Viral genetic determinants of persistent human orthopneumovirus infection. *

Dylan Lawless, PHD[1], Christopher G. McKennan, PhD[2], James D. Chappell, MD[3], Jacques Fellay, MD, PhD [1], and Tina V. Hartert, MD, MPH[3,4]

[1]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America
[3]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America
[4]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Abbreviations

INSPIRE (The INfant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure in Infancy Birth Cohort) Multiple sequence alignment (MSA) Respiratory syncytial virus (RSV)

*This document's private source code is available to co-authors from the GitHub repository, from the Overleaf online editor document, or as MS Word format on box. The completed document will be published on biorxiv or medrxiv before journal submission.

1 Abstract

1. RSV world health. 2. Viral genome sequencing and host contributing features. 3. Control for all know interactions. 4. Here we identify RSV variants that are associated with persistent infection. 5. Conclusion

2 Introduction

Human orthopneumovirus, commonly known as respiratory syncytial virus (RSV), is the single most important respiratory virus resulting in the most significant respiratory morbidity and mortality in infants [1]. By the age of 2 years, nearly all children are infected with RSV at least once [2]. RSV infects primarily the upper and lower respiratory tract epithelium, although has been recovered from non-airway sources [3–8]. Prolonged shedding of RSV, especially in young infants and following first infection, has been demonstrated, with longer average duration of viral shedding using polymerase chain reaction (PCR) to detect RSV [9]. While younger age and first infection are associated with persistence of infection, what isn't understood is whether there are viral factors contributing to prolonged shedding or persistence of RSV in young infants. This is important, as persistent infection, or prolonged shedding may contribute to enhanced transmission. Further, the reservoir of RSV infection is not understood, and it is possible that some RSV strains and/or hosts could serve as a dormant reservoir for infection that is activated by seasonal or other influences [10]. Further, host genetic and viral genetic interactions have never been studied. With increasing adoption of human genomics in parallel with pathogen sequencing, novel methods for genome-to-genome analysis provide opportunities to identify selective pressure between host and pathogen (cite Naret). We have previously investigated to host genetics as a source of acceptability to infection (summarise result) (cite). Several environmental have a significant impact on the risk of infection. Population genetics also contributes important features that must be accounted for during genetics association analysis. These features not only depend on the host genetics (as seen with typical GWAS), but also the viral genetics. RSV strains A and B impart the largest separation within the viral population phylogeny. In this study

we perform genomic analysis of RSV to identify variants that are significantly associated with persistent infection in otherwise healthy children.

3 Methods

3.1 Study population

The protocol and informed consent documents were approved by the Institutional Review Board at Vanderbilt University Medical Center. One parent of each participant in the cohorts used for this study provided written informed consent for participation in this study. The informed consent document explained study procedures, use of data and biospecimens for future studies, including genetic studies.

The study population is a longitudinal birth cohort - The INfant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure in Infancy Birth Cohort (INSPIRE) - specifically designed to capture the first RSV infection during infancy in a term health birth cohort. Additional details of this birth cohort has been previously published [11]. Briefly, the cohort includes 1952 term ($\geq$ 37 weeks gestation), non-low birth weight ($\geq$ 2250 g, 5 lbs), otherwise healthy infants from a population-representative sample from pediatric practices located in a rural, suburban and rural region of the southeastern US during 2012-2014. Infants were born June through December so that they would by design be 6 months of age or less entering their first RSV season. Infant (i.e., the first year of life) RSV infection was ascertained through passive and active biweekly surveillance during each infants' first RSV season (Table 1).

3.2 Biweekly surveillance of RSV infection

To capture all RSV infections of children enrolled in INSPIRE throughout infancy (i.e., the first year of life), we conducted passive and active surveillance during their first RSV season by 1) performing bi-weekly phone, email, and/or in person follow-up, 2) frequently educating and reminding parents to call us at the onset of any acute respiratory symptoms, and 3) approaching all infants who were seen at one of the participating pediatric practices for an unscheduled visit. If an infant

met pre-specified criteria for an acute respiratory infection, we then conducted an in-person respiratory illness visit at which time we administered a parental questionnaire, performed a physical exam, collected a nasal wash, and (in infants seen during an unscheduled visit) completed a structured medical chart review. Nasal sample collections were assessed by reverse transcription-quantitative PCR for RSV [Jim to provide reference]. At one year of age infants underwent blood draw for RSV serology to determine infection status during infancy. [Plots with CT-value from Tina, either as supplemental or first mention later so not to pollute the order]. Infants with positive PCR separated by more than 15 or more days were annotated during analysis as "persistent or repeat infection" (Figure 1).

3.3 Descriptive analyses

Descriptive analyses of the cohort were conducted using R 4.0.5 (available at: http://www.r-project.org). Pearson or Wilcoxon tests were used for comparing infants with and without persistent RSV infection. The main descriptive features are provided in Table 1. [Consider plotting these also in the Hmisc Harrell style as seen in his book, RegerssionModellingStrategies2015].

3.4 RSV whole-genome sequencing

RSV whole-genome sequencing of this study population has been previously described [12]. Briefly, RNA was extracted at J. Craig Venter Institute (JCVI) (https://www.jcvi.org) in Rockville, MD from nasal wash samples which were RSV PCR positive and collected during a respiratory illness visit triggered through biweekly surveillance of symptoms. Four forward reverse transcription (RT) primers were designed and four sets of PCR primers were manually picked from primers designed across a consensus of complete RSV genome sequences using JCVI's automated primer design tool, [13]. cDNA was generated from 4 µL undiluted RNA, using the pooled forward primers and SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA, USA). 100 ng of pooled DNA amplicons were sheared to create 400-bp libraries, which were pooled in equal volumes and cleaned. For samples requiring extra coverage, in addition

to Ion Torrent sequencing, Illumina libraries were prepared using the Nextera DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA, USA). Sequence reads were sorted by barcode, trimmed, and de novo assembled using CLC Bio's *clc_novo_assemble* program, and the resulting contigs were searched against custom, full-length RSV nucleotide databases to find the closest reference sequence. All sequence reads were then mapped to the selected reference RSV sequence using CLC Bio's *clc_ref_assemble_long* program [14]. Curated assemblies were validated and annotated with the viral annotation software called Viral Genome ORF Reader, VIGOR 3.0 (https://sourceforge.net/projects/jcvi-vigor/files/), before submission to GenBank as part of the Bioproject accession PRJNA225816 (https://www.ncbi.nlm.nih.gov/bioproject/225816) [15] and PRJNA267583 (https://www.ncbi.nlm.nih.gov/bioproject/267583).

3.5 Viral Sequence alignment

The NCBI-tools Tbl2asn (https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/) was used in the creation of sequence records for submission to GenBank (https://www.ncbi.nlm.nih.gov/genbank/). A total of 350 viral sequences in *.sqn* file format were used for downstream analysis.

We computed a phylogenetic tree for each gene, as follows. NCBI-tools asn2fsa (https://www.huge-man-linux.net/man1/asn2fsa.html) was used to to convert to fasta format. Each sample consisted of 11 sequence segments (NS1, NS2, N, P, M, M2-1, M2-2, SH, G, F, and L) as shown in Figure 1. These were separated and repooled to create 11 single fasta file for each gene containing all 350 samples. Sequences were checked that they also be at least 90for the corresponding gene in order to minimize the loss of aligned positions when computing the phylogenetic tree. Each of the eleven resulting sets was aligned with MAFFT v7 (https://mafft.cbrc.jp/alignment/software/) [16], using default parameters. The sequence of the orthologous gene from the bovine orthopneumovirus (GenBank:NC_001989) was added to each set as an outgroup.

IQ-Tree (https://www.iqtree.org) [17] was used with per-gene multiple sequence alignment (MSA) files for estimating maximum-likelihood phylogenies. Examining the sequences with an alignment viewer showed that X sequences

had frame shift variants but which did not affect the regions included in our testing criteria.

Viral sequence data and clinical information was was merged and cleaned with R. Clinical IDs matching more than one viral sequence IDs were used to label "repeat or persistent" infections. Genetic variantion was quantified in these samples and for subsequent analysis, only the first viral sequence was included for association testing. Strain A and B typing had been completed previously and labels were included to annotate each sample accordingly.

The cohort-specific variant frequency per position was calculated; residues were counted and ranked by frequency, with the most frequent residue defined as reference (REF) or alternative (ALT). Positions with at least one ALT were checked for potential misalignment or other sources of error. Variants positions were selected for association analysis, while non-variant position were ignored.

A number host features have been previously shown to influence infection susceptibility and were therefore included as covariates in our analysis (cite Rosas-Salazar). Six samples were excluded due to insufficient covariate data, resulting in 344 test samples. Of these, 36 were from the same patients ("persistent or repeat" infection) of which half (18) were included for association testing; 326 samples total.

3.6 Population structure

The genetic distances to nearest neighbors were computed based on phylogenetic trees generated with MAFFT. [Other methods also used but not pertinent; include some info from the code]. Principal component analysis (PCA) and singular value decomposition (SVD) were used in dimensionality reduction for exploratory data analysis of viral phylogeny. R package *factoextra* was used for PCA, and to visualise eigenvalues and variance. R package *caret* was used to analyse genetic correlations.

3.7 Association testing

Viral amino acids (genotype collapsed into REF/ALT) were tested for association with infection types *single* and *persistent*, including key covariates that are

significantly associated with infection. Analysis was performed using logistic regression with the R stats (3.6.2) *glm* function as a generalized linear model. The model consisted of the binary response (persistent infection Yes/No), and predictors; viral genotype (REF/ALT amino acid), viral PCs 1-5, host sex, and it also accounted for host features that have been previously demonstrated as significantly associated with infection; self-reported race/ethnicity, child-care attendance, living with siblings (cite Rosas-Salazar).

The environmental host covariates did not contribute any significant effect in our model for the candidate-causal association. Two viral PCs were included in our model for accuracy, however the clinical strain labels (A/B) also reflect the same cohort population structure. [Check the % VE from PC screen plot stats to 1 decimal place and list it here.] Bonferroni correction for multiple testing was applied based on the number of independent variants tested. R package *stats* was used for a range of analysis including glm for logistic regressions. R package *MASS* was used to analyse logistic regression model data.

Second infections occurred only in those with strain B. To test if the significantly associated variants were due to population structure, a subset of only strain B was performed.

3.8 Biological interpretation

Some but not all of these methods will be included for our results section. Adjust based on discussion with co-authors. * Infant RSV infection results in decreased barrier function of the airway epithelium.

* Association between INF-gama and RSV amino acid position (w= wild vs A=alternatives) adjusted for covariates. * Wilcox test comparing IFN-gamma, and INF-alpha, between RSV amino acid positions (W= wild type vs A=alternatives [3 combined]).

* Illustration and discuss known protein domains. * Interactions, * PTM * Motifs, * Epitopes, * protein structure. * Define the choice of PDB used. * Multiple organism alignment. * Domain blast.

4 Results

4.1 Cohort characteristics

The INSPIRE cohort consisted of 1,949 enrolled infants (Figure 1.). Of these, 1,220 (63In total, there were 2,093 in-person respiratory illness visits completed and the median (interquartile range [IQR]) number of in-person respiratory illness visits per infant was 1 (1-2). The characteristics of these infants compared with the other RSV infected infants and the entire cohort is shown in Table 1. From the cohort, 344 RSV viral samples from 326 individuals were sequenced (methods). There were 20 infants with RSV positive PCR $\geq$ 15 days apart who we suspected as having either persistent or repeat infection (based on genetic analysis).

Table 1. Cohort characteristics of infants with persistent RSV infection compared with other RSV infection and entire cohort. Infection is defined as RSV sequence positive, with $\geq$ 15 days between testing. Pearson1, Wilcoxon2. [For Tebeb: We will need to recalculate this to represent just the first infection, as this small number infections include repeat infections which likely drives the median higher.]

The relatively small sample size of of our cohort required analysis that targeted only genes which were *a priori* likely to functionally contribute to the clinical phenotype. Therefore, our analysis focused on F and G glycoproteins (citations).

4.2 Population structure

The phylogenetic tree based on G protein is shown in Figure 2 A. One obvious feature causing a separation in genetic diversity is seen due to the G protein partial gene duplication, which has emerged in recent years within RSV-A strains [18]. RSV-B strains with an analogous duplication have existed for two decades, although the mechanisms leading to emergence and clinical implications have not been entirely defined.

We observed persistent infections by viruses from different phylogenetic clades, rather than one specific clade Figure 2 B. A genotype correlation matrix

produced with the R package *caret* and PCA and eignenvalues from package *factoextra* were used to for reducing the dimensionality of sequence data. Figure 2 scree plot. Dimension one accounted for 95.19% cumulative variance explained in our cohort. All other dimensions account for very little variance, which is evenly distributed; no particular F or G protein protein coding sequence separates the cohort. For his reason, in our main analysis, viral population structure is accounted for by the first five PCs. To test for type I errors dues to the population structure between strain A and B, a subset analysis of individual strains was performed to confirm the validity of the combined analysis.

Note for presentation slides: there is no general theoretical reason that the most informative linear function of the predictor variables should lie among the dominant principal components of the multivariate distribution of the predictor variables. However, if there were then we would like to know since it would produce a false positive in this case. Conversely, for example, in a principal component regression you would hope to find the assoc based on PCs.

4.3 Genetic invariance of persistent infection

The duration of RSV shedding duration in Kenyan infants has been reported previously [19]. Based on these finding, infections separated by at least 15 days were expected to be "new" infections. Figure 2 C shows genetic invariance between for viral sequences within the same host for infections separated by at least 15 days. There is a significant difference between the genetic diversity for multiple viral samples from individuals compared to diversity between all other samples from the same viral clades; P-value $= 1.3e - 8$. We therefore, report these cases as persistent infection rather than second infections.

4.4 Variants in G glycoprotein significantly associated with persistent infection

The consensus sequence within the cohort was assigned based on the major allele. Variants at the amino acid level were defined as either REF/ALT and assessed for their association with persistence. The model consisted of the binary response (persistent infection Yes/No), and predictors; viral genotype (REF/ALT

amino acid), viral PCs 1-5, host sex, and it also accounted for host features that have been previously demonstrated as significantly associated with infection; self-reported race/ethnicity, child-care attendance, living with siblings (cite). Analysis was performed using R stats (3.6.2) *glm* function. A significant genetic association was identified for persistent infection after Bonferroni correction multiple testing (threshold ¡.05/23 = .002), as shown in Figure 3 A. Since many variants within RSV coding genes have non-random association due to strong linkage disequilibrium (LD), we reduced the multiple testing burden by retaining proxy variants and removing those with $r^2 \geq .8$. After identifying a significant association with persistent infection, we quantified the correlation of variants in LD with the lead proxy. Clumping was performed with ranking based on minor allele frequency (MAF) and with a cut-off threshold of $r^2 \geq .8$. The association model was repeated for all variants to produce a Manhattan plot with r^2 by color and P-value statistics as shown in Figure 3 B. This shows both G protein p.E123K/D and p.P217T/S/L as candidate causal variants associated with persistent infection, and no other variants in correlation with this association.

To determine whether this association was simply due to population stratification between strains A and B, a subset analysis was performed using independently assessed clinical laboratory strain labels for A and B. Due to the smaller sample size the result no longer passed the significant threshold. However, the same direct of effect indicated that the association was not a false positive.

To assess the possibility of a false positive due to population structure within our cohort, we assessed the magnitude of variance explained (VE) by the lead variant and found it as -0.996% VE for PC1 and -1.66% VE for PC2; a negligible effect that precludes spurious association by allele frequency between populations, as shown in Figure 3 C.

To investigate genetic variance over time we assessed the public viral data repository of NCBI Human orthopneumovirus, taxid:11250 which contained data from 27 unique countries worldwide, sample collection dates as far back as 1956, and 1084 glycoprotein protein sequences after curation. We observed no enrichment for our variants of interest over time; a low frequency was observed in the available samples with no particular features compared to other low frequency variants. However, correlation between the two positions associated

with persistent infection indicates that it does not arise as random mutation event.

4.5 Functional interpretation

We have collected the known features on the protein domain illustration. There are no known features that directly overlap our variants. However, possible binding interactions with host receptors are discussed in (cite). (Reword) Attachment of the virion to the host cell membrane is thought to occur through interaction with heparan sulfate, initiating infection [20–22]. (Reword) Interactions with host CX3CR1, the receptor for the CX3C chemokine fractalkine, have been reported to modulate the immune response and facilitate infection [23–25]. (Reword) Unlike the other paramyxovirus attachment proteins, RSV glycoprotein lacks both neuraminidase and hemagglutinating activities (cite, probable). (Reword) The isoform of secreted glycoprotein G also is also believed to help the virus escape antibody-dependent restriction of replication by acting as an antigen decoy and by modulating the activity of leukocytes bearing Fc-gamma receptors [26]. Interactions have also been identified with protein SH [27] and via the N-terminus with protein M [28]. G protein has been reported to form homo-oligomers (which we will check next for interaction residues. remove this citation if not fruitful) [29]. The mature secreted form of the protein is also reported for amino acid positions 66 – 298. This secreted isoform includes the variants associated with persistent infection in our analysis. [These 2 citations summaries wre copied and have not been read yet] Interacts with the host lectins CD209/DC-SIGN and CD209L/L-SIGN on dendritic cells; these interactions stimulate the phosphorylation of MAPK3/ERK1 and MAPK1/ERK2, which inhibits dendritic cell activation and could participate in the limited immunity against RSV reinfection [30]. Part of a complex composed of F1, F2 and G glycoproteins have been reported to form part of a complex [31].

Bring up the idea of immune response once already inside cell. Tina made the point that initial binding may not be the most important feature. Known neutralization epitopes were not found for our variant site (Jim).

4.6 Other notes

* Criteria for clearing infection; Tina has the details about this. There are a few details that we want to clarify. * RNA virus like HIV persist for life - Reservoirs within host? - environmental? - closer equator less seasonality. - virus may retreat to parts of the world where it can overseason; temp, humidity, etc. * Severity of second infection Characterised URI/LRI and score * Emergence? * Kenya - family sampling every 5 days (tropical medicine funded this, and South Africa Heather Zar)

5 Discussion

We initially performed a host GWAS to potentially identify any common host variant association with susceptibility to infection [32]. Among 1959 enrolled, 5446There were significant differences in environmental factors associated with RSV infection, including child-care (p=0.001), siblings (p=0.002) and ethnicity (p=0.002). GWAS analyses of a subset of 663 participants adjusted for birth month, sex, race, child-care and siblings revealed no significant associations. Multiple testing burden may mask any small genetic effects. Therefore, we estimated narrow sense heritability (h_l) 0 would indicate accumulation of small genetic effects [33]. A normally distributed latent liability variable was used to model the genetic correlation, including covariates The maximum likelihood estimate for h_l was exactly 0. Therefore we found no evidence of host genetic susceptibility due to common variants. The possibility of rare variants causing susceptibility to infection may exist, although this is very unlikely to affect our analysis on the cohort of our sample population.

Accounting for host genetic factors allowed our analysis to focus on the viral genetic features which drive persistence. The possibility of viral mutational immune escape has been reported for infants who struggle to control primary RSV infections, allowing for prolonged viral replication and not previously described viral rebound [34]. We suspected that our variants may either be enriched by selective pressure over time, however inspecting public data from the last two decades shows presence of these variants at low frequencies. Within-host

variation with denovo mutation may allow this variant to present within some individuals but failing to persist within the population, however, we have not been able to conclusively assess this possibility.

Our analysis also consists of the primary host infection for affected children and therefore we do not expect any host immune memory before this first infection, potentially beyond maternal antibody.

A host genetic interaction for asthma has been demonstrated previously [35]. We performed an interaction analysis for the outcome of host asthma, host genetics and pathogen genetics but no significant interaction was found. However, our sample size is unlikely to be sufficient to answer this question, which may be addressed with future studies.

This variants identified in this study appear to be present for children that are less sick. We are formally testing this based on immune response markers. TBD. Persist - variation within host - expand that this could be done and state how but we do not have the data. CT values go up when it is being cleared - first illness CT rarely lower than second CT = reducing virus.

Functional interpretation section:

What Qs does the reader have?

Why is the variant present at low level in population.

Chronic disease.

Airway reprogramming.

Cleared virus may have less of influence than chronic.

Infants without RSV less likely to have asthma.

Infants infected go on to have blunted subsequent antiviral responses.

Chronic stimulation versus immune exhaustion?

Metabolism of airway epithelium, glycolytic pathways.

How would this mutation lead to persistence? - Epitope - Evasion - etc.

Selected in some backgrounds but not very fit?

Not increasing over time. Stable.

Heather Zar - papers

Acute and chronic resp morbidity to give it the spin for CID.

Note that not only the persistent have variants of interest, but many others

also have this variants.

6 Links

6.1 Software

R v4.1.0 was used for data preparation and analysis http://www.r-project.org.

R package *caret* was used to analysis: genetic correlations.

R package *dplyr* was used for data curation.

R package *factoextra* was used for analysis: PCA, and to visualise eigenvalues and variance.

R package *ggplot2* was used for data visualisation.

R package *MASS* was used to analysis: logistic regression model data.

R package *stats* was used for analysis: including glm for logistic regressions.

R package *stringr* was used for data curation.

R package *tidyr* was used for data curation.

asn2fsa https://www.huge-man-linux.net/man1/asn2fsa.html

clc_novo_assemble qiagenbioinformatics.com

Clustal Omega https://www.ebi.ac.uk/Tools/msa/clustalo/

GenBank https://www.ncbi.nlm.nih.gov/genbank/

IQ-Tree https://www.iqtree.org/

MAFFT https://mafft.cbrc.jp/alignment/software/ [16]

Tbl2asn https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/

Viral Genome ORF Reader, VIGOR 3.0 https://sourceforge.net/projects/jcvi-vigor/files/

6.2 Additional Data sources

GenBank:NC_001989 https://www.ncbi.nlm.nih.gov/nuccore/NC_001989

Dataset https://www.ncbi.nlm.nih.gov/bioproject/267583

Dataset https://www.ncbi.nlm.nih.gov/bioproject/225816

J. Craig Venter Institute https://www.jcvi.org

7 Code availability

Public upload of analysis code to GitHub https://github.com/DylanLawless/. Do you want a stand-alone repository that we will abandon, or is it OK in my personal page?

8 On-line supplement Methods:

8.1 Host GWAS for genetic susceptibility to infection

Include the section on sample collection and genotyping array used. These notes are in the raw genotype directory.

To determine whether a genetic susceptibility to infection was evident in our cohort, we perfromed a GWAS analysis of 663 of samples from our cohort [32]. Samples were genotyped using X genotyping array and genotypes were called using Illumina GenomeStudio. Study participants were excluded based on a missing genotype call rate of 10Subject independence was assessed using KING (https://people.virginia.edu/~wc9c/KING/) any samples with a high degree of kinship or duplication (pairwise identify-by-state (IBS) estimated kinship coefficient > 0.18) were removed [36].

Variants were removed for minor allele frequencies < 0.05, missingness > 0.1, and additionally for controls, Hardy-Weinberg Equilibrium (HWE) $P < 1E - 6$. Reported and estimated sex was examined for discrepancy. We compared the genetic ancestry in cases to self-reported ethnicity to check for mislabeling. Genotyping data was phased [SHAPEIT2] https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html and imputed [IMPUTE2] https://mathgen.stats.ox.ac.uk/impute/impute_v2.html using the 1000 Genomes Project phase 3 reference panel. The reference genome build and LD population used was hg19/1000G Nov2014 EUR. Imputation quality was assessed and SNPs with an information score of < 0.8 or minor allele frequency < 0.05 were removed.

GCTA https://cnsgenomics.com/software/gcta/ was used to calculate the genetic relationship matrix (GRM) and to perform principal component analysis (PCA) to quantify population structure [37]. Datasets were merged using PLINK v1.9. SNP positions and identifiers were updated according to

dbNSFP4.0a (hg19) [38]. QC was repeated after merging cases and controls for combined cohort-specific frequencies. Genome-wide association analysis was performed using PLINK version 1.9 for logistic regression with multiple covariates that included the child's birth month, enrollment year (as a marker of RSV season), daycare attendance, the presence of another child less than 6 years of age at home, and 6 ancestry principal components as covariates Population structure was controlled by GRM eigenvectors and analysis covariates consisted of sex, age, and study site.

References

[1] C. B. Hall, G. A. Weinberg, M. K. Iwane, A. K. Blumkin, K. M. Edwards, M. A. Staat, P. Auinger, M. R. Griffin, K. A. Poehling, D. Erdman, C. G. Grijalva, Y. Zhu, and P. Szilagyi. The burden of respiratory syncytial virus infection in young children. *N Engl J Med*, 360(6):588–98, February 2009. ISSN 0028-4793 (Print) 0028-4793. doi: 10.1056/NEJMoa0804877. Edition: 2009/02/07.

[2] W. P. Glezen, L. H. Taber, A. L. Frank, and J. A. Kasel. Risk of primary infection and reinfection with respiratory syncytial virus. *Am J Dis Child*, 140(6):543–6, June 1986. ISSN 0002-922X (Print) 0002-922x. doi: 10.1001/archpedi.1986.02140200053026. Edition: 1986/06/01.

[3] V. Bokun, J. J. Moore, R. Moore, C. C. Smallcombe, T. J. Harford, F. Rezaee, F. Esper, and G. Piedimonte. Respiratory syncytial virus exhibits differential tropism for distinct human placental cell types with Hofbauer cells acting as a permissive reservoir for infection. *PLoS One*, 14 (12):e0225767, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0225767. Edition: 2019/12/04.

[4] H. A. Cubie, L. A. Duncan, L. A. Marshall, and N. M. Smith. Detection of respiratory syncytial virus nucleic acid in archival postmortem tissue from infants. *Pediatr Pathol Lab Med*, 17(6):927–38, November 1997. ISSN 1077-1042 (Print) 1077-1042. Edition: 1997/11/14.

[5] D. Nadal, W. Wunderli, O. Meurmann, J. Briner, and J. Hirsig. Isolation of respiratory syncytial virus from liver tissue and extrahepatic biliary atresia material. *Scand J Infect Dis*, 22(1):91–3, 1990. ISSN 0036-5548 (Print) 0036-5548. doi: 10.3109/00365549009023125. Edition: 1990/01/01.

[6] D. R. O'Donnell, M. J. McGarvey, J. M. Tully, I. M. Balfour-Lynn, and P. J. Openshaw. Respiratory syncytial virus RNA in cells from the peripheral blood during acute infection. *J Pediatr*, 133(2):272–4, August 1998. ISSN 0022-3476 (Print) 0022-3476. doi: 10.1016/s0022-3476(98)70234-3. Edition: 1998/08/26.

[7] F. Rezaee, L. F. Gibson, D. Piktel, S. Othumpangat, and G. Piedimonte. Respiratory syncytial virus infection in human bone marrow stromal cells. *Am J Respir Cell Mol Biol*, 45(2):277–86, August 2011. ISSN 1044-1549 (Print) 1044-1549. doi: 10.1165/rcmb.2010-0121OC. Edition: 2010/10/26.

[8] A. Rohwedder, O. Keminer, J. Forster, K. Schneider, E. Schneider, and H. Werchau. Detection of respiratory syncytial virus RNA in blood of neonates by polymerase chain reaction. *J Med Virol*, 54(4):320–7, April 1998. ISSN 0146-6615 (Print) 0146-6615. doi: 10.1002/(sici)1096-9071(199804)54:4⟨320::aid-jmv13⟩3.0.co;2-j. Edition: 1998/04/29.

[9] P. K. Munywoki, D. C. Koech, C. N. Agoti, N. Kibirige, J. Kipkoech, P. A. Cane, G. F. Medley, and D. J. Nokes. Influence of age, severity of infection, and co-infection on the duration of respiratory syncytial virus (RSV) shedding. *Epidemiol Infect*, 143(4):804–12, March 2015. ISSN 0950-2688 (Print) 0950-2688. doi: 10.1017/s0950268814001393. Edition: 2014/06/06.

[10] L. Hobson and M. L. Everard. Persistent of respiratory syncytial virus in human dendritic cells and influence of nitric oxide. *Clin Exp Immunol*, 151(2):359–66, February 2008. ISSN 0009-9104 (Print) 0009-9104. doi: 10.1111/j.1365-2249.2007.03560.x. Edition: 2007/12/08.

[11] E. K. Larkin, T. Gebretsadik, M. L. Moore, L. J. Anderson, W. D. Dupont, J. D. Chappell, P. A. Minton, R. S. Peebles, Jr., P. E. Moore, R. S. Valet,

D. H. Arnold, C. Rosas-Salazar, S. R. Das, F. P. Polack, and T. V. Hartert. Objectives, design and enrollment results from the Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure Study (INSPIRE). *BMC Pulm Med*, 15:45, April 2015. ISSN 1471-2466. doi: 10.1186/s12890-015-0040-0. Edition: 2015/05/30.

[12] S. A. Schobel, K. M. Stucker, M. L. Moore, L. J. Anderson, E. K. Larkin, J. Shankar, J. Bera, V. Puri, M. H. Shilts, C. Rosas-Salazar, R. A. Halpin, N. Fedorova, S. Shrivastava, T. B. Stockwell, R. S. Peebles, T. V. Hartert, and S. R. Das. Respiratory Syncytial Virus whole-genome sequencing identifies convergent evolution of sequence duplication in the C-terminus of the G gene. *Sci Rep*, 6:26311, May 2016. ISSN 2045-2322. doi: 10.1038/srep26311. Edition: 2016/05/24.

[13] K. Li, S. Shrivastava, A. Brownley, D. Katzel, J. Bera, A. T. Nguyen, V. Thovarai, R. Halpin, and T. B. Stockwell. Automated degenerate PCR primer design for high-throughput sequencing improves efficiency of viral sequencing. *Virol J*, 9:261, November 2012. ISSN 1743-422x. doi: 10.1186/1743-422x-9-261. Edition: 2012/11/08.

[14] QIAGEN Aarhus. White paper on de novo assembly in CLC Assembly Cell 4.0. *digitalinsights*, page 14, June 2016. URL https://digitalinsights.qiagen.com/files/whitepapers/whitepaper-denovo-assembly.pdf. Place: Denmark Publisher: Qiagen.

[15] S. Wang, J. P. Sundaram, and T. B. Stockwell. VIGOR extended to annotate genomes for additional 12 different viruses. *Nucleic Acids Res*, 40(Web Server issue):W186–92, July 2012. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gks528. Edition: 2012/06/07.

[16] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[17] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating

maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1): 268–274, 2015.

[18] AliReza Eshaghi, Venkata R Duvvuri, Rachel Lai, Jeya T Nadarajah, Aimin Li, Samir N Patel, Donald E Low, and Jonathan B Gubbay. Genetic variability of human respiratory syncytial virus a strains circulating in ontario: a novel genotype with a 72 nucleotide g gene duplication. *PloS one*, 7(3):e32807, 2012.

[19] Emelda A Okiro, Lisa J White, Mwanajuma Ngama, Patricia A Cane, Graham F Medley, and D James Nokes. Duration of shedding of respiratory syncytial virus in a community study of kenyan children. *BMC infectious diseases*, 10(1):1–7, 2010.

[20] S Levine, R Klaiber-Franco, and PR Paradiso. Demonstration that glycoprotein g is the attachment protein of respiratory syncytial virus. *Journal of General Virology*, 68(9):2521–2524, 1987.

[21] Steven A Feldman, R Michael Hendry, and Judy A Beeler. Identification of a linear heparin binding domain for human respiratory syncytial virus attachment glycoprotein g. *Journal of virology*, 73(8):6610–6617, 1999.

[22] Steven A Feldman, Susette Audet, and Judy A Beeler. The fusion glycoprotein of human respiratory syncytial virus facilitates virus attachment and infectivity via an interaction with cellular heparan sulfate. *Journal of Virology*, 74(14):6442–6447, 2000.

[23] Sara M Johnson, Beth A McNally, Ioannis Ioannidis, Emilio Flano, Michael N Teng, Antonius G Oomens, Edward E Walsh, and Mark E Peeples. Respiratory syncytial virus uses cx3cr1 as a receptor on primary human airway epithelial cultures. *PLoS pathogens*, 11(12):e1005318, 2015.

[24] Ralph A Tripp, Les P Jones, Lia M Haynes, HaoQiang Zheng, Philip M Murphy, and Larry J Anderson. Cx3c chemokine mimicry by respiratory syncytial virus g glycoprotein. *Nature immunology*, 2(8):732–738, 2001.

[25] Kwang-Il Jeong, Peter A Piepenhagen, Michael Kishko, Joshua M DiNapoli, Rachel P Groppo, Linong Zhang, Jeffrey Almond, Harry Kleanthous, Simon

Delagrave, and Mark Parrington. Cx3cr1 is expressed in differentiated human ciliated airway cells and co-localizes with respiratory syncytial virus on cilia in a g protein-dependent manner. *PloS one*, 10(6):e0130517, 2015.

[26] Alexander Bukreyev, Lijuan Yang, Jens Fricke, Lily Cheng, Jerrold M Ward, Brian R Murphy, and Peter L Collins. The secreted form of respiratory syncytial virus g glycoprotein helps the virus evade antibody-mediated restriction of replication by acting as an antigen decoy and through effects on fc receptor-bearing leukocytes. *Journal of virology*, 82(24):12191–12204, 2008.

[27] HW McL Rixon, G Brown, JT Murray, and RJ Sugrue. The respiratory syncytial virus small hydrophobic protein is phosphorylated via a mitogen-activated protein kinase p38-dependent tyrosine kinase activity during virus infection. *Journal of General Virology*, 86(2):375–384, 2005.

[28] Reena Ghildyal, Dongsheng Li, Irene Peroulis, Benjamin Shields, Phillip G Bardin, Michael N Teng, Peter L Collins, Jayesh Meanger, and John Mills. Interaction between the respiratory syncytial virus g glycoprotein cytoplasmic domain and the matrix protein. *Journal of General Virology*, 86(7): 1879–1884, 2005.

[29] Peter L Collins and Geneviève Mottet. Oligomerization and post-translational processing of glycoprotein g of human respiratory syncytial virus: altered o-glycosylation in the presence of brefeldin a. *Journal of General Virology*, 73(4):849–863, 1992.

[30] Teresa R Johnson, Jason S McLellan, and Barney S Graham. Respiratory syncytial virus glycoprotein g interacts with dc-sign and l-sign to activate erk1 and erk2. *Journal of virology*, 86(3):1339–1347, 2012.

[31] Kit-Wei Low, Timothy Tan, Ken Ng, Boon-Huan Tan, and Richard J Sugrue. The rsv f and g glycoproteins interact to form a complex on the surface of infected cells. *Biochemical and biophysical research communications*, 366 (2):308–313, 2008.

[32] D Lawless, C Rosas-Salazar, T Gebretsadik, K Turi, B Snyder, P Wu, J Fellay, and T Hartert. Genome-wide association study of susceptibility to respiratory syncytial virus infection during infancy. In *EUROPEAN JOURNAL OF HUMAN GENETICS*, volume 28, pages 319–319. SPRINGERNATURE CAMPUS, 4 CRINAN ST, LONDON, N1 9XW, ENGLAND, 2020.

[33] David Golan and Saharon Rosset. Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics*, 27(13): i317–i323, 2011.

[34] Monica E Brint, Joshua M Hughes, Aditya Shah, Chelsea R Miller, Lisa G Harrison, Elizabeth A Meals, Jacqueline Blanch, Charlotte R Thompson, Stephania A Cormier, and John P DeVincenzo. Prolonged viral replication and longitudinal viral dynamic differences among respiratory syncytial virus infected infants. *Pediatric research*, 82(5):872–880, 2017.

[35] Miriam F Moffatt, Ivo G Gut, Florence Demenais, David P Strachan, Emmanuelle Bouzigon, Simon Heath, Erika von Mutius, Martin Farrall, Mark Lathrop, and William OCM Cookson. A large-scale, consortium-based genomewide association study of asthma. *New England Journal of Medicine*, 363(13):1211–1221, 2010.

[36] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010. Number: 22 ISBN: 1460-2059 Publisher: Oxford University Press.

[37] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011. Number: 1 Publisher: Elsevier.

[38] Xiaoming Liu, Chunlei Wu, Chang Li, and Eric Boerwinkle. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, 37(3):235–241, 2016. doi: 10.1002/humu.22932. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.22932. Number: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22932.

9 Tables and Figures

		Persistent RSV N=19	Other RSV N=342	Total N=1949	Test statistic
	Illness age, months (median, IQR)	6 (4, 6)	4 (2, 5)	NA	NA
	Respiratory severity score (median, IQR)	2.0 (1.2, 3.0)	3.0 (2.0, 4.0)	NA	$P = 0.27^1$
Viral strain	RSV A	73%	60%	NA	NA
	RSV B	27%	40%		NA
RSV season	2012-13	68%	54%	44%	NA
	2013-14	32%	46%	56%	NA
Self reported Race	Non-Hispanic Black	11%	16%	18%	NA
	Non-Hispanic White	79%	66%	65%	NA
	Hispanic	0%	9%	9%	NA
	Multi-race/ethnicity/other	11%	8%	9%	NA
Sex	Female	53%	44%	48%	NA
	Male	47%	56%	52%	NA
	Second-hand smoke exposure	58%	44%	47%	NA
Insurance	Medicaid	32%	52%	54%	NA
	Private	68%	47%	45%	NA
	None/unknown	0%	1%	1%	NA
	Daycare				
	Siblings				

Table 1: Caption

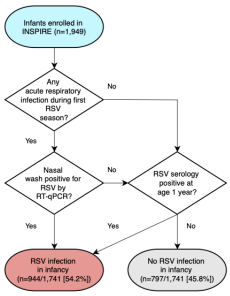

Figure 1: legend.

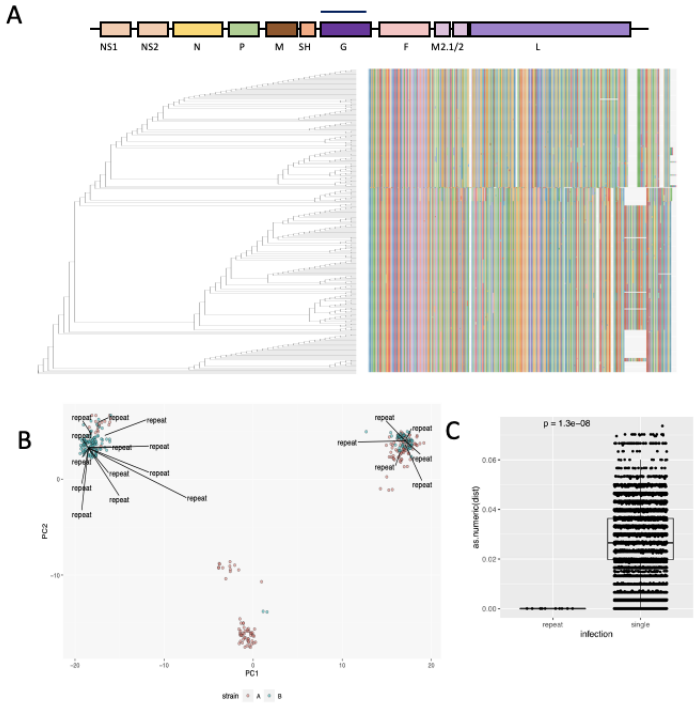

Figure 2: legend.

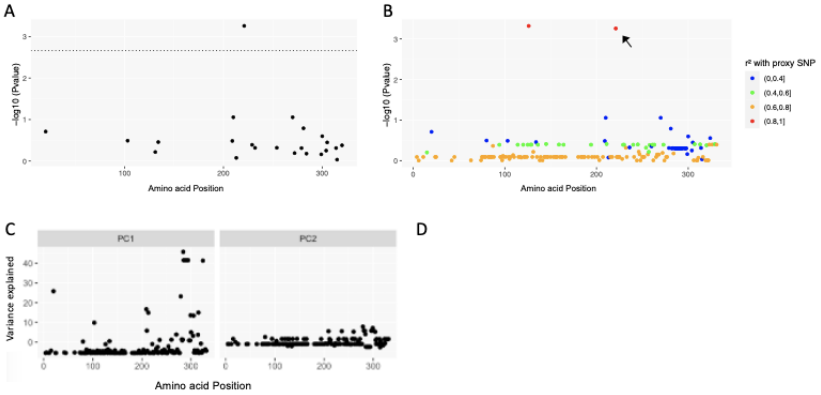

Figure 3: legend.

Family & Domains[1]

Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region[1]	125 – 161	Disordered (Sequence analysis)	Add BLAST		37
Region[1]	187 – 198	Binding to host heparan sulfate (1 Publication)	Add BLAST		12
Region[1]	190 – 298	Disordered (Sequence analysis)	Add BLAST		109

Compositional bias

Feature key	Position(s)	Description	Actions	Graphical view	Length
Compositional bias[1]	125 – 159	Polar residues (Sequence analysis)	Add BLAST		35
Compositional bias[1]	221 – 288	Polar residues (Sequence analysis)	Add BLAST		68

Sites

Feature key	Position(s)	Description	Actions	Graphical view	Length
Site[1]	65 – 66	Cleavage (1 Publication)			2

Topology

Feature key	Position(s)	Description	Actions	Graphical view	Length
Topological domain[1]	1 – 42	Cytoplasmic (1 Publication)	Add BLAST		42
Transmembrane[1]	43 – 63	Helical (Sequence analysis)	Add BLAST		21
Topological domain[1]	64 – 298	Extracellular (1 Publication)	Add BLAST		235

Figure 4: legend.

	0.wt $N = 163$		1.alt $N = 139$		Combined $N = 302$	
V221						
L	0%	$\frac{0}{163}$	1%	$\frac{1}{139}$	0%	$\frac{1}{302}$
P	100%	$\frac{163}{163}$	0%	$\frac{0}{139}$	54%	$\frac{163}{302}$
S	0%	$\frac{0}{163}$	16%	$\frac{22}{139}$	7%	$\frac{22}{302}$
T	0%	$\frac{0}{163}$	83%	$\frac{116}{139}$	38%	$\frac{116}{302}$
strain						
A	85%	$\frac{139}{163}$	15%	$\frac{21}{139}$	53%	$\frac{160}{302}$
B	15%	$\frac{24}{163}$	85%	$\frac{118}{139}$	47%	$\frac{142}{302}$

Tally from clinical label

```
df <- test_set_RefAlt %>%
  filter(var_pos == "V221") %>%
  group_by(cfsubjid, var_pos, var)%>%
  tally() %>%
  mutate(strain = case_when(str_detect(cfsubjid, 'A') == 'A',
                            str_detect(cfsubjid, 'B') == 'B')) %>%
  mutate(genotype = case_when(str_detect(var, 'P') == 'WT',
                              str_detect(var, 'L') == 'ALT',
                              str_detect(var, 'S') == 'ALT',
                              str_detect(var, 'T') == 'ALT',))

df_counts <- x %>% group_by(strain, var, genotype) %>% tally()
df_counts <- rename(df_counts, strain_geno_count = n)

df_counts_geno <- x %>% group_by(var, genotype) %>% tally()
df_counts_geno <- rename(df_counts_geno, geno_count = n)

df_strain_total <- x %>% group_by(strain, genotype) %>% tally()
df_strain_total <- rename(df_strain_total, strain_geno_total = n)

df_geno_total <- x %>% group_by(genotype) %>% tally()
df_geno_total <- rename(df_geno_total, geno_total = n)


df2 <- merge( (merge(df_counts, df_strain_total)), df_geno_total)
```

	N			%		
	WT 163	ALT 140	Combined 303	WT	ALT	Combined
V221						
L	0	1	1	0	1	0
P	163	0	163	100	0	54
S	0	22	22	0	16	7
T	0	117	117	0	84	39
strain						
A	142	31	173	87	22	57
B	21	109	130	13	78	43

Tally from cfsubjid label

Figure 5: legend.