Step I: White Paper Application

Application Guidelines

1. *The application should be submitted electronically per requirements via the web site of any of the NIAID Genomic Sequencing Centers for Infectious Diseases. Include all attachments, if any, to the application.*
2. *There are no submission deadlines; white papers can be submitted at anytime.*
3. *GSC personnel at any of the three Centers can assist / guide you in preparing the white paper.*
4. *Investigators can expect to receive a response within 4-6 weeks after submission.*
5. *Upon approval of the white paper, the NIAID Project Officer will assign the project to a NIAID GSC to develop a management plan in conjunction with the participating scientists.*

White Paper Application

Project Title: Respiratory syncytial virus (RSV) intra-host sequence variation over the course of infection in immune competent and immunocompromised hosts.

Authors: Yonatan Grad, M.D. ,Ph.D., Ruchi Newman, Ph.D.
Primary Investigator Contact:

Name	Yonatan Grad, M.D., Ph.D.
Position	Clinical and Research Fellow
Institution	Brigham and Women's Hospital/MGH
Address	Division of Infectious Diseases Department of Medicine Brigham and Women's Hospital 75 Francis St.
State	Boston, MA
ZIP Code	02115
Telephone	
Fax	
E-Mail	ygrad@partners.org

1. Executive Summary *(Please limit to 500 words.)*

EXECUTIVE SUMMARY

Human respiratory syncytial virus (RSV) is the leading cause of severe viral respiratory disease in infants and young children (*1*), and an important respiratory pathogen in the elderly (*2*)and immunocompromised (*3*). Re-infection can take place throughout life, though in older children and adults infection is usually associated with milder disease. No vaccine for RSV exists, and ribavirin, the single approved therapy for RSV, is toxic, difficult to administer, and of questionable clinical utility. Moreover, while population-wide molecular epidemiology studies have shown that there are multiple co-circulating RSV strains that undergo 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 strains, or the effect of immune status on viral diversity and potential immune-escape mutations.

RNA viruses have high mutation rates, with an estimate of 0.4-1.1 nucleotide errors per genome per round of replication (*4*). Recent technological advances in deep sequencing have yielded insights into the development of viral subpopulations in chronic RNA viral infections such as HCV and HIV, thereby deepening our understanding of viral evolutionary dynamics and identifying the clinically important presence and emergence of drug resistant viruses (*5-6*). Applying these same technologies to acute, self-limited infections with RNA viruses, such as RSV, can make key contributions to understanding viral pathogenesis and the role of immune pressure in generation of viral sequence diversity, generation of and location of immune-escape mutations, as well as assist in the design and application of novel therapeutics and vaccines.

We propose to evaluate RSV sequence diversity and the effects of immune pressure on viral molecular evolution through deep sequencing over the course of infection in immunocompetent and immunocompromised hosts. To do so, we have obtained collections of upper airway samples from individuals of varying immune status with RSV infection. These include multiple samples from a single infant with Severe Combined Immune Deficiency Syndrome (SCID) who had persistent RSV infection continuing over several months before and after institution of an effective

immune response, samples from multiple experimentally infected healthy adults, all inoculated with the same wild-type virus, and samples from multiple previously healthy RSV infected infants who had been naturally infected with RSV. Quantitative RSV load has been obtained from parallel aliquots of all of these samples, and clinical parameters have been characterized in association with each time point.

These studies will provide the first comprehensive view of viral evolutionary dynamics during individual infections, and comparisons among the datasets derived from individuals with varying immune states will provide insight into the interaction between host immune function and viral evolution. Furthermore, this approach will lay the groundwork with which to investigate the evolutionary dynamics of other acute viral infections.

2. Justification

RSV is an enveloped virus with a nonsegmented negative stranded RNA genome in the family *Paramyxoviridae* (*7*). The two major subgroups, A and B, were initially defined on the basis of distinct patterns of reactivity to panels of monoclonal antibodies (*8-9*). The antigenic differences correspond to genetically distinct subgroups, with the most significant antigenic and genetic diversity located in the attachment (G) glycoprotein (*10-11*). The G protein, a type II glycoprotein with mucin-like domains, is also one of the main targets for inducing neutralizing and protective antibodies (*12*). The fusion (F) protein can also induce protective antibodies (*12*); in contrast to antibodies to the G protein, though, monoclonal antibodies to the F protein cross-react with both subgroups (*13*).

Extensive molecular epidemiology studies have shown that RSV epidemics are comprised of multiple strains from both A and B subgroups, with proportions of the subgroups and strains varying with each epidemic (*10, 14-20*). The complex molecular epidemiological patterns include multiple genotypes or lineages co-circulating in the same community and replacement of the predominating genotype over successive epidemics. The genotype distribution patterns also appear to be distinctive regionally. One study looking at the composition of a single epidemic season in five locations in North America demonstrated that the RSV epidemic in each community was composed of a number of distinct RSV genotypes but the predominant strains and overall patterns of circulating genotypes differed among three of the five communities (*21*). Several studies have demonstrated that, in a single geographic area, strains emerge and disappear, reflecting progressive genetic and antigenic change (*22-23*). A survey of several contemporaneous studies has also shown the rapid worldwide appearance of a new variant, in this case an RSV-B strain with a novel 60 nucleotide duplication (*23-26*). Taken together, these data indicate that each regional RSV epidemic is comprised of a complex combination of locally and globally evolving strains. The potential for global spread of novel RSV strains highlights the importance of improved understanding of the mechanisms by which these novel strains are generated.

That RSV can repeatedly re-infect individuals suggests an ineffective and short-lived immune response to RSV and / or viral variation leading to evasion of the immune response. Studies of decay of antibody titers after RSV infection in adults provide evidence that the boost to RSV adaptive immunity post-infection is short term (*27*) and the progressive accumulation of genetic and antigenic change in the G protein and replacement of dominant genotypes in successive epidemics implies viral variation plays at least some role in susceptibility to re-infection. The molecular epidemiological pattern of multiple, co-circulating strains raises further questions about how genotype composition of an epidemic influences future transmission of homologous and heterologous variants. Moreover, the host conditions under which new genotypes are generated remains completely uncharacterized.

Studies of RSV viral loads have shown that RSV dynamics and viral load parallel the symptoms of

disease in immunocompetent adults and children, with mean peak viral loads on the order of 10^6 Plaque Forming Units/ml in acute RSV infection of childhood (*28-29*) followed by a decline to undetectable levels, reflecting immune system mediated clearance. Since RSV is an RNA virus, we anticipate that it has a relatively high mutation rate, estimated at 0.4-1.1 nucleotide errors per genome per round of replication (*4, 30*). The mutation rate of RNA viruses is speculated to be important in escaping the adaptive immune response through generation of antigenic variation. Interestingly, the mechanism by which ribavirin may function as an antiviral agent is through its effect as a viral mutagen, raising the viral mutation rate to lethal levels (*31*). However, the extent and dynamics of intra-host variation of an RNA viral infection has never been described.

To gain understanding of the extent of RSV variation and to characterize the impact of immune pressure, we propose deep sequencing of RSV from respiratory samples at multiple time points during RSV infection from hosts of varying immune status. We aim to characterize the extent of intra-host viral sequence variation during increasing viral replication at the initiation of infection, peak infection as determined by maximum viral loads, and decreasing viral loads during the time of immune mediated resolution of infection. Through collaboration with Dr. John DeVincenzo, we have obtained nasal wash specimens from several uniquely well-characterized populations, including

First Data Set: a patient with essentially no adaptive immune response (Congenital Severe Combined Immune Deficiency Syndrome (SCID), with persistently high viral loads within aliquots of multiple nasal washes obtained over 80 days;

Second Data Set: Immunocompetent, healthy adults experimentally infected with a standardized known-sequence virus RSV with twice daily nasal wash specimens obtained and aliquoted over the subsequent 12 days.

Third Data Set: previously healthy children with naturally-acquired RSV infection and nasal wash samples over multiple (at least three) days of infection;

Making these samples incredibly rich for the purposes of characterizing molecular evolution, both clinical data and quantitative RSV loads from parallel aliquots are available for each sample. The quantification of virus from these samples was obtained directly from the patient using a fresh quantitative culture technique and also by quantitative PCR. With these samples, we will be positioned to provide the first detailed evaluation of intra-host variation, and, through analysis of variation over the course of infection in multiple immune states, be able to study the impact of immunity on the molecular evolution of RSV over the course of infection.

First Data Set. a. Characterize the extent and nature of RSV sequence diversity in a human host without the effect of a adaptive immune response. This important and rich specimen data set is derived from a patient with severe combined immune deficiency syndrome (SCID). This syndrome is characterized by lack of T and B cells and hence a complete lack of ability to mount an adaptive immune response. This experiment of nature provides a unique ability to observe RSV intra-host sequence diversity in a human host, where congenital absence of both main arms of the adaptive immune response is observed for approximately one month, and which is associated with sustained high-level viral replication over a prolonged time (approximately 1 month). Deep sequencing of these specimens will provide a baseline viral sequence variation rate in the absence of adaptive immune pressure. This baseline sequence variation rate will then be used to compare the dynamics and variation seen in immune-competent populations (see items 2 and 3 below).

Samples of quantitatively collected respiratory secretions were obtained serially from this 4-month old boy with severe combined immune deficiency and prolonged, symptomatic RSV infection.

The patient was treated with numerous high doses of palivizumab (monoclonal antibody against the F-surface protein of RSV, and an infusion of human RSV-immune globulin. This evaluation and sample collection occurred over a >80-day period encompassing the times before and after receipt of an allogenic bone marrow transplant (BMT) and subsequent engraftment (*29*). See Figure 1. The patient was followed over approximately twelve weeks, including six weeks prior to BMT, when he received four infusions of palivizumab, then followed by the BMT itself and extending six weeks post-transplant including administration of immunosuppressants for graft-versus-host prophylaxis. The samples from this study include 26 serial nasal washes using standardized quantitative collection techniques. Serial peripheral blood lymphocyte counts and lymphocyte subset counts were also measured indicating times of engraftment (timing of the beginning of a cognate antiviral immune response). Clinical characterization of the infant included his respiratory condition over the entire course.

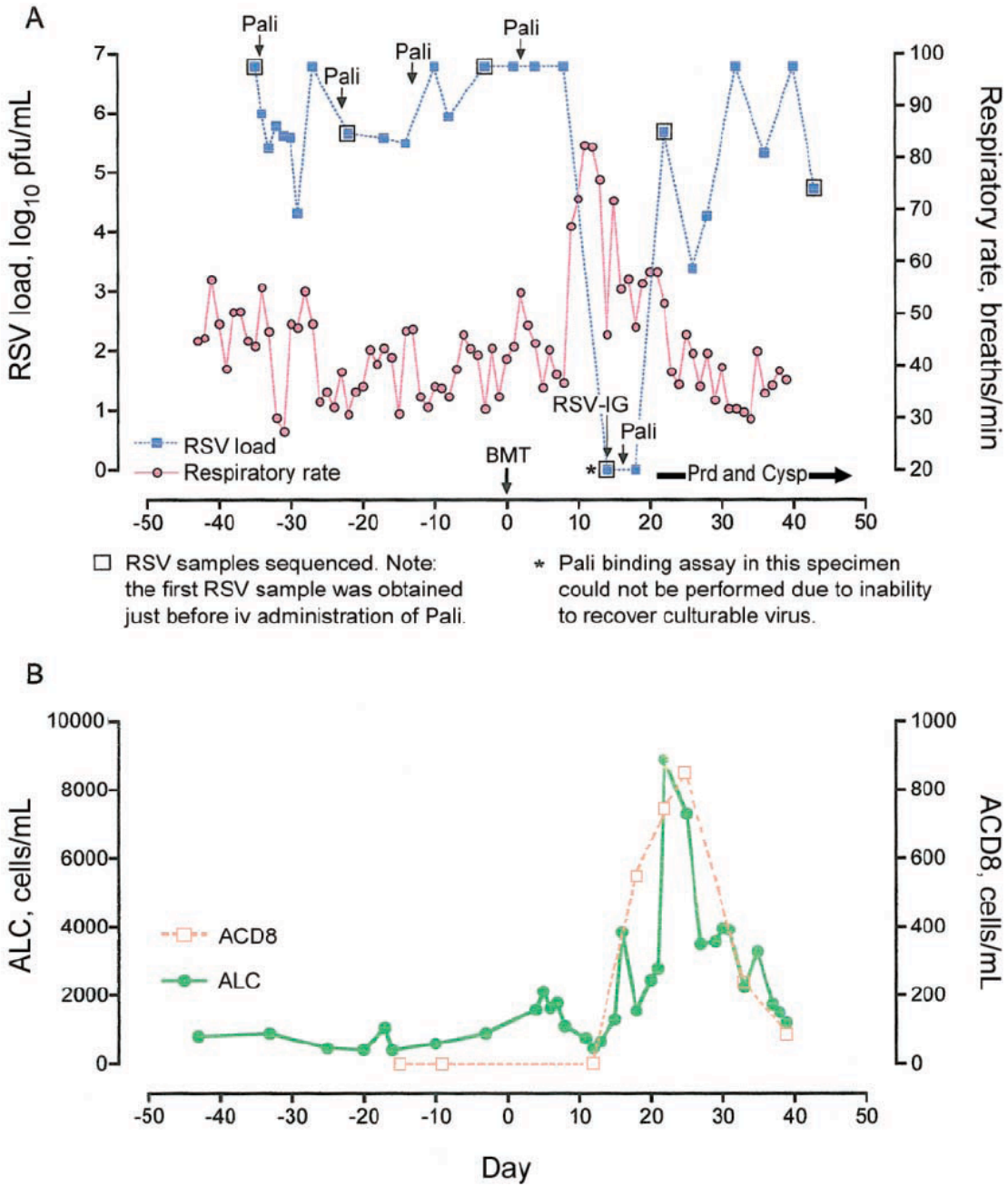

Figure 1 (subject and samples from First Data Set). A. Temporal association between the RSV load in nasal wash specimens, the disease severity (respiratory rate), and the use of various therapeutic interventions in relation to the time of bone marrow transplantation (BMT; day 0). B. Temporal variation in absolute lymphocyte counts (ALCs) and absolute CD8 cell counts (ACD8s) in relation to the time of the BMT (day 0). Cysp, cyclosporine; Pali, palivizumab; Prd, prednisone; RSV-Ig, RSV immune globulin. [From El Saleeby et al., 2005] (29)

The viral loads were on the order of 10^6-10^7 PFU/mL for at least one month, and likely for three months dating back to the diagnosis of RSV infection. The persistent high viral load indicates the inability of this patient's immune system to suppress or clear RSV infection and raises the question of the extent of diversity given the persistently elevated viral loads. One hypothesis is that deep sequencing of RSV samples from this individual will reveal extensive sequence variation reflecting the absence of immune-mediated selection. An alternative hypothesis is that there will be little sequence diversity, reflecting selection of the most fit viral strain over the patient's several months-long RSV infection. In either case, the sequence variability observed will characterize the fitness landscape in the absence of adaptive immunity and provide a baseline to compare populations in the presence of immunity.

First Data Set. b. Characterize the nature and extent of diversity before and after immune-mediated reduction of RSV viral load. One striking finding in this dataset is the reduction in viral load to undetectable levels at the time of increase in absolute lymphocyte and absolute CD8 count, followed by rebound to high levels after the initiation of immunosuppression and decrease in absolute lymphocyte and absolute CD8 count. That the bone marrow donor was the patient's father and likely exposed to the RSV shed by the patient suggests that the clinical response may have been due to mature and expanded effector T cells that escaped the mild T cell depletion that was performed, prior to stem cell infusions. The inverse correlation between the drop in RSV viral load and the rise in symptoms and lymphocyte number implies a link between cell mediated RSV clearance and lung disease.

Analysis of the sequence variation before and after this dramatic decrease in RSV viral loads (two weeks post-BMT) will provide insight into the selective pressure of the immune system as well as possible founder effect mediated by the extreme depletion of the viral population, followed by a bounceback as immune pressure was removed.

First Data Set. c. Evaluate the effects of palivizumab administration on RSV strains, specifically looking at diversity of the F protein. In this study, six RSV isolates were selected from before palivizumab exposure to 50 days after the first of multiple doses of palivizumab. These isolates were passaged twice in HEp-2 cell culture and evaluated for palivizumab binding in an immunofluorescent assay. A segment of the F gene from these 6 cultured supernatants was amplified by PCR and sequenced. The sequenced fragment encompassed the known palivizumab binding site on the F protein, and the technique used for sequencing was the standard Sanger sequencing. Interestingly, the sequence remained constant over all samples analyzed. In the only other study to look for escape mutants in humans receiving palivizumab, functional assays demonstrate continued palivizumab susceptibility in RSV samples from 25 children receiving RSV prophylaxis with monthly palivizumab (*32*)

There are several interpretations to this observation. The first is that absence of amino acid variation could indicate that this segment of the F protein is strongly constrained. Secondly, the concentrations of palivizumab may not have been sufficient to generate selective pressure on the RSV isolates examined due, perhaps, to known difficulty of diffusion of IgG molecules into the respiratory secretions of the upper respiratory tract. However, even if this were the case, the duration of high viral load and presumed high level RSV replication would have been expected to generate variation under a neutral model. Lastly, there may have been multiple strains of RSV, and the obtained sequence may reflect only the dominant strain.

Deep sequencing of full RSV genomes will be able to address the constraint on the F gene as well as extent of variation for the virus under conditions of high viral load for an extended period of time by asking whether variant strains with F gene mutants exist. The constraint hypothesis would predict that the degree of observed mutational variation with the F protein binding sites for palivizumab would be less than that observed at other sites within the genome.

Second Data Set. Characterize the dynamics of RSV intra-host sequence variation in experimentally infected immunocompetent adults. After evaluating the intra-host RSV viral diversity during human infection in the absence of adaptive immune response, it is important to compare the variation seen over the course of an entire infection in healthy immunocompetent adults. A data set from adults who have been experimentally infected with a known and fully sequenced wild-type strain of RSV is being made available to this study by Dr. DeVincenzo. This data set and original aliquots of respiratory secretions were obtained from these healthy adults who were closely monitored in a quarantined setting. This allows for investigation of intra-host viral evolution under pressure of the normal human immune system having had experience with prior natural RSV exposures.

Thirty-five adults were inoculated with the identical GMP-produced preparation of RSV (RSV Memphis 37), which has been fully sequenced. Of these, 27 were defined as infected by quantitative PCR assay, and 21 were defined as infected by quantitative culture. The subjects were observed in quarantine for a total of 13 days. Nasal washes were obtained pre-RSV inoculation and then twice daily post-RSV inoculation on days 1 through 12, and clinical parameters, including daily physical exams and twice-daily symptom score cards, were monitored and recorded. See Figure 2.

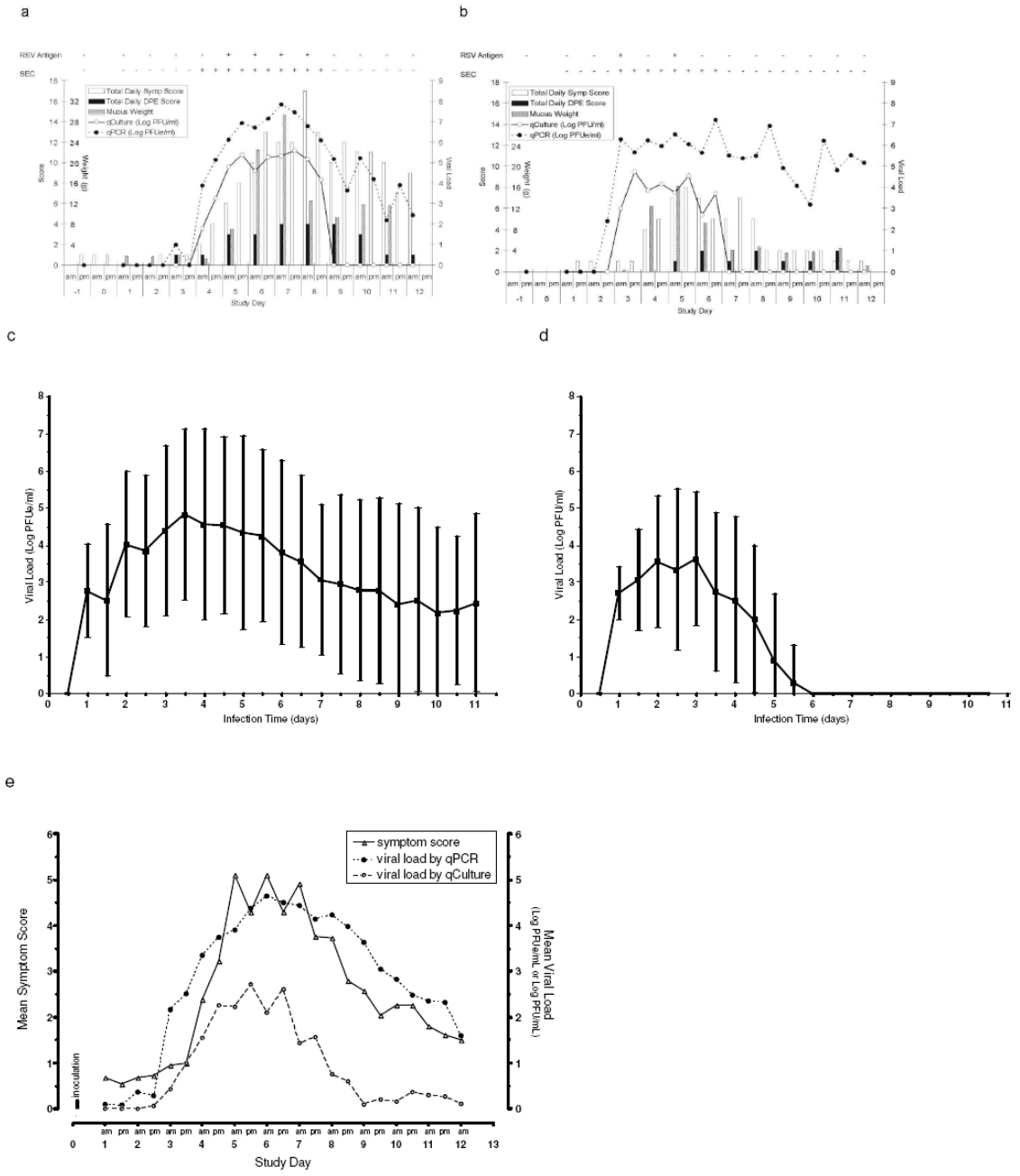

Figure 2. Viral load and disease (subjects from Second Data Set). **(a and b)**: Selected individual volunteer's viral and disease measurements over time. Volunteers were inoculated with RSV on day 0 and evaluated in quarantine until day 12. Twice daily symptom scores (open bars), once daily directed physical exam (DPE) scores (closed bars), daily quantification of nasal mucus weights (hashed bars) are measures of RSV disease severity. Viral testing was performed on nasal washes collected twice daily. The non-quantitative twice-daily spin enhanced culture (SEC) and daily RSV antigen test results are indicated as ±. Quantitative real-time reverse transcriptase PCR (qPCR, dashed line) and quantitative culture (plaque assay on HEp-2 cells, solid line) were measured twice daily. **(c and d)** Mean daily viral loads (qPCR and quantitative culture) with normalized incubation periods. The timing of individual volunteers' viral load curves were normalized by inserting a common incubation period so that their first positive viral measurement occurred arbitrarily on Infection Day 1. Panel c represents viral load measured by qPCR and panel d represents viral load measured by quantitative culture. Error bars represent SD. **(e)** Timing of mean viral load and symptomatic disease. Mean data from all infected volunteers from each collection time point starting from day 1 post-inoculation is shown. No adjustment for the variable incubation periods is included in this panel. This makes the breadth of the curves wider and the magnitude of the curves lower than when incubation periods are normalized (as in c and d). Initial rises in mean viral loads (qPCR and quantitative culture) correlate with the timing of onset of the rise in mean symptom scores. The timing of peak viral load correlates with the occurrence of peak symptom severity. See (*37*).

These samples will allow for the evaluation of viral evolution from a known and sequenced viral strain under pressure from the normal human immune response with prior RSV exposures. We will be equipped to characterize the extent of diversity and the variation of diversity over the course of infection in multiple individuals. Thus, we can compare intra-host viral evolution in multiple individuals from a fixed starting point. Moreover, we can compare the diversity and variation in diversity seen in immunocompetent hosts to the results from an immunocompromised host, as in dataset #1. This data set will uniquely allow us to compare the viral diversity seen during the initial viral load rise (a time during which there is minimal adaptive immune pressure on the virus) vs. the viral diversity observed during the clearance phase of the infection (a time during which adaptive and effective immune responses are present). No other data set and sample set is available world-wide to address this issue since it is logistically nearly impossible to evaluate the early phases of RSV infection during natural infection. Experimental wild-type RSV experimental infections have only been performed in this model system and these unique samples are being made available to us for the purposes of this study (see letter of collaboration from John DeVincenzo).

Third Data Set. Characterize the extent and nature of RSV sequence diversity in immunocompetent RSV-naïve infants over the course of infection. RSV is a winter-time virus in temperate climates of the Northern hemisphere, including in the USA. Therefore, once an infant is identified as being RSV infected, it is also possible to determine, with a high degree of accuracy, whether that infant has ever experienced an RSV infection before, simply by evaluating the date of their infection and their date of birth. We will characterize RSV sequence diversity in this third data set composed of samples collected from previously healthy, immunocompetent, RSV- naïve infants who have been naturally infected by RSV. The first dataset will provide an opportunity to look at the extent of intra-host RSV sequence variation in an immunocompromised infant followed over twelve weeks with high viral loads, thus characterizing RSV diversity in a setting with limited adaptive immune pressure. The second dataset provides an opportunity to query for sequence variation in healthy adults with prior RSV exposure and compare immunocompromised and immunocompetent hosts, and to evaluate sequence diversity during infections all of which start from a single specific virus sequence (Memphis-37 experimental infection). In this third dataset, we will examine the patterns of RSV diversity seen in previously healthy, RSV- naïve (ie, infants with no prior RSV exposures) RSV infected infants evaluated over several days during their infection. As RSV causes severe respiratory tract infection in neonates, it will be important to characterize the patterns of diversity seen in this population. Infants can be characterized as RSV-naïve through knowledge of their date of birth and the knowledge of the predictable yearly seasonality of RSV in the geographic area of patient recruitment (Memphis, TN) (Figure 3)

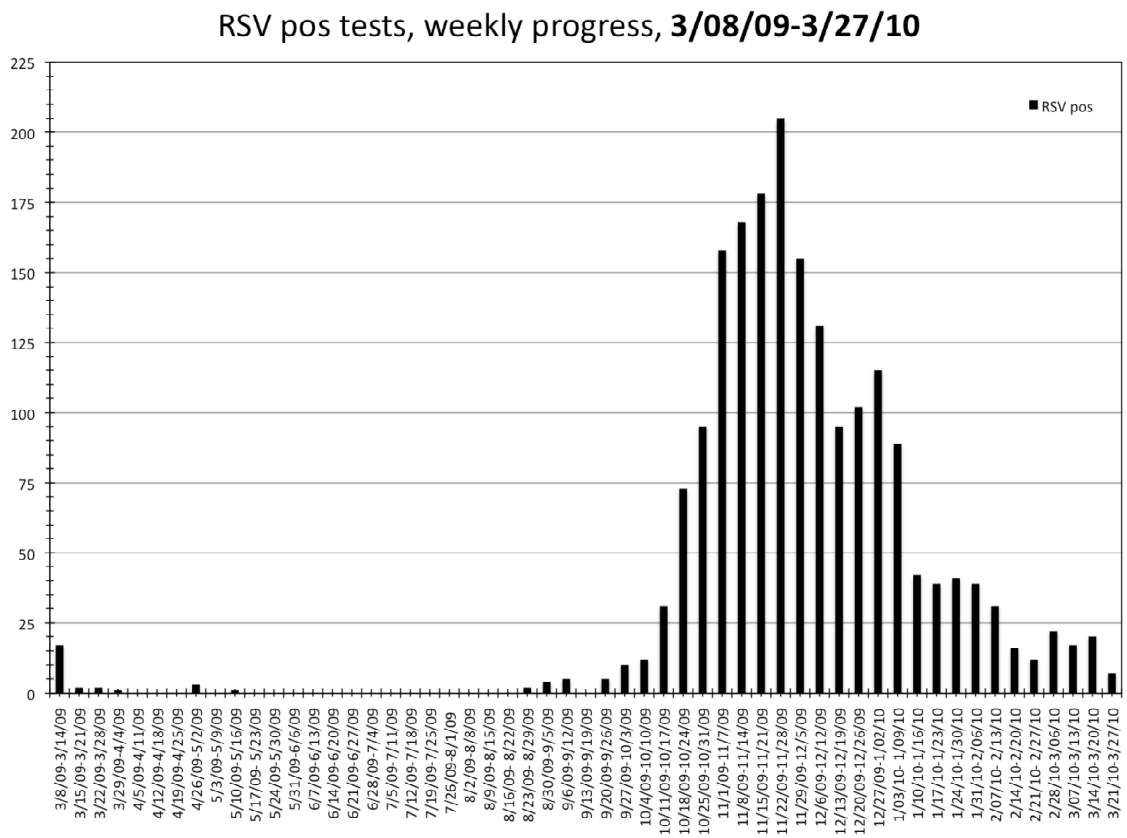

Figure 3. RSV epidemic seasonality in Memphis TN, USA (Characteristics of Third Data Set). Bars represent the number of RSV infected children presenting to the Le Bonheur Children's Hospital (Memphis TN) each week. The clinical samples for Data set 3 were all obtained from this site of enrolment. Predictability of RSV epidemic start and stop dates is very high. Data on numerous years is available upon request (Unpublished data courtesy of Dr. DeVincenzo, Director: Methodist/Le Bonheur Molecular Diagnostics Laboratory, Memphis, TN. USA)

Among the samples in this third dataset, viral loads are noted to increase or stay stable in approximately 20% of the patients during their first 24 hours of evaluation, but to decrease in 80% of the patients during that same first 24 hour sampling period. Consequently, following the viral sequence variation over time with concomitant clinical and viral load data will allow for observation of the impact of immunity on molecular evolution of RSV. It will be particularly interesting to look for differences in infants younger than 3 months as compared to those older than 3 months, as there are data indicating that neonates of this age are unable to produce a antibody repertoire with optimal antibody affinity maturation due to an antibody diversity engine limited by a significantly lower frequency of somatic mutations (*33*).

This third dataset includes data and samples from a total of 219 RSV infected children with mean age of 115 days collected over five RSV seasons. RSV was measured in respiratory secretions by fresh plaque assay over 3 consecutive hospital days (*34*) to determine RSV viral loads and viral clearance. Clinical parameters of disease severity, including duration of hospitalization, need for intensive care, and respiratory failure, were collected. Nasal aspirates include samples from 218 subjects on day 1, 102 subjects on day 2, and 41 subjects on day 3, with mean nasal viral loads decreasing over each day (mean viral load ± SEM were 4.63 ± 0.13 log PFU/mL, 3.63 ±0.18 log PFU/mL, and 2.91 ± 0.33 log PFU/mL on days 1,2, and 3 respectively). Viral load at enrollment was not an independent predictor of hospitalization, but higher viral loads predicted more prolonged hospitalization. Similarly, higher viral load independently predicted the development of

respiratory failure. Importantly, a more rapid rate of RSV clearance (Δ viral load/Δ time)was an independent statistically significant predictor of disease severity.

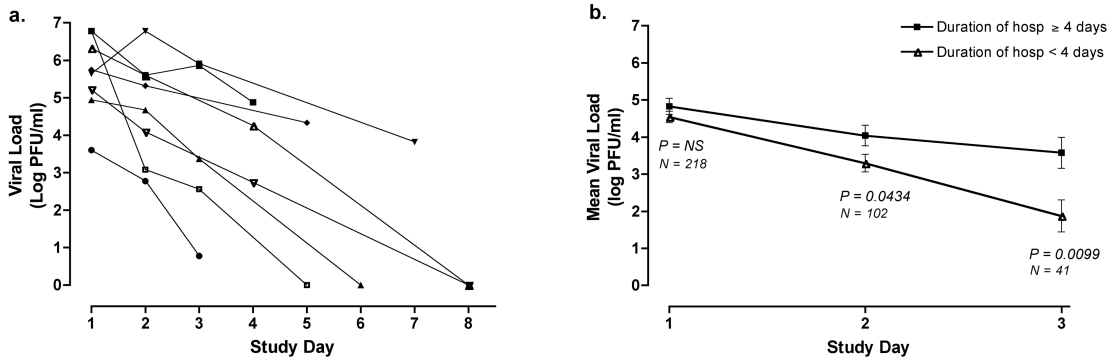

Figure 4. Viral dynamics from naturally infected children (subjects from Third Data Set). **a.** Representative viral load curves from individual subjects. **b.** Mean viral loads in enrolled subjects for first three days since study enrollment. Individual observations are mean RSV viral loads. Error bars represent the standard error of the mean. (Figures and data from DeVincenzo et al., 2005 Presented at International Conference on Respiratory Viral Infections, Curacao)

This dataset provides the opportunity to track intra-host viral populations and the dynamics of viral populations in response to adaptive immune pressure in the first exposure to RSV. In roughly 20% of the study population, viral loads were increasing or stable initially before declining; in the remaining 80%, viral loads were decreasing during the study period. We will aim to look for signatures of immune pressure on RSV; moreover, the varying antibody response in infants < 3 months old and those older than 3 months provides an opportunity to search for a correlate in molecular evolution of RSV.

3. Rationale for Strain Selection

The isolates in this analysis represent common wild-type strains of RSV. They are generally of two origins. RSV has two major strains, RSV-A and RSV-B. RSV-A is the most predominant strain world-wide. The RSV from the immunocompromised patient (First Data Set) is RSV-A. The RSV strain used in the experimental infection of adults (Second Data Set) is Memphis 37. The strain of Memphis 37 is RSV-A and its clade (as defined by sequence of the variable region of the G gene) is the common clade of RSV-A (Yale Classification NH/A2) (*38*). Memphis 37 was isolated and GMP-manufactured in Vero cells from the respiratory secretions of an infant hospitalized for bronchiolitis who had known high viral loads of RSV-A. The isolate was plaque-picked and passaged only five times in tissue culture to prepare the inoculum. The genome of this strain has been sequenced, and represents a starting point for analysis of sequences obtained from samples of experimentally infected subjects. The strains represented in the naturally infected infants (Third Data Set) reflect wild type RSV strains from one geographic area (Tennessee) over a multiple year period, and are a naturally-occurring ratio of predominantly RSV-A with some (approximately 20%) RSV-B strains. Full genome sequencing and characterization of these strains will add greatly to the number of available full genome sequences of RSV. In summary, the viruses selected for sequencing analysis in these data sets are representatives of common, naturally-occurring strains in current circulation.

4a. Approach to Data Production: Data Generation

For the purposes of the analyses proposed in this project, we will use 454 sequencing to attain high-fold coverage of RSV viral populations from each of the samples (see Table 1, below).

Sequencing will be done on RNA prepared from nasal wash specimens. Unadulterated nasal wash specimens were collected directly from patients and placed on ice at the bedside. Aliquots were frozen within 30 minutes of collection and maintained in controlled freezers at -80°C. Fresh aliquots were used for analysis of viral load, both through plaque assays to determine plaque forming units (PFU) and from an additional frozen/thawed aliquot through PCR using a standard curve based on samples with known PFUs to generate PFU equivalents (PFUe). Parallel aliquots have been frozen and stored untouched at -80°C since their collection. Total RNA will be prepared from these aliquots via the Qiagen BioRobot EZ1 Workstation.

The data sets selected for sequencing in this project include RNA samples isolated from respiratory secretions of immunocompromised and immunocompetent RSV infected patients with respective high and low viral loads. Given the difficulties in sequencing samples with low viral loads due to low viral RNA input and increased potential for human nucleic acid contamination, we propose to separate the project in to two phases. In Phase I we would tackle those samples that have ample viral loads and can most likely be sequenced using modifications to our existing protocols for 454 sequencing. Additionally, in Phase I we would we would work on development of a strategy and methods to enrich and sequence vRNA from low viremia samples using respiratory samples spiked with known different amounts of laboratory cultured RSV. When we have a robust protocol that we know will work on the pilot set of low viremia samples, we will proceed to Phase II of the project which will entail sequencing of the remaining 613 samples in Data Set 2 and 3.

Phase I: We will adapt the Broad Institute's current 454 high-throughput viral sequencing pipeline and intra-host diversity sequencing method for HCV, Dengue, WNV and HIV to RSV samples with high viral loads (Sample Set 1 – 26 samples). The existing viral sequencing and assembly pipeline was designed to be robust to viral type and application to RSV should only demand design of a RSV primer panel. The Broad will work closely with the collaborators to design an optimal primer panel using the primer design tools available at the Broad. Once at the Broad, the RNA samples will undergo RT-PCR, construction of barcoded 454 libraries and sequencing with an anticipated target genome coverage of 200X.

Additionally, in anticipation of the known challenges to sequencing the low viremia samples (Data Set 2 and 3), we will work with collaborators to develop a strategy to enrich and sequence vRNA from low viremia samples. Initial studies will use experimental samples generated by spiking respiratory samples with known different amounts of laboratory cultured RSV. These samples will be used for experiments testing various methods to enrich viral for RNA such as RNA amplification, alternative genome specific RT and PCR protocols, addition of nuclease treatment steps, hybrid selection using viral specific baits and oligoarray enrichment. If we are able to identify a method that works well with experimental samples, then a subset of the samples described in Samples Sets 2 and 3 from healthy patients will be tested (12-24 samples in Phase I). Identification of a robust protocol that is successful at generating high quality 454 sequence from the pilot set of samples will be required to proceed to Phase II of this project.

Phase II: We will use the low viremia protocol defined in Phase I of this project to sequence the low viremia samples defined in Data Sets 2 and 3 (613 samples).

Table 1:

First Data Set: RSV infected immunocompromised patient studied over 80 days	
Number of nasal wash samples for deep sequencing:	**26**
Second Data Set: RSV infected healthy adults studied over 12 days	
Number of nasal wash samples for deep sequencing:	**252**
Third Data Set: RSV infected previously healthy neonates studied over three days	
Number of samples from day 1	218
Number of samples from day 2	102
Number of samples from day 3	41
Number of nasal aspirate samples for deep sequencing:	**361**
TOTAL number of samples for deep sequencing:	**639**

4b. Approach to Data Production: Data Analysis

Sample analysis will begin with characterization of sequence diversity. RSV sequence diversity will be described by the number of unique sequences obtained through 454 sequencing and the relative abundance of each sequence, using the overall population size as has been characterized through viral loads (either PFU or ePFU). For sequences from the First and Third Data Set, in which the inoculating sequence is not known, we will determine whether there is an early-occurring single dominant sequence that may represent the inoculating strain, and can use maximum parsimony phylogenetic methods to infer the initial strain sequence. We will determine average pairwise distance for all genome sequences as a measure of diversity. For sequences from the Second Dataset, the experimentally infected healthy adult population, in which we know the sequence of the infecting strain (RSV Memphis 37), we will begin by determining the number of genome sequences with nucleotide and amino acid mutations from this sequence. For the Third Data Set, we will use the full genome sequence and inferred infecting strain sequence to determine whether multiple neonates were infected with the same or similar RSV strains; if so, we can also compare the patterns of diversity manifest by the same viral strains in multiple hosts.

We will then evaluate for evidence of selection. As a first step, we will characterize dN/dS to obtain the position and frequency of synonymous and non-synonymous polymorphisms within each gene. However, it is worth noting that dN/dS will need to be interpreted with much caution, as it is a test originally designed to compare fixed silent and non-silent mutations from divergent species, and not polymorphisms from within a host. We suspect that deep sequencing will reveal many neutral or mildly deleterious sequences. As such, the relationship between the observed dN/dS derived from sequences within a population and the selection coefficients may be uncertain.

In datasets with multiple samples over the course of an individual host's infection, we can use counting approaches and pairwise Hamming distance to search for evidence of selection on the basis of population persistence. For those genes in the RSV genome with sufficient diversity, which we suspect will be the gene coding for the G protein and particularly the antigenic epitopes, we can characterize the fraction of sequences with 0-*x* number of polymorphisms and describe the persistence and diversity of the polymorphisms over time. Through maximum likelihood phylogenetic constructions we can visualize the polymorphisms that persist and contribute to the overall population diversity. We will take particular interest in comparing sequence variation occurring within the regions coding for known human dominant T-Cell epitopes (*39*) and will, among other comparisons, evaluate the diversity within the first phases of the infection (while viral load is still increasing) to later time points within the individual's infection (while viral load is

declining).

Contingent on sufficient diversity, we can also consider using the Bayesian probabilistic algorithm BEAST (*40*) to infer phylogeny. For the First Data Set (extending until the time of BMT), we can apply a continuous-time coalescent model assuming constant population size. For the Second and Third Data Sets (and the section of dataset 1 post-BMT), we can adjust the coalescent model to account for increasing and decreasing populations. As we will be able to track viral populations quantitatively, we will also be able to evaluate where the models may fail to fit the data and explore the aspects of the models that are violated to gain additional understanding both of the dynamics of RSV intrahost viral evolution and ways to improve modeling.

If there are sufficient numbers of variants, we can search for presence of epistasis, demonstrated by co-occurring nucleotide/amino acid substitutions observed at a statistically significant rate both within and across hosts and immune populations.

Analyses of the location of polymorphisms and their co-occurrence will advance understanding of immune pressure and RSV genome epistatic phenomena. To date, molecular epidemiology studies have focused narrowly on the G protein. Most changes in the G protein are localized at an ectodomain containing two hypervariable segments separated by a highly conserved region between amino acids 164 and 176. Molecular epidemiology studies have suggested specific amino acids of the G protein that are subject to positive selection, and have hypothesized a "flip-flop" model in which the available amino acid repertoire at these sites is restricted (*41*). We will be able to evaluate the extent of variation at these proposed sites of selection and at recognized T-cell receptor immunodominant epitope sites (*39*) over the course of individual infections, while also characterizing the extent of variation of other genes. As an additional specific hypothesis to test, we will be equipped to clarify whether administration of Palivizumab generated immune pressure by looking in dataset #1 for any changes in F protein sequence variation in temporal association with Palivizumab doses.

By making publicly available the RSV subpopulation sequences over the course of infection, this study will provide the raw material for other investigators to study aspects of viral molecular evolution within a host. We are unaware of any other comparable data set that includes clinical information and quantitative RSV viral loads sampled over the course of infection; the addition of deep sequencing will make this an invaluable data set for hypothesis generation with regard to RSV viral evolution and, more broadly, RNA viral evolution; immune pressure and selection for viral genetic and antigenic change; and the relationship between viral evolution in individual hosts and population-wide molecular epidemiology. Moreover, deeper understanding of viral change in response to selective pressures more generally may help guide the development of novel therapeutics.

REFERENCES

1. C. B. Hall, in *Principles and practice of infectious diseases,* G. L. Mandell, Bennett, J.E., Dolin, R., Ed. (Churchill Livingstone Elsevier, Philadelphia, PA, 2010), vol. 2, pp. 2207-2221.
2. A. R. Falsey, P. A. Hennessey, M. A. Formica, C. Cox, E. E. Walsh, *N Engl J Med* **352**, 1749 (Apr 28, 2005).
3. Y. J. Kim, M. Boeckh, J. A. Englund, *Semin Respir Crit Care Med* **28**, 222 (Apr, 2007).
4. J. W. Drake, J. J. Holland, *Proc Natl Acad Sci U S A* **96**, 13910 (Nov 23, 1999).

5. A. M. Tsibris *et al.*, *PLoS One* **4**, e5683 (2009).
6. B. B. Simen *et al.*, *J Infect Dis* **199**, 693 (Mar 1, 2009).
7. P. L. C. Collins, R.M.; Murphy, B.R. , in *Feilds virology,* P. M. H. D.M. Knipe, D.E. Griffin, R.A. Lamb, M.A. Martin, B. Roizman, and S.E. Straus, Ed. (Lippincott Williams & Wilkins, Philadelphia, PA, 2001), pp. 1443-1485.
8. L. J. Anderson *et al.*, *J Infect Dis* **151**, 626 (Apr, 1985).
9. M. A. Mufson, C. Orvell, B. Rafnar, E. Norrby, *J Gen Virol* **66 (Pt 10)**, 2111 (Oct, 1985).
10. P. A. Cane, *Rev Med Virol* **11**, 103 (Mar-Apr, 2001).
11. P. R. Johnson, Jr. *et al.*, *J Virol* **61**, 3163 (Oct, 1987).
12. M. Connors, P. L. Collins, C. Y. Firestone, B. R. Murphy, *J Virol* **65**, 1634 (Mar, 1991).
13. G. Taylor, E. J. Stott, J. Furze, J. Ford, P. Sopp, *J Gen Virol* **73 (Pt 9)**, 2217 (Sep, 1992).
14. L. J. Anderson, R. M. Hendry, L. T. Pierik, C. Tsou, K. McIntosh, *J Infect Dis* **163**, 687 (Apr, 1991).
15. P. A. Cane, D. A. Matthews, C. R. Pringle, *J Clin Microbiol* **32**, 1 (Jan, 1994).
16. E. H. Choi, H. J. Lee, *J Infect Dis* **181**, 1547 (May, 2000).
17. O. Garcia *et al.*, *J Virol* **68**, 5448 (Sep, 1994).
18. C. B. Hall *et al.*, *J Infect Dis* **162**, 1283 (Dec, 1990).
19. T. C. Peret, C. B. Hall, K. C. Schnabel, J. A. Golub, L. J. Anderson, *J Gen Virol* **79 (Pt 9)**, 2221 (Sep, 1998).
20. J. Reiche, B. Schweiger, *J Clin Microbiol* **47**, 1800 (Jun, 2009).
21. T. C. Peret *et al.*, *J Infect Dis* **181**, 1891 (Jun, 2000).
22. P. A. Cane, C. R. Pringle, *J Virol* **69**, 2918 (May, 1995).
23. K. T. Zlateva, P. Lemey, E. Moes, A. M. Vandamme, M. Van Ranst, *J Virol* **79**, 9157 (Jul, 2005).
24. P. D. Scott *et al.*, *J Med Virol* **74**, 344 (Oct, 2004).
25. A. Trento *et al.*, *J Gen Virol* **84**, 3115 (Nov, 2003).
26. A. Trento *et al.*, *J Virol* **80**, 975 (Jan, 2006).
27. A. R. Falsey, H. K. Singh, E. E. Walsh, *J Med Virol* **78**, 1493 (Nov, 2006).
28. J. DeVincenzo *et al.*, *Proc Natl Acad Sci U S A* **107**, 8800 (May 11, 2010).
29. C. M. El Saleeby, J. Suzich, M. E. Conley, J. P. DeVincenzo, *Clin Infect Dis* **39**, e17 (Jul 15, 2004).
30. R. Belshaw, A. Gardner, A. Rambaut, O. G. Pybus, *Trends Ecol Evol* **23**, 188 (Apr, 2008).
31. M. Vignuzzi, J. K. Stone, R. Andino, *Virus Res* **107**, 173 (Feb, 2005).
32. J. P. DeVincenzo *et al.*, *J Infect Dis* **190**, 975 (Sep 1, 2004).
33. J. V. Williams, J. H. Weitkamp, D. L. Blum, B. J. LaFleur, J. E. Crowe, Jr., *Mol Immunol* **47**, 407 (Dec, 2009).
34. C. El Saleeby, A. J. Bush, L. M. Harrison, J. A. Aitken, J. DeVincenzo, *In review.*, (2010).
35. C. M. El Saleeby *et al.*, *J Pediatr* **156**, 409 (Mar, 2010).
36. S. M. Perkins *et al.*, *J Clin Microbiol* **43**, 2356 (May, 2005).
37. J. P. Devincenzo *et al.*, *Am J Respir Crit Care Med*, (Jul 9, 2010).
38. Alvarez R *et al.*, *Antimicrob Agents Chemother.* **53**, 9 (Sep 2009).
39. M. Olson and S. Varga, *Vaccines* **7,** 8. (2008).
40. A. J. Drummond, A. Rambaut, *BMC Evol Biol* **7**, 214 (2007).
41. V. F. Botosso *et al.*, *PLoS Pathog* **5**, e1000254 (Jan, 2009).

5. Community Support and Collaborator Roles:

Collaborators on this project include the following:

Yonatan Grad, MD, PhD. Clinical and Research Fellow, Division of Infectious Diseases, Brigham and Women's Hospital / Massachusetts General Hospital. Visiting Scientist, Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health.

Marc Lipsitch, DPhil. Professor of Epidemiology, Director of the Center for Communicable Disease Dynamics, Harvard School of Public Health. Associate Member, Broad Institute.

John DeVincenzo, MD. Professor of Pediatrics, Division of Infectious Diseases, University of Tennessee School of Medicine. Professor of Molecular Sciences, School of Graduate Health Sciences, University of Tennessee. Medical Director, Molecular Diagnostics and Virology Laboratories, Methodist / Le Bonheur Children's Hospitals, Memphis, Tennessee.

Ruchi Newman, Ph.D., Research Scientist, The Broad Institute of MIT and Harvard.

Matthew Henn, Ph.D., Director of Viral Genomics, The Broad Institute of MIT and Harvard.

6. Availability & Information of Strains:

Samples are stored in Dr DeVincenzo's laboratory in Memphis, Tennessee. If this white paper receives funding, samples will be shipped to the Broad Institute.

Each sample has corresponding extensive clinical data as well as quantitative RSV viral loads as described above and in more details in the references below. These data have already been published, or are under pending journal review and are available for community use.

7. Compliance Requirements:

7a. Review NIAID's Reagent, Data & Software Release Policy:

NIAID supports rapid data and reagent release to the scientific community for all sequencing and genotyping projects funded by NIAID GSC. It is expected that projects will adhere to the data and reagent release policy described in the following web sites.

http://www3.niaid.nih.gov/research/resources/mscs/data.htm

http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html

<Each Center to include their website that describes/points to the guidelines>

Once a white paper project is approved, NIAID GSC will develop with the collaborators a detailed data and reagent release plan to be reviewed and approved by NIAID.

Accept ☒ Decline ☐

7b. Public Access to Reagents, Data, Software and Other Materials:

All sequences and read files generated under this proposal will be submitted to the archive at NCBI/NLM/NIH on a weekly basis. These data will also include information on templates and quality values for each sequence.

Genome assemblies will be made available via GenBank and the Broad Institute web site. Assembled contings and scaffolds will be deposited in GenBank within 45 calendar days of completing assemblies. If it is determined that the final assembly can be significantly improved, an updated record will be deposited in the appropriate part of Gen Bank when complete.

Annotation data will be made available via GenBank and the Broad Institute web site after consistency checks and quality control have been completed by the GSCID and collaborators. Assuming no significant errors are detected during the validation process, annotation will be released within 45 calendar days of being generated.

7c. Research Compliance Requirements

Upon project approval, NIAID review of relevant IRB/IACUC documentation is required prior to commencement of work. Please contact the GSC Principal Investigator(s) to ensure necessary documentation are filed for / made available for timely start of the project.

Investigator Signature:

Investigator Name: **Yonatan Grad** **Date:** **January 7, 2010**

Blank Last Page