Within-Host Whole-Genome Deep Sequencing and Diversity Analysis of Human Respiratory Syncytial Virus Infection Reveals Dynamics of Genomic Diversity in the Absence and Presence of Immune Pressure

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1128/JVI.00038-14

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:25123559

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://
Within-Host Whole-Genome Deep Sequencing and Diversity Analysis of Human Respiratory Syncytial Virus Infection Reveals Dynamics of Genomic Diversity in the Absence and Presence of Immune Pressure

Yonatan H. Grad, Ruchi Newman, Michael Zody, Xiao Yang, Ryan Murphy, James Qu, Christine M. Malboeuf, Joshua Z. Levin, Marc Lipsitch, John DeVincenzo

Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA; Division of Infectious Diseases, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, USA; Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA; Department of Pediatrics, University of Tennessee School of Medicine, Memphis, Tennessee, USA; Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA; Department of Microbiology, Immunology, and Molecular Biology, University of Tennessee School of Medicine, Memphis, Tennessee, USA; Children’s Foundation Research Institute, LeBonheur Children’s Hospital, Memphis, Tennessee, USA.

ABSTRACT

Human respiratory syncytial virus (RSV) is the leading cause of lower respiratory tract disease in infants and young children and an important respiratory pathogen in the elderly and immunocompromised. While population-wide molecular epidemiology studies have shown multiple cocirculating RSV genotypes and revealed antigenic and genetic change over successive seasons, little is known about the extent of viral diversity over the course of an individual infection, the origins of novel variants, or the effect of immune pressure on viral diversity and potential immune-escape mutations. To investigate viral population diversity in the presence and absence of selective immune pressures, we studied whole-genome deep sequencing of RSV in upper airway samples from an infant with severe combined immune deficiency syndrome and persistent RSV infection. The infection continued over several months before and after bone marrow transplant (BMT) from his RSV-immune father. RSV diversity was characterized in 26 samples obtained over 78 days. Diversity increased after engraftment, as defined by T-cell presence, and populations reflected variation mostly within the G protein, the major surface antigen. Minority populations with known palivizumab resistance mutations emerged after its administration. The viral population appeared to diversify in response to selective pressures, showing a statistically significant growth in diversity in the presence of pressure from immunity. Defining escape mutations and their dynamics will be useful in the design and application of novel therapeutics and vaccines. These data can contribute to future studies of the relationship between within-host and population-wide RSV phylodynamics.

IMPORTANCE

Human RSV is an important cause of respiratory disease in infants, the elderly, and the immunocompromised. RSV circulating in a community appears to change season by season, but the amount of diversity generated during an individual infection and the impact of immunity on this viral diversity has been unclear. To address this question, we described within-host RSV diversity by whole-genome deep sequencing in a unique clinical case of an RSV-infected infant with severe combined immunodeficiency and effectively no adaptive immunity who then gained adaptive immunity after undergoing bone marrow transplantation. We found that viral diversity increased in the presence of adaptive immunity and was primarily within the G protein, the major surface antigen. These data will be useful in designing RSV treatments and vaccines and to help understand the relationship between the dynamics of viral diversification within individual hosts and the viral populations circulating in a community.

Human respiratory syncytial virus (RSV) is the leading cause of severe viral respiratory disease in infants and young children (1) and is a significant cause of morbidity and mortality in the elderly (2) and immunocompromised (3, 4). Reinfection can take place throughout life, although in older children and healthy adults infection is usually associated with milder disease (5, 6). Population-wide molecular epidemiology studies have shown that there are multiple cocirculating RSV genotypes (where genotypes have been defined by the sequence of the variable region of the attachment [G] protein) that undergo a gradual replacement in dominant genotypes over successive seasons (7–9). Studies of RSV evolution in response to selective pressures have demonstrated escape mutations in vitro in the G protein (10, 11), and analyses of clinical collections of RSV sequences have found population-genetic evidence suggesting positive selection within the G protein (12–14). These studies support the hypothesis that the variation in circulating RSV genotypes reflects response to immune pressure. However, little is known about the extent of viral diversity over the course of an individual infection or the effect of immune status on viral diversity. Additionally, studies to date have focused almost exclusively on variations within small segments of the G protein, as these appear to be the most variable
regions of the genome. As the G protein is a surface protein primarily responsible for virus attachment to host cells and an immunomodulatory secreted form also exists (15), understanding the natural diversity of the G protein in the absence of specific immune pressure within a host is important. Likewise, understanding the diversity of the rest of the genome and the relationship of this variation to immune pressure may aid in development of therapeutics and guide vaccine targets and strategies.

Recent technological advances in deep sequencing have established a platform with which we are able to generate full-genome assemblies without the aid of a reference and to investigate the diversity of viral subpopulations over the course of individual infections. Deep sequencing studies of chronic RNA viral infections, such as HIV infections, have expanded our understanding of viral evolutionary dynamics in response to immunity and drug resistance (16, 17). Applying these same technologies to acute, self-limited infections with RNA viruses such as RSV can aid in understanding viral pathogenesis and the role of immune pressure in generation of viral sequence diversity. The assessment of the relationship between immune pressure and viral diversity and adaptation will allow for deeper understanding of the origins and dynamics of viral population diversity (18).

To study the within-host diversity of RSV and the role of immune pressure, we investigated the viral populations from an RSV-infected infant with severe combined immune deficiency (SCID) who subsequently underwent a bone marrow transplant (19). The infant developed progressive respiratory symptoms over the month prior to hospitalization. Subsequently, nasal aspirate samples were collected prospectively and viral loads quantified under a research protocol designed to study RSV pathogenesis. Treatment with 4 infusions of palivizumab had no clinical impact, under a research protocol designed to study RSV pathogenesis. To determine the host and virus composition of each library, reads were aligned to viral and host reference genomes as previously described (20). The reference genomes used were the concatenation of the human genome assembly (GRCh37/hg19), human rRNA sequences (NR_003286.1, NR_003287.1, V00589.1, NR_003285.2, gi|251831106648–1601, and gi|251831106671–3229), and the RSV-A sequence (M74568.1). The reads were assembled using the VICUNA assembly program (21). Variant calling was performed using V-Phaser 2 (22) to compare reads from each time point to the consensus sequence from the first time point. The diversity of sites was visualized using WebLogo 3.3 (23).

Calculation of information content. For each time point, the population diversity at each amino acid position was identified using V-Phaser 2 (22). For variant positions in which the dominant amino acid was no more than 95% of the overall population, the Shannon diversity index \( H_i \) was determined with the following equation: \( H_i = -\sum P_{i,\alpha} \log_2 P_{i,\alpha} \), where \( P_{i,\alpha} \) is the relative frequency of an amino acid at position \( i \). The average diversity index for the RSV population at each time point was then determined by summing \( H_i \) over all variant positions as defined above and dividing by the length of the consensus genome sequence from the first time point, which was used as the reference for read-based mapping.

Analysis of average diversity. The average diversity indices were plotted by time. To determine whether diversity differed before and after the rise in lymphocyte count at day +12, the data were fitted to the linear regression model \( Y = \beta_0 + \beta_1 A + \beta_2 (A \times I) \), where \( A \) is the explanatory variable, the \( \beta \) terms are the regression coefficients, and \( I \) is an indicator for the final six time points (after the rise in lymphocyte count).

Comparison of within-host and population-wide diversity. To determine population-wide diversity, RSV-A full-genome nucleotide sequences were downloaded from GenBank (see Table S1 in the supplemental material). Each gene sequence was extracted, visualized, and translated to protein sequence in SeaView (24). The Shannon diversity index was determined as described above, ignoring sites containing degenerate nucleotide calls. For the purposes of comparison of the population-wide diversity to within-host diversity, the maximum diversity at each site from all of the within-host samples was used.

Nucleotide sequence accession numbers. Sequences determined in the course of this work have been placed in GenBank under accession numbers KP973319 to KP973340.

RESULTS AND DISCUSSION

Deep sequencing of RSV samples. Of the 26 samples, 21 yielded high depth of coverage across the entire coding region of the genome, one yielded lower depth of coverage, and four had far fewer reads aligning across the genome (Table 1). Subsequent de novo assembly of the viral reads resulted in assemblies that covered 100%, ~90%, and <60% of the coding region, respectively (see Fig. S1 in the supplemental material). The 22 assemblies that covered >90% of the genome were largely contiguous, with few in-
and

Patterns of diversity prior to and in the presence of transplant-related immune response. The observation of diversity in the RSV population from day 35 before transplant (D − 35) to D + 8 (Fig. 1), prior to the rise in ALC/ACD8 and cellular immunity, could be due to neutral drift and/or to selection within the host under innate immune or other immune pressures. In addition, some or all of the diversity seen at these time points may have been present in the inoculum infecting the patient, as can be seen with other RNA viruses, such as HIV (25). Notably, most sites that appear heterogeneous prior to day +12 show no diversity subsequently (Fig. 2, G protein aa 140, 142, 150, 152, and 158). This likely reflects the impact of the RSV population bottleneck and, thereby, a reduction in genetic variation at the time of posttransplant proliferation of lymphocytes.

The increase in diversity after day +12 is attributable to an increase in the number of variant sites rather than an increase in the amount of diversity at sites that previously had small amounts of diversity (Fig. 2). Sites with markedly increased diversity after day +12 include aa 78 and 102 in NS2, P-207, M-50, M-53, and M-54; numerous sites in the transmembrane and proximal ect-
odomain of the G protein, as well as G-192 and F-268; and aa 184 and 1280 in the L protein. Despite diversity demonstrating minority variants prior to the rise of lymphocyte count, no switch in amino acid majority predominance occurs. After day 12, the predominant amino acids switch at multiple sites (e.g., NS2-102, P-207, G-56, G-60, G-292, F-268, and L-1280) (Fig. 2).

Several amino acid sites, including M2-31, G-229, and G-292, show diversity that persists across all time points. In the case of G-229, one subpopulation contains a premature stop codon, truncating a large segment of the carboxy (C) terminus. The functional implications of the other variants are unclear, but the persistent polymorphisms at all three sites have at least two possible causes: balancing selection driven by a pressure other than adaptive immunity, or sampling of populations from more than one distinct subpopulation that do not directly compete with one another, thereby maintaining polymorphisms.

Diversity in the G protein. The greatest diversity observed in the samples appears in the G protein, which contains 45% of the variant amino acid sites across all samples. The diversity in the G protein is consistent with its role as a major surface antigen of RSV and relaxed selective constraints. The G protein product has two forms: a transmembrane form (Gm), which functions as the attachment glycoprotein for RSV, and a secreted form (Gs), which is cleaved just downstream from the transmembrane domain and is believed to function as an antigen decoy to allow evasion of antibody targeting (26, 27) and as a modulator of dendritic cell function (28).

The diversity within the G protein is notable for several features. First, variants in G protein appear concentrated in three regions (Fig. 3). These are (i) aa 33 to 97, extending across the transmembrane domain and into the subsequent proximal ectodomain; (ii) aa 135 to 166, which includes the mucin-like domain and extends slightly into the central conserved domain; and (iii) aa 204 to 292, which includes the mucin-like domain downstream from the central conserved domain. Second, the diversity in these regions shows patterns that appear to change based on the presence or absence of host functional lymphocytes. Diversity in variable region ii is seen predominantly before the rise in ACD8/ALC (Fig. 2). Diversity in variable region iii appears across all time points, and diversity in variable region i is seen predominantly after the rise in ACD8/ALC. Third, the within-host diversity seen in the G protein contains features that have not been reported in epidemiological studies of RSV-A (discussed below), raising questions about the extent to which consensus sequences mask true viral population diversity and the role of G protein variants in RSV quasi-species fitness in transmission and infection.

The within-host diversity of the transmembrane domain is unexpected, as the transmembrane domain is frequently highly conserved in molecular epidemiology studies of the RSV-A G protein (Fig. 3) (13). However, escape mutations in the transmembrane domain have been demonstrated to emerge under selection by a neutralizing antibody targeting the G protein (10). In that example, reduced amounts of full-length G protein were detected on the cell surface and incorporated into the virion, and increased
amounts of the secreted form were seen in the cell supernatant. These results could reflect that mutations in the transmembrane domain lead to diminished incorporation of the G protein into the virion, or that the mutations lead to greater susceptibility to cleavage resulting in Gs. The observation, then, of transmembrane domain variants indicates an adaptive response to the pressure of posttransplant immunity, either through diminished levels of Gm leading to decreased antibody binding or elevated levels of Gs functioning as decoys and decreasing the amount of available soluble antibody.

Similarly, while a number of truncated G proteins have been observed in RSV-B in vivo (29), we are unaware of reports of truncated G proteins in RSV-A in vivo. Studies of RSV escape from antibody-mediated selection have shown that premature stop codons in the C-terminal region can lead to antibody escape in cell culture (11). The presence of premature stop codons at every time point, however, suggests that the emergence of C-terminally truncated G protein is not due solely to immune pressure. In addition, the presence of minority populations with premature stop codons in the first mucin-like domain, prior to the heparin binding domain, during the pretransplant time points raises questions about whether the variant-containing viruses are viable or fit. While in vitro and mouse in vivo experiments have demonstrated that RSV lacking the G protein can still infect cells, on the whole such virus appears to infect less efficiently and replicate more slowly (30, 31).

Diversity in the F protein. In the initial study of these samples, consensus sequencing of the palivizumab binding domain of the fusion (F) protein (aa 186 to 315) revealed no changes across the sample set despite multiple administrations of palivizumab (19). In contrast, deep sequencing of the viral populations (Fig. 2) shows that minority populations contain known palivizumab escape mutations at aa 268 (N268I [32]), 272 (both K272E and K272N [33, 34]), and 275 (S275F [33, 34]); haplotype analysis demonstrates that the variant alleles at positions 268, 272, and 275 occurred in separate populations. The presence of these minority populations only after day +22 (13 days after the start of palivizumab infusions) suggests that palivizumab exerted a selective pressure in this population, consistent with what has been observed rarely in palivizumab-treated populations (35).

The N268I variant at day +39 is the only F protein variant that becomes dominant. It is notable that the N268I variant does not appear until day +28, after multiple administrations of palivizumab, single-dose polyclonal RSV-Ig, and lymphocyte infusion at the time of transplant. The more commonly epidemiologically observed palivizumab escape variants at aa 272 and 275 appear during the pre-engraftment period but remain minority alleles despite multiple rounds of palivizumab administration, likely reflecting levels of palivizumab insufficient to generate enough selective pressure on the viral population to overcome the presumed fitness cost of the variants. It is also notable that the clinical samples evaluated all were obtained from the upper respiratory tract (nasal aspirates), and the concentrations of functional RSV-neu-
Palivizumab immune pressure measured in the upper tract may not have been great enough to create and sustain a consensus amino acid mutation within this anatomic site. The additional mutations observed at aa 97, 176, 269, 446, and 526 predominantly appear after the rise in ACD8 and may reflect the influence of specific or nonspecific immune responses. The impact of these additional mutations on F protein function and viral fitness are uncertain but bear further investigation as functional immune epitopes.

**Diversity in other regions of the genome.** Variant sites observed in the other 9 genes in the genome that arise after bone marrow transplant and subsequent increase in ACD8/ALC may reflect sites undergoing immune selection. Notable sites include aa 78 and 102 in the nonstructural protein 2 (NS2; thought to modulate the immune response through interaction with RIG-I and decreased TRAF3 and STAT2); aa 53 and 207 in the phosphoprotein (P protein), thought to be a key component of the viral RNA-dependent RNA polymerase complex); aa 50, 53, and 54 of the matrix (M) protein; and aa 184 and 1280 in the polymerase (L protein). Whether these variants emerged due to selection from specific or nonspecific immunity, genetic hitchhiking (when the frequency of a variant changes due to linkage with a positively associated variant elsewhere in the genome), or neutral drift is unclear. Comparison of the patterns of diversity in these genes in the within-host sample and the samples from molecular epidemiological studies shows no clear overlap (see Fig. S2A to I in the supplemental material).

While the rate of nonsynonymous versus synonymous substitutions (dN/dS) can be helpful in specific situations in determining evidence of positive selection, dN/dS is difficult to interpret for deep sequencing data of a population in which variants have not yet fixed (37). Interpretability is further undermined by not being able to ascribe linkage across full gene sequences: synonymous sites might be polymorphic due to hitchhiking with nonsynonymous sites in a way that cannot be ascertained due to the short sequence read lengths. Moreover, even without this problem, in general, positive selection need not lead to elevated values of dN/dS for within-population polymorphisms (37). Thus, in the absence of a clear theoretical basis for interpreting the value of the dN/dS statistic in the context of these deep sequencing data, we do not present it and agree with the need for further study and development of methods that permit interpretation of evidence of evolutionary pressure in similar deep sequencing data sets.

**Conclusions.** This study presents the first whole-genome deep sequencing analysis of clinical RSV samples, and the specific aspects of this chosen clinical case offer a unique opportunity to evaluate the impact of differing levels of immune function on the nature and extent of within-host viral population diversity. We observe baseline diversity in the initial set of samples prior to bone marrow transplant and increasing diversity correlating with the presence of donor lymphocytes. The initial diversity may reflect a mixed inoculum, drift, or anatomically distinct populations that are captured in the upper respiratory clinical sampling procedure. The data demonstrate increasing diversity in the presence of adaptive immunity, though it is unclear to what degree the diversity is due to specific versus nonspecific inflammatory processes.

Through the detailed description of pathogen variation over the course of an individual infection, this study raises a number of questions regarding RSV pathogenesis and evolution. Do the patterns of variation observed in this example represent the unique circumstance of longstanding RSV infection and the combination of bone marrow transplant and immunosuppression, or are they representative of general RSV dynamics during individual infections? Do the variable sites which emerge after the rise in lympho-

**FIG 3** Comparison of within-host and population-level G protein diversity. Black lines indicate the Shannon diversity index at each residue in the population-wide analysis, and dashed red lines indicate the maximum Shannon diversity index across all time points in the within-host samples. See “Comparison of within-host and population-wide diversity” in Materials and Methods for a description of the analytical approach using the sequences reported in Table S1 in the supplemental material.
cyte count reflect specific or nonspecific immune pressure, other selective pressures, or neutral drift? Does the abundance of variants in the G protein transmembrane domain indicate greater production of Gs as an antibody decoy or another process, and does the rare appearance of transmembrane domain variants in population-wide surveys of G protein reflect the limited ability of these variants to transmit and cause new infections? Are the variable sites seen in genes coding for nonsurface antigens evidence of selection in these genes, and if so, through what mechanism? Additional deep sequencing whole-genome studies over the course of individual infections in immune-naive and immune-experienced immunocompetent hosts will help address many of these questions.

ACKNOWLEDGMENTS

This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN272200900018C. M.L. was supported by grant U54GM088558 from the National Institute of General Medical Sciences. Internal funds from the DeVincenzo laboratory were also used for sample processing and technique development.

We thank Tiffany Poon and Jennifer Kelliher for their technical assistance.

REFERENCES


