# Viral genetic determinants of prolonged respiratory syncytial virus infection among infants in a healthy term birth cohort. *

Dylan Lawless, PhD[epfl], Christopher G. McKennan, PhD[penn], Thomas Junier, PhD[sib], Zhi Ming Xu, MSc[epfl], Suman Das, PhD[**], Larry J Anderson[emoryPed], Tebeb Gebretsadik[bioVan], Meghan Shilts[**], Christian Rosas-Salazar[**], James D. Chappell, MD[pedVan], Jacques Fellay, MD, PhD [epfl], and Tina V. Hartert, MD, MPH[pedVan,medVan]

[epfl]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[penn]Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

[sib]Swiss Institute of Bioinformatics, Vital-IT Group, Switzerland

[pedVan]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

[medVan]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

[emoryPed]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, United States of America

[bioVan]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

[**]Missing

*This document's source code is available to co-authors from the GitHub repository and from the overleaf online editor document. All code and supplemental live data results at https://github.com/DylanLawless/inspire2022lawless.github.io. The PDF will be published on medrxiv before submission. Please note the URL is on my personal domain temporarily; this will be switched for release. For eLife, authors are asked to agree to publish their work under the terms of the Creative Commons Attribution license CC-BY-4.0.

### Abbreviations

ALT (alternative); CI (confidence interval); GWAS (genome-wide association study); G (glycoprotein); H (hemagglutinin); HN (hemagglutinin-neuraminidase); IFN (interferon); IQR (interquartile range); INSPIRE (The INfant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure in Infancy Birth Cohort); LD (linkage disequilibrium); MSA (multiple sequence alignment); OR (odds ratio); PCR (polymerase chain reaction); PCA (Principal component analysis); REF (reference); RT (reverse transcription); SVD (singular value decomposition); SNP (single nucleotide polymorphism); VE (variance explained); MSA (multiple sequence alignment); RSV (respiratory syncytial virus).

### Notice of Prior Presentation

The results of the genome wide association study analyses included in this manuscript were presented during the European Society of Human Genetics Conference in June 2020 in Berlin, Germany, which was held remotely.

## 1 Abstract

**Background**: Respiratory syncytial virus (RSV) is primarily associated with acute respiratory infection, however, many RNA viruses can establish prolonged or persistent infection in some infected individuals.

**Objectives**: To identify viral genetic variants associated with "prolonged infection" and determine if there are host genetic risk alleles for first RSV infection risk.

**Methods**: In a population-based cohort study of healthy term infants, RSV infection was determined by biweekly surveillance for RSV and and 1-year RSV serology. Using RSV whole-genome sequencing, viral amino acids (genotype) were tested for association with a priori defined "prolonged" infant RSV infection adjusting for host features associated with increased infection risk. We tested the association of infant RSV infection risk with severe RSV and childhood asthma-associated SNPs.

**Results**: A significant viral genetic association in the RSV G protein p.E123K/D and p.P218T/S/L were the candidate causal variants associated with "prolonged" infection after Bonferroni correction for multiple testing. These variants were associated exclusively with upper respiratory tract infection, and on average, milder clinical infection compared with other circulating variants (results). We found no evidence of host genetic risk of RSV infection.

**Conclusions**: While we found no evidence of host genetic susceptibility to first RSV infection during infancy, we identified a novel RSV viral variant associated with prolonged infection in healthy infants. As the capacity of RSV for chronicity and its viral reservoir are not

understood, these results are of fundamental interest in understanding host genetic and viral genetic contributions that may underlie the development of chronic respiratory morbidity.

## 2 Introduction

Respiratory syncytial virus (RSV), a human orthopneumovirus, is the single most important respiratory virus to infant global health, resulting in significant morbidity and mortality in infants [1]. By the age of two to three years, nearly all children are infected with RSV at least once [2]. RSV is a seasonal mucosal pathogen that infects primarily the upper and lower respiratory tract epithelium, although it has been recovered from non-airway sources [3–8]. RSV is primarily associated with acute respiratory infection, however, many RNA viruses can establish prolonged or persistent infection in some infected individuals. [refs] 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 is not understood is whether there are viral factors contributing to prolonged shedding or persistence of RSV in young infants. This is important, as prolonged infection, or prolonged shedding may contribute to enhanced transmission and developmental changes to the early life airway epithelium. 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].

The objectives of this study were to identify viral genetic variants associated with "prolonged infection" and determine if there are host genetic risk alleles for first RSV infection risk. These questions are of fundamental interest in understanding host genetic and viral genetic contributions that may underlie the development of chronic respiratory morbidity.

To determine if there is evidence of host genetic factors for risk of infant RSV infection we tested the association of $\approx$ 60 severe RSV and childhood asthma-associated single nucleotide polymorphisms (SNPs) in a population-based cohort of term healthy infants. We also conducted a host GWAS to identify common variants associated with infant RSV infection, and narrow sense heritability to test for small cumulative effects. We a priori defined the clinical entity of "prolonged" infection during infancy as those with repeat positive PCR separated by 15 or more days and who repeatedly met pre-specified criteria for an acute respiratory infection. RSV genome sequencing was done on all isolates meeting illness criteria with positive RSV PCR. Viral amino acids (genotype) of the F and G glycoprotein were tested for association with "prolonged" infection adjusting for host features associated with increased infection risk. We focused our analyses on the surface F (fusion) and G (attachment) proteins of RSV as they have been implicated in pathogenesis (refs a,b), and both are targets for neutralizing antibodies during infection (refs c,d). Lastly, to determine if the variants of interest were en-

riched by selective pressure over time, we used public data from the past three decades to assess variant frequency over time.

# 3 Methods

## 3.1 Study population

The protocol and informed consent documents were approved by the Institutional Review Board at Vanderbilt University Medical Center (#111299). One parent of each participant in the cohort 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 healthy birth cohort. Additional details of this birth cohort have 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 of pediatric practices located in a rural, suburban, and urban regions 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.

## 3.2 Biweekly surveillance of RSV infection

Infant (i.e., the first year of life) RSV infection was ascertained through passive and active biweekly surveillance during each infants' first RSV season and RSV serology (Table 1). 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 completed a structured medical chart review in infants seen during an unscheduled visit. Viral identification in nasal samples was done by reverse transcription-quantitative PCR for RSV [12]. [Plots with CT-value from Tina, either as supplemental or first mention later so not to pollute the order]. We use the term "prolonged infection" for infections among infants with positive PCR separated by 15 or more days and repeatedly met pre-specified criteria for an acute respiratory infection (Figure 1). Viral genetic analyses were then conducted on this set of infants.

### 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 prolonged RSV infection. The main descriptive features are provided in Table 1.

### 3.4 Host DNA Collection and Genotyping

One-year blood samples were selected based on availability of DNA among a random group of children and genotyped with the MEGA microarray (Illumina, CA, United States) at the University of Washington DNA Sequencing and Gene Analysis Center as per their standard manual of procedures.

### 3.5 Genetic Analyses of RSV Infection in Infancy

A GWAS was performed on 621 children with available DNA for the association between host genotype and RSV infection during infancy. Due to sample size constraints we also restricted our sub-analysis to the $\approx$ 60 LRTI- and childhood asthma-associated host SNPs identified in Pividori et al. [13]; Janssen et al. [14]; Pasanen et al. [15], respectively, to test the association of infant RSV infection risk as defined by RSV infection detected through biweekly surveillance or RSV serology with known childhood asthma- or RSV bronchiolitis-associated single nucleotide polymorphisms. To address power, we additionally evaluated the accumulation of small genetic effects that would go undetected in a GWAS by estimating the heritability of RSV infection.

For GWAS analyses, the initial round of data quality control was performed on individual populations (self-reported as White, Black, and Hispanic) using PLINK version 1.9 [16]. Subjects with a missing genotype call rate of 5% were removed. The single nucleotide polymorphism (SNP) minor allele frequency (MAF) threshold was set for cohorts as MAF $> 0.01$, 0.03, 0.08 for White, Black, and Hispanic, respectively [17]. The groups were merged for a total of 1,086,830 variants and a genotyping rate of 0.78. Subject independence was assessed to prevent spurious associations. However, no probable relatives or duplicates were detected based on pairwise identify-by-state. Reported and estimated sex was also examined for discrepancies. Next, a second round of quality control on the combined dataset was conducted, which removed 74 samples due to genotype missingness and 399,991 variants with a genotype rate $< 0.1$. Samples were checked for departure from Hardy-Weinberg equilibrium (HWE) ($P < 1e^{-6}$) to uncover features of selection, population admixture, cryptic relatedness, or genotyping error. This was only performed on controls to prevent removal of genuine genetic associations that can be associated with this measurement, removing 6,024 variants. No variants

had a MAF $< 0.01$ after merging. SNP positions and identifiers were compared and updated according to dbNSFP4.0a (hg19) with 289 variants removed due to a missing coordinate and SNPs identifier [17]. This resulted in an analysis-ready dataset of 680,526 variants from 621 children (509 and 112 with and without RSV infection in infancy, respectively) with a total genotyping rate of 0.98. No genomic inflation was evident with an estimated lambda (based on median chi-squared test) equal to 1. We then used genome-wide complex trait analysis (GCTA) software (https://cnsgenomics.com/software/gcta/) to calculate the genetic relationship matrix and performed principal component analysis to account for population structure [18]. 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 $\leq 6$ years of age at home, and 6 ancestry principal components as covariates [16]. Due to the multiple testing burden likely precluding our ability to identify small genetic effects in our GWAS, we conducted additional heritability analyses.

For the heritability analyses, we used the method described by Golan et al to estimate the narrow-sense heritability of RSV infection infancy on the latent liability scale ($h_l^2$), which, if $>$ 0, would indicate an accumulation of small genetic effects [19]. We estimated $h_l^2$ to be exactly 0, suggesting that, if present, infant RSV infection-related genetic signals are both small and sparse.

### 3.6 RSV whole-genome sequencing

RSV whole-genome sequencing of this study population has been previously described [20]. 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, [21]. 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 [22]. Curated assemblies were validated and annotated with the viral annotation software called Vi-

ral 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) [23] and PRJNA267583 (https://www.ncbi.nlm.nih.gov/bioproject/267583).

### 3.7 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 files for each gene containing all 350 samples. Sequences were checked so that they would also be at least 90% as long as the maximum length for 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/) [24], 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) [25] was used with per-gene multiple sequence alignment (MSA) files for estimating maximum-likelihood phylogenies. Examining the sequences with an alignment viewer showed that a small number of sequences had frame-shift variants but these did not affect the regions included in our testing criteria.

Viral sequence data and clinical information was merged and cleaned with R. Clinical IDs matching more than one viral sequence ID were used to re-identify samples from the same individual as "prolonged" infections. Genetic variation was quantified in these samples and for subsequent analysis only the first viral sequence was included for association testing. Typing of strain A and B 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) and alternative (ALT) for variants. Positions with at least one ALT were checked for potential misalignment or other sources of error. Variant positions were selected for association analysis, while non-variant position were ignored.

A number of host features have been previously shown to influence infection susceptibil-

ity 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 ("prolonged or repeat" infection) of which half (18) were included for association testing; 326 samples total.

### 3.8 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.

### 3.9 Association testing

Viral amino acids (genotype collapsed into REF/ALT) were tested for association with infection types *single* and *prolonged*, 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 (prolonged 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. Five viral PCs were included in our model to account for population structure. 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.

### 3.10 Biological interpretation

Infant RSV infection results in decreased barrier function of the airway epithelium [26]. Association between INF-$\gamma$ and RSV amino acid position (W=wild type versus A=alternatives)

was adjusted for the same covariates as the main analysis. Wilcox test comparing interferon (IFN)-$\gamma$, and INF-$\alpha$, between RSV amino acid positions (W= wild type vs A=alternatives [3 combined]). Protein structures were analysed with data sourced from RCSB PDB https://www.rcsb.org. Protein function and domains were assessed using UniProt (https://www.uniprot.org) for P03423 (GLYC_HRSVA) (strain A2) and O36633 (GLYC_HRSVB) (strain B1) in gff format; https://www.uniprot.org/uniprot/P03423 and https://www.uniprot.org/uniprot/O36633, respectively. Interactions, PTM, motifs, and epitopes were assessed from literature. Protein features were assessed using data from NCBI (https://www.ncbi.nlm.nih.gov/ipg/NP_056862.1) and via sequence viewer with O36633.1 Human respiratory syncytial virus B1, (https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=O36633.1) The variant effect on protein was assessed for evidence about structure, family and domains, features, and location; including defining protein chain, topological domain, transmembrane, site, disulfide bonds, glycosylation, alternative sequence, mutagenesis, compositional bias, helix, turn, regions, and additional annotation notes on protein functions for positions that were; disordered, binding to host heparan sulfate, cleavage, helical, mature secreted glycoprotein G, extracellular, polar residues, cytoplasmic, and missing in secreted isoform.

**IFN response section?**

# 4 Results

## 4.1 Cohort characteristics

The INSPIRE cohort consisted of 1,949 enrolled infants (Figure 1). Of these, 1,220 ($\sim$ 63%) had $\geq$ 1 in-person respiratory illness visit(s). In 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). From the cohort, 344 RSV viral samples from 326 individuals were sequenced. There were 19 infants with RSV-positive PCR $\geq$ 15 days apart who met the a priori definition of prolonged infection with viral genetic analysis used to determine if these represented the same or new virus. Table 1 lists the cohort characteristics of infants with prolonged RSV infection compared with other RSV infection and the entire cohort. Prolonged infection was a priori defined as RSV sequence positive with $\geq$ 15 days between testing and meeting criteria for acute respiratory infection.

The relatively small sample size 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 the F and G glycoprotein.

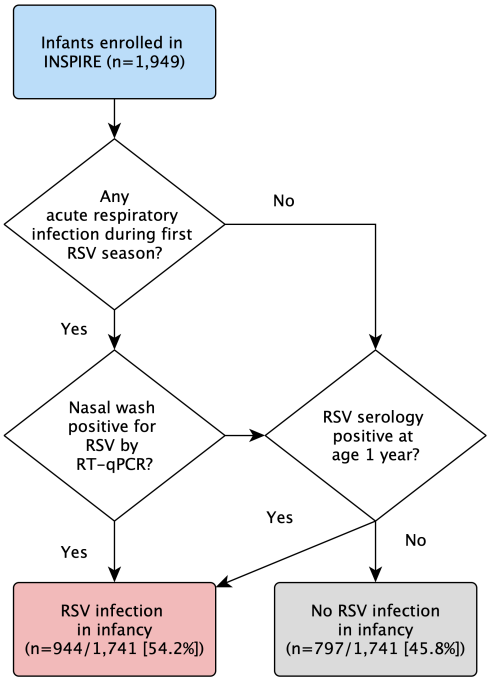


Figure 1: **Cohort characteristics of infants with prolonged RSV infection compared with other RSV infection and entire cohort**. Prolonged infection is defined as RSV sequence positive, with $\geq$ 15 days between testing and meeting criteria for acute respiratory infection. Respiratory severity score (median, IQR). Test statistic $P = 0.27$. Pearson, Wilcoxon.

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

Table 1: **Cohort characteristics of infants with prolonged RSV infection compared with other RSV infection and entire cohort**. Prolonged infection is defined as RSV sequence positive, with ≥15 days between testing. Respiratory severity score (median, IQR) Test statistic $P = 0.27$[1]. Pearson[1], Wilcoxon[2].

## 4.2 Host Genetic Analyses

We explored whether RSV infection in infancy (or the lack thereof) is a natural assignment (quasi-random) event and, unlike the severity of early-life RSV infection, [27] not determined by host genetics. For the candidate-SNP analysis, we considered childhood asthma- and RSV bronchiolitis-associated SNPs identified in Pividori et al. [13]; Janssen et al. [14]; Pasanen et al. [15]. The former is the largest childhood asthma GWAS to date, and, as far as we are aware, the latter 2 represent the most comprehensive studies of RSV bronchiolitis-associated SNPs. To further reduce the multiple testing burden, we only analyzed SNPs with MAF$\geq$ 0.1 in at least one of the White, Black, or Hispanic ethnicity groups. The associations between the genotype at the resulting 54 SNPs (50 childhood asthma- and 4 RSV bronchiolitis-associated SNPs) and RSV infection in infancy in our data are given in **Figure** 2 which suggests that the genotype at these SNPs have little to no effect on RSV infection in infancy. We further investigated the possibility that we were underpowered to observe associations with these SNPs by pooling information across SNPs to estimate the average genetic effect size. In brief, we computed a z-score for each SNP, where the average (across SNPs) squared z-score $\bar{G}$ is proportional to the average squared genetic effect on RSV infection in infancy. As $\bar{G}$ is an average of $p = 54$ approximately independent statistics, it is approximately $N(n\mu^2 + 1,2/p)$ where $n = 621$ is the sample size and $\mu^2$ is a function of the average squared genetic effect on RSV infection in infancy. Using the genetic effect estimates from Pividori et al. [13]; Janssen et al. [14]; Pasanen et al. [15] we calculated that we would have 80% power to reject the global null hypothesis of no genetic effect on RSV infection in infancy at any of these SNPs (i.e. $\mu^2 = 0$) if, on average across the 54 SNPs, the genetic effect on RSV infection in infancy was at least 61% as large as those estimated in the aforementioned 3 studies. We found $\bar{G}$=1.00 in our data, which corresponds to a p-value of 0.50. This result indicates that the genetic effect on RSV infection in infancy is zero or small at SNPs a priori likely to be associated with RSV infection.

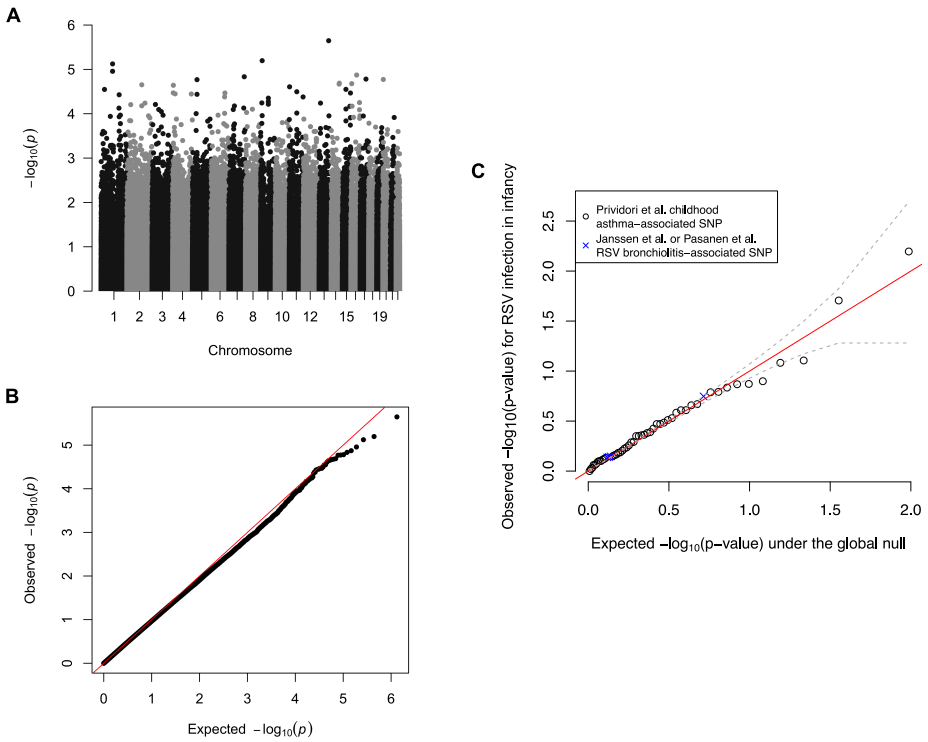


Figure 2: **Genetic analyses of RSV infection in infancy**. (A) The Manhattan plot shows no genome-wide significant associations (p-value threshold of $5e^{08}$). (B) The Q-Q plot demonstrates that the observed p-values are congruent with those expected under the null hypothesis that RSV infection in infancy is independent of genotype. (C) The association between the 54 selected childhood asthma- or RSV bronchiolitis-associated SNPs and RSV infection in infancy in our data. The solid red line is the identity line, and the dashed grey lines are $/pm1$ standard deviation around the expected -log10(p-value). The results suggest that the genotype at these SNPs have little to no effect on RSV infection in infancy. Definition of abbreviations: RSV = Respiratory syncytial virus, SNP = Single nucleotide polymorphism.

## 4.3 Population structure

A summary of protein coding genes in RSV is illustrated in Figure 3 A. Our analysis focused on G protein, as indicated. The phylogenetic tree based based on multiple sequence alignment (MSA) of amino acid G protein sequences is shown in Figure 3 B. 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 [28]. 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 repeat or prolonged infections by viruses from different phylogenetic clades, rather than one specific clade (Figure 3 C). A genotype correlation matrix and PCA eignenvalues were used for reducing the dimensionality of sequence data. 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 protein coding sequence separated the cohort. Three principal components (PC) are shown in Figure 3 C. For his reason, in our main analysis, viral population structure is accounted for by the first five PCs. To test for type I errors due 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 downstream.

## 4.4 Genetic invariance of prolonged infection

The duration of RSV shedding duration in Kenyan infants has been reported previously [29]. Based on these findings, infections separated by at least 15 days were expected to be "new" infections. Figure 3 D (panel [i]) summarises every pairwise genetic distance between every viral sequence. Sequence from the same clades have the smallest distance; panel [ii] shows genetic invariance between viral sequences within the same host for infections separated by at least 15 days. There was no genetic variation in repeat/prolonged viral sequence within individuals versus significantly increased genetic diversity between any of the most closely related sequences (P-value = 0.008). Panel [iii] shows distances between all possible pairs within clades (P-value = $1.3e^{-8}$). We therefore refer to these cases as prolonged infection rather than second infections.

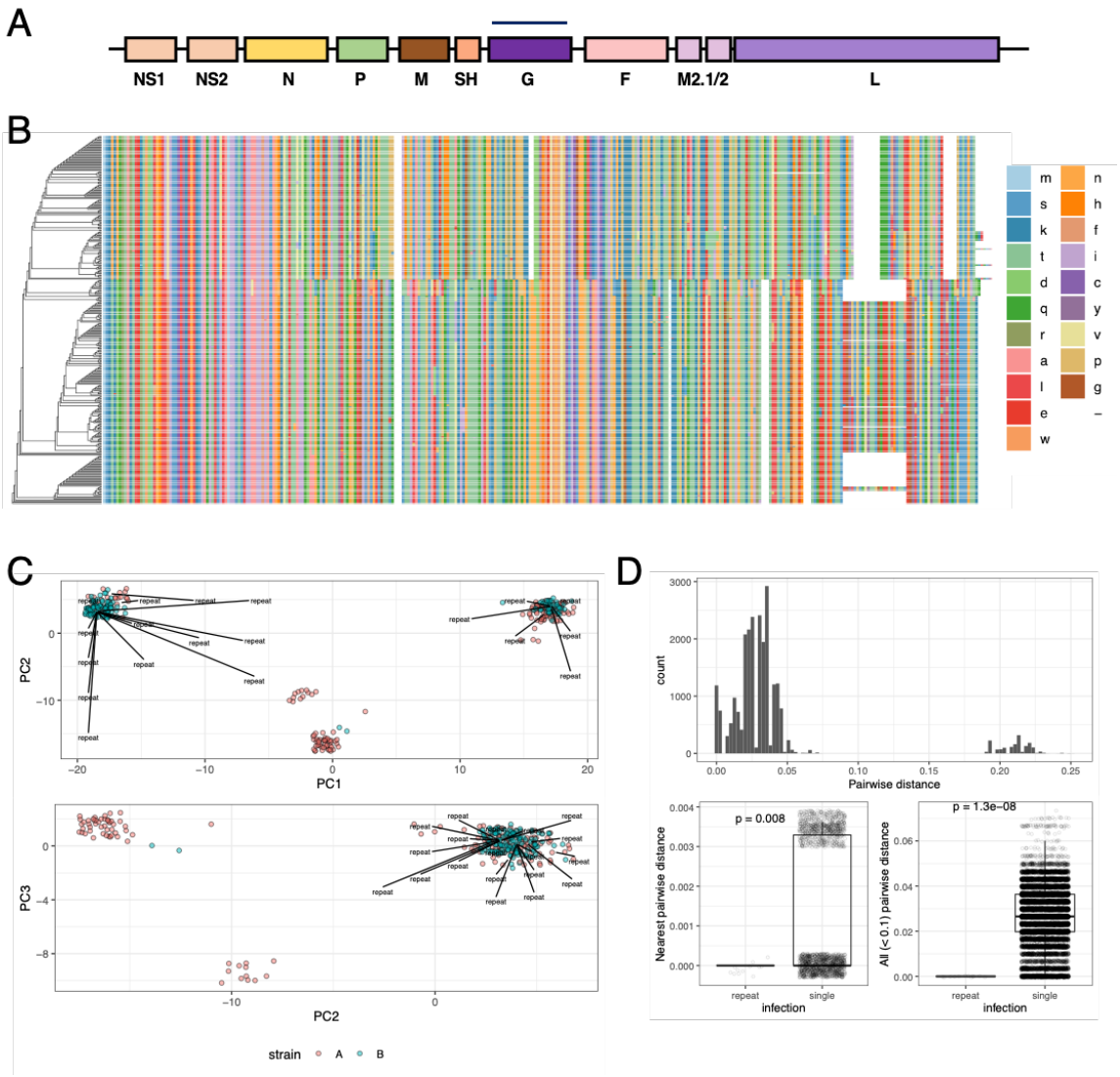


Figure 3: **Population structure**. (A) Protein coding genes in RSV. (B) Phylogenetic tree based based on multiple sequence alignment (MSA) of amino acid G protein sequences. (C) Principal component analysis (PCA) PCs1-3 with labels indicating repeat/prolonged infections from different phylogenetic clades. (D) Panel [i] summarises every pairwise genetic distance between every viral sequence. Genetic invariance in repeat/prolonged infections separated by at least 15 days compared to other genetic variation within clades (panel [ii]) and within all possible pairs (panel [iii]).

## 4.5 Variants in G glycoprotein significantly associated with prolonged 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 (prolonged infection Yes/No), and predictors; viral genotype (REF/ALT amino acid), viral PCs 1-5, host sex, and host features that have been previously demonstrated as significantly associated with infection; self-reported race/ethnicity, child-care attendance, or living with siblings [30]. Analysis was performed using R stats (3.6.2) *glm* function. A significant genetic association was identified for prolonged infection after Bonferroni correction for multiple testing (threshold for number independent variants $< 0.05/23 = 0.002$), as shown in Figure 4 A. Since many variants within RSV coding genes have non-random association due to selection, like linkage disequilibrium (LD) in human GWAS, we reduced the multiple testing burden by retaining proxy variants and removing those with $r^2 \geq 0.8$.

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. The same direction of effect indicated that the association was not a false positive, although the smaller sample size means that sub-analysis result no longer passes the significant threshold.

To assess the possibility of a false positive due to population structure within our cohort, we assessed the magnitude of variance explained (VE) at every amino acid position. Figure 4 B (panel [i]) shows the variance explained by each individual variant in PCs1-5. The values are illustrated according to protein position in panels [ii-iii]. The lead association variant had $-0.996\%$ VE for PC1 and $-1.66\%$ VE for PC2; a negligible effect that precludes spurious association by allele frequency between populations.

After identifying a significant association with prolonged infection, we quantified the correlation of variants with the lead proxy. Clumping was performed with ranking based on minor allele frequency (MAF) and with a cut-off threshold of $r^2 \geq 0.8$ (Supplemental Figure S1). The association model was repeated for all variants to produce a LocusZoom-style Manhattan plot with $r^2$ by color and P-value statistics as shown in Figure 4 C. This shows both G protein p.E123K/D and p.P218T/S/L as candidate causal variants associated with prolonged infection, and no other variants in correlation with this association.

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

### 4.6 Functional interpretation

The main features of RSV surface glycoprotein are illustrated in Figure 4 D. The variant associated with prolonged infection in our cohort are seen in the extracellular region. There are no known mechanistic features that directly overlap. Figure 4 D (structure) shows the possible site for initiation of infection by interaction of heparan sulfate and host cell membrane (p.187-198) [31–33]. Protein structure evidence from PDB was insufficient to determine an effect on conformation. Paramyxoviridae cell attachment proteins (G protein in RSV) span the viral envelope and project from the surface as spikes (1-43 cytoplasmic, 43-63 helical, 64-298 extracellular). Interactions have been identified with protein SH [34] and via the N-terminus with protein M [35] (Figure 4 D). G protein has been reported to form homo-oligomers (which we will check next for interaction residues. remove this citation if not fruitful) [36]. Known neutralization epitopes were not found at these positions. Figure 4 D (Features) shows the isoform of mature secreted glycoprotein G is reported for amino acid positions p.66 – 298, including the variants of interest.

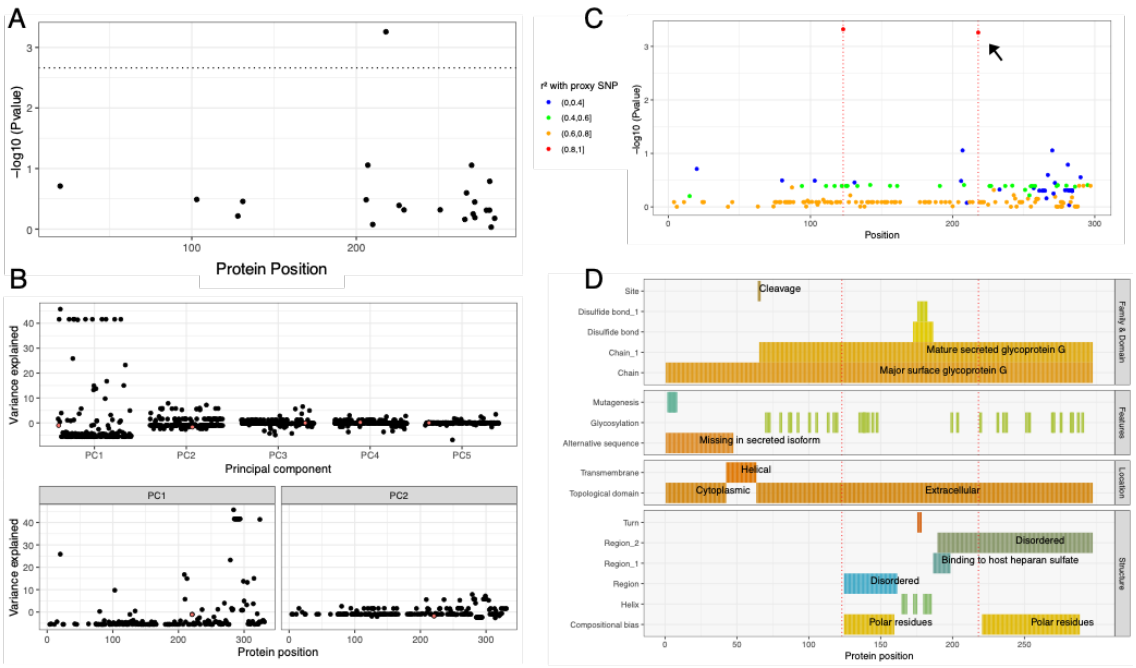


Figure 4: **Genetic association with prolonged infection**. (A) Amino acid association with prolonged infection after multiple testing correction (significant threshold shown by dotted line). (B) Variance explained (VE) within cohort. The effect of each variant on cohort structure is shown for PCs1-2. A large % VE for a significantly associated variant would indicate a false positive. (C) Variants in strong correlation were clumped for association testing using proxies for $r^2 \geq 0.8$. One significant association was identified (shown in A); the $r^2$ values for all other variants show a single highly correlated variant with the lead proxy (red). (D) Evidence for biological interpretation for every amino acid position is summarised

### 4.7 Clinical and interferon response

These variants were not associated with more severe infection in patients. Infections were on average less severe compared with other circulating variants, and all were upper respiratory tract infection (Table 1). We also assessed the association of nasal cytokines in nasal wash samples during acute infection as anti-viral immune response biomarkers with the variants of interest. IFN$\alpha$ and IFN$\gamma$ were associated with strain A/B. Since the viral strains correlated with the presence of the variants of interest, co-linearity meant that it was not possible to attribute these variants as the cause of a differential interferon response. However, the more pronounced antiviral response is unlikely to be due to the variants associated with prolonged infection and more likely due to other features that separate strain A and B; an observation that has been anecdotally reported previously (do we have a reference to strain B being more pathogenic than strain A? - rephrase if the reports are anecdotal or common knowledge).

## 5 Discussion

In this study of term healthy infants in whom we identified their first RSV infection through surveillance, we conducted a viral genetic association study and identified variants associated with prolonged infection. The variants were not associated with severe disease, had not arisen as a recent mutation, and have persisted in the population at a low prevalence for decades. We hypothesized that they might induce dampened anti-viral immune responses making it more difficult to be cleared. However, we were unable to demonstrate so using nasal cytokine data. Since these variants are co-linear with strain A and B, the differences in antiviral response are most likely strain-dependent rather than due to these particular variants. We do not expect any host immune memory before this first infection, potentially beyond maternal antibody.

While this study has a number of significant strengths, including being one of few population-based studies of the first RSV infection during infancy regardless of symptoms among a term healthy infant population, there are limitations which must be considered. This cohort was not designed to study persistence of infection, and repeat sampling following initial RSV infection was only done based on symptoms and likely missed some prolonged infections in this population. This resulted in a small number of prolonged infections to study. The small sample size also required focus on surface proteins for their prior probability of association and interpretation. A larger cohort may in future perform a whole genome analysis.

A host genetic interaction for asthma has been demonstrated previously [37]. 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.

Genomic analysis of both human and pathogen in tandem is being increasingly adopted. Application of the same statistical methods that are commonly used for human GWAS while extending the boundary from host genome to pathogen genome can allow for robust analysis of novel genotype-phenotype associations. In this study we combined the outcome of host phenotype and covariates with the pathogen genotype as a predictor. Cohort sizes as little as 2-3 times larger than that reported here are capable of detecting signals in genome-to-genome analysis [38]. While we have not overtly assessed patients for rare monogenic variants that may cause an underlying immunodeficiency, our enrollment criteria included only patients who were otherwise healthy. We have also previously demonstrated that there is no evidence for host genetic susceptibility due to common variants for risk of infection with RSV during infancy (cite). Accounting for host genetic factors allowed our analysis to focus on the viral genetic features which may 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 [39].

We suspected that the variants of interest 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 *de novo* mutation may allow variants to present within some individuals but failing to persist within the population, however, we have not been able to conclusively assess this possibility. Why these variants, which appear quite stable, have persisted at low frequency in the population for decades is uncertain. It is possible that a virus that persists or is cleared less rapidly may have a greater impact on epigenetic changes and reprogramming of the developing airway epithelium, or results in a low-level chronic stimulation or immune exhaustion. We have previously demonstrated that infants infected with RSV in their first year of life have dampened subsequent anti-viral immune responses in early childhood (cite Frontiers in Immunol Chirkova T et al. under review) as well as changes in airway epithelial cell metabolism [26].

Known neutralization epitopes were not found for our variant site (Jim). Attachment of the virion to the host cell membrane is thought to occur through interaction with heparan sulfate as shown in Figure 4 (p.187-198), thereby initiating infection [31–33]. Specifically, interactions between viral G protein and host CX3CR1, the receptor for the CX3C chemokine fractalkine, have been reported to modulate the immune response and facilitate infection [40–42]. CX3CR1 is well known as a coreceptor for HIV-1. Variations in *CX3CR1* are known to have important effects on the susceptibility to HIV-1 infection and the hosts' potential for controlling infection [cite, ask Jacques]. In general, viral envelope glycoproteins bind to specific cellular receptors and initiate fusion with the host cell membrane, which allows the penetration of the viral genome into host cells. The negative-strand RNA family of Paramyxoviridae rely on a pair of binding and fusion functions for infection, mediated by one or multiple envelope glycoproteins, which span the viral envelope and project from the surface as spikes (p.1-43 cytoplasmic, 43-63 helical, 64-298 extracellular). These proteins are generally desig-

nated as either hemagglutinin (H), hemagglutinin-neuraminidase (HN), or glycoprotein (G). Haemagglutination activity which is responsible for the binding of virus (i.e. Influenza) to sialic acid on the surface of target cells. Haemagglutination and neuraminidase (HN) activity cleaves sialic acid on the cell surface, preventing viral particles from reattaching to previously infected cells. Unlike the other paramyxovirus attachment proteins, RSV glycoprotein lacks hemagglutinin-neuraminidase (HN) activities. It uses the attachment protein with neither haemagglutination nor neuraminidase activity, designated as G (glycoprotein), which are found in henipaviruses. [43–47] The mature isoform of secreted glycoprotein G is 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 [48]. This secreted isoform is reported for amino acid positions p.66 – 298, and includes the variants associated with prolonged infection in our analysis.

Increasingly efficient vaccine development combined with the growing availability of large genomic and functional data sources provides opportunities for controlling the severity of illness, delaying age of infection, and preemptively monitoring for increased viral pathogenicity. As a disease which causes significant morbidity and mortality, each advance in preventative measures can have lasting consequences for global health.

# 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 for 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/

GCTA https://cnsgenomics.com/software/gcta/

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

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

MAFFT https://mafft.cbrc.jp/alignment/software/ [24]

NextAlign https://github.com/nextstrain/nextclade

PLINK http://zzz.bwh.harvard.edu/plink/

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

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

RCSB PDB https://www.rcsb.org

UniProt https://www.uniprot.org

### 6.2 Data sources

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.

GenBank:NC_001989 Bovine orthopneumovirus, complete genome https://www.ncbi.nlm.nih.gov/nuccore/NC_001989.

Reference data https://www.ncbi.nlm.nih.gov/gene/?term=1489824. G attachment glycoprotein [Human orthopneumovirus]; ID: 1489824; Location: NC_001781.1 (4675..5600); Aliases: HRSVgp07.

Reference data https://www.ncbi.nlm.nih.gov/gene/?term=37607642. G attachment glycoprotein [Human orthopneumovirus]; ID: 37607642; Location: NC_038235.1 (4673..5595); Aliases: DZD21_gp07.

Reference data for all public NCBI Virus https://www.ncbi.nlm.nih.gov/labs/virus/vssi/ for species: Human orthopneumovirus; genus: orthopneumovirus; family: Pneumoviridae.

Reference data https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Human%20orthopneumovirus,%20taxid:11250 - contains sequence data for Virus Lineage ss=Human orthopneumovirus, taxid:11250 nucleotide: 26'965, protein: 53'804, RefSeq Genomes: 2.

Reference https://www.ncbi.nlm.nih.gov/protein/NP_056862.1

GCF_002815475.1 (release 2018-08-19) Nucleotide Accessions: NC_038235.1, protein: Y_009518856.1

Reference https://www.ncbi.nlm.nih.gov/protein/YP_009518856.1

GCF_000855545.1 (release 2015-02-12) Nucleotide Accessions: NC_001781.1, protein: NP_056862.1 (strain B1).

## 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 [49]. 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 [50].

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 [51]. Datasets were merged using PLINK v1.9. SNP positions and identifiers were updated according to dbNSFP4.0a (hg19) [52]. 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] Emma K Larkin, Tebeb Gebretsadik, Martin L Moore, Larry J Anderson, William D Dupont, James D Chappell, Patricia A Minton, R Stokes Peebles, Paul E Moore, Robert S Valet, et al. Objectives, design and enrollment results from the infant susceptibility to pulmonary infections and asthma following rsv exposure study (inspire). *BMC pulmonary medicine*, 15(1):1–12, 2015.

[13] Milton Pividori, Nathan Schoettler, Dan L Nicolae, Carole Ober, and Hae Kyung Im. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *The Lancet Respiratory Medicine*, 7(6): 509–522, 2019.

[14] Riny Janssen, Louis Bont, Christine LE Siezen, Hennie M Hodemaekers, Marieke J Ermers, Gerda Doornbos, Ruben van't Slot, Ciska Wijmenga, Jelle J Goeman, Jan LL Kimpen, et al. Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. *Journal of Infectious Diseases*, 196(6):826–834, 2007.

[15] Anu Pasanen, Minna K Karjalainen, Louis Bont, Eija Piippo-Savolainen, Marja Ruotsalainen, Emma Goksör, Kuldeep Kumawat, Hennie Hodemaekers, Kirsi Nuolivirta, Tuomas Jartti, et al. Genome-wide association study of polymorphisms predisposing to bronchiolitis. *Scientific reports*, 7(1):1–9, 2017.

[16] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[17] 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.

[18] 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.

[19] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Andrew R Connelly, Brian M Jeong, Mackenzie E Coden, Jacob Y Cao, Tatiana Chirkova, Christian Rosas-Salazar, Jacqueline-Yvonne Cephus, Larry J Anderson, Dawn C Newcomb, Tina V Hartert, et al. Metabolic reprogramming of nasal airway epithelial cells following infant respiratory syncytial virus infection. *Viruses*, 13(10):2055, 2021.

[27] Emma K Larkin and Tina V Hartert. Genes associated with rsv lower respiratory tract infection and asthma: the application of genetic epidemiological methods to understand causality. *Future virology*, 10(7):883–897, 2015.

[28] 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.

[29] 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.

[30] Caroline Breese Hall, Joyce M Geiman, Robert Biggar, David I Kotok, Patricia M Hogan, and R Gordon Douglas Jr. Respiratory syncytial virus infections within families. *New England journal of medicine*, 294(8):414–419, 1976.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Jacques Fellay and Vincent Pedergnana. Exploring the interactions between the human and viral genomes. *Human genetics*, 139(6):777–781, 2020.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] Toru Takimoto, Garry L Taylor, Helen C Connaris, Susan J Crennell, and Allen Portner. Role of the hemagglutinin-neuraminidase protein in the mechanism of paramyxovirus-cell membrane fusion. *Journal of virology*, 76(24):13028–13033, 2002.

[44] Etienne Malvoisin and T Fabian Wild. Measles virus glycoproteins: studies on the structure and interaction of the haemagglutinin and fusion proteins. *Journal of General Virology*, 74(11):2365–2372, 1993.

[45] XL Hu, Ranjit Ray, and Richard W Compans. Functional interactions between the fusion protein and hemagglutinin-neuraminidase of human parainfluenza viruses. *Journal of Virology*, 66(3):1528–1534, 1992.

[46] CM Horvath, RG Paterson, MA Shaughnessy, R Wood, and RA Lamb. Biological activity of paramyxovirus fusion proteins: factors influencing formation of syncytia. *Journal of virology*, 66(7):4564–4569, 1992.

[47] Tatiana Bousse, Toru Takimoto, Wendy L Gorman, Tatsufumi Takahashi, and Allen Portner. Regions on the hemagglutinin-neuraminidase proteins of human parainfluenza virus type-1 and sendai virus important for membrane fusion. *Virology*, 204(2):506–514, 1994.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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 Supplemental

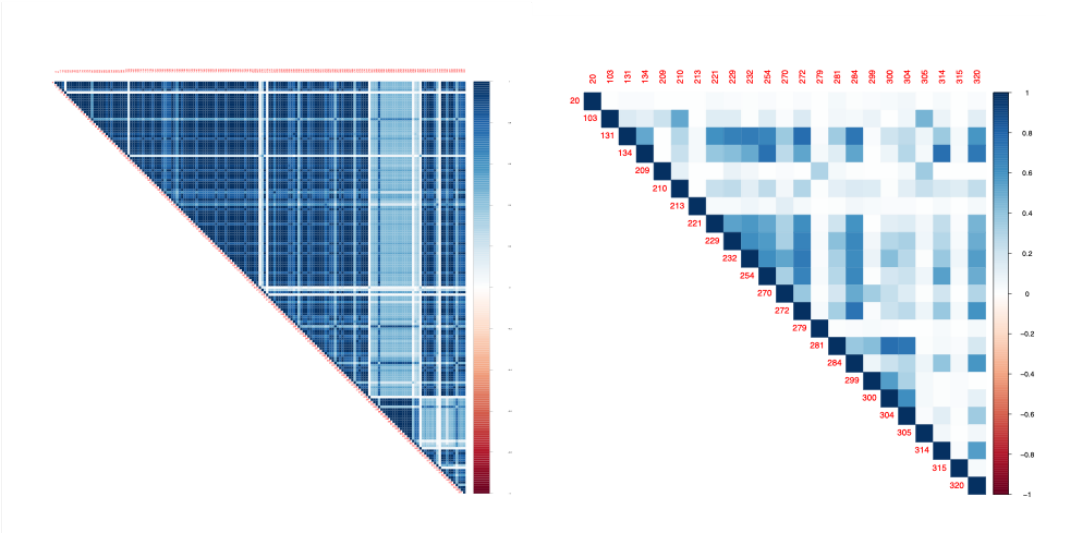


Figure S1: **Supplemental: Variant clumping for reduction in association testing**. [Left] Correlation between all positions. [Right] Correlation between proxy variants are clumping to remove $r^2 \geq 0.8$.

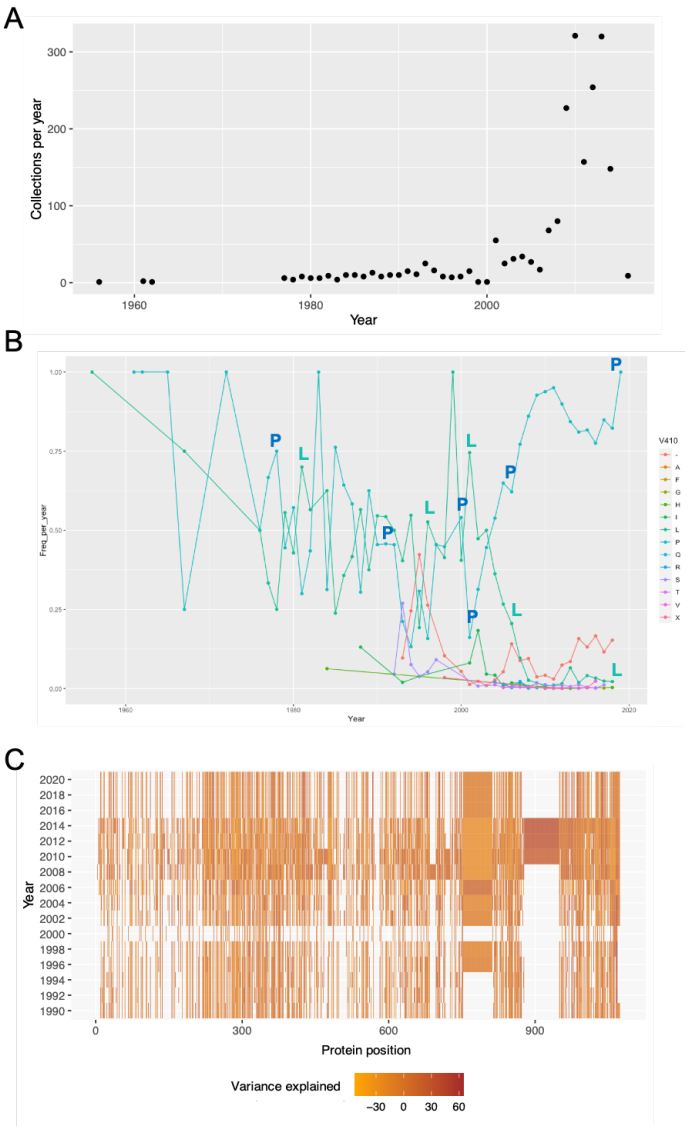


Figure S2: **Supplemental: Publicly available RSV sequence data for $>$ 30 years.** (A) Global sample collection per year. (B) Variant associated with prolonged infection tracked in public data. (C) % variance explained per year for all G protein amino acid variants from 1990-2022