

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

Dylan Lawless, PhD¹, Christopher G. McKennan, PhD², Suman Das, PhD³, Thomas Junier, PhD⁴, Zhi Ming Xu, MPhil¹, Larry J Anderson, MD⁵, Tebeb Gebretsadik, MPH⁶, Meghan Shilts, MHS, MS⁷, Emma Larkin, PhD⁸, Christian Rosas-Salazar, MD, MPH⁸, James D. Chappell, MD⁹, Jacques Fellay, MD, PhD^{1,10}, and Tina V. Hartert, MD, MPH^{7,9}

¹Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, ²Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, ³Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, ⁴Swiss Institute of Bioinformatics, Vital-IT Group, Switzerland, ⁵Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, United States of America, ⁶Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, ⁷Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, ⁸Division of Allergy, Immunology, and Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, ⁹Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, ¹⁰Biomedical Data Science Center, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland.

¹ Running title

² RSV variants and prolonged infection

³ Key points

⁴ Using a comprehensive computational analysis of viral and host genetics we identified a novel
⁵ RSV variant associated with prolonged infection and no evidence supporting host genetic
⁶ infection susceptibility, findings important to understanding RSV contribution to chronic
⁷ disease and viral endemicity.

⁸ Abbreviations

⁹ ALT (alternative); CI (confidence interval); GWAS (genome-wide association study); G (gly-
¹⁰ coprotein); H (hemagglutinin); HN (hemagglutinin-neuraminidase); IFN (interferon); IQR (in-

11 terquartile range); INSPIRE (The INFant Susceptibility to Pulmonary Infections and Asthma
12 Following RSV Exposure); LD (linkage disequilibrium); LRTI (lower respiratory tract infec-
13 tion); MAF (minor allele frequency); MFI (median fluorescence intensity); MSA (multiple
14 sequence alignment); OR (odds ratio); PCR (polymerase chain reaction); PCA (Principal
15 component analysis); REF (reference); RT (reverse transcription); SVD (singular value decom-
16 position); SNP (single nucleotide polymorphism); VE (variance explained); MSA (multiple
17 sequence alignment); RSV (respiratory syncytial virus).

18 **Notice of Prior Presentation**

19 The results of the host genome wide association study analyses included in this manuscript
20 were presented during the European Society of Human Genetics Conference in June 2020 in
21 Berlin, Germany, which was held remotely [1].

22 **Ethics Statement for Human Subjects Research**

23 The protocol and informed consent documents were approved by the Institutional Review
24 Board at Vanderbilt University Medical Center (#111299). One parent of each participant
25 in the cohort study provided written informed consent for participation in this study. The
26 informed consent document explained study procedures and use of data and biospecimens for
27 future studies, including genetic studies.

28 **Competing interests**

29 All authors have completed a conflict of interest form (COI). Summary of any COI: There
30 were no COI. Funding was supplied from National Institutes of Health and Swiss National Sci-
31 ence Foundation. U19 AI 095227, UG3/UH3 OD023282, UL1 TR002243, SNSF IZSEZ0_191968
32 (TVH), SNSF 310030L_197721 (JF), X01 HLG244 RS&G (EL).

33 **Summary**

34 A comprehensive computational statistical analysis of both host and viral genetics provided
35 compelling evidence for RSV viral persistence in healthy human infants, a finding of signifi-
36 cant importance to understanding the impact of RSV on chronic disease and viral endemicity.

37 **1 Abstract**

38 **Background:** Respiratory syncytial virus (RSV) is primarily associated with acute respiratory
39 infection. However, many RNA viruses establish persistent infection. The objective was
40 to identify RSV variants associated with prolonged infection.

41 **Methods:** Among healthy term infants we identified those with prolonged RSV infection and
42 conducted 1) a viral GWAS using RSV whole-genome sequencing to determine the relationship
43 between viral genotypes and prolonged infant RSV infection, 2) a human GWAS to test
44 the dependence of first year RSV infection risk on the genotype, 3) an analysis of all viral public
45 sequence data, 4) an assessment of the local immunological responses, and 5) a summary of
46 all the major functional data for the identified viral variant. Analyses were adjusted for viral
47 and human population structure and host factors associated with infection risk.

48 **Results:** We identified two variants, p.E123K/D and p.P218T/S/L, in the RSV G protein
49 that were associated with prolonged infection after correction for multiple testing (p_{adj} value
50 = 0.01). We found no evidence of host genetic risk for RSV infection. The RSV variant positions
51 approximate sequences that could bind a putative viral receptor, heparan sulfate.

52 **Conclusions:** Using an in-depth comprehensive computational analysis of both viral and
53 host genetics we identified a novel RSV viral variant associated with prolonged infection in
54 healthy infants and no evidence supporting host genetic susceptibility to RSV infection. As
55 the capacity of RSV for chronicity and its viral reservoir are not defined, these findings are
56 important to understanding the impact of RSV on chronic disease and viral endemicity.

57 **2 Introduction**

58 Human orthopneumovirus, formerly known (and frequently still referred to) as Respiratory
59 syncytial virus (RSV), results in significant global morbidity and mortality [2]. By the age of
60 two to three years, nearly all children have been infected with RSV at least once [3]. RSV is a
61 seasonal mucosal pathogen that primarily infects upper and lower respiratory tract epithelium,
62 although it has been recovered from non-airway sources [4–9]. While RSV is mainly associated
63 with acute respiratory infection, many RNA viruses can establish prolonged or persistent infection
64 in some infected individuals [10]. Prolonged shedding of RSV, especially in young infants
65 and following first infection, has been demonstrated, with longer average duration of viral
66 shedding when polymerase chain reaction (PCR) is used to detect RSV [11]. While younger
67 age and first infection are associated with protracted infection [3; 12], it is not known whether
68 specific viral factors contribute to prolonged RSV infection in infants. This is important, as
69 prolonged infection may contribute to enhanced transmission and developmental changes to
70 the early life airway epithelium. Further, the reservoir of RSV infection is not understood,

71 and it is possible that some RSV strains sustain a low level of ongoing viral circulation in the
72 community until seasonal or other influences favor epidemic spread [13].

73 The objectives of this study were therefore to determine if there exist host genetic risk
74 alleles for RSV infection and to identify viral genetic variation associated with prolonged
75 infection. These motivating questions are of fundamental interest in understanding viral and
76 host genetic contributions that may underlie the development of chronic respiratory morbidity
77 due to RSV, including asthma.

78 **3 Methods**

79 **3.1 Study population**

80 The protocol and informed consent documents were approved by the Institutional Review
81 Board at Vanderbilt University Medical Center (#111299). One parent of each participant
82 in the cohort study provided written informed consent for participation in this study. The
83 informed consent document explained study procedures and use of data and biospecimens for
84 future studies, including genetic studies.

85 Among healthy term infants in a cohort specifically designed to capture first RSV infec-
86 tion we identified those with prolonged RSV infection and conducted 1) a viral GWAS using
87 RSV whole-genome sequencing to determine the relationship between viral genotypes and
88 prolonged infant RSV infection, 2) a human GWAS to test the dependence of first year RSV
89 infection risk on the genotype, 3) an analysis of all viral public sequence data, 4) an assess-
90 ment of the local immunological RSV responses, and 5) a summary of all the major functional
91 data for the identified viral variant. Full details of the methods are included in the Supple-
92 ment, sections 8.1 - 8.13.

93 **4 Results**

94 **4.1 Cohort characteristics**

95 The INSPIRE cohort consisted of 1,949 enrolled infants among whom there were 2,093 in-
96 person respiratory illness visits completed during winter virus season, November - March, of
97 each year (Figure S1); the median (interquartile range [IQR]) number of in-person respira-
98 tory illness visits per infant during this surveillance window was 1 [1; 2]. There were 344 RSV
99 PCR-positive samples from 325 individuals which were sequenced. Prolonged infection was *a*
100 *priori* defined as repeatedly meeting criteria for acute respiratory infection with RSV PCR

¹⁰¹ positive nasal samples \geq 15 days between testing. There were 19 infants who met the defini-
¹⁰² tion of prolonged infection with available viral sequencing used to confirm clonality of original
¹⁰³ and subsequent virus detections. The mean RSV CT value of first infections was 25.9 ± 7.1 ,
¹⁰⁴ and second detection was 31.6 ± 5.4 . The mean number of days between detections was 25
¹⁰⁵ \pm 25 days (Figure S2). Table 1 lists the cohort characteristics of infants with prolonged RSV
¹⁰⁶ infection compared with other RSV infection and the entire cohort.

	Prolonged RSV infection N=19	RSV infection N=342	Total N=1949
Age in months at first illness (median, IQR)	6 (4, 6)	4 (2, 5)	NA
Illness respiratory severity score (median, IQR)	2.0 (1.2, 3.0)	3.0 (2.0, 4.0)	NA
RSV season			
2012-13	68%	54%	44%
2013-14	32%	46%	56%
Self reported race			
Non-Hispanic Black	37%	13%	18%
Non-Hispanic White	63%	69%	65%
Hispanic	0%	10%	9%
Multi-race/ethnicity/other	0%	8%	8%
Female sex	53%	44%	48%
Second-hand smoke exposure	21%	23%	47%
Health insurance			
Medicaid	68%	48%	54%
Private	32%	51%	45%
None/unknown	0%	1%	1%
Daycare and/or siblings	84%	78%	66%

Table 1: **Characteristics of infants with prolonged RSV infection compared with other RSV infection and the entire cohort.** Prolonged infection is defined as repeatedly RSV PCR-positive with ≥ 15 days between testing and meeting criteria for acute respiratory infection. *Presence of sibling or another child ≤ 6 years of age at home.

107 **4.2 Host genetic analyses**

108 We explored whether RSV infection in infancy is a natural assignment (quasi-random) event
109 and, unlike severity of early-life RSV infection [14], occurs independently of host genetics.
110 For the candidate SNP analysis, we considered childhood asthma- and RSV LRTI-associated
111 SNPs identified in Pividori et al. [15]; Janssen et al. [16]; Pasanen et al. [17]. The first is the
112 largest childhood asthma GWAS to date, and, to our knowledge, the latter 2 represent the
113 most comprehensive studies of RSV LRTI-associated SNPs. To further reduce the multiple
114 testing burden, we only analysed SNPs with MAF ≥ 0.1 in at least one of the White, Black,
115 or Hispanic ethnicity groups. Associations between genotype at the resulting 54 SNPs (50
116 childhood asthma- and 4 RSV LRTI-associated SNPs) and RSV infection in infancy in our
117 data are given in Figure 1. The data are consistent with little to no effect of genotype at these
118 SNPs on RSV infection in infancy.

119 We further investigated the possibility that the analysis was underpowered to identify
120 associations with these SNPs by pooling information across SNPs to estimate the average
121 genetic effect size [18]. We estimated the narrow-sense heritability of RSV infection during
122 infancy on the latent liability scale (h_l^2), which, if > 0 , would indicate an accumulation of
123 small genetic effects. We estimated h_l^2 to be exactly 0, suggesting that, if present, infant RSV
124 infection-related genetic signals are both small and sparse (Supplemental section 8.6).

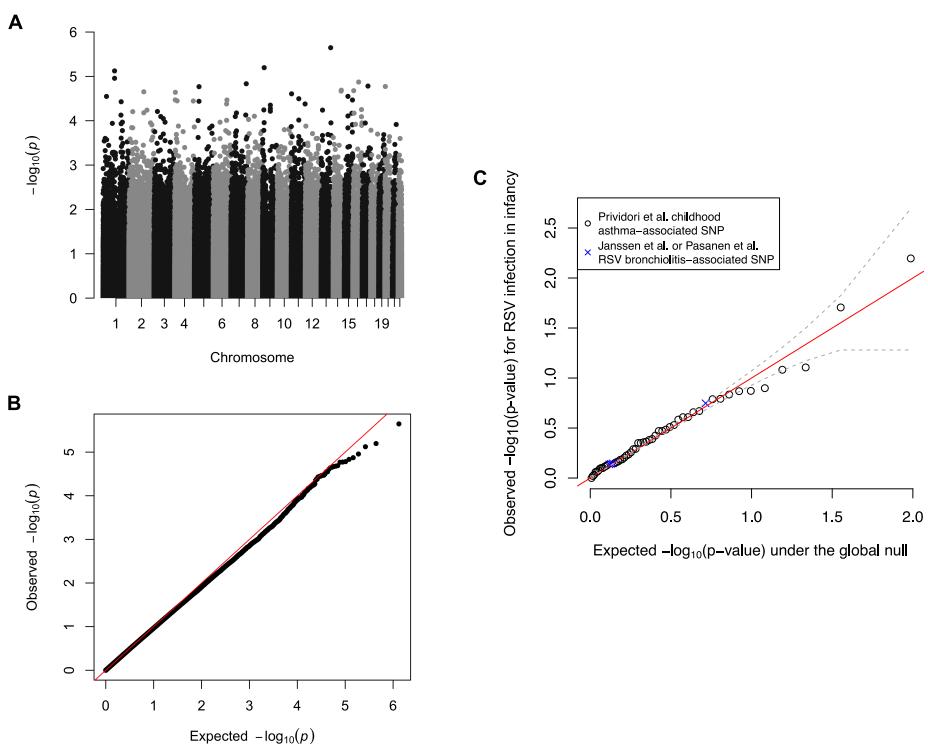


Figure 1: **Genetic analyses of RSV infection in infancy.** (A) The Manhattan plot shows no genome-wide significant associations (p value threshold of $5e^{-8}$). (B) The Q-Q plot demonstrates that the observed p values are congruent with those expected under the null hypothesis that RSV infection in infancy is independent of host genotype. (C) The association between the 54 selected childhood asthma- or RSV LRTI-associated SNPs and RSV infection in infancy in our data. The identity line is shown in red, and the dashed grey lines are ± 1 standard deviation around the expected $-\log_{10}(p$ value). RSV: respiratory syncytial virus; SNP: single nucleotide polymorphism.

125 **4.3 Population structure**

126 summary of protein coding genes in RSV is illustrated in Figure 2 A. Our analysis focused
127 on F and G protein. The phylogenetic tree based on multiple sequence alignment (MSA) of
128 G protein amino acid sequences is shown in Figure 2 B. One obvious feature causing a sepa-
129 ration in genetic diversity is G protein partial gene duplication, which has emerged in recent
130 years within RSV-A strains [19]. RSV-B strains with a homologous duplication have existed
131 for two decades, although the selection process leading to emergence and clinical implications
132 have not been entirely defined.

133 PCA was used for reducing the dimensionality of sequence data, where PC1 accounted for
134 95.19% of cumulative variance, and variance attributed to other PCs was roughly uniformly
135 distributed (Figure 2 C). We observed prolonged infections by viruses from different phyloge-
136 netic clades, rather than one specific clade (Figure 2 C), indicating that these results are not
137 confounded by latent clade membership.

138 **4.4 Genetic invariance of prolonged infection**

139 The duration of RSV shedding in Kenyan infants has been reported previously [13]. Based on
140 these findings, infection events separated by at least 15 days with symptoms were expected to
141 be “new” infections [13]. Figure 2 D panel [i] summarizes every pairwise genetic distance be-
142 tween every viral sequence, where small distances indicate pairs with closely related sequences.
143 Panels [ii] and [iii], which summarize the difference in sequence similarity distributions be-
144 tween viruses from the same host and different hosts, show that RSV sequences corresponding
145 to initial and subsequent viral detections are nearly identical. These results support the con-
146 clusion that such cases are prolonged (i.e., failure to clear) infections rather than new infec-
147 tions.

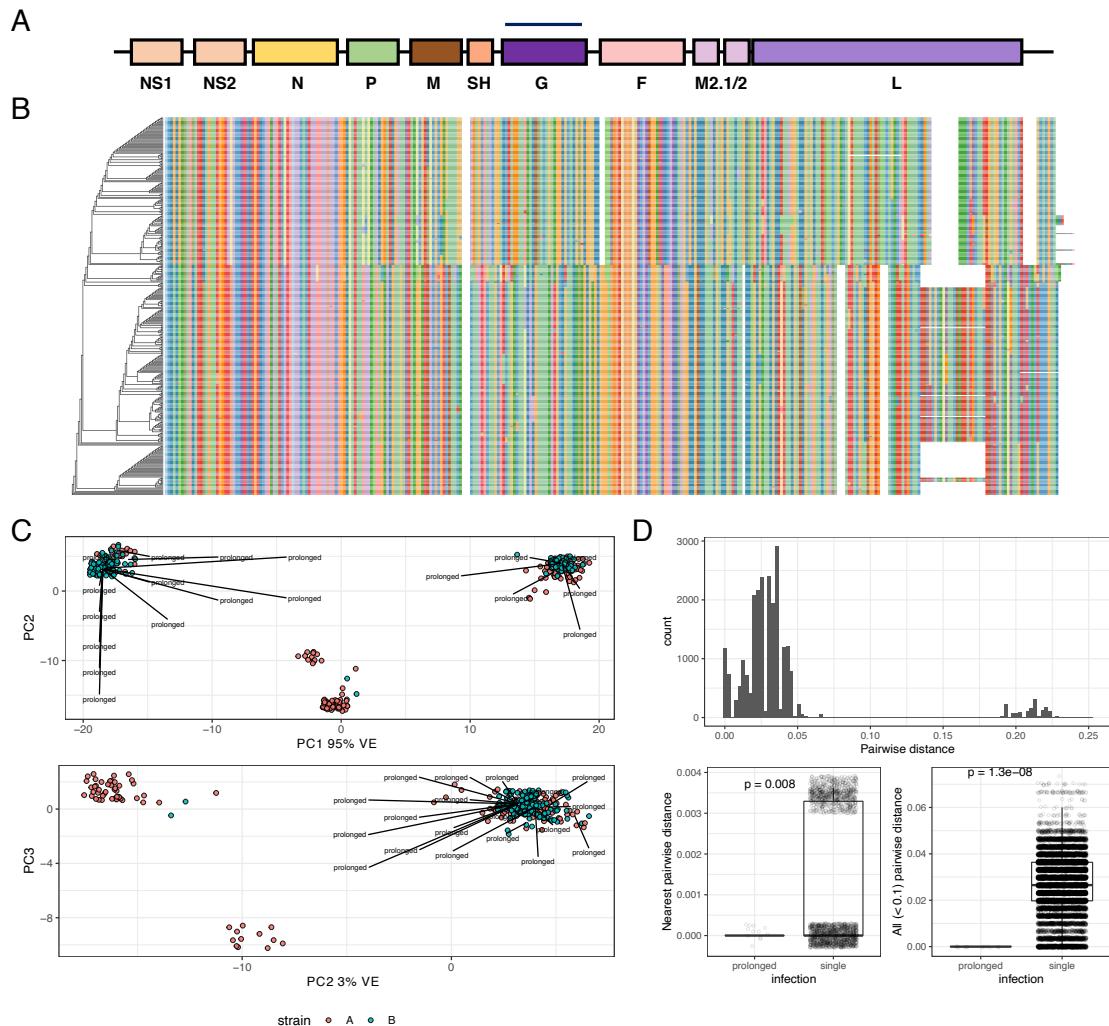


Figure 2: Viral population structure. (A) Linear map of the RSV genome. (B) Phylogenetic tree based on multiple sequence alignment MSA of G protein amino acid sequences. Color; amino acids. (C) Principal component (PC) analysis. PCs1-3 with labels indicating prolonged infections from different phylogenetic clades. (D) Panel [i] summarises every pairwise genetic distance between every viral sequence. Genetic invariance in prolonged infections separated by at least 15 days compared to other genetic variation within the most closely related sequences (panel [ii]) and within all possible closely related pairs (panel [iii]). VE (variance explained). Jitter applied for visualisation.

148 **4.5 Variants in G glycoprotein significantly associated with prolonged**
149 **infection**

150 The consensus sequence within the cohort was assigned based on the major allele. Variants at
151 the amino acid level were defined as either reference (REF) or alternative (ALT) and assessed
152 for their association with prolonged infection. The model consisted of (A) the binary response
153 (prolonged infection Yes/No), and (B) predictors; (1) viral genotype (REF/ALT amino acid),
154 (2) viral PCs 1-5, (3) host sex, and host features that have been previously demonstrated
155 as significantly associated with infection; (4) self-reported race/ethnicity, (5) child-care at-
156 tendance, or living with another child \leq 6 years of age at home [20]. A significant genetic
157 association was identified between prolonged infection and the lead variant after Bonferroni
158 correction for multiple testing (threshold for number independent variants $< 0.05/23 = 0.002$),
159 as shown in Figure 3 A, p value = 0.0006.

160 To determine whether this association was simply due to population stratification between
161 strains A and B, a subset analysis was performed using independently assessed clinical labo-
162 ratory strain labels for A and B. The same direction of effect indicated that the association
163 was not a false positive, although in this significantly smaller sub-analysis the result was not
164 significant.

165 To assess the possibility of a false positive association due to population structure within
166 our cohort, we assessed the magnitude of variance explained (VE) at every amino acid posi-
167 tion. Figure 3 B (panel [i]) shows the variance explained by each amino acid in PCs1-5. The
168 cumulative proportion of variance for PCs 1-5 was 99.5% (PC1 = 95%, PC2 = 3%). The val-
169 ues are illustrated according to protein position in panels [ii-iii]. The lead association variant
170 had 0.603% VE for PC1 and 0.458% VE for PC2, a negligible effect that precludes spurious
171 association by allele frequency between populations.

172 After identifying a significant viral genetic association with prolonged infection, we quan-
173 tified the correlation of variants with the lead proxy. Clumping was performed with ranking
174 based on MAF and with a cut-off threshold of $r^2 \geq 0.8$ (Supplemental Figure S3). The as-
175 sociation model was repeated for all variants, defining protein p.E123K/D and p.P218T/S/L
176 as candidate causal variants associated with prolonged infection as shown in Figure 3 C. No
177 other variants were correlated with this outcome.

178 To determine whether p.E123K/D and p.P218T/S/L variant genotypes are novel and po-
179 tentially influence viral fitness, we searched the public viral data repository of NCBI Human
180 orthopneumovirus, taxid:11250, which contained data from 27 countries worldwide, sample
181 collection dates from 1956 onward, and 1084 glycoprotein protein sequences after curation.
182 The variants were present at a low and stable frequency, without obvious temporal enrichment
183 (Supplemental Figure S4). Thus, while historical data reveal no positive selective advantage
184 attached to p.E123K/D and p.P218T/S/L, longstanding circulation and linkage in prolonged

¹⁸⁵ RSV infection suggest that these polymorphisms are present in the viral inoculum and do not
¹⁸⁶ arise through recurrent mutational events.

¹⁸⁷ Due to multiple testing correction according to our analysis plan, an association also orig-
¹⁸⁸ inally identified in F protein was rejected and therefore omitted from further discussion. For
¹⁸⁹ posterity, the variant position was p.N116S (relative to strain A GenBank: AMN91253.1).

¹⁹⁰ **4.6 Functional interpretation**

¹⁹¹ Cell-attachment proteins of paramyxoviruses (G protein in RSV) span the viral envelope
¹⁹² and form spike-like projections from the virion surface. RSV G protein is a type II integral
¹⁹³ membrane protein consisting of 298 amino acid residues comprising N-terminal cytoplasmic
¹⁹⁴ (p.1-43), transmembrane helical (p.43-63), and extracellular (p.64-298) domains (Figure 3 D).
¹⁹⁵ RSV G protein ectodomain also exists in a soluble secreted form, p.66 – 298, which functions
¹⁹⁶ in immune evasion [21–23]. G protein interacts with the small hydrophobic (SH) protein [24]
¹⁹⁷ and, via the N-terminus, with matrix (M) [25] protein. It has also been reported to form
¹⁹⁸ homo-oligomers [26]. The variant amino acid positions associated with prolonged infection
¹⁹⁹ reside in a portion of the G protein ectodomain of unassigned specific function and linearly
²⁰⁰ non-contiguous with sequences that bind cell-surface heparan sulfate, which likely promotes
²⁰¹ RSV cell-attachment (p.187-198) [21–23]. In addition, these positions do not contribute to
²⁰² known neutralization epitopes on G protein. Information available in PDB was insufficient
²⁰³ to infer effects of p.E123K/D and p.P218T/S/L on local or regional protein structure. The
²⁰⁴ potential effect on glycosylation is indeterminate. Figure 3 D illustrates the position of these
²⁰⁵ variants relative to summarised known functional features.

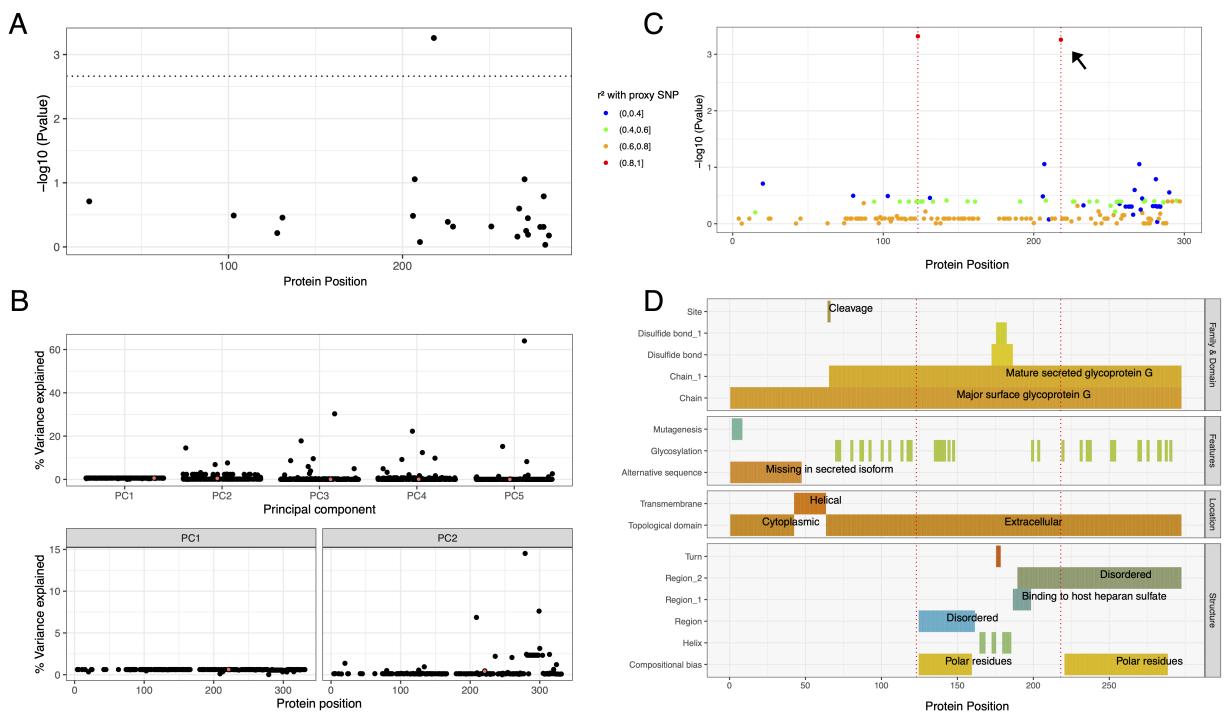


Figure 3: Viral genetic association with prolonged infection. (A) Amino acid association with prolonged infection after multiple testing correction (significant threshold shown by dotted line). (B) Variance explained (VE) within cohort. The effect of each variant on cohort structure is shown for PCs 1-2. The small % VE for a significantly associated lead variant supports a true 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), identifying p.E123K/D and p.P218T/S/L. (D) Evidence for biological interpretation for every amino acid position is summarised. Dotted red lines indicate the positions at p.123, p.218.

206 **4.7 Host response**

207 Prolonged infections associated with G protein variants p.E123K/D and p.P218T/S/L were
208 on average less severe compared with other circulating variants, and all were limited to the
209 upper respiratory tract (Table 1). Therefore, we analysed nasal wash samples collected dur-
210 ing acute RSV infection for a panel of cytokines involved in antiviral immune responses and
211 observed differential IFN α and IFN γ levels segregating according to viral antigenic group—A
212 or B. Both cytokines were elevated in group B infections compared to group A. The groups
213 A and B median (lower-and upper-quartile) values were 9.5 (3-22.5) and 12.6 (4.1-25.8) me-
214 dian fluorescence intensity (MFI), respectively, for IFN α and 3.6 (1-7) and 4 (2-7.4) MFI,
215 respectively, for IFN γ (group A, n = 149; group B, n = 103). As prolonged infections with
216 p.E123K/D and p.P218T/S/L genotypes were exclusively group B, the dichotomous relation-
217 ship of IFN α and IFN γ levels to antigenic group precluded evaluation of G protein variants as
218 independent predictors of IFN α and IFN γ production.

219 **5 Discussion**

220 In this study of term healthy infants, we found no evidence of host genetic susceptibility
221 to RSV infection during infancy. This allowed our analysis to focus on elucidation of viral
222 drivers of prolonged infection. A significant viral genetic association in the RSV G protein,
223 p.E123K/D and p.P218T/S/L, with prolonged infant RSV infection was identified. These vari-
224 ants were not associated with severe disease, and public data reveal their consistent presence
225 at low frequencies over the past 30 years, without evidence of enrichment by positive selec-
226 tive pressure over time. The two variants we identified in G are correlated with non-random
227 association analogous to LD in the human diploid genome and therefore not likely random
228 mutations, but instead co-inherited in the infecting inoculum. This suggests an evolution-
229 ary benefit and raises the question of why such variants have maintained a stable but low
230 frequency in the human population for at least four decades. These strains are a potential
231 reservoir, emerging seasonally in response to immune, environmental, or other forces. Alterna-
232 tively, the polymorphisms might recurrently arise de novo during infection of some individuals
233 but are poorly transmissible because of suboptimal fitness. The possibility of viral mutational
234 immune escape has been reported for infants who struggle to control primary RSV infections,
235 allowing for prolonged viral replication and not previously described viral rebound [27].

236 The RSV variants associated with prolonged infection in our cohort, G p.E123K/D and
237 p.P218T/S/L, lie in the extracellular region, and there are no known mechanistic features that
238 directly overlap, although it is possible that variant positions approximate sequences that bind
239 a putative viral receptor, heparan sulfate [22], in the G protein three-dimensional structure. G
240 protein amino acid positions 123 and 218 are not part of known antibody neutralization epi-

topes or CD8+ cytotoxic T-cell epitopes (Figure 3 D). In addition to heparan sulfate, interactions between viral G protein and CX3CR1, the receptor for the CX3C chemokine fractalkine, have been reported to modulate the immune response and facilitate infection [21–23; 28–30]. Furthermore, the mature secreted isoform of G protein (p.66-298) is thought to facilitate viral antibody evasion by acting as an antigen decoy and modifying the activity of leukocytes bearing Fc-gamma receptors [31]. Our findings raise the interesting prospect that G protein variants associated with prolonged infection alter a key interaction at the immune interface between pathogen and host.

Although this study was not designed to define mechanisms underlying the association of G protein variants with prolonged infection, these sequence changes might dampen antiviral immune responses and thereby delay viral clearance. Although we observed differences in the acute antiviral response between subjects with resolved and prolonged infection, specifically increased levels of types 1 and 2 IFN in nasal secretions, we could not make causal inference about variant sequences because of confounding by co-linearity of these polymorphisms with RSV antigenic group. Results of nasal cytokine analysis are nevertheless consistent with a contemplated role for altered immune responses in extended infections by G protein variant strains [32]. It is also possible that strains harbouring G protein p.E123K/D and p.P218T/S/L variants are cleared more slowly and foster an immune environment of low-level chronic stimulation or exhaustion. We previously demonstrated that infants infected with RSV in their first year of life have damped subsequent antiviral immune responses in early childhood [33] as well as changes in airway epithelial cell metabolism [34].

While this study has a number of significant strengths, including one of few population-based surveillance studies of first RSV infections during infancy among term healthy infants, our findings are also subject to some limitations. First, this study was not designed with the primary intention to examine infection duration, and additional sampling following initial RSV infection was triggered by a repeat acute respiratory illness. Asymptomatic prolonged infections would therefore not have been captured. Second, our study cohort was small, necessitating focus on viral surface glycoproteins, F and G, due to their variability and importance in host immunity. A larger cohort with serial sampling would be required to diminish the impact of co-linearity of viral genotypes with antigenic groups and to perform informative viral whole genome analysis. Genome-wide information might elucidate other determinants of prolonged infection or pathogen fitness that mediate and/or modulate effects of phenotype-driving variations. Third, again due to small sample size, we could only investigate host genetic risk for infection, not prolonged infection. While we have not specifically assessed subjects for rare monogenic variants that may underlie immunodeficiency, our enrolment criteria included only infants who were term and otherwise healthy. While we performed an interaction analysis for the outcome of host asthma, host genetics, and pathogen genetics and found no significant interaction, our sample size is unlikely sufficient to exclude such an interaction. Lastly, while we do not expect a role for immune memory in these first-in-life RSV infections,

280 we cannot exclude modulatory effects of maternal antibody, which we did not measure. De-
281 spite these limitations, the results are novel and represent an in-depth comprehensive compu-
282 tational statistical analysis of both host and viral genetics providing compelling evidence for
283 RSV viral strain persistence in healthy human infants, a finding of significant importance to
284 understanding the impact of RSV on chronic disease and viral endemicity.

285 In summary, we identified a novel RSV viral variant associated with prolonged infection in
286 healthy infants, but no evidence of host genetic susceptibility to infant RSV infection. Under-
287 standing host and viral mechanisms that contribute to prolonged infection will be important
288 in crafting strategies to control the short- and long-term impact of RSV infection. The iden-
289 tification of RSV variants associated with prolonged infection might also improve vaccine
290 design, particularly if these variants stimulate robust immunity or, in contrast, escape the
291 immune response or induce immunopathologic conditions. The growing availability of large
292 genomic and functional data sources provides opportunities for advancing our understanding
293 of the pathogenesis of infant RSV infection, defining the contribution of viral genetic variants
294 to acute and chronic disease, and informing the development of effective vaccines. As neither
295 the capacity of RSV for prolonged infection in immunocompetent hosts nor a viral reservoir
296 has been delineated, these results are of fundamental interest in understanding viral and host
297 genetic contributions that may promote prolonged infection and influence development of
298 chronic respiratory morbidity.

299 6 Links

300 6.1 Software

301 R v4.1.0 was used for data preparation and analysis <http://www.r-project.org>.
302 R package *caret* was used for analysis: genetic correlations.
303 R package *dplyr* was used for data curation.
304 R package *factoextra* was used for analysis: PCA, and to visualise eigenvalues and variance.
305 R package *ggplot2* was used for data visualisation.
306 R package *MASS* was used to analysis: logistic regression model data.
307 R package *stats* was used for analysis: including *glm* for logistic regressions.
308 R package *stringr* was used for data curation.
309 R package *tidyR* was used for data curation.
310 asn2fsa <https://www.huge-man-linux.net/man1/asn2fsa.html>
311 clc_novo_assemble qiagenbioinformatics.com
312 Clustal Omega <https://www.ebi.ac.uk/Tools/msa/clustalo/>
313 dbNSFP (database) <http://database.liulab.science/dbNSFP> [35]
314 GCTA <https://cnsgenomics.com/software/gcta/> [36]
315 GenBank <https://www.ncbi.nlm.nih.gov/genbank/>

316 IQ-Tree <https://www.iqtree.org/> [37]
317 KING <https://people.virginia.edu/~wc9c/KING/> [38]
318 MAFFT <https://mafft.cbrc.jp/alignment/software/> [39]
319 NextAlign <https://github.com/nextstrain/nextclade>
320 PLINK <http://zzz.bwh.harvard.edu/plink/> [40]
321 Tbl2asn <https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/>
322 Viral Genome ORF Reader, VIGOR 3.0 <https://sourceforge.net/projects/jcvi-vigor/files/>
323 RCSB PDB <https://www.rcsb.org>
324 UniProt <https://www.uniprot.org>

326 **6.2 Data sources**

327 Dataset <https://www.ncbi.nlm.nih.gov/bioproject/267583>.
328 Dataset <https://www.ncbi.nlm.nih.gov/bioproject/225816>.
329 J. Craig Venter Institute <https://www.jcvi.org>.
330 GenBank:NC_001989 *Bovine orthopneumovirus*, complete genome https://www.ncbi.nlm.nih.gov/nuccore/NC_001989.
331 Reference data <https://www.ncbi.nlm.nih.gov/gene/?term=1489824>. G attachment glyco-
332 protein [*Human orthopneumovirus*]; ID: 1489824; Location: NC_001781.1 (4675..5600);
333 Aliases: HRSVgp07.
334 Reference data <https://www.ncbi.nlm.nih.gov/gene/?term=37607642>. G attachment gly-
335 coprotein [*Human orthopneumovirus*]; ID: 37607642; Location: NC_038235.1 (4673..5595);
336 Aliases: DZD21_gp07.
337 Reference data for all public NCBI Virus <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/> for species: *Human orthopneumovirus*; genus: *Orthopneumovirus*; family: *Pneu-
338 moviridae*.
339 Reference data https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Human%20orthopneumovirus,%20taxid:11250 - con-
340 tains sequence data for Virus Lineage ss=*Human orthopneumovirus*, taxid:11250 nu-
341 cleotide: 26'965, protein: 53'804, RefSeq Genomes: 2.
342 Reference https://www.ncbi.nlm.nih.gov/protein/NP_056862.1
343 GCF_002815475.1 (release 2018-08-19) Nucleotide Accessions: NC_038235.1, protein: Y_009518856.1
344 Reference https://www.ncbi.nlm.nih.gov/protein/YP_009518856.1
345 GCF_000855545.1 (release 2015-02-12) Nucleotide Accessions: NC_001781.1, protein: NP_056862.1
346 (strain B1).

350 **7 Code availability**

351 Analysis code is available at <https://github.com/DylanLawless/inspire2022lawless>.
352 github.io.

353 **References**

- 354 [1] D Lawless, C Rosas-Salazar, T Gebretsadik, K Turi, B Snyder, P Wu, J Fellay, and
355 T Hartert. Genome-wide association study of susceptibility to respiratory syncytial
356 virus infection during infancy. In *European Journal of Human Genetics*, volume 28, pages
357 319–319. Springer Nature Campus, 4 Crinan St, London, N1 9XW, England, 2020.
- 358 [2] C. B. Hall, G. A. Weinberg, M. K. Iwane, A. K. Blumkin, K. M. Edwards, M. A. Staat,
359 P. Auinger, M. R. Griffin, K. A. Poehling, D. Erdman, C. G. Grijalva, Y. Zhu, and
360 P. Szilagyi. The burden of respiratory syncytial virus infection in young children. *N
361 Engl J Med*, 360(6):588–98, February 2009. ISSN 0028-4793 (Print) 0028-4793. doi:
362 10.1056/NEJMoa0804877. Edition: 2009/02/07.
- 363 [3] W Paul Glezen, Larry H Taber, Arthur L Frank, and Julius A Kasel. Risk of primary
364 infection and reinfection with respiratory syncytial virus. *American journal of diseases of
365 children*, 140(6):543–546, 1986.
- 366 [4] V. Bokun, J. J. Moore, R. Moore, C. C. Smallcombe, T. J. Harford, F. Rezaee, F. Esper,
367 and G. Piedimonte. Respiratory syncytial virus exhibits differential tropism for distinct hu-
368 man placental cell types with Hofbauer cells acting as a permissive reservoir for infection.
369 *PLoS One*, 14(12):e0225767, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0225767.
370 Edition: 2019/12/04.
- 371 [5] H. A. Cubie, L. A. Duncan, L. A. Marshall, and N. M. Smith. Detection of respiratory
372 syncytial virus nucleic acid in archival postmortem tissue from infants. *Pediatr Pathol
373 Lab Med*, 17(6):927–38, November 1997. ISSN 1077-1042 (Print) 1077-1042. Edition:
374 1997/11/14.
- 375 [6] D. Nadal, W. Wunderli, O. Meurmann, J. Briner, and J. Hirsig. Isolation of respiratory
376 syncytial virus from liver tissue and extrahepatic biliary atresia material. *Scand J Infect
377 Dis*, 22(1):91–3, 1990. ISSN 0036-5548 (Print) 0036-5548. doi: 10.3109/00365549009023125.
378 Edition: 1990/01/01.
- 379 [7] D. R. O'Donnell, M. J. McGarvey, J. M. Tully, I. M. Balfour-Lynn, and P. J. Openshaw.
380 Respiratory syncytial virus RNA in cells from the peripheral blood during acute infection.
381 *J Pediatr*, 133(2):272–4, August 1998. ISSN 0022-3476 (Print) 0022-3476. doi: 10.1016/
382 s0022-3476(98)70234-3. Edition: 1998/08/26.

- 383 [8] F. Rezaee, L. F. Gibson, D. Piktel, S. Othumpangat, and G. Piedimonte. Respiratory
384 syncytial virus infection in human bone marrow stromal cells. *Am J Respir Cell Mol*
385 *Biol*, 45(2):277–86, August 2011. ISSN 1044-1549 (Print) 1044-1549. doi: 10.1165/rcmb.
386 2010-0121OC. Edition: 2010/10/26.
- 387 [9] A. Rohwedder, O. Kemerer, J. Forster, K. Schneider, E. Schneider, and H. Werchau.
388 Detection of respiratory syncytial virus RNA in blood of neonates by polymerase chain
389 reaction. *J Med Virol*, 54(4):320–7, April 1998. ISSN 0146-6615 (Print) 0146-6615. doi:
390 10.1002/(sici)1096-9071(199804)54:4<320::aid-jmv13>3.0.co;2-j. Edition: 1998/04/29.
- 391 [10] Richard E Randall and Diane E Griffin. Within host rna virus persistence: mechanisms
392 and consequences. *Current opinion in virology*, 23:35–42, 2017.
- 393 [11] P. K. Munywoki, D. C. Koech, C. N. Agoti, N. Kibirige, J. Kipkoech, P. A. Cane, G. F.
394 Medley, and D. J. Nokes. Influence of age, severity of infection, and co-infection on the
395 duration of respiratory syncytial virus (RSV) shedding. *Epidemiol Infect*, 143(4):804–12,
396 March 2015. ISSN 0950-2688 (Print) 0950-2688. doi: 10.1017/s0950268814001393. Edition:
397 2014/06/06.
- 398 [12] Bindya Bagga, L Harrison, P Roddam, and JP DeVincenzo. Unrecognized prolonged viral
399 replication in the pathogenesis of human rsv infection. *Journal of Clinical Virology*, 106:
400 1–6, 2018.
- 401 [13] Emelda A Okiro, Lisa J White, Mwanajuma Ngama, Patricia A Cane, Graham F Medley,
402 and D James Nokes. Duration of shedding of respiratory syncytial virus in a community
403 study of kenyan children. *BMC infectious diseases*, 10(1):1–7, 2010.
- 404 [14] Emma K Larkin and Tina V Hartert. Genes associated with rsv lower respiratory tract
405 infection and asthma: the application of genetic epidemiological methods to understand
406 causality. *Future virology*, 10(7):883–897, 2015.
- 407 [15] Milton Pividori, Nathan Schoettler, Dan L Nicolae, Carole Ober, and Hae Kyung Im.
408 Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma:
409 genome-wide and transcriptome-wide studies. *The Lancet Respiratory Medicine*, 7(6):
410 509–522, 2019.
- 411 [16] Riny Janssen, Louis Bont, Christine LE Siezen, Hennie M Hodemaekers, Marieke J Ermers,
412 Gerda Doornbos, Ruben van't Slot, Ciska Wijmenga, Jelle J Goeman, Jan LL Kimpen, et al. Genetic susceptibility to respiratory syncytial virus bronchiolitis is predom-
413 inantly associated with innate immune genes. *Journal of Infectious Diseases*, 196(6):
414 826–834, 2007.
- 415 [17] Anu Pasanen, Minna K Karjalainen, Louis Bont, Eija Piippo-Savolainen, Marja Ruot-
416 salainen, Emma Goksör, Kuldeep Kumawat, Hennie Hodemaekers, Kirsi Nuolivirta, Tuo-
417 mas Jartti, et al. Genome-wide association study of polymorphisms predisposing to
418 bronchiolitis. *Scientific reports*, 7(1):1–9, 2017.

- 420 [18] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring
421 the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111
422 (49):E5272–E5281, 2014.
- 423 [19] AliReza Eshaghi, Venkata R Duvvuri, Rachel Lai, Jeya T Nadarajah, Aimin Li, Samir N
424 Patel, Donald E Low, and Jonathan B Gubbay. Genetic variability of human respiratory
425 syncytial virus strains circulating in ontario: a novel genotype with a 72 nucleotide g
426 gene duplication. *PloS one*, 7(3):e32807, 2012.
- 427 [20] Caroline Breese Hall, Joyce M Geiman, Robert Biggar, David I Kotok, Patricia M Hogan,
428 and R Gordon Douglas Jr. Respiratory syncytial virus infections within families. *New
429 England journal of medicine*, 294(8):414–419, 1976.
- 430 [21] S Levine, R Klaiber-Franco, and PR Paradiso. Demonstration that glycoprotein g is the
431 attachment protein of respiratory syncytial virus. *Journal of General Virology*, 68(9):
432 2521–2524, 1987.
- 433 [22] Steven A Feldman, R Michael Hendry, and Judy A Beeler. Identification of a linear
434 heparin binding domain for human respiratory syncytial virus attachment glycoprotein g.
435 *Journal of virology*, 73(8):6610–6617, 1999.
- 436 [23] Steven A Feldman, Susette Audet, and Judy A Beeler. The fusion glycoprotein of human
437 respiratory syncytial virus facilitates virus attachment and infectivity via an interaction
438 with cellular heparan sulfate. *Journal of Virology*, 74(14):6442–6447, 2000.
- 439 [24] HW McL Rixon, G Brown, JT Murray, and RJ Sugrue. The respiratory syncytial virus
440 small hydrophobic protein is phosphorylated via a mitogen-activated protein kinase p38-
441 dependent tyrosine kinase activity during virus infection. *Journal of General Virology*, 86
442 (2):375–384, 2005.
- 443 [25] Reena Ghildyal, Dongsheng Li, Irene Peroulis, Benjamin Shields, Phillip G Bardin,
444 Michael N Teng, Peter L Collins, Jayesh Meanger, and John Mills. Interaction between
445 the respiratory syncytial virus g glycoprotein cytoplasmic domain and the matrix protein.
446 *Journal of General Virology*, 86(7):1879–1884, 2005.
- 447 [26] Peter L Collins and Geneviève Mottet. Oligomerization and post-translational processing
448 of glycoprotein g of human respiratory syncytial virus: altered o-glycosylation in the
449 presence of brefeldin a. *Journal of General Virology*, 73(4):849–863, 1992.
- 450 [27] Monica E Brint, Joshua M Hughes, Aditya Shah, Chelsea R Miller, Lisa G Harrison,
451 Elizabeth A Meals, Jacqueline Blanch, Charlotte R Thompson, Stephania A Cormier,
452 and John P DeVincenzo. Prolonged viral replication and longitudinal viral dynamic
453 differences among respiratory syncytial virus infected infants. *Pediatric research*, 82(5):
454 872–880, 2017.

- 455 [28] Sara M Johnson, Beth A McNally, Ioannis Ioannidis, Emilio Flano, Michael N Teng,
456 Antonius G Oomens, Edward E Walsh, and Mark E Peeples. Respiratory syncytial virus
457 uses cx3cr1 as a receptor on primary human airway epithelial cultures. *PLoS pathogens*,
458 11(12):e1005318, 2015.
- 459 [29] Ralph A Tripp, Les P Jones, Lia M Haynes, HaoQiang Zheng, Philip M Murphy, and
460 Larry J Anderson. Cx3c chemokine mimicry by respiratory syncytial virus g glycoprotein.
461 *Nature immunology*, 2(8):732–738, 2001.
- 462 [30] Kwang-II Jeong, Peter A Piepenhagen, Michael Kishko, Joshua M DiNapoli, Rachel P
463 Groppo, Linong Zhang, Jeffrey Almond, Harry Kleanthous, Simon Delagrange, and Mark
464 Parrington. Cx3cr1 is expressed in differentiated human ciliated airway cells and co-
465 localizes with respiratory syncytial virus on cilia in a g protein-dependent manner. *PloS
466 one*, 10(6):e0130517, 2015.
- 467 [31] Alexander Bukreyev, Lijuan Yang, Jens Fricke, Lily Cheng, Jerrold M Ward, Brian R
468 Murphy, and Peter L Collins. The secreted form of respiratory syncytial virus g glyco-
469 protein helps the virus evade antibody-mediated restriction of replication by acting as an
470 antigen decoy and through effects on fc receptor-bearing leukocytes. *Journal of virology*,
471 82(24):12191–12204, 2008.
- 472 [32] Megan E Schmidt and Steven M Varga. Modulation of the host immune response by
473 respiratory syncytial virus proteins. *Journal of Microbiology*, 55(3):161–171, 2017.
- 474 [33] Tatiana Chirkova, Christian Rosas-Salazar, Tebeb Gebretsadik, Samadhan J Jadhao,
475 James D Chappell, R Stokes Peebles Jr, William D Dupont, Dawn C Newcomb, Sergejs
476 Berdnikovs, Peter J Gergen, et al. Effect of infant rsv infection on memory t cell responses
477 at age 2-3 years. *Frontiers in immunology*, 13:826666, 2022.
- 478 [34] Andrew R Connelly, Brian M Jeong, Mackenzie E Coden, Jacob Y Cao, Tatiana
479 Chirkova, Christian Rosas-Salazar, Jacqueline-Yvonne Cephus, Larry J Anderson,
480 Dawn C Newcomb, Tina V Hartert, et al. Metabolic reprogramming of nasal airway
481 epithelial cells following infant respiratory syncytial virus infection. *Viruses*, 13(10):2055,
482 2021.
- 483 [35] Xiaoming Liu, Chunlei Wu, Chang Li, and Eric Boerwinkle. dbnsfp v3. 0: A one-stop
484 database of functional predictions and annotations for human nonsynonymous and splice-
485 site snvs. *Human mutation*, 37(3):235–241, 2016.
- 486 [36] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for
487 genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):
488 76–82, 2011.
- 489 [37] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree:
490 a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.
491 *Molecular biology and evolution*, 32(1):268–274, 2015.

- 492 [38] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and
493 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
494 Press.
495
- 496 [39] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software
497 version 7: improvements in performance and usability. *Molecular biology and evolution*, 30
498 (4):772–780, 2013.
- 499 [40] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira,
500 David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink:
501 a tool set for whole-genome association and population-based linkage analyses. *The
502 American journal of human genetics*, 81(3):559–575, 2007.
- 503 [41] E. K. Larkin, T. Gebretsadik, M. L. Moore, L. J. Anderson, W. D. Dupont, J. D. Chappell,
504 P. A. Minton, R. S. Peebles, Jr., P. E. Moore, R. S. Valet, D. H. Arnold, C. Rosas-
505 Salazar, S. R. Das, F. P. Polack, and T. V. Hartert. Objectives, design and enrollment
506 results from the Infant Susceptibility to Pulmonary Infections and Asthma Following
507 RSV Exposure Study (INSPIRE). *BMC Pulm Med*, 15:45, April 2015. ISSN 1471-2466.
508 doi: 10.1186/s12890-015-0040-0. Edition: 2015/05/30.
- 509 [42] Emma K Larkin, Tebeb Gebretsadik, Martin L Moore, Larry J Anderson, William D
510 Dupont, James D Chappell, Patricia A Minton, R Stokes Peebles, Paul E Moore,
511 Robert S Valet, et al. Objectives, design and enrollment results from the infant sus-
512 ceptibility to pulmonary infections and asthma following rsv exposure study (inspire).
513 *BMC pulmonary medicine*, 15(1):1–12, 2015.
- 514 [43] Seyhan Boyoglu-Barnum, Sean O Todd, Tatiana Chirkova, Thomas R Barnum, Kelsey A
515 Gaston, Lia M Haynes, Ralph A Tripp, Martin L Moore, and Larry J Anderson. An anti-g
516 protein monoclonal antibody treats rsv disease more effectively than an anti-f monoclonal
517 antibody in balb/c mice. *Virology*, 483:117–125, 2015.
- 518 [44] Alexander Bukreyev, Lijuan Yang, and Peter L Collins. The secreted g protein of hu-
519 man respiratory syncytial virus antagonizes antibody-mediated restriction of replication
520 involving macrophages and complement. *Journal of virology*, 86(19):10880–10884, 2012.
- 521 [45] Larry J Anderson, P Bingham, and JC Hierholzer. Neutralization of respiratory syncytial
522 virus by individual and mixtures of f and g protein monoclonal antibodies. *Journal of
523 virology*, 62(11):4232–4238, 1988.
- 524 [46] Joan O Ngwuta, Man Chen, Kayvon Modjarrad, M Gordon Joyce, Masaru Kanekiyo,
525 Azad Kumar, Hadi M Yassine, Syed M Moin, April M Killikelly, Gwo-Yu Chuang, et al.
526 Prefusion f-specific antibodies determine the magnitude of rsv neutralizing activity in
527 human sera. *Science translational medicine*, 7(309):309ra162–309ra162, 2015.

- 528 [47] S. A. Schobel, K. M. Stucker, M. L. Moore, L. J. Anderson, E. K. Larkin, J. Shankar,
529 J. Bera, V. Puri, M. H. Shilts, C. Rosas-Salazar, R. A. Halpin, N. Fedorova, S. Shrivastava,
530 T. B. Stockwell, R. S. Peebles, T. V. Hartert, and S. R. Das. Respiratory Syncytial
531 Virus whole-genome sequencing identifies convergent evolution of sequence duplication
532 in the C-terminus of the G gene. *Sci Rep*, 6:26311, May 2016. ISSN 2045-2322. doi:
533 10.1038/srep26311. Edition: 2016/05/24.
- 534 [48] K. Li, S. Shrivastava, A. Brownley, D. Katzel, J. Bera, A. T. Nguyen, V. Thovarai,
535 R. Halpin, and T. B. Stockwell. Automated degenerate PCR primer design for high-
536 throughput sequencing improves efficiency of viral sequencing. *Virol J*, 9:261, November
537 2012. ISSN 1743-422x. doi: 10.1186/1743-422x-9-261. Edition: 2012/11/08.
- 538 [49] QIAGEN Aarhus. White paper on de novo assembly in CLC Assembly Cell 4.0. *digita-*
539 *linsights*, page 14, June 2016. URL <https://digitalinsights.qiagen.com/files/whitepapers/whitepaper-denovo-assembly.pdf>. Place: Denmark Publisher: Qiagen.
- 540 [50] S. Wang, J. P. Sundaram, and T. B. Stockwell. VIGOR extended to annotate genomes for
541 additional 12 different viruses. *Nucleic Acids Res*, 40(Web Server issue):W186–92, July
542 2012. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gks528. Edition: 2012/06/07.
- 543 [51] Christian Rosas-Salazar, Zheng-Zheng Tang, Meghan H Shilts, Kedir N Turi, Qilin Hong,
544 Derek A Wiggins, Christian E Lynch, Tebeb Gebretsadik, James D Chappell, R Stokes
545 Peebles Jr, et al. Upper respiratory tract bacterial-immune interactions during respiratory
546 syncytial virus infection in infancy. *Journal of Allergy and Clinical Immunology*, 149(3):
547 966–976, 2022.

549 **8 Supplemental methods**

550 **8.1 Study population**

551 The protocol and informed consent documents were approved by the Institutional Review
552 Board at Vanderbilt University Medical Center (#111299). One parent of each participant
553 in the cohort study provided written informed consent for participation in this study. The
554 informed consent document explained study procedures and use of data and biospecimens for
555 future studies, including genetic studies.

556 The study population is a longitudinal birth cohort, the INFant Susceptibility to Pul-
557 monary Infections and Asthma Following RSV Exposure (INSPIRE), specifically designed
558 to capture the first RSV infection in term healthy infants. Additional details of this birth
559 cohort have been previously published [41]. Briefly, the cohort included 1949 term (≥ 37
560 weeks gestation), non-low birth weight (≥ 2250 g, 5 lbs), otherwise healthy infants from a

561 population-representative sample of pediatric practices located in rural, suburban, and urban
562 regions of the south-eastern US during 2012-2014. Infants were born June through December;
563 per study design, they were 6 months of age or less entering their first RSV season.

564 **8.2 Biweekly surveillance of RSV infection**

565 Infant (i.e., first year of life) RSV infection was ascertained through passive and active bi-
566 weekly surveillance during each infant's first RSV season and RSV serology (Table 1). If an
567 infant met pre-specified criteria for an acute respiratory infection, we conducted an in-person
568 respiratory illness visit at which time we administered a parental questionnaire, performed a
569 physical exam, collected a nasal wash, and completed a structured medical chart review for
570 infants seen during an unscheduled visit. RSV RNA in nasal samples was detected by reverse-
571 transcription quantitative PCR [42]. We *a priori* defined the clinical entity of "prolonged"
572 infection during infancy as repeatedly meeting pre-specified criteria for an acute respiratory
573 infection accompanied by repeatedly positive RSV PCR separated by 15 or more days (Figure
574 S1) [13].

575 **8.3 Descriptive analyses**

576 Descriptive analyses of the cohort were conducted using R 4.0.5. Pearson or Wilcoxon tests
577 were used for comparing infants with and without prolonged RSV infection. The main descriptive
578 features are provided in Table 1.

579 **8.4 Host DNA collection and genotyping**

580 One-year blood samples were selected based on availability of DNA among a subset of chil-
581 dren with RSV infection and a random group of those without infection, and were genotyped
582 with the Multi-Ethnic Global Array microarray (Illumina, CA, United States) at the Univer-
583 sity of Washington DNA Sequencing and Gene Analysis Center (Seattle, WA, United States).

584 **8.5 Host genetic analyses of RSV infection in infancy**

585 To determine whether host genetic factors are associated with infant RSV infection risk, we
586 examined single nucleotide polymorphisms (SNPs) previously shown to alter infant RSV in-
587 fection severity or childhood asthma risk [15–17]. We also conducted a host GWAS to iden-
588 tify common variants associated with infant RSV infection, and examined narrow sense heri-
589 tability to test for small cumulative effects. The GWAS was performed on 621 children with
590 available DNA for the association between host genotype and RSV infection during infancy.

591 Due to sample size constraints, we restricted our sub-analysis to the 54 host SNPs previously
592 associated with RSV lower respiratory tract infection or childhood asthma [15–17]. We ad-
593 ditionally evaluated the accumulation of small genetic effects that would go undetected in a
594 GWAS by estimating the narrow sense heritability of RSV infection.

595 For GWAS analyses, the initial round of data quality control was performed on individ-
596 ual populations (self-reported as White, Black, and Hispanic) using PLINK version 1.9 [40].
597 Subjects with a missing genotype call rate above 5% were removed. The SNP minor allele
598 frequency (MAF) threshold was set at > 0.01, 0.03, and 0.08 for White, Black, and Hispanic,
599 respectively [36]. The groups were merged for a total of 1,086,830 variants and a genotyping
600 rate of 0.78. Subject independence was assessed using KING (<https://people.virginia.edu/~wc9c/KING/>) to prevent spurious associations. However, no probable relatives or dupli-
601 cates were detected based on pairwise identify-by-state. We compared the genetic ancestry in
602 cases to self-reported ethnicity to check for mislabelling. Reported and estimated sex was also
603 examined for discrepancy. A second round of quality control on the combined dataset was con-
604 ducted, which removed 74 samples due to genotype missingness and 399,991 variants with a
605 genotyping rate < 0.1. Variants were checked for departure from Hardy-Weinberg equilibrium
606 (HWE) ($P < 1e^{-6}$) to uncover features of selection, population admixture, cryptic related-
607 ness, or genotyping error. This was only performed on controls to prevent removal of genuine
608 genetic associations that can be associated with this measurement; 6,024 variants were re-
609 moved. No variants had a MAF MAF < 0.01 after merging. SNP positions and identifiers
610 were compared and updated according to dbNSFP4.0a (hg19) with 289 variants removed due
611 to a missing coordinate and SNPs identifier [35]. This resulted in an analysis-ready dataset
612 of 680,526 variants from 621 children (509 with and 112 without RSV infection in infancy),
613 yielding a total genotyping rate of 0.98. No genomic inflation was evident with an estimated
614 lambda (based on median chi-squared test) equal to 1. We then used genome-wide complex
615 trait analysis (GCTA) software (<https://cnsgenomics.com/software/gcta/>) to calculate
616 the genetic relationship matrix and performed principal component analysis (PCA) to account
617 for population structure [36]. Genome-wide association analysis was performed using PLINK
618 version 1.9 for logistic regression with multiple covariates consisting of the child's birth month,
619 enrolment year (as a marker of RSV season), daycare attendance, presence of another child \leq
620 6 years of age at home, sex, and 6 ancestry principal components (PCs) [40].

622 As the multiple testing burden likely precluded identification of small genetic effects in our
623 GWAS, we conducted an additional heritability analysis using the method described by Golan
624 et al. [18] to estimate narrow-sense heritability of RSV infection during infancy on the latent
625 liability scale (h_l^2), which, > 0 , would indicate an accumulation of small genetic effects. We
626 estimated h_l^2 to be exactly 0, suggesting that, if present, infant RSV infection-related genetic
627 signals are both small and sparse.

628 **8.6 Host genetic analyses for known associations**

629 We further investigated the possibility that the analysis was underpowered to identify asso-
630 ciations with reported childhood asthma- and RSV LRTI-associated SNPs [15–17]. This was
631 done by pooling information across SNPs to estimate the average genetic effect size. In brief,
632 we computed a z-score for each SNP, where the average (across SNPs) squared z-score \bar{G} is
633 proportional to the average squared genetic effect on RSV infection in infancy. As \bar{G} is an av-
634 erage of $p = 54$ approximately independent statistics, it is approximately $N(n\mu^2 + 1, 2/p)$, where
635 $n = 621$ is the sample size and μ^2 is a function of the average squared genetic effect on RSV
636 infection in infancy. Using the genetic effect estimates from Pividori et al. [15]; Janssen et al.
637 [16]; Pasanen et al. [17], we calculated that we would have 80% power to reject the global
638 null hypothesis of no genetic effect at any of these SNPs (i.e., $\mu^2 = 0$) if, on average across the
639 54 SNPs, the genetic effect on RSV infection in infancy was at least 61% as large as those esti-
640 mated in the aforementioned 3 studies. The z-score \bar{G} is proportional to the average squared
641 genetic effect on RSV infection in infancy. We found $\bar{G}=1.00$ in our data, which corresponds
642 to a p value of 0.50. This result indicates that the genetic effect on RSV infection in infancy is
643 zero or small at SNPs likely to be associated with RSV infection *a priori*.

644 **8.7 Host acute local immune response**

645 Nasal wash samples collected at the time of acute infant RSV infection were profiled to mea-
646 sure the acute host response to infection using Luminex xMap multianalyte bead assays (Mil-
647 liplex Human Cytokine/Chemokine Panel II MAGNETIC Premixed 23 Plex Kit, EMD Milli-
648 pore; and Cytokine 30-Plex Human Panel, Life Technologies Corporation). These data were
649 used to test the host nasal interferon (IFN) response and the viral variant associated with
650 prolonged infection.

651 **8.8 RSV whole-genome sequencing**

652 RSV genome sequencing was performed on all specimens from subjects meeting illness criteria
653 and with positive RSV PCR. Viral amino acid variants (genotype) of the F and G glycopro-
654 tein were tested for association with prolonged infection adjusting for host features associated
655 with increased infection risk. The relatively small sample size of our cohort required analysis
656 that targeted only genes which were *a priori* likely to functionally contribute to the clinical
657 phenotype. Therefore, our analysis focused on the surface F (fusion) and G (attachment) pro-
658 teins of RSV as they have been implicated in pathogenesis [43; 44], and both are targets for
659 neutralizing antibodies during infection [45; 46]. Lastly, to determine if the variants of inter-
660 est were enriched by selective pressure over time, we used public data from the past several
661 decades to assess variant frequency over time.

662 RSV whole-genome sequencing of this study population has been previously described [47].
663 Briefly, RNA was extracted at J. Craig Venter Institute (JCVI) (<https://www.jcvi.org>) in
664 Rockville, MD from nasal wash samples which were RSV PCR positive and collected during
665 a respiratory illness visit triggered through biweekly surveillance of symptoms. Four forward
666 reverse-transcription (RT) primers were designed and four sets of PCR primers were manually
667 picked from primers designed across a consensus of complete RSV genome sequences using
668 JCVI's automated primer design tool [48]. cDNA was generated from 4 µL undiluted RNA,
669 using the pooled forward primers and SuperScript III Reverse Transcriptase (Thermo Fisher
670 Scientific, Waltham, MA, USA). 100 ng of pooled DNA amplicons were sheared to create
671 400-bp libraries, which were pooled in equal volumes and cleaned with Ampure XP reagent
672 (Beckman Coulter, Inc., Brea, CA, USA). Sequencing was performed on the Ion Torrent PGM
673 using 316v2 or 318v2 chips (Thermo Fisher Scientific).

674 For samples requiring extra coverage, in addition to Ion Torrent sequencing, Illumina li-
675 braries were prepared using the Nextera DNA Sample Preparation Kit (Illumina, Inc., San
676 Diego, CA, USA). Sequence reads were sorted by barcode, trimmed, and assembled de novo
677 using CLC Bio's *clc_novo_assemble* program, and the resulting contigs were searched against
678 custom, full-length RSV nucleotide databases to find the closest reference sequence. All se-
679 quence reads were then mapped to the selected reference RSV sequence using CLC Bio's
680 *clc_ref_assemble_long* program [49]. Curated assemblies were validated and annotated with the
681 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
682 accession PRJNA225816 (<https://www.ncbi.nlm.nih.gov/bioproject/225816>) [50] and
683 PRJNA267583 (<https://www.ncbi.nlm.nih.gov/bioproject/267583>).

685 8.9 Viral sequence alignment

686 The NCBI-tools Tbl2asn (<https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/>) was used
687 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
688 analysis.

689 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 convert sequences to fasta
690 format. Each sample consisted of 11 sequence segments (NS1, NS2, N, P, M, M2-1, M2-2, SH,
691 G, F, and L) as shown in Figure S1. These were separated and repooled to create 11 single
692 fasta files for each gene containing all 350 samples. Sequences were checked for at least 90%
693 coverage of the corresponding gene to minimize loss of aligned positions when computing
694 the phylogenetic tree. Each of the eleven resulting sets was aligned with MAFFT v7 (<https://mafft.cbrc.jp/alignment/software/>) [39], using default parameters. The sequence of the

698 orthologous gene from *Bovine orthopneumovirus* ([GenBank:NC_001989](#)) was added to each set
699 as an outgroup.

700 IQ-Tree (<https://www.iqtree.org>) [37] was used with per-gene multiple sequence align-
701 ment (MSA) files based on amino acid sequence for estimating maximum-likelihood phylo-
702 genies using protein substitution model. Examining the sequences with an alignment viewer
703 showed that a small number of sequences had frame-shift variants but these did not affect the
704 regions included in our testing criteria.

705 Viral sequence data and clinical information were merged and cleaned with R. Clinical
706 IDs matching more than one viral sequence ID were used to re-identify samples from the
707 same individual as prolonged infections. Genetic variation was quantified in these samples,
708 and for subsequent analysis, only the first viral sequence was included for association testing.
709 Antigenic grouping of strain A and B had been completed previously and labels were included
710 to annotate each sample accordingly.

711 The cohort-specific variant frequency per position was calculated; residues were counted
712 and ranked by frequency with the most frequent residue defined as reference (REF) and al-
713 ternative (ALT) for variants. Positions with at least one ALT were checked for potential mis-
714 alignment or other sources of error. Variant positions were selected for association analysis,
715 while non-variant positions were ignored.

716 A number of host features have been previously shown to influence infection susceptibility
717 and were therefore included as covariates in our analysis [51]. Six samples were excluded due
718 to insufficient covariate data, resulting in 344 test samples. Of these, 38 were from the same
719 patients (prolonged infection) of which half (19) were included for association testing. Thus,
720 the test set was comprised of single samples collected from 325 individuals.

721 8.10 Viral population structure

722 The genetic distances to nearest neighbors were computed based on phylogenetic trees gener-
723 ated with MAFFT. PCA and singular value decomposition (SVD) were used in dimensionality
724 reduction for exploratory data analysis of viral phylogeny. The R package factoextra was
725 used for PCA and to visualise eigenvalues and variance. R package caret was used to analyse
726 genetic correlations.

727 8.11 Viral variant association testing

728 Viral amino acids (genotype collapsed into REF/ALT) were tested for association with infec-
729 tion types (i.e., resolved and prolonged) including key covariates that alter infection risk. To
730 reduce the multiple testing burden, proxy amino acid variants were identified by performing

731 clumping with ranking based on MAF and with a cut-off threshold of $r^2 \geq 0.8$ (Supplemental
732 Figure S3). Since many variants within RSV coding genes have non-random association
733 due to selection, analogous to linkage disequilibrium (LD) in human GWAS, we reduced the
734 multiple testing burden by retaining proxy variants and removing those with $r^2 \geq 0.8$. Anal-
735 ysis was performed using logistic regression with the R stats (3.6.2) glm function. The model
736 consisted of the binary response (prolonged infection Yes/No) and predictors viral genotype
737 (REF/ALT amino acid, including multi-allelic non-REF collapsed into ALT), viral PCs 1-5,
738 host sex, and host features that have been previously demonstrated as significantly associated
739 with infection: daycare attendance, living with siblings and self-reported race/ethnicity [51].

740 Environmental host covariates did not contribute significant effect in our model for can-
741 didate causal association. For this reason, in our main analysis, viral population structure
742 was accounted for by the first five PCs. The Bonferroni correction for multiple testing was
743 applied based on the number of variants tested. For the significant association found by proxy
744 amino acid variants, the association test was repeated for all clumped variants to produce
745 a LocusZoom-style Manhattan plot containing r^2 by color and p value statistics. R package
746 stats was used for a range of analyses including glm for logistic regressions. R package MASS
747 was used to analyse logistic regression model data. To test if the significantly associated vari-
748 ants were due to population structure, we re-estimated models using the subset of individuals
749 infected with RSV strain B to confirm validity of combined analysis.

750 8.12 Public viral sequence data

751 We gathered publicly available sequence data to further assess variants of interest. We used
752 the public viral data repository of NCBI ([https://www.ncbi.nlm.nih.gov/labs/virus/
753 vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Human%20orthopneumovirus,
754 %20taxid:11250](https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Human%20orthopneumovirus,%20taxid:11250)) to retrieve information using search criteria that follow. Virus: Human
755 orthopneumovirus (HRSV), taxid:11250. Proteins: attachment glycoprotein. Host: Homo (hu-
756 mans), taxid:9605. Collection dates: Jan 1, 1956 onward. Nucleotide and protein sequence
757 data was collected, which contained data from 27 countries and 1084 glycoprotein protein se-
758 quences after curation. Sequence and meta data were merged. Multiple sequence alignment
759 was performed to find consensus relative positions for all sequences. Regions of interest were
760 then extracted and re-annotated with their correct amino acid positions matching the refer-
761 ence sequence. Summary statistics were generated, including number of samples, collection
762 date, geo-location, variant frequency, and strain. for the specified amino acid (Supplemental
763 Figure S4).

764 **8.13 Biological interpretation**

765 As infant RSV infection stimulates an acute antiviral response and also results in decreased
766 barrier function of the airway epithelium [34], we tested for association between host inter-
767 feron (IFN) response and the amino acid (REF/ALT) identified as the viral variant associated
768 with prolonged infection. A Wilcoxon test was performed to compare IFN- γ , and IFN- α , be-
769 tween RSV amino acid positions, with adjustment for the same covariates as in the main
770 analysis. Protein structures were analysed with data sourced from RCSB PDB [https://www.](https://www.rcsb.org)
771 [rcsb.org](https://www.rcsb.org). Protein function and domains were assessed using UniProt ([https://www.uniprot.](https://www.uniprot.org)
772 [org](https://www.uniprot.org)) for P03423 (GLYC_HRSVA) (strain A2) and O36633 (GLYC_HRSVB) (strain B1) in
773 gff format (<https://www.uniprot.org/uniprot/P03423> and <https://www.uniprot.org/uniprot/036633>, respectively). Interactions, post-translational modifications, motifs, and epi-
774 topes were assessed from the literature. Protein features were assessed using data from NCBI
775 (https://www.ncbi.nlm.nih.gov/ipg/NP_056862.1) and via sequence viewer with O36633.1
776 human RSV B1, (<https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=036633.1>). Po-
777 tential effects of amino acid variation on protein structure and function were considered ac-
778 cording to available information on a broad range of biological and biochemical features, in-
779 cluding native conformation (secondary, tertiary, and quaternary), domains and topology,
780 disulfide bonds, glycosylation, interactions with other viral proteins and host-cell factors,
781 proteolytic cleavage sites, normal patterns of intra-and/or extra-cellular distribution, and
782 secretion status.
783

784 **9 Supplemental figures**

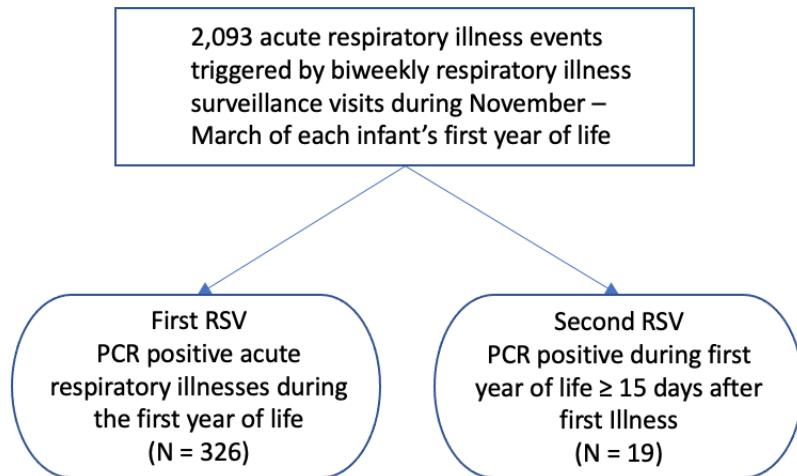


Figure S1: Supplemental: INFant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure (INSPIRE). The study population is a longitudinal birth cohort specifically designed to capture the first RSV infection in term healthy infants. Prolonged infection was a priori defined as repeatedly meeting criteria for acute respiratory infection with RSV PCR positive nasal samples ≥ 15 days between testing.

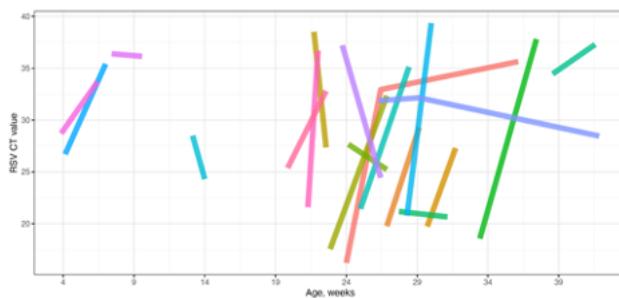


Figure S2: Supplemental: Infant RSV prolonged infections. Each line represents an infant in the study, and line start and end correspond to clinical respiratory illness sampling timepoints. CT values are inversely related to viral RNA abundance.

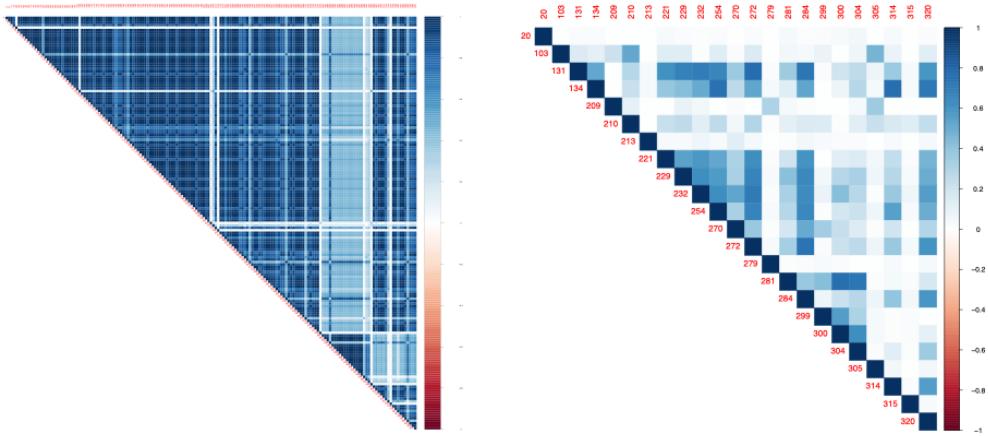


Figure S3: Supplemental: Variant clumping for reduction in association testing.
 [Left] Correlation between all positions. [Right] Correlation between proxy variants after clumping to remove $r^2 \geq 0.8$. Values indicate relative amino acid positions within MSA. r^2 indicated by color scale.

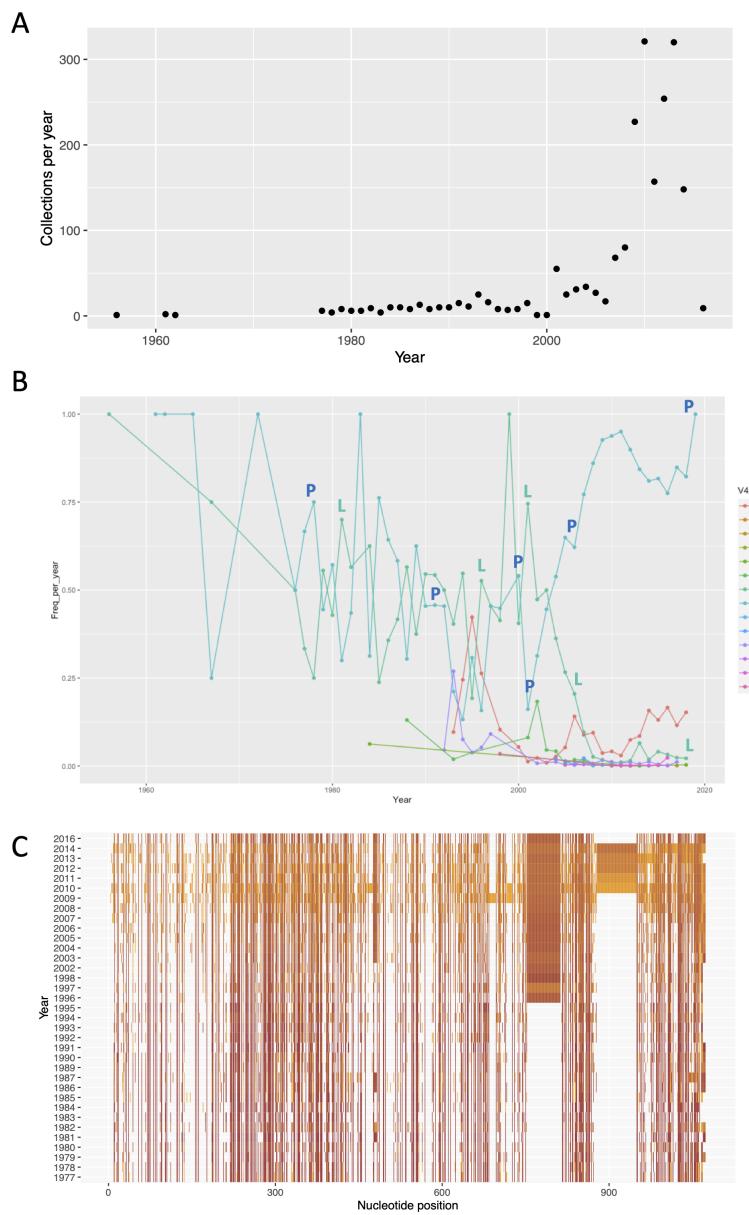


Figure S4: Supplemental: Publicly available RSV sequence data for > 30 years.
 (A) Global sample collection per year. (B) Variant associated with prolonged infection tracked in public data. The lead proxy SNP, p.P218T/S/L is illustrated here (relative amino acid positive 410 in MSA). The major alleles (proline, leucine) are seen for group A/B, with minor alleles (serine, threonine) generally at low frequency <10%. (C) % variance explained per year for all G protein amino acid variants from 1990-2022.