamics are important for regional and local outbreaks, very few have tried to elucidate those dynamics, and those studies have relied on solely on the short sequences of the G gene (Hibino, et al., 2018; Giallonardo, et al., 2018). This is also the first effort to link characteristics of the three surface proteins to patient outcomes, as previous studies have focused on RSV genotypes or patient characteristics (Hibino, et al., 2016; Yoshihara, et al., 2016; Panayiotou, et al., 2014).

With the estimation from phylogenetic modeling, we estimated that the tMRCA of RSV A was 1963.40 and the mean substitution rate was 7.43×10^{-4} substitutions/site/year. Zou and colleagues reported the tMRCA for the GA2 and GA7 genotypes, which they estimated to be 1965 (95% HPD: 1970-1975) with a substitution rate of 2.3×10^{-3} (Zou, et al., 2016). The phylogenetic tree generated by Zou and colleagues appears to have a similar shape to the tree we generated for RSV A. The tree shape for RSV A is also similar to the phylogenetic trees generated in the Elawar, Schobel, and Agoti studies (Elawar, et al., 2017; Schobel, et al., 2016; Agoti, et al., 2017). The tMRCA from the Zou tree appears to be around 1965, which is within our 95% BCI (1946.15, 1969.60). The huge differences in the substitution rates may be due to the differences in the data examined. Zou and colleagues just looked at G gene data, which is shorter and includes a hypervariable region, while our study looked at all the currently available full genome data, which may result in a more accurate substitution rate over the full length of the genome. Agoti and colleagues found a substitution rate of 4.95×10^{-3} in their study of household transmission dynamics when looking at full genome data (Agoti, et al., 2017). A study conducted by Giallonardo and colleagues (2018) also looked at full sequence data, and their estimate of the substitution rate using BEAST was 7.48×10^{-4} substitutions/site/year for RSV A and 7.39×10^{-4} substitutions/site/year for RSV B. Their estimate for RSV A is very similar to the one we have reported here. Since few studies have focused on RSV B, this study, in addition to the Giallonardo study, provides preliminary estimates of the evolutionary dynamics of RSV B with tMRCA estimated as 1963.80 and the mean substitution rate as 8.34×10^{-4} substitutions/site/year.

The results from the global transmission modeling showed that for both RSV A and B, ECA was an important source of transmission events for many regions, including Houston and Chile. Although the rate of transmissions from ECA to UC was high for RSV A, it was not found to be significant for RSV A. This may be due to the scarce samples from ECA over years, resulting in insufficient statistical power to detect the significance, though the transmission rates showed the potential close relationships between these two locations. For Houston, LAC and MEN were also important sources of RSV A and B transmission. A potential mechanism for RSV global introductions is air travel, which provides pathogens a quick way to reach all corners of the globe. RSV is mainly transmitted through fomites and droplets (Kutter, et al., 2018). Although RSV does not spread as efficiently as an aerosol (e.g., influenza), transmission could still occur in a small, enclosed space, such as an airplane. In 2017, 54.2 million people flew into Houston area airports, and of these passengers, 11.2 million came from international airports (Houston Airport System, 2018). Houston is the fourth largest city in the United States and is a key business center for the global oil and gas industry. At Nuevo Paudahuel, the main airport for Santiago, Chile, close to 50% of the 1.23 million passengers are from international origins (Aeropuerto de Santiago, 2015). The large number of people travelling from around the globe facilitates many potential RSV introductions.

Studies of influenza have indicated that air travel patterns paired with information on seasonal peaks of influenza outbreak help explain its global transmission patterns (Kenah, et al., 2011). A similar phenomenon could be happening with RSV. The peak of RSV in Texas is usually in December or January, with the number of cases starting to increase in September

or October (Texas Department of State Health Services, 2018). This peak season overlaps with peaks in the ECA and MEN regions, which may explain why these transmission routes are supported (Hendaus, et al., 2018; Sricharoenchai, Palla, & Sanicas, 2016). The peak in LAC varies since the region includes countries in the northern and southern hemisphere with both temperate and tropical climates. Within the region, RSV peaks in Mexico would overlap with the peaks in Houston, but the peaks for Peru, Chile, and Argentina would occur in June, July, and August (Rodriguez-Auad, et al., 2012; Sricharoenchai, Palla, & Sanicas, 2016), while the tropical locations in LAC with year-round outbreak may facilitate the seasonality of both hemispheres. Peaks in Brazil occur over a wide range of time periods that correspond to the rainy season, with slight overlap with the peaks in Texas (Sricharoenchai, Palla, & Sanicas, 2016). The peak season overlap does not fully explain transmission from ECA to Chile since they have opposite peak seasons, but the intermediate links in the tropical areas may contribute the transmission between LAC and Houston. Further studies are needed to investigate this connection.

Many studies have taken a genotype-specific approach to RSV phylogeny and focus on the evolution of each genotype. Zou and colleagues were one of the first to try to determine the global transmission of RSV A, but they did not investigate transmission routes (Zou, et al., 2016). Their analysis from genotypes GA2 and GA7 supported the potential of substantial international movement of RSV lineages although it was not detected in their current dataset. Our results of RSV global transmission patterns, which focus on the evolution of all genotypes, support the idea from Zou and Elawar, as well as studies of local transmissions (Zou, et al., 2016; Elawar, et al., 2017). Therefore, our study with a more

complete RSV global dataset adds more evidence to the hypothesis that international transmission events are key to RSV transmission patterns as many routes are statistically supported in our study.

With the clinical disease severity data from Houston and Chile, we identified a significant association between the RSV A tree topologies and disease severity. We did not identify any significant associations in RSV B, which may be due to the small sample size of the clinical dataset. It is important to note that the clinical RSV dataset only includes the main circulating genotypes (two RSV A genotypes: GA2 and ON, and one RSV B genotype: BA), so it may not represent the complete picture of RSV epidemics. Although we did not analyze the data by genotype, the clusters associated with more severe outcomes are in the ON clade for RSV A. This association between the ON genotype and increased severity was also reported in the Hibino and Yoshihara studies, although both of these studies compared ON to NA1, and not to GA2. Since we had only one genotype of RSV B, we cannot compare it to the studies that took a genotypic approach.

Our preliminary analysis on disease severity provides some insights on data collection and future directions. It would be ideal to collect a larger number of samples for both RSVA and B from patients on both ends of the severity spectrum, those who did not go to the hospital and those who had to be admitted to the ICU. It would also be better to have multiple genotypes represented within the sample for easier comparison to previous studies that related RSV genotype to disease severity. In our analysis, we did not include data on patient characteristics and other confounding variables that could impact outcome severity (e.g., patient age or whether they have co-morbidities) or the patients' viral loads. While this

preliminary analysis indicates that there are genetic characteristics associated with disease severity, further analyses should take viral load and patient characteristics into account to control for these potential confounders. Further studies should also conduct a scan of the genomes for clades significantly associated with severe outcomes to determine what variations are associated with more severe outcomes. Identifying critical mutations that lead to more severe outcomes, and determining how these outcomes are correlated with patient genetics, can help inform antiviral medication and vaccine development.

This study has some limitations. First, the global data have obvious sampling biases. The United States and Canada (UC) region are heavily overrepresented in the dataset, while other regions may only have very scarce samples over countries or over years. For example, data from Latin American and the Caribbean (LAC) and Sub-Saharan Africa (AF) represent only one or two countries within that region, data from Europe and Central Asia (ECA) mainly have samples only for a few years, and South Asia (SA) is not represented at all in the RSV B dataset. The US in general is heavily overrepresented across both datasets and across all years, although a large portion of the data from the US comes for a few cohort studies conducted in localized areas, and so may not be representative of the US as a whole. To reduce the selection bias in the data, we have subsampled the data by country and year and conducted sensitivity analysis with multiple subsampled datasets. Another key limitation is related to using geo-regions to infer transmission dynamics. World Bank regions may be represented by only one or two countries within that region, so the data may not reflect the true dynamics of transmission; that is, we cannot determine if the routes are direct, neither can we determine the intermediate stops within each of those regions. However, inferences

drawn from using geo-regions may help us identify key regions where surveillance efforts need to be increased. Furthermore, some of the disease severity data may not be accurate. For example, it is unlikely that someone was admitted to the hospital and the ICU for a respiratory infection but was not given oxygen, but a few of these cases were reported. Without direct access to patient records, it is impossible to verify the medical treatments required by each patient. Another limitation of the severity data is the small sample size for RSV B, which means we are likely underpowered to detect an association between disease severity and tree topology if an association does exist. Further studies, with more data are needed to determine if the associations for RSV B are similar to RSV A.

Better surveillance systems to routinely collect global representative samples are needed in many parts of the world in order to gain an unbiased understanding of global transmission dynamics. A system proposed by the WHO would help collect more complete datasets that are more representative of the viral variation around the world (WHO, 2017). This in turn would lead to more robust inferences on transmission dynamics to determine the importance of within region persistence and global transmission routes. Further studies of global transmission dynamics with a more representative dataset could be improved with more sophisticated population coalescent models. For disease severity, studies need to be conducted with sequence scans to identify key mutations associated with increased severity.

CONCLUSION

In summary, though data limitations exist, with rigorous subsampling strategy, phylodynamic modeling and clustering analysis, we were able to depict the global

transmission patterns of RSV. The ECA region was an important source of RSV A and B

transmission for Chile and Houston, so increased surveillance in this region may help predict

the main genotype of seasonal outbreaks in Chile and Houston, and eventually help guide

vaccine design. Even before a vaccine is available, information about transmission dynamics

can be used to help control the spread of the disease. There was an association between tree

topology and disease severity for RSV A. More data and further studies are needed to

identify which genetic variants are associated with this increased severity.

APPENDICES

Appendix A: Distribution of Global RSV Samples

Figure A1. Distribution of Sequences by Geo-Region for RSV A(a) and RSV B(b)

a) RSV A Sample Distribution by Geo Region
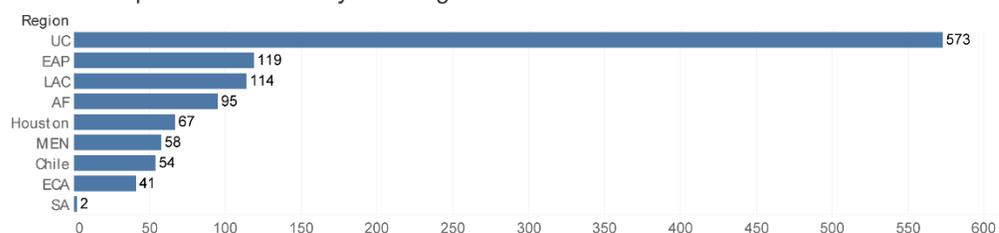
b) RSV B Sample Distribution by Geo Region
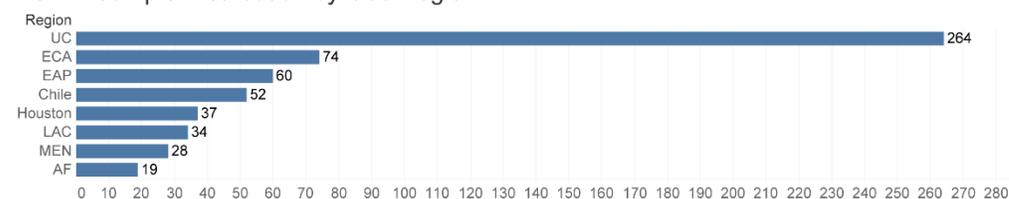
The number of RSV A (a) and B (b) sequences from each of the adapted World Bank regions included in the full dataset.

Appendix B: Transmission Rates for RSV and Level of Support

Table B1. Rates and Level of Support for RSV A Transmission

From	To	Mean Rate*	Median Rate	HPD lower	HPD upper	Bayes Factor
Europe & Central Asia	Latin America & the Caribbean	2.666	2.554	1.211	4.362	>100
East Asia & the Pacific	United States & Canada	2.556	2.460	1.061	4.399	>100
Europe & Central Asia	East Asia & the Pacific	2.086	1.983	0.644	3.486	>100
East Asia & the Pacific	Middle East & North Africa	1.678	1.578	0.567	3.036	>100
Latin America & the Caribbean	Chile	1.500	1.387	0.345	2.860	>100
Chile	Europe & Central Asia	1.476	1.395	0.072	2.915	38.003
Europe & Central Asia	Houston	1.438	1.366	0.427	2.597	>100

Europe & Central Asia	Sub-Saharan Africa	1.343	1.276	0.473	2.319	>100
Europe & Central Asia	Middle East & North Africa	0.972	0.915	0.110	1.878	53.576
Middle East & North Africa	Houston	0.841	0.746	0.044	1.786	34.560
Middle East & North Africa	East Asia & the Pacific	0.832	0.708	0.051	1.877	71.613
Houston	Middle East & North Africa	0.787	0.651	0.005	1.896	35.657
Chile	East Asia & the Pacific	0.767	0.679	0.008	1.736	5.120
East Asia & the Pacific	Chile	0.711	0.623	0.054	1.570	13.105
Europe & Central Asia	Chile	0.666	0.599	0.026	1.473	8.213
Middle East & North Africa	Latin America & the Caribbean	0.616	0.510	0.007	1.556	5.250
Latin America & the Caribbean	Houston	0.584	0.485	0.011	1.487	4.499
East Asia & the Pacific	Houston	0.515	0.425	0.019	1.252	7.493
East Asia & the Pacific	Latin America & the Caribbean	0.497	0.420	0.010	1.178	6.512
East Asia & the Pacific	South Asia	0.468	0.405	0.034	1.063	>100
Sub-Saharan Africa	Houston	0.425	0.338	0.011	1.034	75.713
South Asia	East Asia & the Pacific	0.345	0.230	0.001	0.945	10.678
South Asia	Latin America & the Caribbean	0.296	0.184	0.000	0.933	23.900
South Asia	Houston	0.289	0.202	0.000	0.858	>100
South Asia	Chile	0.269	0.201	0.000	0.809	>100
South Asia	Europe & Central Asia	0.256	0.187	0.001	0.746	5.626
Latin America & the Caribbean	South Asia	0.192	0.116	0.002	0.827	18.069
Chile	South Asia	0.192	0.155	0.001	0.710	3.867
Europe & Central Asia	South Asia	0.187	0.174	0.031	0.437	>100
*Ordered by the value of rates.						

Table B2. Rates and Level of Support for RSV B Transmission

From	To	Mean Rate*	Median rate	HPD lower	HPD upper	BAYES_FACTOR
Europe & Central Asia	United States & Canada	2.854	2.745	1.272	4.618	>100
Europe & Central Asia	East Asia & the Pacific	2.815	2.723	1.204	4.547	>100
Europe & Central Asia	Latin America & the Caribbean	1.620	1.553	0.619	2.685	>100
Houston	United States & Canada	1.584	1.484	0.343	2.979	>100
Europe & Central Asia	Middle East & North Africa	1.398	1.323	0.515	2.439	>100
United States & Canada	Chile	1.270	1.195	0.452	2.262	>100
Middle East & North Africa	Houston	0.751	0.661	0.017	1.608	23.435
Europe & Central Asia	Houston	0.652	0.586	0.045	1.353	89.947
East Asia & the Pacific	United States & Canada	0.642	0.538	0.000	1.618	3.947
Europe & Central Asia	Chile	0.632	0.592	0.092	1.260	>100
United States & Canada	Houston	0.628	0.562	0.107	1.324	>100
Latin America & the Caribbean	Houston	0.571	0.491	0.029	1.321	41.055
Europe & Central Asia	Sub-Saharan Africa	0.500	0.452	0.072	1.023	>100
United States & Canada	Europe & Central Asia	0.454	0.408	0.050	0.950	21.471
*Ordered by the value of rates.						

REFERENCES

Aeropuerto de Santiago. (2015). Passenger Traffic Statistics. Retrieved from:

https://www.nuevopudahuel.cl/passenger_traffic_statistics?language=en

Amand, C., Tong, S., Kieffer, A., Kyaw, M.H. (2018). Healthcare resource use and economic

burden attributable to respiratory syncytial virus in the United States: a claims

database analysis. BMC Health Services Research, 18:294.

Bahl, J., Krauss, S., Kühnert, D., Fourment, M., Raven, G., et al. (2013). Influenza A Virus

Migration and Persistence in North American Wild Birds. *PLOS Pathogens 9*(8):

e1003570. https://doi.org/10.1371/journal.ppat.1003570

Bielejec, F., Baele, G., Vrancken, B., Suchard M.A., Rambaut, A., Lemey, P. (2016).

SpreaD3: interactive visualisation of spatiotemporal history and trait evolutionary

processes. *Molecular Biology & Evolution, 33*(8), 2167-2169.

Bont, L., Checchia, P. A., Fauroux, B., Figueras-Aloy, J., Manzoni, P., Paes, B., …

Carbonell-Estrany, X. (2016). Defining the Epidemiology and Burden of Severe

Respiratory Syncytial Virus Infection Among Infants and Children in Western

Countries. *Infectious Diseases and Therapy*, *5*(3), 271–298.

http://doi.org/10.1007/s40121-016-0123-0

Borchers, A.T., Chang, C., Gershwin, M.E., Gershwin, L.J. (2013). Respiratory syncytial virus—a comprehensive review. Clinical Reviews in Allergy & Immunology, 45(3), 331-79. http://doi.org/10.1007/s12016-013-8368-9

Bose, M. E., He, J., Shrivastava, S., Nelson, M. I., Bera, J., Halpin, R. A., … Henrickson, K. J. (2015). Sequencing and Analysis of Globally Obtained Human Respiratory Syncytial Virus A and B Genomes. *PLoS ONE*, *10*(3), e0120098. http://doi.org/10.1371/journal.pone.0120098

Breese Hall, C., Weinberg, G.A., Iwane, M.K., Blumkin, A.K., Edwards, K.M., Staat, M.A., Auinger, P., Griffin, M.R… (2009). The Burden of Respiratory Syncytial Virus Infection in Young Children. *New England Journal of Medicine, 360,* 588-98. http://doi.org/10.1056/NEJMoa0804877

Centers for Disease Control and Prevention. (2017). Respiratory Syncytial Virus Infection: RSV Transmission. Retrieved from https://www.cdc.gov/rsv/about/transmission.html

Chanock, R.M., Parrott, R.H., Vargosko, A.J., Kapikian, A.Z., Knight, V., & Johnson, K.M. (1962). Respiratory Syncytial Virus. *American Journal of Public Health, 52*(6), 918-25.

Cui, G., Qian, Y., Zhu, R., Deng, J., Zhao, L., Sun, Y., & Wang, F. (2013). Emerging human respiratory syncytial virus genotype ON1 found in infants with pneumonia in Beijing, China. *Emerging Microbes & Infections, 2*, e22. http://doi.org/10.1038/emi.2013.19

47

de-Paris, F., Beck, C., de Souza Nunes, L., Machado, A. B. M. P., Paiva, R. M., da Silva Menezes, D., … Barth, A. L. (2014). Evaluation of respiratory syncytial virus group A and B genotypes among nosocomial and community-acquired pediatric infections in southern Brazil. *Virology Journal*, *11*, 36. http://doi.org/10.1186/1743-422X-11-36

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, *161*(3), 1307–1320.

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. http://doi.org/10.1093/molbev/mss075

Drummond, A.J., Ho, S.Y.W., & Phillips, M.J., Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. PLOS Biology 4(5): e88. https://doi.org/10.1371/journal.pbio.0040088

Duvvuri, V. R., Granados, A., Rosenfeld, P., Bahl, J., Eshaghi, A., & Gubbay, J. B. (2015). Genetic diversity and evolutionary insights of respiratory syncytial virus A ON1 genotype: global and local transmission dynamics. *Scientific Reports*, *5*, 14268. http://doi.org/10.1038/srep14268

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high

throughput. *Nucleic Acids Research, 32*(5), 1792–1797.

http://doi.org/10.1093/nar/gkh340

Espinosa, Y., San Martín, C., Torres, A. A., Farfán, M. J., Torres, J. P., Avadhanula, V., …

Tapia, L. I. (2017). Genomic Loads and Genotypes of Respiratory Syncytial Virus:

Viral Factors during Lower Respiratory Tract Infection in Chilean Hospitalized

Infants. *International Journal of Molecular Sciences*, *18*(3), 654.

http://doi.org/10.3390/ijms18030654

Esposito S, Piralla A, Zampiero A, Bianchini S, Di Pietro G, et al. (2015) Characteristics and

Their Clinical Relevance of Respiratory Syncytial Virus Types and Genotypes

Circulating in Northern Italy in Five Consecutive Winter Seasons. PLOS ONE, 10(6):

e0129369. https://doi.org/10.1371/journal.pone.0129369

Falsey, A.R., Hennessey, P.A., Formica, M.A., Cox, C., & Walsh, E.E. (2005). Respiratory

syncytial virus infection in elderly and high-risk adults. *New England Journal of

Medicine, 352*(17), 1749-59.

Giallonardo, F.D., Kok, J., Fernandez, M., Carter, I., Geoghegan, J.L., Dwyer, D.E., Holmes,

E.C., Eden, J.S. (2018). Evolution of Human Respiratory Syncytial Virus (RSV) over

Multiple Seasons in New South Wales, Australia. Viruses, 10(476).

Hall, T. (2005). BioEdit: Biological sequence alignment editor for Win95/98/NT/2K/XP. Retrieved from http://www.mbio.ncsu.edu/BioEdit/bioedit.html

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22(2), 160-74.

Hause, A. M., Henke, D. M., Avadhanula, V., Shaw, C. A., Tapia, L. I., & Piedra, P. A. (2017). Sequence variability of the respiratory syncytial virus (RSV) fusion gene among contemporary and historical genotypes of RSV/A and RSV/B. *PLoS ONE*, *12*(4), e0175792. http://doi.org/10.1371/journal.pone.0175792

Hendaus, M.A., Alhammadi, A.H., Chandra, P., Muneer, E., & Khalifa, M.S. (2018). Indentifying agents triggering bronchiolitis in the State of Qatar. *International Journal of General Medicine,* 11, 143-149.

Hibino, A., Saito, R., Taniguchi, K., Zaraket, H., Shobugawa, Y., Matsui, T., … for the Japanese HRSV Collaborative Study Group. (2018). Molecular epidemiology of human respiratory syncytial virus among children in Japan during three seasons and hospitalization risk of genotype ON1. *PLoS ONE*, *13*(1), e0192085. http://doi.org/10.1371/journal.pone.0192085

Houston Airport System. (2018). Statistical Report: 2017 Calendar Year Summary. Retrieved

from: https://d14ik00wldmhq.cloudfront.net/media/filer_public/6f/28/6f28c41d-b124-

482e-bc55-f672475a4d4c/cy17_report_final.pdf

Kass, R.E. & Raftery A.E. (1995). Bayes factors. *Journal of the American Statistical

Association, 90*(430):773–95.

Kenah, E., Chao, D.L., Matrajt, L., Halloran, M.E., & Longini Jr., I.M. (2011). The global

transmission and control of influenza. *PLoS One, 6*(5), e19515.

Kutter, J.S., Spronken, M.I., Fraaij, P.L., Fouchier, R.A.M., Herfst, S. (2018). Transmission

routes of respiratory viruses among humans. *Current Opinion in Virology, 28,* 142-

151.

Lemey, P., Rambaut, A., Drummond, A.J., & Suchard, M.A. (2009). Bayesian

Phylogeography Finds Its Roots. *PLOS Computational Biology, 5(9*): e1000520.

https://doi.org/10.1371/journal.pcbi.1000520

Levitz, R., Wattier, R., Phillips, P., Solomon, A., Lawler, J., Lazar, I., Weibel, C., & Kahn,

J.S. (2012). Induction of IL-6 and CCL5 (RANTES) in human respiratory epithelial

(A549) cells by clinical isolates of respiratory syncytial virus is strain specific.

Virology Journal, 9(190). https://doi.org/10.1186/1743-422X-9-190

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J.,

    Adair, T… Global and regional mortality from 235 causes of death for 20 age groups

    in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.

    Lancet, 380(9859), 2095-128. http://doi.org/10.1016/S0140-6736(12)61728-0

Luchsinger, V., Ampuero, S., Palomino, M.A., Chnaiderman, J., Levican, J., Gaggero, A., &

    Larranaga, C.E. (2014). Comparison of virological profiles of respiratory syncytial

    virus and rhinovirus in acute lower tract respiratory infections in very young Chilean

    infants, according to their clinical outcome. Journal of Clinical Virology, 61(1), 138-

    44, http://doi.org/10.1016/j.jcv.2014.06.004

McLellan, J. S., Ray, W. C., & Peeples, M. E. (2013). Structure and Function of RSV

    Surface Glycoproteins. *Current Topics in Microbiology and Immunology*, *372*, 83–

    104. http://doi.org/10.1007/978-3-642-38919-1_4

Minin, V. N., Bloomquist, E. W., & Suchard, M. A. (2008). Smooth Skyride through a

    Rough Skyline: Bayesian Coalescent-Based Inference of Population

    Dynamics. *Molecular Biology and Evolution*, *25*(7), 1459–1471.

    http://doi.org/10.1093/molbev/msn090

Minin, V.N. & Suchard, M.A. (2008). Counting labeled transitions in continuous-time

    Markov models of evolution. *Journal of Mathematical Biology, 56*(3), 391-412.

Ogra, P.L. (2004). Respiratory syncytial virus: The virus, the disease and the immune response. *Paediatric Respiratory Reviews, 5*(Suppl A), S119-26.

Panayiotou, C., Richter, J., Koliou, M., Kalogirou, N., Georgiou, E., & Christodoulou, C. (2014). Epidemiology of respiratory syncytial virus in children in Cyprus during three consecutive winter seasons (2010-2013): age distribution, seasonality and association between prevalent genotypes and disease severity. *Epidemiology and Infection, 142*(11), 2406-11.

Parker, J., Rambaut, A. & Pybus, O.G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infection, Genetics, and Evolution, 8*(3), 239-46.

Pickles, R. J., & DeVincenzo, J. (2015). RSV and its propensity for causing bronchiolitis. *The Journal of Pathology*, *235*(2), 266–276. http://doi.org/10.1002/path.4462

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., & Suchard, M.A. (2018). Tracer v1.7, Available from http://beast.community/tracer

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*, *2*(1), vew007. http://doi.org/10.1093/ve/vew007

Rodriguez-Auad, J., Nava-Frias, M., Casasola-Flores, J., Johnson, K.M., Nava-Ruiz, A., Perez-Robles, V., & Caniza, M.A. (2012). The epidemiology and clinical characteristics of respiratory syncytial virus in children at a public pediatric referral hospital in Mexico. *International Journal of Infectious Diseases, 16(*7), e508-513.

Schobel, S. A., Stucker, K. M., Moore, M. L., Anderson, L. J., Larkin, E. K., Shankar, J., … Das, S. R. (2016). Respiratory Syncytial Virus whole-genome sequencing identifies convergent evolution of sequence duplication in the C-terminus of the G gene. *Scientific Reports*, *6*, 26311. http://doi.org/10.1038/srep26311

Shi, T., McAllister, D. A., O'Brien, K. L., Simoes, E. A. F., Madhi, S. A., Gessner, B. D., … RSV Global Epidemiology Network. (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. *Lancet*, *390*(10098), 946–958. http://doi.org/10.1016/S0140-6736(17)30938-8

Sricharoenchai, S., Palla, E., & Sanicas, M. (2016). Seasonality of respiratory syncytial virus-lower respiratory tract infection (RSV-LRTI) in children in developing countries. *Journal of Human Virology & Retrovirology, 3(*1), 00076.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313.

Tabatabai, J., Prifert, C., Pfeil, J., Grulich-Henn, J., & Schnitzler, P. (2014). Novel

    Respiratory Syncytial Virus (RSV) Genotype ON1 Predominates in Germany during

    Winter Season 2012-13. PLOS ONE, 9(10), e109191.

    https://doi.org/10.1371/journal.pone.0109191

Tahsin, T., Weissenbacher, D., O'Connor, K., Magge, A., Scotch, M., & Gonzalez-

    Hernandez, G. (2017). GeoBoost: accelerating research involving the geospatial

    metadata of virus GenBank records. *Bioinformatics, btx799*.

    https://doi.org/10.1093/bioinformatics/btx799

Tan, L., Coenjaerts, F. E. J., Houspie, L., Viveen, M. C., van Bleek, G. M., Wiertz, E. J. H.

    J., … Lemey, P. (2013). The Comparative Genomics of Human Respiratory Syncytial

    Virus Subgroups A and B: Genetic Variability and Molecular Evolutionary

    Dynamics. *Journal of Virology*, *87*(14), 8213–8226.

    http://doi.org/10.1128/JVI.03278-12

Texas Department of State and Health Services. (2018). Respiratory Syncytial Virus.

    Retrieved from: http://www.dshs.state.tx.us/idcu/disease/rsv/

Thompson, T. M., Roddam, P. L., Harrison, L. M., Aitken, J. A., & DeVincenzo, J. P.

    (2015). Viral Specific Factors Contribute to Clinical Respiratory Syncytial Virus

    Disease Severity Differences in Infants. *Clinical Microbiology (Los Angeles,*

    Calif.), *4*(3), 206. http://doi.org/10.4172/2327-5073.1000206

Tran, D. N., Pham, T. M. H., Ha, M. T., Tran, T. T. L., Dang, T. K. H., Yoshida, L.-M., …
Ushijima, H. (2013). Molecular Epidemiology and Disease Severity of Human
Respiratory Syncytial Virus in Vietnam. *PLoS ONE*, *8*(1), e45436.
http://doi.org/10.1371/journal.pone.0045436

World Bank Group. (2018). Where We Work. Retrieved from
http://www.worldbank.org/en/where-we-work

World Health Organization. (2017). WHO Global RSV surveillance pilot- objectives.
Retrieved from http://www.who.int/influenza/rsv/rsv_objectives/en/

Yoshihara, K., Le, M. N., Okamoto, M., Wadagni, A. C. A., Nguyen, H. A., Toizumi, M., …
Yoshida, L.-M. (2016). Association of RSV-A ON1 genotype with Increased
Pediatric Acute Lower Respiratory Tract Infection in Vietnam. *Scientific Reports*, *6*,
27856. http://doi.org/10.1038/srep27856

Zou, L., Yi, L., Wu, J., Song, Y., Huang, G., Zhang, X., … Lu, J. (2016). Evolution and
Transmission of Respiratory Syncytial Group A (RSV-A) Viruses in Guangdong,
China 2008–2015. *Frontiers in Microbiology*, *7*, 1263.
http://doi.org/10.3389/fmicb.2016.01263