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: Examination of Differential Occupancy in early innate immunity in an infection time course.

Author: Shannon McWeeney on behalf of PNWRCE & MMIC (Systems Virology)

Primary Investigator Contact:

Name	Shannon McWeeney
Position	Associate Professor, Biostatistics and Bioinformatics
Institution	OHSU
Address	3181 SW Sam Jackson Park Rd CR145
State	OR
ZIP Code	97239
Telephone	503-494-8347
E-Mail	mcweeney@ohsu.edu

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

Biological Relevance for Transcription Factor Selection

The main pathway for the induction of type I interferons (IFN) by viruses is through the recognition of viral RNA by cytosolic receptors and the subsequent activation of interferon regulatory factor 3 (IRF-3), which drives IFN-alpha/beta transcription. IRF-3 has been shown to be a target by viruses to interfere with type I IFN induction. We propose to examine changes in chromatin occupancy for the transcription factor IRF-3 pre and post infection for both Influenza (A/VN/1203/04) and SARS (SARS-CoV) in Calu-3 cells. This will provide us with a genome-wide list of targets for IRF-3 and allow us to identify temporal changes in occupancy, as well as to compare common changes in occupancy between the two viruses. Another member of the early innate immunity signaling pathway, STAT-1 will also be examined to allow comparisons across key transcription factors for the same temporal window. (Figure 1).

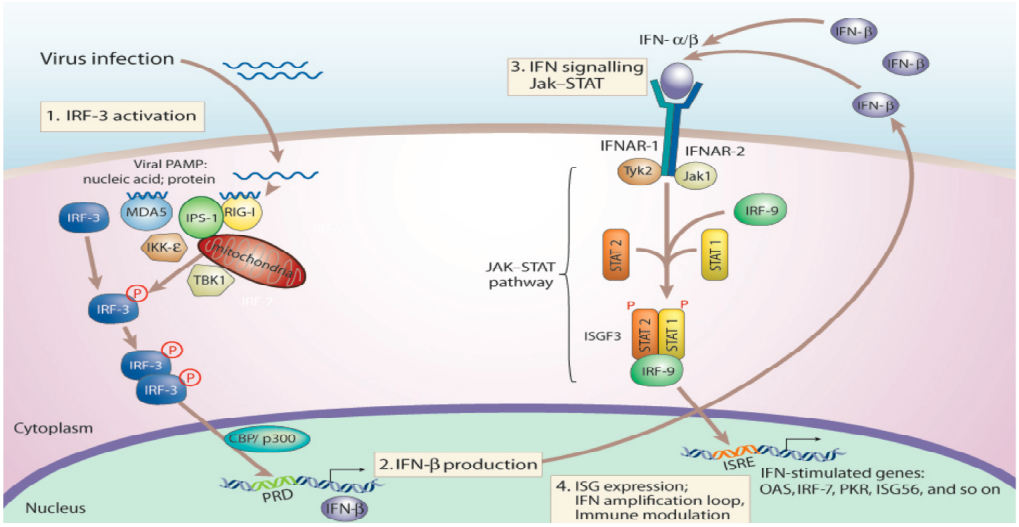

Figure 1. Graphical representation of early innate immune signaling which highlights the role of the two transcription factors, IRF-3 and STAT-1 that have been selected for ChIP-seq experiments (Figure courtesy of Michael Gale).

Experimental Data Supporting Choice of Transcription Factors

IRF3

The role of IRF3 in SARS and Influenza has been well documented. Early work by Talon et al (2000, J Virology) showed that double-stranded RNA (dsRNA) binding protein NS1 of influenza virus can prevent the potent antiviral interferon response by inhibiting the activation of interferon regulatory factor 3 (IRF-3), a key regulator of IFN-alpha/beta gene expression. IRF-3 activation and, as a consequence, IFN-beta mRNA induction are inhibited in wild-type (PR8) influenza virus-infected cells but not in cells infected with an isogenic virus lacking the NS1 gene (delNS1 virus). This suggested that inhibition of IRF-3 activation by a dsRNA binding protein significantly contributes to the virulence of influenza A viruses and possibly to that of other viruses. Dr. Kawoka, one of the co-PIs on the System Virology contract, found that the amino acid S42 of NS1 is critical for the H5N1 influenza virus to antagonize host cell interferon induction and for the NS1 protein to prevent the double-stranded RNA-mediated activation of the NF-kappaB pathway and the IRF-3 pathway (Jiao et al 2008, J Virology). In SARS, a study by Devaraj et al (2007, J Biol Chem) identified a papain-like protease, PLpro, which interacts with IRF-3 and inhibits the phosphorylation and nuclear translocation of IRF-3, thereby disrupting the activation of type I IFN responses through either Toll-like receptor 3 or retinoic acid-inducible gene I/melanoma differentiation-associated gene pathways. This work suggested that the regulation of IRF-3-dependent innate antiviral defenses by PLpro may contribute to the establishment of SARS-CoV infection. Given the large number of publications supporting the role of IRF3, we focus our experimental data on the rationale for STAT1.

STAT1

Dr. Baric, one of the co-PIs on the System Virology demonstrated (Freidman et al (2010) Plos Pathogens) that SARS-CoV Urbani and rMA15 viruses induce severe end stage lung disease by a STAT1 dependent mechanism that is independent of IFN receptor type I, II and III signaling. The data point to a novel mechanism by which STAT1 function regulates disease severity in the lung following SARS-CoV induced acute lung injury. In Urbani virus infected lungs, minimal changes were seen in 129 WT, IFNAR1−/−, IFNAGR−/− mice for all 4 cytokines examined (Figure 2A, B, C and D). However, in STAT1 −/− mice, significant increases in protein expression patterns were detected across different time points, peaking between 9 and 15 days post-infection. TNFα protein levels increased steadily from day 2 through day 29 post-infection while IL6 and IFNγ protein levels peaked at day 9, before reducing to levels seen in mock infected animals. MCP1 protein levels were increased between days 5 through 22 but diminished by day 29 post-infection.

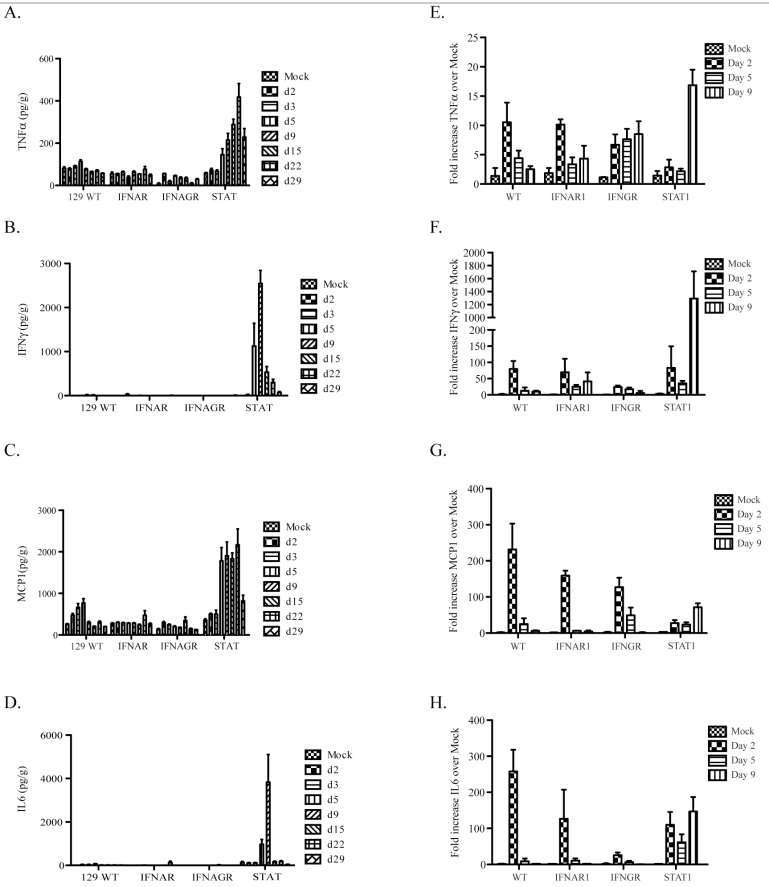

Figure 2. A–D. A Cytometric bead array was used to quantify the protein levels of cytokines in lung homogenates after infection with the Urbani virus. Lung samples were homogenized and protein levels of (A) TNFα, (B) IFNγ, (C) MCP1 and (D) IL6 were quantified. E–F. Real-time PCR analysis was performed on RNA from infected mouse lungs. Mice were harvested at days 2, 5 and 9 post-infection and lung RNA isolated by Trizol. cDNA was made from 1ug of RNA of each and used in real-time analysis and the results for 3 mice per group were combined. Fold increase over mock-infected lungs of each cytokine is graphed for (E) TNFα, (F) IFNγ, (G) MCP1 and (H) IL6.

Preliminary data from the Systems Virology contract highlights STAT1 regulation in Infuenza as well. Network inference using weighted co-expression gene network analysis (WCGNA) was conducted on microarray time series data for influenza in Calu-3 cells (Figure 3). Modules (sub-networks) were validated based on significant topological overlap (compared to the null distribution) and functionally annotated. One of the most significant modules (color-coded by navy blue in dendorgram in Figure 3) was enriched for functional annotations related to immune response and response to virus. In-silico analysis of the upstream regions of the transcripts in this module found significant enrichment for the STAT1 transcription regulatory element (TRE) (FDR adjusted p-value = 0.004). This module was also significantly enriched for members of the JAK-STAT pathway. Addition of the chIP-seq data with the microarray, proteomics (in progress) and network modeling, would allow us to further elucidate the role of STAT1 and delineate key targets involved in early response. We note that the

corresponding data in SARS is forthcoming (hybridizations are underway).

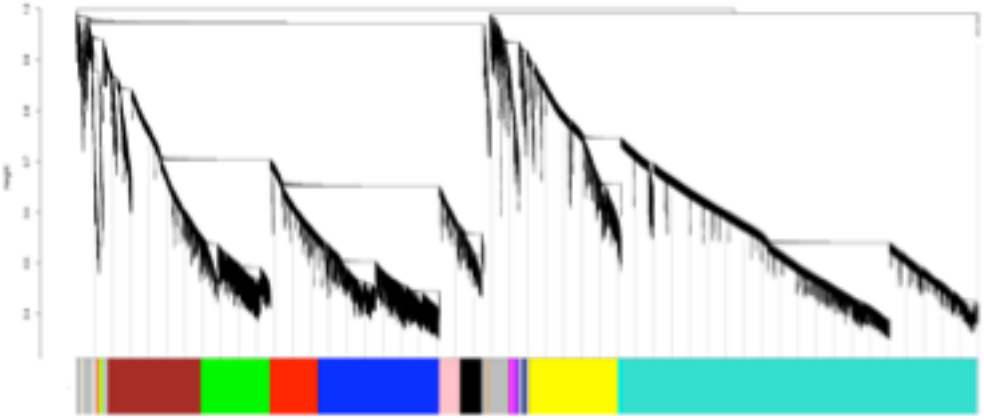

Figure 3. Dendrogram from ICL004 Influenza time series in Calu-3 cells. Transcripts are represented as vertical lines and are arranged in branches according to topological overlap (TO) similarity using hierarchical clustering. The dynamic treecut algorithm was used to automatically detect 20 highly connected sub-networks or "modules" represented as colors below the dendrogram. Modules are summarized as "eigengenes" by taking the first principal component of module member's intensity values.

2. Justification

Influenza virus and SARS-CoV represent serious ongoing threats to the public health and economy. New technological approaches are needed to protect global health. The data generated in this proposal will enhance ongoing systems biology modeling in two large-scale NIAID-funded efforts. This systems level approach allows us to gain a deeper understanding of how these viruses cause disease and to provide new targets for the rational development of therapies against these and future threats.

3. Rationale for Strain-Selection

Influenza virus. One of the best known attributes of influenza virus is its ability to occasionally generate strains that can cause global pandemics. Three such pandemics have occurred over the last century, the worst of which, the 1918 "Spanish flu," resulted in approximately 500,000 deaths in the U.S. and over 50 million worldwide . The potential for the emergence of deadly new subtypes of influenza virus was dramatically evident in the 1997 outbreak of H5N1 avian influenza in Hong Kong, in which 6 of 18 infected individuals died. Fortunately, this virus did not spread readily by person-to-person contact and the outbreak remained isolated. Of particular concern is the growing number of cases of direct avian-to-human transmission, which since mid-2003 has resulted in the infection of 340 individuals and 208 deaths (WHO Report on the cumulative number of confirmed human cases of avian influenza A/H5N1; December 14, 2007). Whether naturally occurring or genetically engineered, highly lethal forms of influenza virus also make for potential bioterrorism agents. Influenza virus, and H5N1 viruses in particular, are therefore a significant public health concern and additional research is necessary to better understand viral pathogenesis and to improve rapid diagnostic, therapeutic, and vaccine strategies.

SARS-CoV. The SARS-CoV epidemic was a particularly informative harbinger of the relationship between highly pathogenic respiratory viruses, human health, and modern civilization. Nearly 5 years after the emergence of SARS-CoV, effective licensed vaccines and drugs do not exist to protect the public health.

4a. Approach to Data Production: Data Generation

Cell line

We use a human lung epithelial cell line as the primary experimental system from which to generate high-throughput data that will be used for computational modeling and iterative rounds of experimental validation and model refinement. To provide a common host cell background for our modeling efforts, and to strengthen our ability to make comparisons of the host response to different viruses, it is preferable that all investigators work with the same type of cells. Therefore, we have selected Calu-3 cells, a human airway epithelial cell line, as the best candidate for a cell line that can be infected with influenza virus and SARS-CoV (see preliminary data Figure 4). The literature supports their use for a broad range of respiratory viruses, including influenza and SARS-CoV. All research projects and cores use cells from the same stocks and passage numbers from thawing are recorded, not to exceed 15 passages from thawing. Biological cell replicates are defined as cells from different passage numbers.

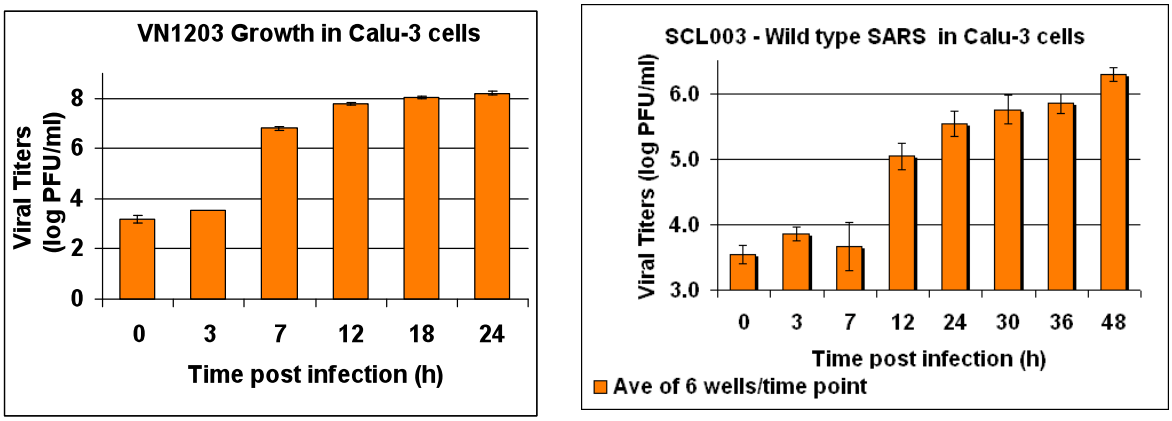

Figure 4. Influenza (left) and SARS (right) viral growth curves in calu-3 cell lines. Similar patterns were seen for both viral genomic and viral mRNA.

Experimental Design for Chip-Seq

We wish to compare temporal changes in occupancy in Calu-3 cells for SARS, Influenza and a mock control pre and post-infection. Based on our preliminary Calu-3 experiments for these viruses, we would examine 8 time points (0, 1.5, 3, 7, 12, 24, 36 hours) to examine changes over the course of infection. Host response (based on microarray data) to SARS begins later than in Influenza-infected cells and our time point selection reflects this We will use 3 biological replicates for each experimental timepoint along with

accompanying input chromatin and IgG controls for each replicate.

Key questions we wish to investigate are the following:

(1) What is the earliest temporal window in which we can detect differences in chromatin occupency? This is key as it will help us refine the window for intial host response. **We hypothesize that we will see changes in occupency (compared to mock) as early as 1.5 in influenza and potentially as early as 3 hours in SARS (earlier then when we are able to detect differential expression)**. To what degree are the targets between the two transcription factors unique or overlapping? To what degree do we see virus specific differences with regard to target genes? With chromatin occupency are we seeing the same targets later, virus-specific targets or both? **We hypothesize that the majority of the differences will be overlapping but shifted with respect to time but propose that there may be a larger proportion of unique targets for STAT1 for SARS.**

(2) To what extent do changes in chromatin occupency correlate with downstream transcriptional and proetomic responses (data generated by the NIAID contract)? We note that there is a clear temporal shift in host response between the two viruses and differences with regard to differentially expressed genes as well. **We hypothesize that we will see a high correlation with the expression data. However, we propose that a number of targets will be identified for which we do not detect diffferences in expression.** Using a network modeling approach, we will be able to elucidate modulation of host response by the virus and how this manifests with regard to binding, expression and translation.

(3) What percentage of the target nodes and hubs identified from our network inference and modeling are targets of IRF-3 and Stat1? We have noted that the hub genes are often not differentially expressed based on transcriptional data. **However, we hypothesize that many of the hub genes may be targets of key transcription factors (either IRF-3, STAT1 or other transcription factors detected in-silico).**

ChIP protocol (UW)

ChIPs will follow procedures described in detail in previous publications (Impey et al. 2004, Cha-Molstad et al., 2005). ChIP-Seq libraries will be generated using modifications to the standard Solexa Genomic DNA sequencing protocol.

Amendment: As discussed in March phone call with Broad, we will review current protocols used by the Broad to assist the UW group with library construction and update this section based on those conversations.

4b. Approach to Data Production: Data Analysis

For initial determination of binding positions for each replicate, we will use the window tag density (WTD) method described in Kharchenko et al. 2008. We have worked with this algorithm in the past (Bottomly et al. 2010) and have found that it performs well. Briefly, after dividing the genome into windows a binding score is calculated as a function of positive and negative strand reads for each experimental sample. This score will be corrected after subtraction of the adjusted read counts observed for that window in the corresponding input chromatin sample. As an extension to this approach, a false discovery rate will be computed relative to the corresponding input corrected IgG sample. We will examine the read count of the union of the significant peaks from WTD for differential

occupancy using an empirical bayesian method which has been shown to effectively model SAGE and RNA-Seq data (Robinson et al. 2007, Robinson et al. 2008, Robinson et al. 2010). This method is designed to work with count data where overdispersion is likely an issue and there are relatively few replicates present. Although this method provides some facilities for normalization, we will also explore the issue in more detail. As the question of differential occupancy is still nascent, Dr. McWeeney will provide methodological support to provide extensions to this framework during the project period. Dr. McWeeney has been developing computational workflows and methodology for pre-NGS chip-seq data for the last five years (Impey et al 2004, Yochum et al 2007a, Yochum et al 2007b, Agarwal et al 2007, Keller et al 2007), as well as NGS (Walter et al 2009, Bottomly et al 2010) making her expertise in this area invaluable.

Given our interest in differential occupancy, the use of a input control is important because regions exist that have read counts that deviate from an expected uniform background distribution. We have observed such suspect regions with large number of reads in centromeric and repetitive regions. Similar phenomena have been observed by others such Ji et al 2008, Jothi et al. 2008 and have been more thoroughly characterized by Kharchenko et al. 2008 . For this reason, we are requesting an input control be included. Similarly the inclusion of an IgG control will allow us to more accurately assess our rate of false discovery by providing us with an estimate of the number of peaks of specified size we would see by chance from experimental or technical artifacts.

We will examine the sequences surrounding the set of peaks using a de novo motif finder to look for the presence of both motifs previously thought to be associated with the transcription factor of interest (to assist in refining the binding site) and others that could be acting as co-regulaters (to assist in our modeling and validation efforts). An example of this approach is our recent work in which we found canonical TCF4 motifs as well as the unexpected overrepresentation of AP-1 motifs in the regions surrounding beta-catenin peaks (Figure 5, Bottomly et al 2010). We will perform ChIP assays to validate binding of predicted transcription factors that appear to be co-localizing with IRF-3 and STAT1

uding these transcription factors in our modeling predictions.

nt of ChIP-Seq peaks to genes is by no means straight forward we will
ıt, distance and effect size into a score for each gene and for each
as an extension of the methodology of Ouyang et al. 2009. We will
is information into our existing co-expression networks in order to examine the impact of increased or decreased occupancy on variation in transcriptional levels over the timecourse. This information will be further related to both proteomics and metabolomics data to get a better idea of downstream effects and inform our dynamic modeling efforts.

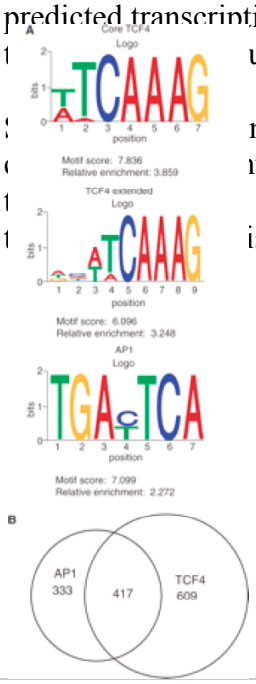

Figure 5. Overrepresented motifs found in the regions surrounding significant beta-catenin peaks in colorectal carcinoma cells. Both expected TCF4 motifs as well as AP-1

motifs were observed providing some evidence for the joint regulation of targets by TCF4, AP-1 and beta-catenin. Figure from Bottomly et al (2010).

In summary, the chIP-seq data generated by this collaboration will assist the NIAD funded efforts by (1) assisting us in refining the window of early innate host-response (2) providing direct binding evidence that will be integrated with our proteomics and transcriptomics data to assist in our network inference and modeling (3) allow us to identify other co-regulators via in-silico sequence analysis of binding regions (4) assist in prioritizaiton of target nodes for validation and (5) allow us to compare our results between the two viruses to identify underlying mechanisms of host response to infection as well as potential virus-specific respsonses.

5. Community Support and Collaborator Roles:

The data from this project would be utilized in modeling efforts in the Mathematical Modeling and Informatics Core (MMIC) within the NIAID funded Systems Virology contract (Katze, PI) and the Bioinformatics and Biostatistics Core within the Pacific NW Regional Center of Excellence (Nelson and Katze, co-PIs), both of which have relevant systems biology projects for NIAID Priority Pathogens, highly pathogenic avian influenza virus and severe acute respiratory syndrome associated coronavirus (SARS-CoV) This data would complement existing microarray, proteomics and metabolomics experiments that are already being generated under these mechanisms. Both the contract and the PNWRCE are fully supportive of this application. The libraries will be constructed at the University of Washington (UW) and Dr. McWeeney's group at OHSU will conduct the data analysis. Dr. McWeeney is the Director of the Biostatistics and Bioinformatics Core for the PNWRCE and co-leader of the MMIC.

6. Availability & Information of Strains:

Samples are available via the existing funded projects in the PNWRCE and Systems Virology contract. .

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:

The data would be provided to the modeling efforts in both the Systems Virology contract and the Pacific NW Regional Center of Excellence. The data would also be disseminated publicly via the Systems Virology portal (http://www.systemsvirology.org). We will amend the existing data sharing plan and timeline for data release in conjunction with NIAID.

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: **Shannon McWeeney**

Blank Last Page