SANDIA REPORT

SAND2013-0935
Unlimited Release
Printed Feb 2013

Identification of Host Response Signatures of Infection

Steven S. Branda, Anupama Sinha, Zachary W. Bent

Prepared by
Sandia National Laboratories
Livermore, California 94551-0969

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576–8401
Facsimile: (865)576–5728
E–Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553–6847
Facsimile: (703)605–6900
E–Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2013-0935
Unlimited Release
Printed Feb 2013

Identification of Host Response Signatures of Infection

Steven S. Branda[1], Anupama Sinha[2], Zachary W. Bent[2]

[1]Biotechnology & Bioengineering Department
[2]Systems Biology Department
Sandia National Laboratories
7011 East Ave, P.O. Box 969
Livermore, CA 94551-0969

Abstract

Biological weapons of mass destruction and emerging infectious diseases represent a serious and growing threat to our national security. Effective response to a bioattack or disease outbreak critically depends upon efficient and reliable distinguishing between infected vs healthy individuals, to enable rational use of scarce, invasive, and/or costly countermeasures (diagnostics, therapies, quarantine). Screening based on direct detection of the causative pathogen can be problematic, because culture- and probe-based assays are confounded by unanticipated pathogens (*e.g.*, deeply diverged, engineered), and readily-accessible specimens (*e.g.*, blood) often contain little or no pathogen, particularly at pre-symptomatic stages of disease. Thus, in addition to the pathogen itself, one would like to detect infection-specific host response signatures in the specimen, preferably ones comprised of nucleic acids (NA), which can be recovered and amplified from tiny specimens (*e.g.*, fingerstick draws). Proof-of-concept studies have not been definitive, however, largely due to use of sub-optimal sample preparation and detection technologies. For purposes of pathogen detection, Sandia has developed novel molecular biology methods that enable selective isolation of NA unique to, or shared between, complex samples, followed by identification and quantitation *via* Second Generation Sequencing (SGS). The central hypothesis of the current study is that variations on this approach will support efficient identification and verification of NA-based host response signatures of infectious disease. To test this hypothesis, we re-engineered Sandia's sophisticated sample preparation pipelines, and developed new SGS data analysis tools and strategies, in order to pioneer use of SGS for identification of host NA correlating with infection. Proof-of-concept studies were carried out using specimens drawn from pathogen-infected non-human primates (NHP). This work provides a strong foundation for large-scale, highly-efficient efforts to identify and verify infection-specific host NA signatures in human populations.

Acknowledgements

We thank Matt Reed and Katie Overheim (Lovelace Respiratory Research Institute, Albuquerque, NM) for providing us with the pathogen-infected NHP specimens that were the focus of this study. We thank Owen D. Solberg, Joseph S. Schoeniger, Kelly P. Williams, and Sidney P. Elmer for bioinformatics support, and Stanley A. Langevin and Deanna Curtis for molecular biology support. Finally, we thank our co-workers at Sandia National Laboratories (SNL) who helped shape this project through formal and informal feedback and advice.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Overview of the Problem and Our Solution

Infectious disease surveillance and outbreak mitigation require rapid, accurate, and reliable means of distinguishing infected *vs* healthy individuals, to enable rational use of countermeasures (diagnostics, therapies, quarantine). Screening populations based on direct detection of the causative pathogen can be problematic, because readily-accessible specimens such as blood often contain little or no pathogen, particularly at pre-symptomatic stages of disease. However, host response to the pathogen is rapid, robust, and evident in blood throughout the course of infection[1]. Thus, screening populations based on host response biomarkers in blood is an attractive approach, especially if the biomarkers are nucleic acids (NA), as these can be efficiently recovered from tiny specimens (*e.g.*, fingerstick draws) and detected with tremendous sensitivity and specificity *via* PCR. Proof-of-concept studies have not been definitive, however, largely because use of sub-optimal sample preparation and detection technologies has precluded comparative analysis of clinical specimens with sufficient sensitivity, specificity, and throughput.

In the context of the RapTOR Grand Challenge Project (142042; 10/1/09-9/30/12), Sandia National Labs (SNL) developed new methods and technologies for: 1) Selective isolation of NA that are unique to, or shared between, clinical specimens; and 2) Highly efficient preparation of NA for Second Generation Sequencing (SGS). In the current study, we used this sample preparation pipeline and SGS to carry out screens for NA biomarkers of infection, focusing on a relatively simple test case: A set of blood specimens drawn from three NHP infected with a biodefense-related pathogen (*Yersinia pestis*, the causative agent of the plague). Each blood specimen was processed to yield white blood cell (WBC) and plasma fractions, RNA was extracted from each fraction, and cDNA was generated from each RNA sample using Peregrine, a newly-developed method for preparation of SGS-ready cDNA libaries from total RNA samples[2]. Each library was sequenced directly. Additionally, in most cases, an aliquot of the library was molecularly suppressed: Hydroxyapatite (HAC) mediated normalization was used to deplete highly-abundant NA[3]; negative capture was used to deplete NA complementary to the probe, which was derived from a different specimen from the same animal; and positive capture was used to enrich for NA complementary to the probe, which was derived from the pathogen itself (*Y. pestis*)[4]. Using this multi-pronged approach, we identified a number of promising candidates for NA biomarkers of infection. In addition to these candidate biomarkers, this work has provided valuable new insight into mammalian host response to a bacterial pathogen of biodefense relevance.

1.2 Background

Historically, the search for disease biomarkers has centered on proteins; in infectious disease research, for instance, cytokines have received particular attention. However, cytokine profiles have relatively low information content (<1000 varieties, ~1000X abundance range, in human

blood), and searches for other proteinaceous biomarkers are constrained by technological limitations (*e.g.*, most specimen fractionation → mass spectrometry approaches show low sensitivity, throughput, and capacity for quantitation). In contrast, NA profiles are information-rich (*e.g.*, mRNA: >100,000 varieties, >10,000X abundance range, in human WBC), and with microarrays it became feasible to conduct high-throughput searches for NA biomarkers of disease. A decade of attempts yielded some notable successes[1,5-7], including diagnostics in clinical use today (*e.g.*, MammaPrint, Oncotype DX). Success rates might have been higher were it not for the low sensitivity and narrow dynamic range of microarrays, which precludes reliable detection of low-abundance NA and accurate quantitation of NA at each end of the abundance spectrum; these constraints cripple attempts to identify NA present at low levels in one specimen (*e.g.*, healthy blood) and high in another (*e.g.*, infected blood). The better tool for this job is SGS, which offers far more sensitive and accurate quantitation across the full spectrum of abundances. In recent years, brute-force SGS of clinical specimens (*e.g.*, conventional RNA-Seq) has led to successful identification of disease biomarkers.[1] However, this approach is slow, labor- and compute-intensive, and expensive, primarily because sequencing bandwidth is dominated by NA that are highly abundant (*e.g.*, in Ref. 8, 75% of reads derived from 7% of expressed transcripts) and/or present at similar levels regardless of disease state (in most experiments, >90% of transcripts are *not* differentially expressed to a statistically significant degree). We hypothesized that selective depletion of highly-abundant and/or consistently-expressed NA from SGS libraries generated from specimen sets would enable us to focus sequencing bandwidth on the potentially informative (less-abundant, differentially-expressed) NA, thereby harnessing the full power of SGS for efficient discovery of NA biomarkers of infectious disease.

2 AEROSOL INFECTION OF NHP WITH *Y. PESTIS*

2.1 Background

Our collaborators at Lovelace Respiratory Research Institute (LRRI) had previously established a well-characterized model of *Y. pestis* infection of NHP (cynomolgus macaques; *Macaca fascicularis*).[9-10] In this model, *Y. pestis* CO92 is delivered in aerosol form at a target dose of 50-150 LD_{50} (*i.e.*, 50-150 times the dose that lethal for 50% of the animals). Previous work at LRRI has shown that the LD_{50} for *Y. pestis* CO92 in this aerosol-delivery model of NHP infection is 55-66 colony forming units (CFU). Once the pathogen is delivered to the airways, the animal typically succumbs to disease within 3-6 days (they are euthanized prior to natural death, in order to minimize suffering). Prior work with this infection model had shown that the earliest sign of systemic infection is fever, which can be detected at ~60 hrs post-exposure; however, consistent detection of fever requires continuous monitoring of each animal's temperature *via* telemetry (intermittent measurement and/or use of a rectal thermometer is not sufficient), and even so there are animals that do not spike temperature and yet clearly show systemic infection in autopsy investigations.[9-10] Other signs of infection (*e.g.*, cardiac and respiratory distress, increased levels of circulating cytokines) are observed only inconsistently and at lates stages of disease (≥72 hrs post-exposure).[9-10] Thus, our collaborators were eager to help us discover molecular signatures of infection in peripheral blood specimens using their NHP infection model. For this purpose, LRRI added three animals to a much larger cohort that they used for study of *Y. pestis* infection. Blood specimens and cage-side observations of the animals were provided for our analysis; unfortunately, due to cost considerations, additional specimens (*e.g.*, necroscopy tissue) and/or data (*e.g.*, cytokine levels) could not be provided by LRRI.

2.2 NHP Infection and Blood Collection

Each animal was held off of feed the night before exposure, and anaesthetized using Telazol (2-6 mg/kg) *via* intramuscular injection. The animals were kept warm under anaesthesia using delta phase heating pads, and their respiration continuously monitored. *Y. pestis* CO92 was suspended in 6-10 ml of brain/heart infusion broth, and aerosolized using a Collison nebulizer. A target dose of 50 LD_{50} of *Y. pestis* CO92 was delivered to each anaesthetized animal in 3.5-5.0 L of aerosol inhaled from a head-only exposure box, adjusting the ventilation rate according to volume measurements made *via* real-time plethysmography. Samples of the aerosol stream were collected for measurement of pathogen content and calculation of the actual dose delivered. Exposure was followed by 3-5 min of fresh air, and the animals' heads were decontaminated (wiped with Amphyl disinfectant solution) prior to their removal from the cabinet. The animals were then returned to their home cages and monitored until they fully recovered from the anesthesia (upright and moving around in their cages).

Cage-side observations were made at 12-hr intervals starting at two days pre-exposure (*i.e.*, Day -2). These observations, as well as assessment of morbidity and mortality, were recorded at each timepoint. Moribund animals were defined as those demonstrating seizures, severe depression, or coma; respiratory distress or severe dyspnea; persistent recumbency and weakness; and unresponsiveness to touch or external stimuli. Moribund animals were immediately euthanized

(25 mg/kg of ketamine followed by 87 mg/kg of pentobarbital and 11 mg/kg of phenytoin *via* intravenous injection).

Peripheral blood specimens (3-5 ml) were collected from anaesthetized animals (5-10 mg/kg of ketamine) *via* venipuncture at: 1) A pre-exposure timepoint; 2) 24 hr intervals starting at 2 days post-exposure; and 3) At terminal sacrifice. The blood was collected into lavender-capped Vacutainer tubes (BD Bioscience) containing the anticoagulant K_2EDTA. The blood was then subjected to centrifugation at 2000 *x g* for 15 min at room temperature (RT). The upper layer (plasma) was recovered into 4 volumes of TRIzol LS (Invitrogen), and the middle layer (buffy coat; white blood cells; WBC) into 4 volumes of TRIzol (Invitrogen); the bottom layer (red blood cells) was discarded. The plasma and WBC fractions (~8 ml and ~2 ml total volume, respectively) were immediately frozen and stored at -80°C.

Tables 1 and 2 summarize key information from the NHP infection and blood collection phase of this study. Pre-exposure blood specimens ("pre-bleeds"; heretofore referred to as "Day 0" bleeds) were collected and ~10 days later (on three consecutive days, after the animals had habituated to the laboratory for ~60 days) the animals were exposed to *Y. pestis* CO92. Post-exposure specimens were collected on Days 2 and 3, as well as on Day 4 for two of the animals (A06861 and A06845). Back-calculating the actual doses delivered revealed a range of 16-87 LD_{50}, which nicely bracketed the target dose of 50 LD_{50}. Cage-side observations revealed little until Day 4 (the final collection day) with the exception of animal A07701, which showed signs of illness on the night of Day 3.

		Pre-Bleed		Y. pestis CO92 Challenge					Post-Bleed 1			Post-Bleed 2			Post-Bleed 3			Sacrifice		
Animal	Sex	Date	Study Day	Date	Study Day	Target Dose (CFU, LD_{50})	Actual Dose CFU (LD_{50})	Bodyweight (kg)	Date	Study Day	Bodyweight (kg)	Date	Study Day	Bodyweight (kg)	Date	Study Day	Bodyweight (kg)	Date	Study Day	Circumstances
A06861	F	6-17-11	-10 (60)	6-27-11	0 (70)	2740 (50)	4780 (87)	2.85	6-29-11	2 (72)	2.78	6-30-11	3 (73)	ND	7-1-11	4 (74)	2.64	7-2-11	5 (75)	found dead
A06845	F	6-17-11	-11 (59)	6-28-11	0 (70)	2740 (50)	871 (16)	2.5	6-30-11	2 (72)	ND	7-1-11	3 (73)	2.18	7-2-11	4 (74)	2,16	7-2-11	4 (74)	moribund
A07701	F	6-17-11	-12 (58)	6-29-11	0 (70)	2740 (50)	1590 (29)	2.65	7-1-11	2 (72)	2.36	7-2-11	3 (73)	ND	ND	ND	ND	7-3-11	4 (74)	moribund

Table 1. Summary of Key Features of Study in which NHP were Infected with *Y. pestis* CO92.

Cage-Side Observations														
Animal	Study Day -2 (68)		Study Day -1 (69)		Study Day 0 (70)		Study Day 1 (71)		Study Day 2 (72)		Study Day 3 (73)		Study Day 4 (74)	
	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
A06861	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NO STOOL	NO STOOL	NO STOOL	NO STOOL	NORMAL	NORMAL	NORMAL	NO STOOL	NORMAL
A06845	SCANT STOOL	SCANT STOOL	SCANT STOOL	SCANT STOOL	NO STOOL	SCANT STOOL	SCANT STOOL	SCANT STOOL	NORMAL	NORMAL	NORMAL	NO STOOL	HUNCHED, SCANT STOOL, MINIMAL URINE OUTPUT, HYPOACTIVE	ND
A07701	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	HUNCHED, REDUCED APPETITE	BRUISING-SLIGHT, HUNCHED, DEHYDRATED-SLIGHT, REDUCED APPETITE, NO STOOL	BRUISING-SLIGHT, HUNCHED, DEHYDRATED-SLIGHT, REDUCED APPETITE, NO STOOL, NO URINE, UNSTEADY, HYPOACTIVE

Table 2. Summary of Cage-Side Observations of NHP Infected with *Y. pestis* CO92.

2.3 Transfer of Blood Specimens from LRRI to SNL

TRIzol a well characterized, powerful, and nearly universal inactivator of pathogens,[11] so it was anticipated that recovery of plasma and WBC into TRIzol LS and TRIzol (respectively) would render the specimens sterile. However, it was necessary to verify their sterility prior to their removal from the LRRI Biological Safety Level 3 (BSL-3) facility and transfer to SNL. For these tests, an aliquot from each specimen, equivalent to 10% of the total volume of the specimen (*i.e.*, 800 µl plasma or 200 µl WBC), was used to inoculate Tryptic Soy Broth (TSB) at 1:10 dilution (*i.e.*, 7.2 ml or 1.8 ml TSB, respectively, and the cultures incubated at 37°C with agitation for 4 days; at this point, each culture was examined for turbidity by two trained technicians, and used to inoculate TSB plates that were further incubated at 37°C for 3 days and examined for bacterial growth.

For 21 of the 22 specimens collected, no bacterial growth was detected in the liquid culture or on the agar plate. In the remaining case (A07701 Day 3 WBC), a single bacterial colony was detected on the TSB agar; this was thought to be due to environmental contamination of the agar, a hypothesis that was supported when re-testing of the specimen showed that it failed to support bacterial growth in liquid medium or on agar. Thus, LRRI determined that all of the 22 specimens collected were in fact sterile, and these were allowed to be removed from the BSL-3 facility and shipped to SNL.

3 PREPARATION OF BLOOD RNA/CDNA FOR SEQUENCE ANALYSIS

3.1 Background

In the course of the RapTOR Grand Challenge Project, we developed new sample preparation methods and technology for: 1) Extraction of total RNA from WBC and plasma; 2) Highly efficient and reliable preparation of SGS-compatible cDNA libraries from total RNA samples; 3) Molecular normalization of cDNA libraries using HAC; and 4) Molecular capture of target cDNA species for their selective enrichment or depletion in SGS libraries. We used these methods and technologies in different combinations to assemble a series of sample preparation pipelines that offered several different perspectives on each blood specimen. In this way we were able to cast a broad net for candidate molecular signatures of infection: Our RNA extraction and cDNA synthesis methods provided access to both polyadenylated [messenger RNA (mRNA)] and non-polyadenylated [*e.g.*, microRNA (miRNA)] RNA species; normalization provided access to the less-abundant library constituents; and capture provided access to constituents unique to, or shared between, libraries derived from each specimen.

3.2 RNA Extraction and Quantitation

Addition of TRIzol (or TRIzol LS) to the blood fractions by LRRI had already effectively lysed all of the cells (bacterial and mammalian) in each specimen, and preserved the RNA species as well. The frozen samples received from LRRI were thawed at 37°C, then mixed with chloroform (200 µl for every 1 ml of TRIzol in the sample) and vigorously vortexed. The mixtures were incubated at RT for 5 min, then centrifuged at 14,000 *x g* at 4°C for 15 min in order to achieve phase separation. A portion (1 ml) of the aqueous phase was transferred to a new Eppendorf tube, mixed 1:1 with 100% nuclease-free ethanol (Sigma), and used in the Direct-zol kit (Zymo Research) for RNA extraction, following the manufacturer's instructions. The purified total RNA was eluted in 10 µl of sterile nuclease-free water and stored at -80°C. Aliquots from each RNA sample were analyzed for quantity and purity using a NanoDrop 2000 (Thermo-Fisher).

Table 3 summarizes the yield and purity measurements for the total RNA extracted from each blood fraction.

The WBC yields averaged 3.90 µg (range: 0.38-9.42 µg), and purities averaged 2.00 (range: 1.94-2.15). These yields were sufficient for all of the planned experimental manipulations, with the exception of that from NHP-B Day 0 (0.38 µg), which was too low to support generation of capture probe in sufficient quantities to be effective.

The plasma yields were measured only in the case of NHP-A fractions; they were found to be exceedingly low (average: 0.06 µg; range: 0.03-0.12 µg), and so it was decided to forego measurement of yield from the NHP-B and NHP-C fractions, in order to conserve material for preparation of cDNA libraries. For the same reason we did not measure the purity of RNA extracts from plasma.

Animal	Blood Fraction	Blood Fraction	RNA Yield (µg)	RNA Purity (A_{260}/A_{280})
NHP-A (A06845)	Day 0	WBC	1.10	1.99
		plasma	0.05	nd
	Day 2	WBC	3.04	2.02
		plasma	0.04	nd
	Day 3	WBC	9.42	1.98
		plasma	0.03	nd
	Day 4	WBC	7.02	1.99
		plasma	0.12	nd
NHP-B (A06861)	Day 0	WBC	0.38	2.15
		plasma	nd	nd
	Day 2	WBC	2.63	1.99
		plasma	nd	nd
	Day 3	WBC	5.24	2.00
		plasma	nd	nd
	Day 4	WBC	8.53	1.98
		plasma	nd	nd
NHP-C (A07701)	Day 0	WBC	1.91	1.99
		plasma	nd	nd
	Day 2	WBC	2.58	1.94
		plasma	nd	nd
	Day 3	WBC	1.01	2.02
		plasma	nd	nd

Table 3. Summary of Yields and Purities of Total RNA Extracts from NHP Blood Fractions.

3.3 First Strand cDNA Synthesis and Quantitation

Preparation of cDNA libraries for SGS involves RNA-templated reverse transcription to generate the first strand of cDNA; synthesis of the second strand of cDNA; and addition of SGS platform-specific (in our case, Illumina) adapters to the ends of the cDNA, which can be accomplished during cDNA synthesis or afterwards *via* ligation. We have developed a new method ("Peregrine") for fast, simple, sensitive, and cost-effective preparation of representative, strand-specific cDNA libraries from both polyadenylated and non-polyadenylated RNA.[2] For this study, we used Peregrine to prepare SGS-ready cDNA libraries from all of the total RNA samples. This section outlines the first steps of Peregrine; following sections outline its later steps, and variations in which molecular suppression (normalization and/or capture) is applied.

For WBC total RNA, 200 ng were subjected to random fragmentation in 20 µl reactions, through addition of 2 µl of 10X NEBNext RNA fragmentation buffer (New England Biolabs) and incubation at 94°C for 3 min, followed by immediate cooling on ice and addition of 2 µl of NEBNext RNA fragmentation stop solution (New England Biolabs). The fragmented RNA was re-purified using the Zymo RNA Clean and Concentrator-5 system (Zymo Research), following the manufacturer's general procedure (for recovery of fragments ≥17 nt) and eluting in 6 µl of nuclease-free water.

For plasma total RNA, no fragmentation step was necessary, as the vast majority of species were already <500 nt in length.

3.5 µl of fragmented WBC RNA (25-50 ng) or non-fragmented plasma RNA (4-16 ng) were mixed with 1 µl of 25 mM primer PP_RT, incubated at 65°C for 2 min, and then immediately cooled on ice. While on ice, 4.5 µl of a master mix containing 2 µl of SMARTScribe 5X First-Strand Buffer, 0.25 µl of 20 mM DTT, 1 µl of 10 mM dNTP mix, 0.25 µl of RiboGuard RNase inhibitor, and 1 µl of SMARTScribe Reverse Transcriptase (all products from Takara) were added, and the mixture incubated at 25°C for 3 min followed by 42°C for 1 min. At this point, 1 µl of 12 mM template-switching oligo PP_TS were added while the reaction mixture remained in the thermocycler, and incubation continued at 42°C for 1 hr. The reaction was then terminated through incubation at 70°C for 10 min. The reaction products (first strand of cDNA) were purified using 18 µl (1.8X volumes) of Agencourt AMPure XP beads (Beckman Coulter Genomics) and eluted in 25-50 µl of nuclease-free water, following the manufacturer's instructions.

A new qPCR-based assay[2] was used to determine the number of PCR cycles required for production and optimal amplification of high-quality double-stranded (ds) cDNA libraries from first strand cDNA synthesis reaction products. After diluting the first strand cDNA at 1:10 in nuclease-free water, 1 µl of the dilution was combined with 5 µl of SsoFast EvaGreen SuperMix (Bio-Rad), 3 µl of nuclease-free water, 0.5 µl of 10 mM primer PP_P1, and 0.5 µl of 10 mM primer PP_P2. The assays were run in quadruplicate on a CFX96 qPCR machine (Bio-Rad), using the following cycle parameters: 95°C for 45 sec, followed by 25 cycles of 95°C for 5 sec and 60°C for 30 sec. The cycle number at which fluorescence intensity exceeded the detection threshold [*i.e.*, the cycle threshold (Ct)] was identified as optimal (maximal yield of SGS-ready cDNA with minimal over-amplification bias) for the 1:10 dilution tested; this number minus three cycles (to compensate for the dilution) was designated the optimal amplification cycle number when returning to the first strand cDNA synthesis reaction products.

3.4 Preparation of Non-Suppressed cDNA Libraries for SGS

For all 22 of the blood specimens collected, we prepared non-suppressed cDNA libraries for SGS. In these cases, to generate the second strand of cDNA and add Illumina-compatible adapters to the ends of the cDNA, 10 µl of the first strand cDNA synthesis reaction products (see Section 3.3) were mixed with 1 µl of 10 mM primer PP_A, 1 µl of 10 mM ScriptSeq Index PCR Primer (PP_I), 12.5 µl of nuclease-free water, 25 µl of Premix E from the FailSafe PCR system, and 0.5 µl of FailSafe Enzyme mix (all products from Epicentre), and subjected to the following PCR conditions: 94°C for 1 min, followed by 10-14 cycles (determined by qPCR result; see Section 3.3) of 94°C for 30 sec, 55°C for 30 sec, and 68°C for 3 min, and a final extension at 68°C for 7 min. The reaction products (ds cDNA libraries) were purified using 0.8X volumes of Agencourt AMPure XP beads (Beckman Coulter Genomics), which enriched for products of 200-500 bp as previously described;[12] each size-selected ds cDNA library was eluted in 20 µl of nuclease-free water, following the manufacturer's instructions.

3.5 Preparation of Normalized cDNA Libraries for SGS

To generate the second strand of cDNA for normalization treatment, 10 µl of the first strand cDNA synthesis reaction products (see Section 3.3) were combined with 1 µl of 10 mM primer PP_P1, 1 µl of 10 mM primer PP_P2, 12.5 µl of nuclease-free water, 25 µl of Premix E from the FailSafe PCR system, and 0.5 µl of FailSafe Enzyme Mix (all products from Epicentre), and subjected the mixture to the following PCR conditions: 94°C for 1 min, followed by 10-14 cycles (determined by qPCR result; see Section 3.3) of 94°C for 30 sec, 55°C for 30 sec, and 68°C for 3 min. After a final extension at 68°C for 7 min, the reaction products (ds cDNA libraries) were purified using the Zymo DNA Clean and Concentrator-5 kit (Zymo Research) and eluted in 10 µl of nuclease-free water. The concentration of each cDNA library was measured using a NanoDrop 2000 (Thermo-Fisher).

Then 5.5 µl (100-200 ng) of the ds cDNA library was added to 2.5 µl of 4X Hybridization Buffer (Nimblegen; Roche) (final = 1X) + 2 µl of 100% formamide (final = 20% vol/vol), and the 10-µl reaction incubated at 98°C for 3 min to denature the ds cDNA. The temperature was then reduced to 68°C to allow reannealing of complementary strands, and the incubation continued for 5 hrs.

HAC-mediated normalization of reannealed cDNA libraries was accomplished through use of spin columns. 1 g of HAC gel (BioGel HTP DNA grade medium; Bio-Rad) was hydrated with Buffer A (10 mM sodium phosphate pH 7 + 20% formamide), then loaded as a slurry into the spin column cartridge (Pierce, catalog #89879) held within a 2-ml microcentrifuge tube, maintaining all at 50°C. After allowing the slurry to settle for several minutes, the spin column was centrifuged at 1100 *x g* for 10 sec, then washed twice with 180 µl of Buffer A at 50°C, centrifuging at 1100 *x g* for 30 sec and discarding the flow-through. At this point the reannealed cDNA library (10 µl) was loaded onto the column and allowed to incubate at 50°C for 5 min. After centrifuging at 1100 *x g* for 10 sec and discarding the flow-through, the column was washed twice with 180 µl of Buffer A at 50°C, centifuging for 30 sec each time. Then 30 µl of Buffer B at 50°C was added to the column, and after incubation at 50°C for 5 min, the ss cDNA eluate was recovered by centrifugation for 30 sec. The second strand of cDNA, and addition of Illumina-compatible adapters, was carried out as described in Section 3.4.

3.6 Preparation of Target-Enriched/Depleted cDNA Libraries for *via* Molecular Capture

Generation of the second strand of cDNA to produce target libraries for capture was achieved using the protocol described in the first paragraph of Section 3.5.

Capture probes were generated from NHP WBC RNA as follows.

200 ng of total WBC RNA were subjected to random fragmentation in 20 µl reactions, through addition of 2 µl of 10X NEBNext RNA fragmentation buffer (New England Biolabs) and incubation at 94°C for 3 min, followed by immediate cooling on ice and addition of 2 µl of NEBNext RNA fragmentation stop solution (New England Biolabs). The fragmented RNA was re-purified using the Zymo RNA Clean and Concentrator-5 system (Zymo Research), following

the manufacturer's general procedure (for recovery of fragments ≥17 nt) and eluting in 10 µl of nuclease-free water.

3.5 µl (25-50 ng) of fragmented WBC RNA were mixed with 1 µl of 25 mM primer Probe cDNA II, incubated at 65°C for 2 min, and then held at 4°C. While at 4°C, 4.5 µl of a master mix containing 2 µl of SMARTScribe 5X First-Strand Buffer, 0.25 µl of 100 mM DTT, 1 µl of 100 mM dNTP mix, 0.25 µl of RiboGuard RNase inhibitor, and 1 µl of SMARTScribe Reverse Transcriptase (all products from Takara) were added, and the mixture incubated at 25°C for 3 min followed by 42°C for 1 min. At this point, 1 µl of 12 mM primer Probe cDNA I were added while the reaction mixture remained in the thermocycler, and incubation continued at 42°C for 1 hr. The reaction was then terminated through incubation at 70°C for 10 min. After adding 10 µl of nuclease-free water, the reaction products (first strand of cDNA) were purified using 36 µl (1.8X volumes) of Agencourt AMPure XP beads (Beckman Coulter Genomics) and eluted in 45 µl of nuclease-free water, following the manufacturer's instructions.

First stand cDNA products were quantified using our qPCR assay, as described in the last paragraph of Section 3.3.

The second strand of cDNA was generated using the procedure described in the first paragraph of Section 3.5, substituting primers Probe PCR I and Probe PCR II. In general, we set up 4 reactions *per* probe, using all of the ss cDNA library generated in the previous step, and the ds cDNA products were combined and loaded onto the same column for clean-up using the Zymo DNA Clean and Concentrator-5 kit (Zymo Research), eluting in 20 µl of nuclease-free water. The concentration of each ds cDNA library was measured using a NanoDrop 2000 (Thermo-Fisher).

At this point, the ds cDNA library was labeled with biotin, using the Bio Prime DNA labeling system (Invitrogen #18094-011). 5 µl (100 ng) of ds cDNA were combined with 20 µl of 2.5X random primers on ice, then denatured by incubation at 98°C for 5 min and immediately cooling on ice again. Then 5 µl of 10X dNTP mix and 19 µl of nuclease-free water were added, followed by 1 µl of Klenow fragment, and the reaction transferred to 37°C for incubation for 1 hr. The biotinylated probe was then re-purified using the PCR Purification kit (QIAGEN).

Capture probes were generated from *Y. pestis* CO92 genomic DNA using the Bio Prime DNA labeling system, following the procedure described in the previous paragraph.

20 ng of target ds cDNA library were mixed with 2000 ng (100X) of biotinylated capture probe, and lyophilized at 60°C for 1 min per µl of mixture. The dried mixture was then resuspended in 10 µl of Roche Hybridization Buffer [5 µl of 2X Hybridization Buffer + 3 µl of nuclease-free water + 2 µl of Component A (formamide)] and incubated at RT for 10 min. The mixture was then denatured at 95°C for 5 min and re-annealed at 60°C for ~16 hrs.

During the re-annealing period, the capture columns were prepared.

For negative capture, 100 µl of NeutrAvidin agarose resin (Pierce) were added to the spin column cartridge (Pierce, catalog #89879) held within a 2-ml microcentrifuge tube at 50°C. The column was centrifuged at 3500 rpm at 50°C for 30 sec, then washed twice with 100 µl of 100 mM PBS at 50°C, discarding the flow-through after each spin. The column was then blocked by adding 100 µl of pre-warmed (50°C) 100 ng/µl COT DNA + 1 µg/µl BSA in nuclease-free water

and incubating at 50°C for 2 min. The washes were repeated, and to the prepared column at 50°C we added the probe-annealed library. After incubating at 50°C for 5 min, the column was centrifuged at 3500 rpm for 1 min at 50°C, and the flow-through ("capture-depleted" library) collected.

For positive capture, 100 µl of monomeric avidin agarose resin (Pierce) were added to the spin column cartridge held within a 2-ml microcentrifuge tube at 50°C. The column was centrifuged at 3500 rpm at 50°C for 30 sec, then washed twice with 100 µl of 100 mM PBS at 50°C, discarding the flow-through after each spin. To the prepared column at 50°C we added the probe-annealed library. After incubating at 50°C for 5 min, the column was centrifuged at 3500 rpm for 1 min at 50°C, and the flow-through ("capture-depleted" library) collected. The column was then washed twice with 100 µl of 100 mM PBS at 50°C, the bound cDNA incubated with 50 µl of pre-warmed (50°C) nuclease-free water for 5 min, and the eluate recovered by centrifugation at 3500 rpm at 50°C for 1 min.

4 SEQUENCING STRATEGY, STATISTICS, AND DATA PROCESSING

4.1 Background

The transcriptomes of mammalian cells can consist of $>10^5$ different RNA species, in relative abundances that can differ by $>10^5$-fold.[13] To fully characterize the transcriptome of a population of cells, the depth of sequencing (*i.e.*, the number of SGS reads *per* library) must be sufficient for robust quantitation of biologically relevant transcripts that are present at low abundance in the sample. On the other hand, SGS is an expensive endeavor, and processing and analysis of SGS datasets is labor-intensive and time-consuming; sequencing in exhaustive depth is rarely necessary, desirable, or feasible. In most cases (as that at hand), one must balance sequencing depth (number of reads *per* library) *vs.* breadth (number of libraries sequenced).

For transcriptional profiling of mammalian cells and tissues at substantial sequencing depth, a general rule of thumb is that ≥25M successfully mapped reads *per* library is a reasonable starting point. We set this as our target sequencing depth for non-suppressed WBC cDNA libraries, anticipating that it would support robust analysis of global trends in expression.

Our previous work with molecular suppression of mammalian cell/tissue cDNA libraries[2-4] indicated that, largely through depletion of rRNA, which typically constitutes ≥80% of RNA species in the transcriptome,[14] ~10X reduced sequencing depth (*i.e.*, 2.5M mapped reads *per* library) is sufficient for comparably robust analysis of global trends in expression; therefore, we sought to meet or exceed this target sequencing depth for suppressed WBC cDNA libraries.

Our previous work with human blood plasma cDNA libraries indicated that they are considerably less complex than libraries derived from mammalian cells/tissues, such that a sequencing depth of ≥1M mapped reads *per* library is typically sufficient for robust analysis of plasma transcriptomes. Accordingly, we set this as our target sequencing depth for non-suppressed plasma cDNA libraries.

4.2 SGS Data Generation and Quality Control Filtering

We generated SGS data from our libraries using our in-house Illumina MiSeq, or the Illumina HiSeq at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley (http://qb3.berkeley.edu/gsl/Home.html). Libraries were multiplexed (*i.e.*, multiple libraries, each bearing a unique barcode, were mixed together for loading into a single lane of the sequencer) and loaded into the MiSeq at 8 pM or the HiSeq at 10-12 pM.

Raw sequence files were demultiplexed using the CASAVA v1.8 pipeline (Illumina). The fastq sequence files were further processed with our custom qfilter.pl perl script, which trims low-quality bases, detects and trims internal barcodes and primer fragments, masks low-complexity sequence, and removes any sequence with an overall quality or length below acceptable thresholds. First, internal barcodes and 3' and 5' tails with minimal quality scores were trimmed off. At this and subsequent trimming steps, a length test was applied; sequences below a

minimum length (default 30 bp) were rejected. Each remaining unique sequence was passed through three filters that do not query quality. In the first filter, sequences of primers used in library construction were identified and trimmed in the following way. Primer "parts" of length 14 nt were collected, first taking the 14-mer DNA oligo sequence from both the 3' and 5' end of each primer, unless this sequence was homopolymeric, in which case it was substituted with the 14-mer taken from an internal position such that only 6 nt of the homopolymer were included. The reverse complement of each 14-mer was then added to the primer part list. In the second filter, sequences with any remaining positions uncalled (*i.e.*, called as "N") were rejected. Then, in the final filter, dustmasker (from the NCBI C++ Toolkit) was used to identify low-complexity sequences; sequences were rejected when masking left less than 30 bp, otherwise, low-complexity sequences were allowed to remain. Returning to individual reads and their quality strings, the quality markings were converted to a 2-40 score scale, the average score for all remaining positions was calculated, and the read was rejected if the average was below a default threshold of 30.

Reads passing our in-house quality filter were further analyzed using two different bioinformatics pipelines, as described in the next two sections.

4.3 SGS Read Mapping for Host Transcriptomics: DNAnexus Pipeline

The DNAnexus pipeline is run using commercially-available software that carries out conventional RNA-Seq analysis of pre-identified contributors to metagenomic libraries. For the present study, we used the DNAnexus pipeline to characterize the host-derived RNA species represented in the SGS datasets generated from the NHP blood fractions.

In the DNAnexus pipeline, reads are first quality-filtered by the software, using its default settings; the reads fed into the pipeline have already been quality filtered (see Section 4.2), so this redundant filtering has essentially no impact on our datasets.

The pipeline then maps each read to the *M. mulatta* genome (rheMac2; MGSC Merged 1.0; 2.86 Gb), allowing a pre-determined number of mismatches in sequence alignment. The reference genome (*M. mulatta*) is not from the NHP species used in our study (*M. fascicularis*), though it shows 99.2-99.7% DNA sequence identity to *M. fascicularis*.[15,16] For this reason we used the default mapping setting, allowing the software to go forward with reads "mapped confidently" (according to the software annotation), as opposed to demanding that reads map perfectly (*i.e.*, no mismatches in alignment).

The pipeline then identifies the gene to which each read maps, and counts the number of reads mapping to each gene in the genome. These "hit" counts are then normalized with respect to the length of the gene's sequence, generating a "reads *per* kilobase *per* million mapped reads" (RPKM) score, and these scores are further normalized by their root mean square (RMS).

4.4 SGS Read Mapping for Non-Host Phylogenetics: RapTOR Pipeline

A major thrust of the RapTOR Grand Challenge LDRD project was to develop an in-house bioinformatics pipeline ("Raptor") for phylogenetic analysis of SGS datasets generated from metagenomic libraries. For the present study, we used the Raptor pipeline to characterize the non-host-derived RNA species represented in the SGS datasets generated from the NHP blood fractions.

In the Raptor pipeline, reads that pass filter are first aligned to sequences from a reference genome, or set of reference genomes, representing the host. In this case, we used the *M. mulatta* genome augmented with all available RefSeq gene records for *M. fascicularis*, as well as *M. fascicularis* rRNA sequences assembled from primary data reported in the literature and publically-available databases. Reads that do not map to the host reference genome(s) are then aligned against a series of additional sequence sets representing host-derived repetitive sequences, fungal and bacterial rRNA, fungal and bacterial non-rRNA transcripts, and viral transcripts/genomes. This first stage of read mapping is carried out by a tool called Bowtie 2,[17] using settings for sensitive local alignment. This is followed by a second stage of read mapping against fungal, bacterial, and viral transcripts (non-rRNA) is carried out using Bowtie 2 set for sensitive global alignment, and then a final stage of read mapping against all sequences deposited in RefSeq is carried out using BLASTN. Each stage of read mapping can be tailored for stringency; in our study, only perfect alignments over a pre-defined sequence length were considered hits.

Once the reads are mapped, the pipeline retrieves the genome identity and locus for each read, and this information is used to identify the taxonomic hits for each read. A custom lowest common ancestor (LCA) algorithm condenses all of the (possibly many) taxonomic hits for each read, and reports only the taxonomic rank common to all of them. In this way, Raptor is conservative in making taxonomic assignments; when combined with the demand for perfect alignment, one can have extremely high confidence in the taxonomic hit counts.

5 HOST TRANSCRIPTOMICS RESULTS

5.1 Transcriptional Profiles of Non-Suppressed WBC

Log_{10}-log_{10} scatter plots comparing global transcriptional expression from Day 0 WBC *vs* Day 2, 3, or 4 WBC confirmed that while most transcripts showed no or modest deviation from the slope = 1 line, a small number of transcripts were clearly induced over time (significantly above the line), and only a few were clearly repressed over time (significantly below the line) (Figure 1). These dramatic changes in expression were most notable in comparison of Day 0 *vs* Day 3 expression. The effect was similar across all three NHP.

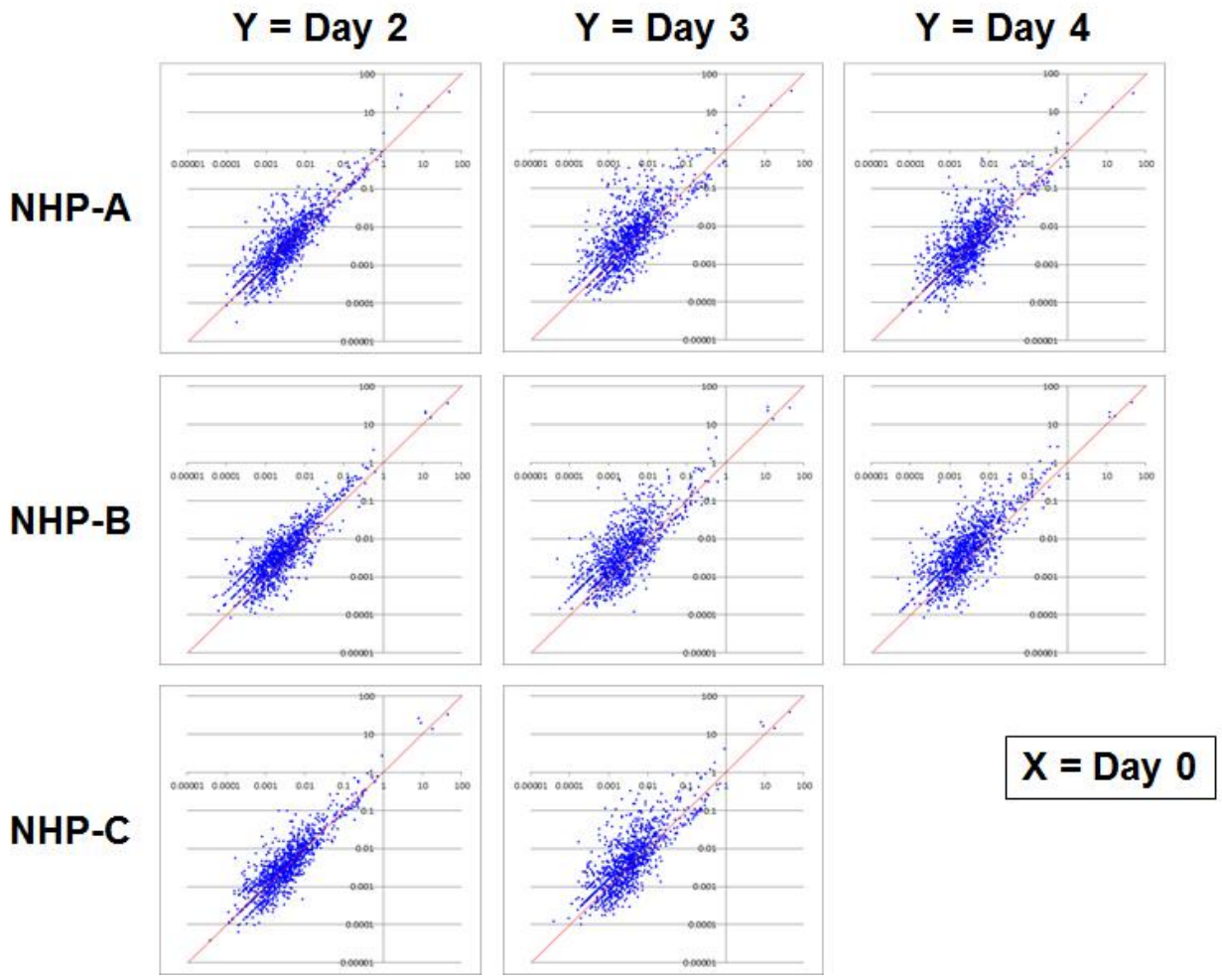

Figure 1. Log_{10}-log_{10} Scatter Plots Comparing Global Transcriptional Expression from Day 0 WBC (X-Axis) *vs* Day 2, 3, or 4 WBC (Y-Axis). Points falling well off of the line indicate genes that are differentially expressed.

As a first-pass analysis of the nature of the differentially expressed transcripts, we identified those that showed a ≥3-fold change in expression that was consistent throughout the timecourse (*i.e.*, Day0/Day2, Day0/Day3, and Day0/Day 4 ratios were all ≤ 0.33 or ≥3.0). Using these criteria, we identified 226 differentially expressed genes. The vast majority of these (200) were induced, rather than repressed. 73 of the genes (72 induced, 1 repressed) were differentially expressed in more than one NHP, and 36 of them (all induced) were differentially expressed in all three NHP.

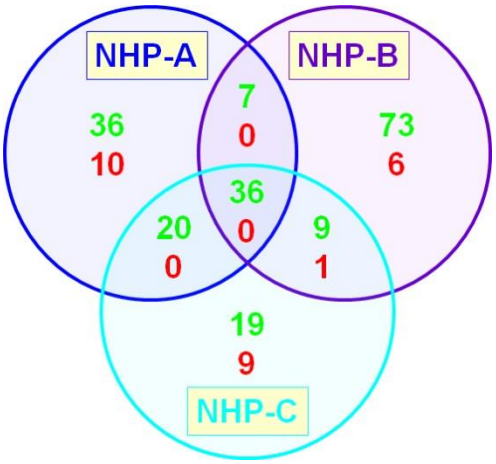

Figure 2. Genes Showing ≥3-Fold Change in Expression Consistently Across the Timecourse. Green numbers indicate induced genes, red numbers indicate repressed genes.

We carried out a more comprehensive analysis of differential expression for comparison. We identified genes showing a Pearson correlation of response between NHP for Days 0-3 of 0.90, 0.95, or 0.99, and showing at least one difference in average response between two days (Day 0 → Day 2, Day 2 → Day 3, or Day 0 → Day 3) that was significant at the 0.05, 0.01, or 0.005 level based on t-test for difference of the means. We did not include Day 4 data in this analysis because our focus was identification of early signs of infection, gene expression almost uniformally decreased in the Day 3 → Day 4 transition (presumably an end-game expression pattern), and we were missing Day 4 data for NHP-C. Using this approach we identified 219 genes that showed a Pearson correlation of response across NHP of ≥0.90 and a change in expression that was significant at ≤0.05 level. As expected, this set of genes shrank with application of more stringent cut-off criteria, such that only 37 genes showed a Pearson correlation of response across NHP of ≥0.995 and a change in expression that was significant at ≤0.005 level.

		Significance of Expression Change (p-value < X)		
		0.05	0.01	0.005
Pearson Correlation of Responses Across Three NHPs	0.90	219	162	126
	0.95	149	112	87
	0.995	59	44	37

Table 4. Genes Identified as Differentially Expressed in Non-Suppressed WBC cDNA Libraries, Enumerated as a Function of Different Selection Criteria. Shown are genes with similar expression across the three NHP (Pearson correlation of ≥ 0.90 to 0.995) and at least one change in expression over the timecourse (D0 → D2, D2 →D3, and/or D0 → D3) that was statistically significant (p-value ≤ 0.05 to 0.005).

To facilitate interpretation of the expression changes and the effects of imposing different cut-off criteria, we plotted the Day 0 → Day 2 (X-axis) and Day 2 → Day 3 (Y-axis) changes in expression (log_2) for genes meeting each combination of cut-off criteria (Figures 3 & 4). We found that our insistence on the high correlation of response across the three NHP biased against the second quadrant (*i.e.*, Day 0 → Day 2 repression, followed by Day 2 → Day 3 induction); in fact, it was entirely empty except on the axes. This is because when the Day 0 → Day 2 repression and Day 2 → Day 3 induction are mean subtracted, as Pearson does, they turn into small deviations on either side of the axis relative to the standard deviations of each data point, and the correlation coefficient is small and noisy. There is also a lesser bias against these due to the significance test (Day 0 and Day 3 values tend to overlap). This effect also works to impoverish the fourth quadrant (*i.e.*, Day 0 → Day 2 induction, followed by Day 2 → Day 3 repression). In any case, these plots indicate that most of the genes identified through this approach showed induced expression throughout the timecourse (*i.e.*, fell into the first quadrant).

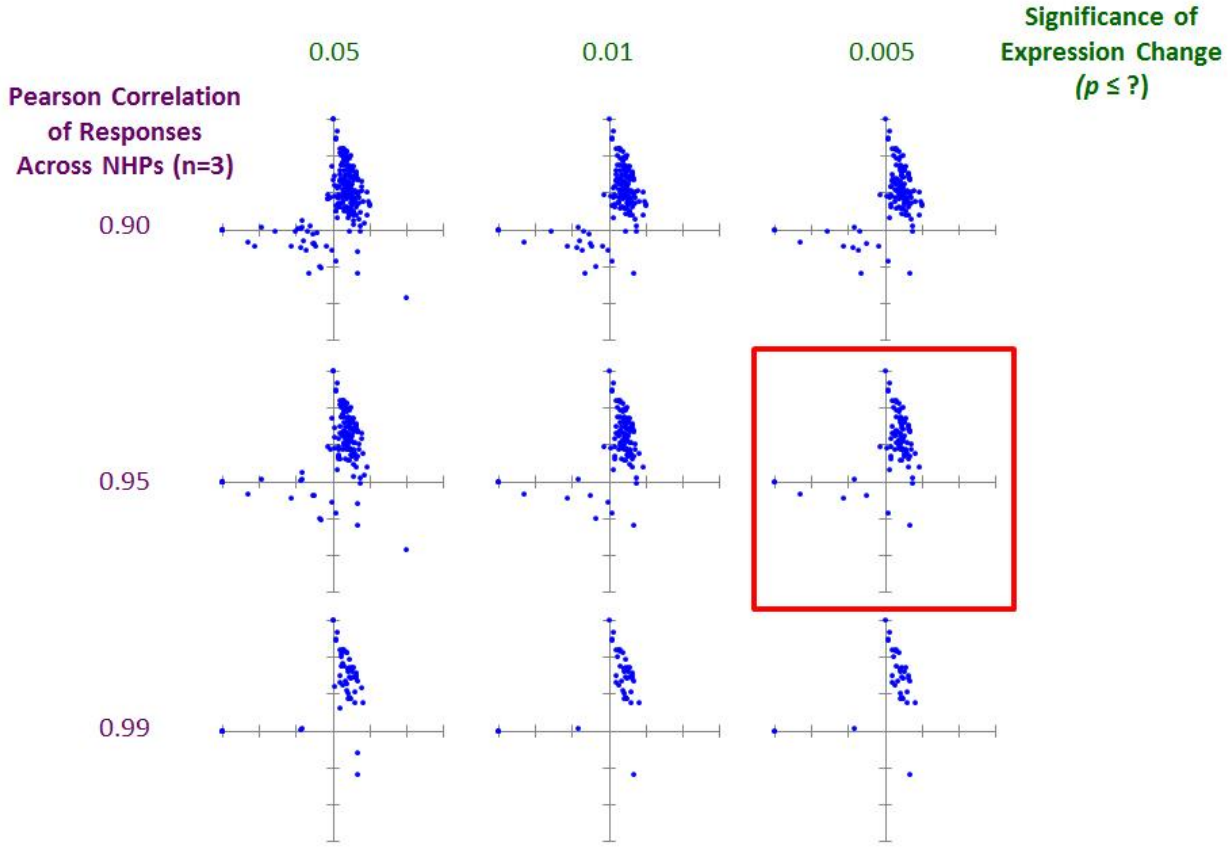

Figure 3. Scatter Plots of Genes Showing WBC Expression Patterns That Are Similar Across Three NHP (Pearson Correlation ≥0.90) and Including a Statistically Significant Change in Expression (P-Value ≤0.05) Over the Timecourse (D0→D3). Each datapoint represents an individual transcript and its expression level relative to that at D0. X-axis = D0→D2; y-axis = D2→D3. Both axes are log_2. Red box indicates the data further analyzed in Figure 4 (below).

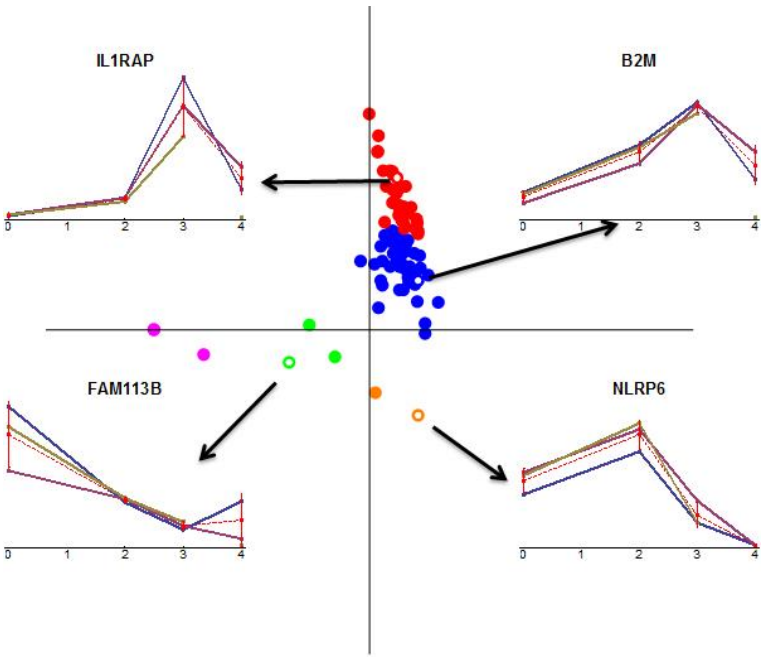

Figure 4. Hierarchical Clustering of Genes Showing WBC Expression Patterns That Are Similar Across Three NHP (Pearson Correlation ≥0.95) and Including a Statistically Significant Change in Expression (P-Value ≤0.005) Over the Timecourse (D0→D3). As in Figure 3, each datapoint represents an individual transcript and its expression level relative to that at D0. X-axis = D0→D2; y-axis = D2→D3. Both axes are $\log_2$. Five discrete clusters were identified. Traces of expression over the timecourse are shown for four representative genes.

We carried out a preliminary analysis of the biological functions represented by the genes that showed differential expression. As expected, Toll-like receptor signaling pathway components were well represented, though it was surprising that so many genes encoding the receptors themselves were differentially expressed. Genes involved in cytokine/chemokine signaling, inflammation, and anti-oxidation/stress response were also well represented, as expected. Genes involved in cell morphology/exocytosis and cell adhesion likely reflect the transition of circulating monocytes and macrophages to their activated states. Differential expression of iron metabolism genes likely reflects the host *vs* pathogen battle over iron stores in the body. A large set of genes of unknown function were differentially expressed; understanding their role in responding to and/or resolving infection will be of great interest in future studies.

Toll-like Receptor	Cytokine/Chemokine	Inflammation	Cell Morphology/ Exocytosis	Cell Adhesion	Iron Metabolism	Anti-Oxidation/ Stress Response	Unknown
SOCS3	CAMP	AIF1	CDC42EP3	EMR2	CYBRD1	TXN	ARL4A
TLR1	CCR1	FCAR	RAB3D	SELL	NFE2	GLRX	GPR160
TLR2	IL1B	LST1	RHOG	SIGLEC5		IDO1	KCNJ2
TLR4	IL1R2	NLRP6	FLOT1			LOC574097	LILRAD
TLR5	IL1RAP		MYL6			SEPX1	TYROBP
TLR6	IRF7		TMEM127				UBTD1
TLR8	TRIB1						C6H5orf32
MYD88	JUNB						FAM49A
	OSM						LBH
							LILRAE
							LOC704474
							NECAP1
							SRSF12
							STAC3

Table 5. Functional Categorization of Genes Identified as Differentially Expressed in Non-Suppressed WBC cDNA Libraries. The 36 genes identified as differentially expressed in all three NHP using simple criteria (Figure 2), and the 37 genes identified using the most rigorous criteria (Table 4), were assigned functional categories on the basis of reports in the literature. A total of 51 genes were analyzed; 26 were identified only using the simple criteria (blue), 15 were identified using only the rigorous criteria (purple), and 10 were identified using either approach (black).

5.2 Profiles of WBC cDNA Libraries After Molecular Suppression

We applied several different types of molecular suppression to cDNA libraries generated from WBC transcripts.

HAC-mediated normalization was used to reduce representation of high-abundance transcripts in the final cDNA libraries sequenced. We applied HAC-mediated normalization to cDNA libraries generated from each of the WBC fractions.

In some cases, HAC-mediated normalization was followed by a second, capture-mediated suppression step. Capture probes generated from Day 0 WBC ["Cap(D0)"] were used to deplete complementary sequences in Day 3/4 WBC cDNA libraries; this enriched for cDNA unique to (or much more abundant in) the Day 3/4 cDNA libraries (*i.e.*, transcripts induced upon infection). Capture probes generated from Day 3/4 WBC ["Cap(D3/4)"] were used to deplete complementary sequences in Day 0 WBC cDNA libraries; this enriched for cDNA unique to (or much more abundant in) the Day 0 cDNA libraries (*i.e.*, transcripts repressed upon infection). Finally, capture probes generated from *Y. pestis* gDNA ["Cap(Yp)"] were used to capture and concentrate cDNA derived from *Y. pestis* transcripts. All of these combinations of suppression (HAC-mediated normalization followed by one of the three versions of capture-mediated suppression) were applied to the cDNA libraries generated from NHP-A WBC fractions. HAC+Cap(D0) was also applied to cDNA libraries generated from NHP-C WBC fractions.

We found that HAC-mediated normalization greatly improved the quality of cDNA libraries, leading to much lower numbers of rejected reads. Addition of a capture step using Cap(D0) or Cap(D3/4) probes further improved cDNA library quality.

Reads mapping to "microbial rRNA" were also reduced upon suppression; however, further inspection of those mapping assignments revealed that the vast majority were actually host (NHP) rRNA sequences. The problem stemmed from the fact that a few NHP rRNA sequences were not represented in the reference sequence set that we used as a filter; thus, some NHP rRNA sequences passed the "host rRNA" filter and were designated "other" (or "microbial") rRNA.

Suppression also led to higher levels of reads that passed the quality filter but failed to align with any sequence in our reference database ("Unhit"). It is not yet clear whether these reads derive from previously uncharacterized microbes - perhaps minority species which are not efficiently sequenced unless the majority species are depleted *via* molecular suppression.

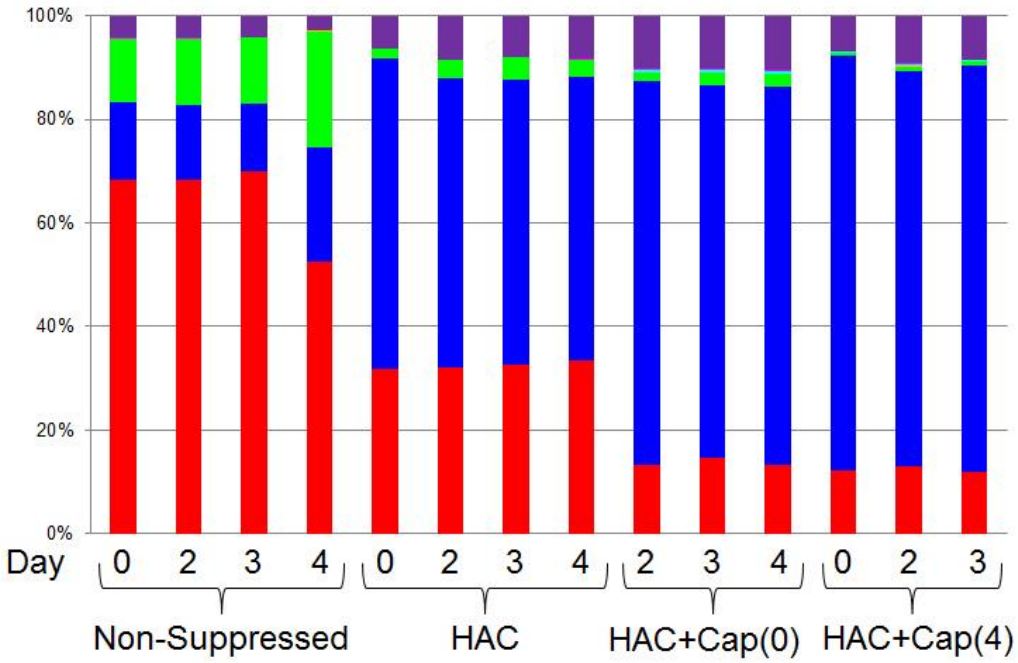

Figure 5. Impact of Suppression on WBC cDNA Read Mapping Results. Reads from each NHP-A WBC cDNA library were categorized as rejected (red), passing filter but non-mapping (purple), or mapping to host (blue), microbial rRNA (green), viral (orange), bacterial (aqua), or fungal (pink) reference sequences.

Scatterplots comparing non-suppressed *vs* HAC-normalized WBC cDNA libraries indicated that virtually all but the most highly abundant transcripts were enriched after suppression. The effect was surprisingly uniform, such that the relative abundances of transcripts were well maintained. These results suggest that HAC-mediated normalization was strongly selective for only the most abundant transcripts - particularly those expressed from rRNA genes - and did not impose strong biases on other, less abundant transcripts.

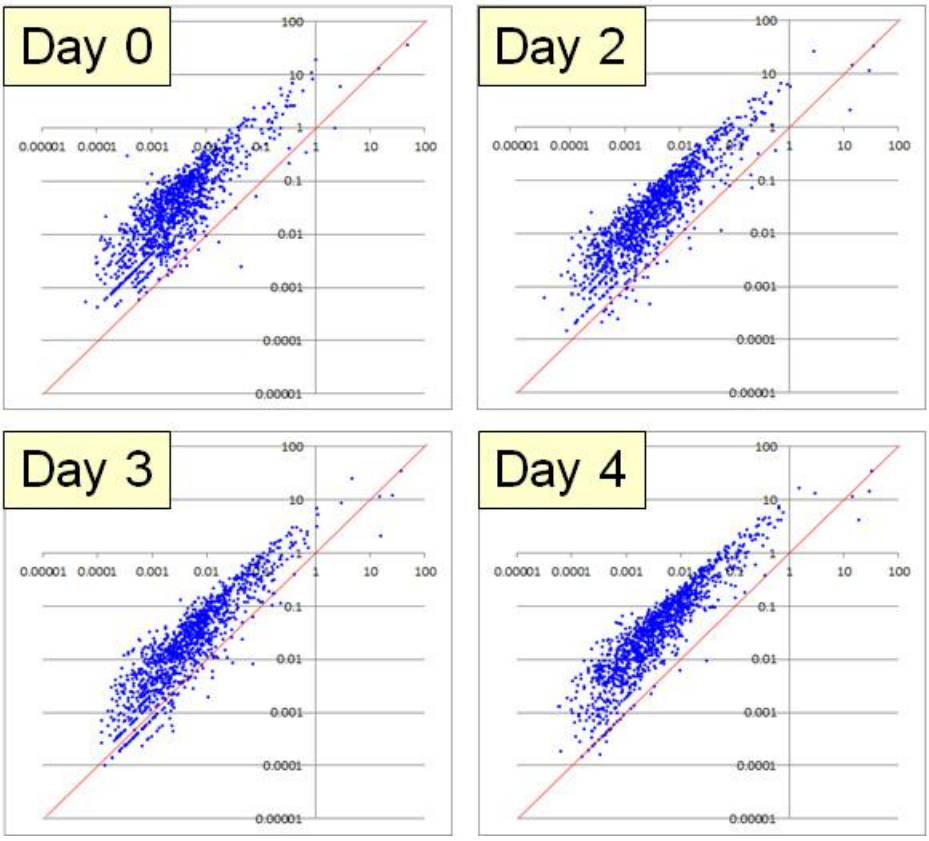

Figure 6. Log_{10}-Log_{10} Scatter Plots Comparing Transcript Levels in Non-Suppressed (X-Axis) *vs* HAC-Normalized (Y-Axis) WBC cDNA Libraries from NHP-A. Global shift of datapoints to positions above the red line indicates that HAC-mediated normalization enriched for most transcripts.

The selective effect of HAC-mediated normalization on only the most abundant transcripts was further evident in a plot of the SGS bandwidth allocated to different abundance tiers. In non-suppressed cDNA libraries, ~80% of the reads mapped to the most highly expressed transcripts (top 1%). In contrast, in HAC-normalized cDNA libraries, < 20% of the reads mapped to these transcripts, and the remainder were broadly distributed across the other abundance tiers.

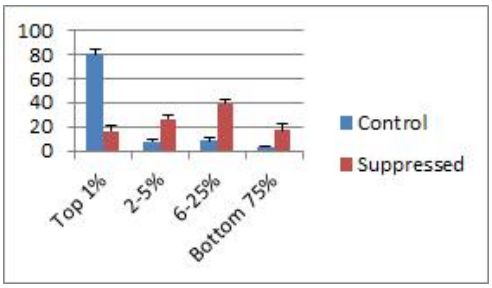

Figure 7. Impact of HAC-mediated Normalization on Representation of High-Abundance Transcripts in WBC cDNA Libraries from NHP-A. Percentage of SGS reads (y-axis) mapping to transcripts as a function of their abundance, in non-suppressed ("control"; blue) *vs* HAC-normalized ("suppressed"; red) cDNA libraries.

Applying our simple criteria for differential expression (≥3-fold change that was consistent throughout the timecourse) to the non-suppressed *vs* HAC-normalized NHP-A WBC cDNA libraries, we identified 134 differentially expressed genes total. 56 of these were identified as differentially expressed in both the non-suppressed and the HAC-normalized cDNA libraries; the vast majority of these (52 genes) showed induced expression. 47 genes were identified as differentially expressed in only the non-suppressed cDNA libraries; again, the vast majority of these (41 genes) showed induced expression. 31 genes were identified as differentially expressed in only the HAC-normalized cDNA libraries; in sharp contrast to the other cases, the vast majority of these (29 genes) showed repressed expression. Line plots of the expression levels of the 134 genes revealed that those identified as differentially expressed in the HAC-normalized libraries tended to be poorly expressed in general; it seems that in enriching for low-abundance transcripts, HAC-normalization enabled consistent measurement of the transcripts even as their levels decreased due to repression. Thus, HAC-normalization improved sensitivity in detecting infection-associated gene repression.

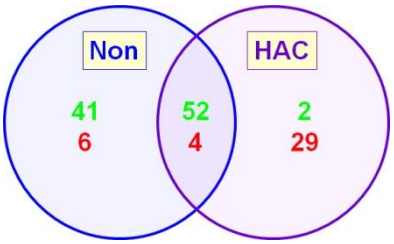

Figure 8. Genes Showing ≥3-fold Change in Expression Consistently Across the Timecourse in Non-Suppressed *vs* HAC-Normalized WBC cDNA Libraries from NHP-A. Green numbers indicate induced genes, red numbers indicate repressed genes.

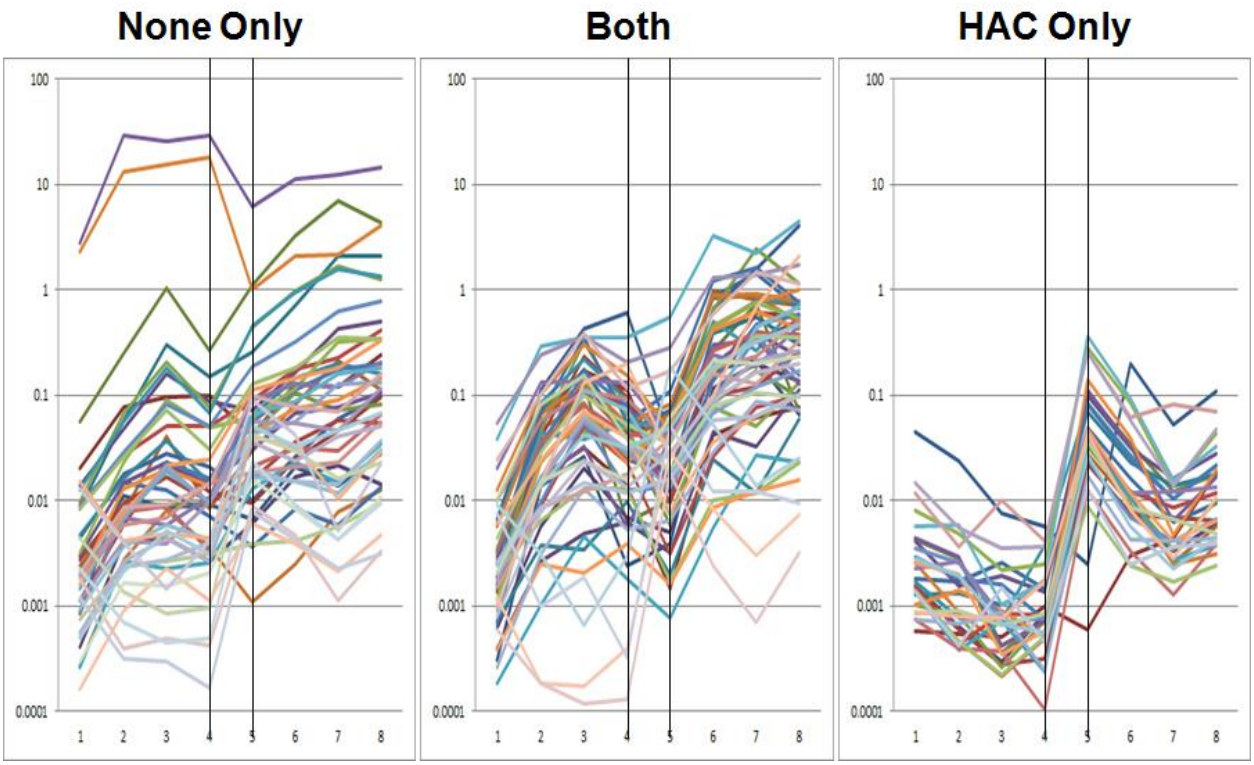

Figure 9. Line Plots of Expression Levels Over Timecourse for Genes Identified as Differentially Expressed in Non-Suppressed and/or HAC-Normalized WBC cDNA Libraries from NHP-A. Traces indicate expression levels (y-axis, $\log_{10}$ scale) of genes showing ≥3-fold change in expression consistently across the timecourse in non-suppressed (left) *vs* HAC-normalized (right) WBC cDNA libraries; transitions shown are D0→D2 (1→2 & 5→6), D2→D3 (2→3 & 6→7), and D3→D4 (3→4 & 7→8). Note that the traces connecting 4→5 are artifacts of the line plot; they simply connect the non-suppressed traces (left) to the HAC-normalized traces (right).

5.3 Summary

The analyses carried out to date indicate that robust immune responses at the transcriptional level were elicited in the WBC of our *Y. pestis* infected NHP. The differentially expressed genes included many previously identified in other studies of host response to infection, though there were surprising numbers of Toll-like receptor genes and genes of unknown function induced upon infection. Molecular suppression reduced the contribution of rRNA and other high-abundance transcripts, enabling efficient sequencing of other transcriptome constituents. This approach was particularly valuable in improving the sensitivity with which genes repressed upon infection could be identified as differentially expressed.

6 NON-HOST PHYLOGENETICS RESULTS

6.1 Bacterial Transcripts Detected in WBC Fractions

We detected appreciable levels of transcripts from a wide variety of bacterial species in WBC fractions. *Y. pestis* hits were not detected in Day 0 (as expected) or in Day 2 samples. In Day 3 samples they were not detected unless suppression was used; even so, the levels remained very low [3 hits in non-suppressed, 31 hits in HAC, 59 hits in HAC+Cap(0)], accounting for no more than 0.0012% of mapped reads in the library. Many of the other bacterial species detected are frequently associated with human skin (*e.g.*, *Propionibacterium acnes*, *Staphylococcus epidermidis*, *S. aureus*) or the gastrointestinal tract *(e.g.*, *Escherichia coli*, *Salmonella enterica*, *Enterococcus faecalis*). It is not clear whether the bacterial transcripts detected originated with the blood specimens themselves, or were introduced during collection of the specimens or preparation of the libraries. Suppression enriched the libraries for bacterial content by ~2-fold, with the exception of the Day 0 sample, which did not show enrichment of bacterial content following HAC-mediated normalization. In general, suppression did not dramatically change the relative abundances of bacterial constituents, with the notable exception of *Taylorella equigenitalis*, which was selectively diminished with suppression.

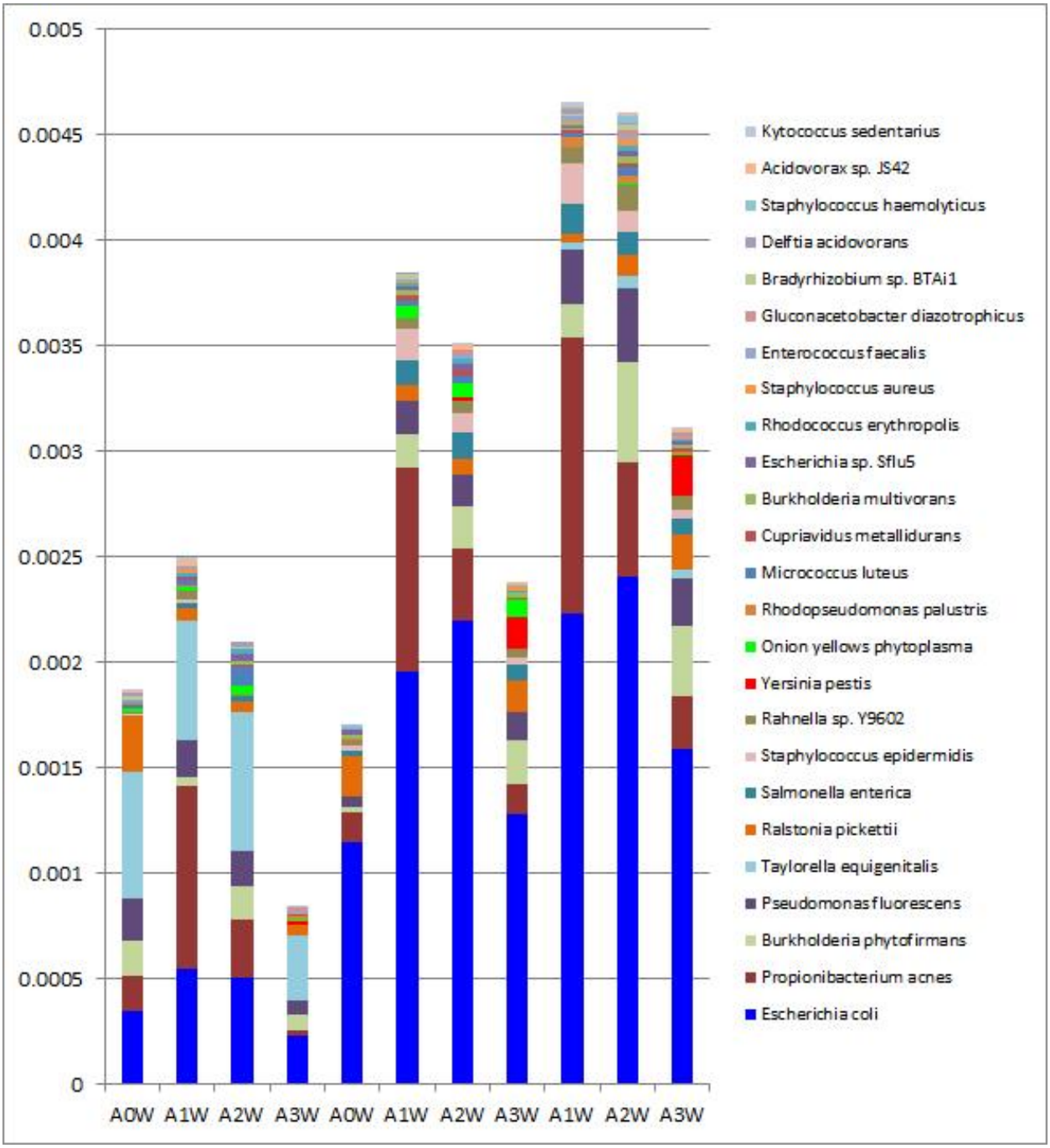

Figure 10. Proportions of Reads Mapping to the Top 25 Most Prevalent Bacterial Species Represented in NHP-A WBC cDNA Libraries. Results from non-suppressed (first four bars), HAC-normalized (middle four bars), and HAC+Cap(D0) (last three bars) are shown. Y-axis = Percentage of mapped reads.

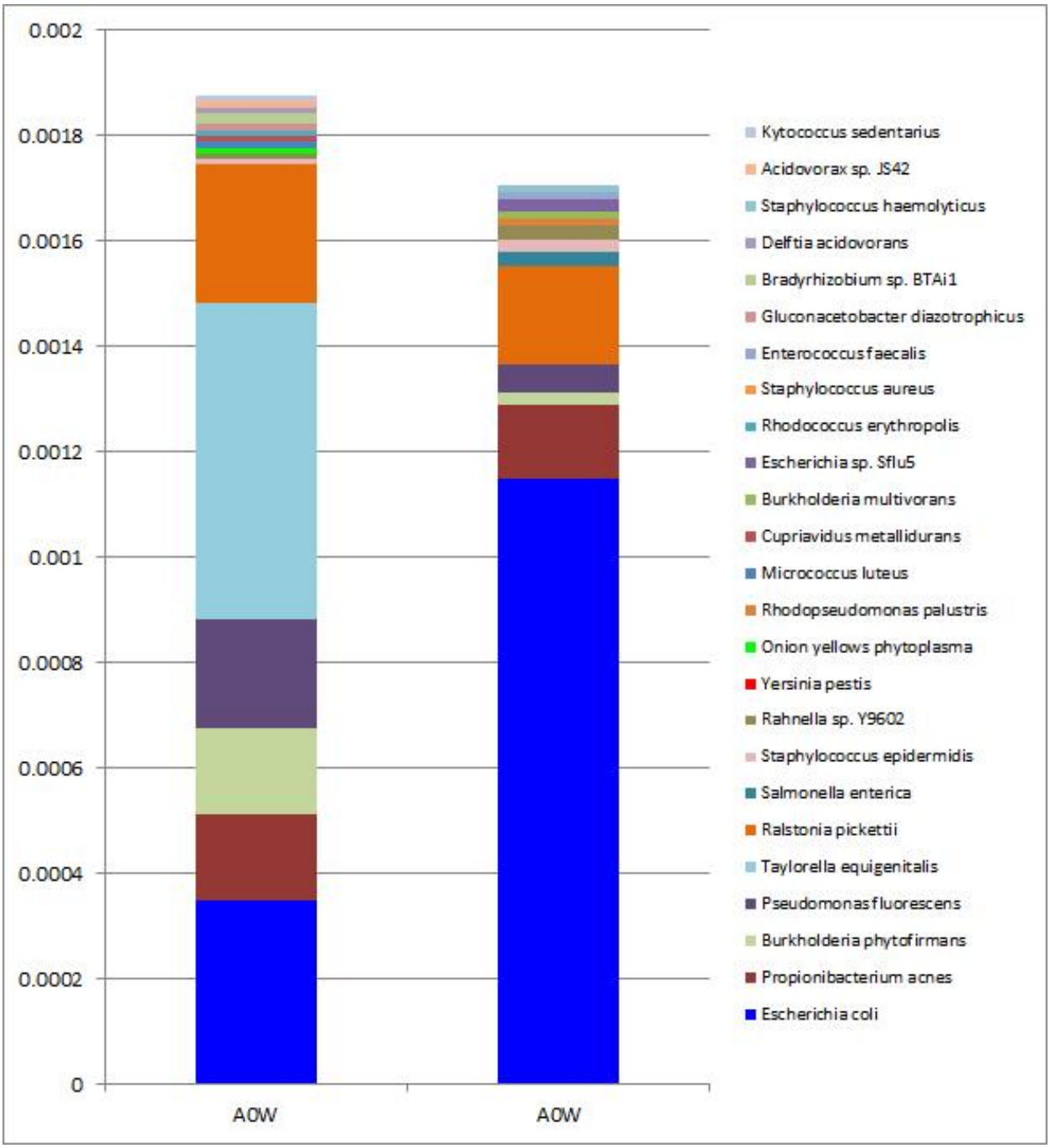

Figure 11. Proportions of Reads Mapping to the Top 25 Most Prevalent Bacterial Species Represented in NHP-A Day 0 WBC cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Right). Y-axis = Percentage of mapped reads.

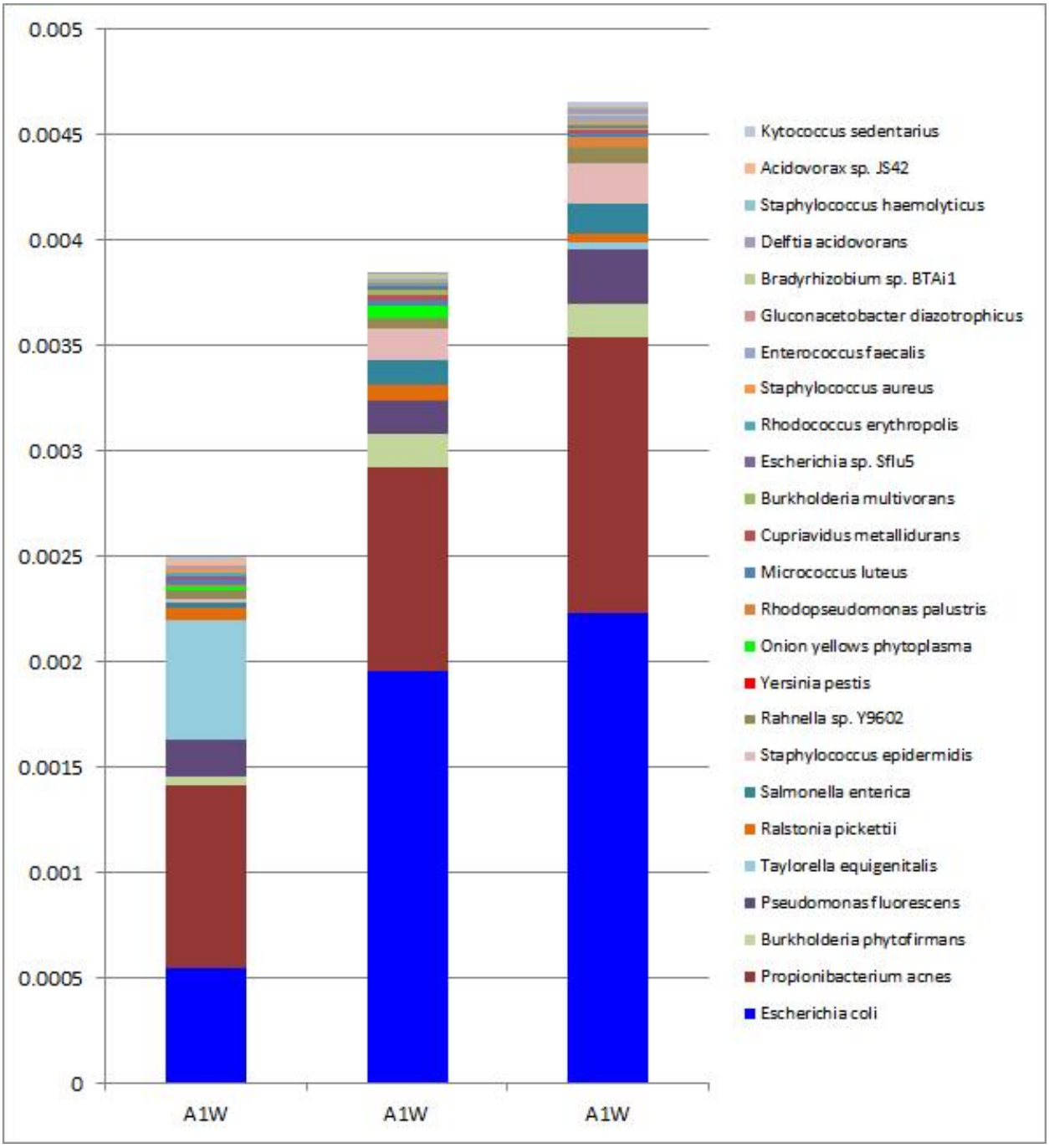

Figure 12. Proportions of Reads Mapping to the Top 25 Most Prevalent Bacterial Species Represented in NHP-A Day 2 WBC cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Middle) *vs* HAC+Cap(D0) (Right). Y-axis = Percentage of mapped reads.

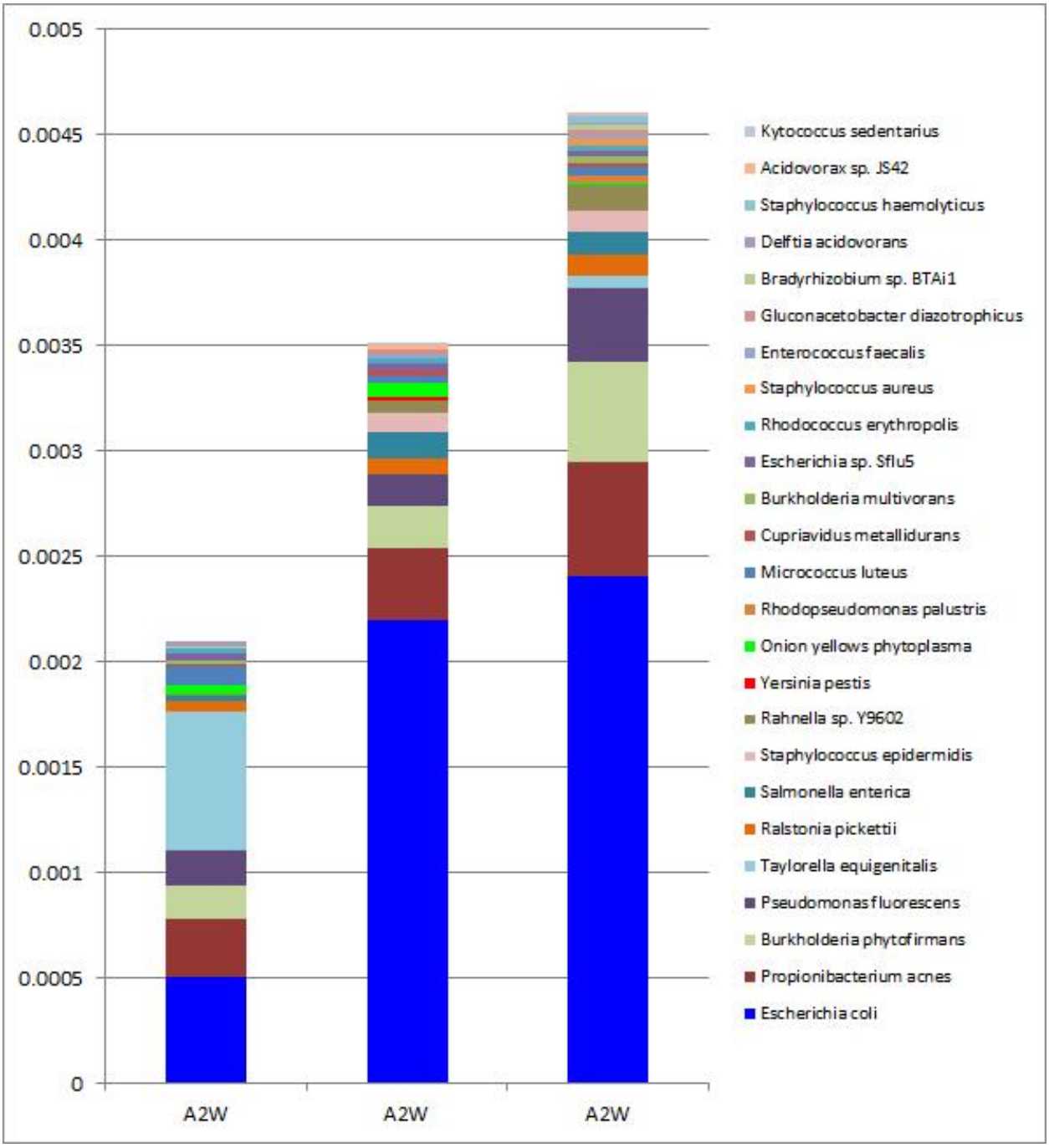

Figure 13. Proportions of Reads Mapping to the Top 25 Most Prevalent Bacterial Species Represented in NHP-A Day 3 WBC cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Middle) *vs* HAC+Cap(D0) (Right). Y-axis = Percentage of mapped reads.

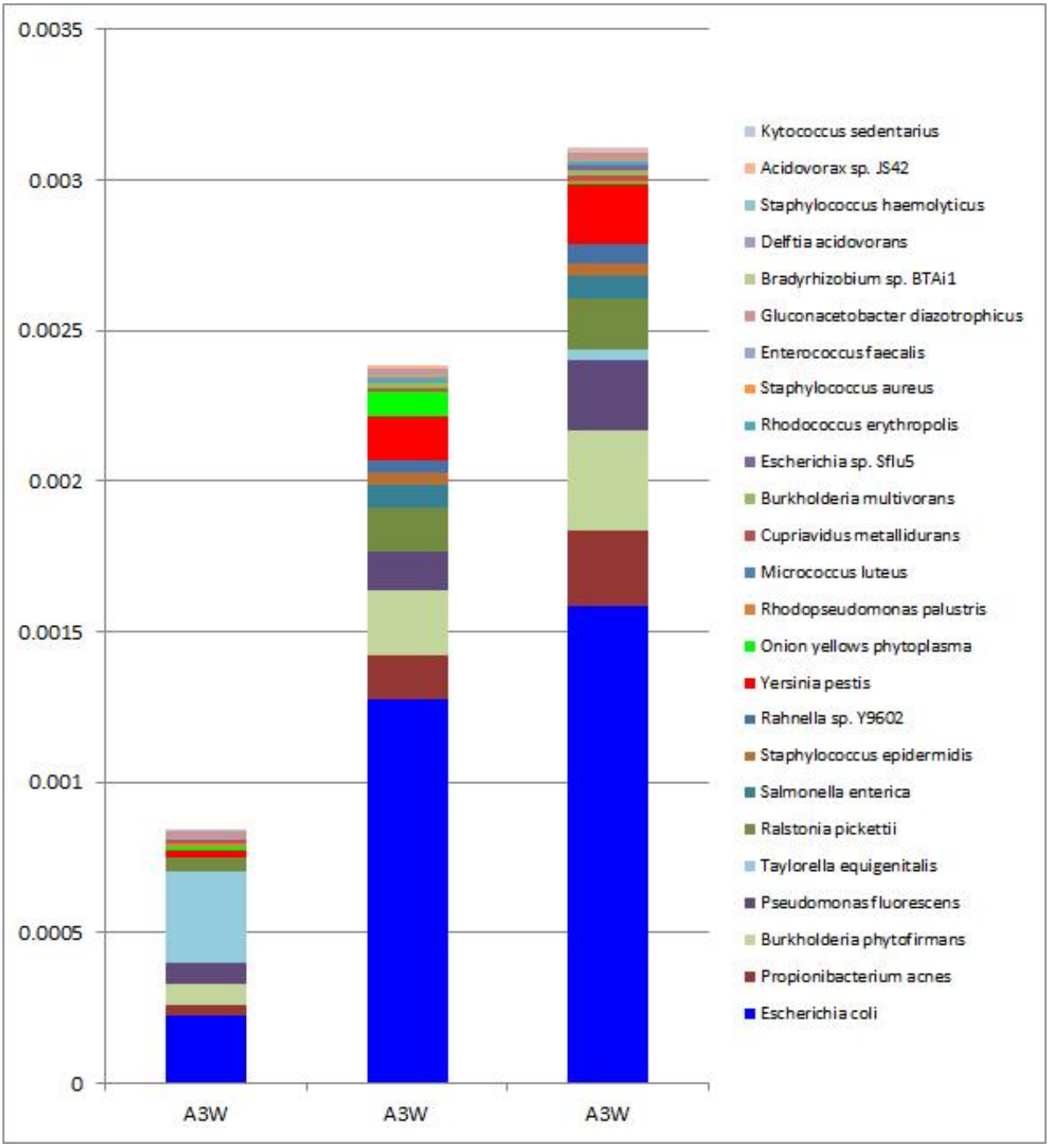

Figure 14. Proportions of Reads Mapping to the Top 25 Most Prevalent Bacterial Species Represented in NHP-A Day 4 WBC cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Middle) *vs* HAC+Cap(D0) (Right). Y-axis = Percentage of mapped reads.

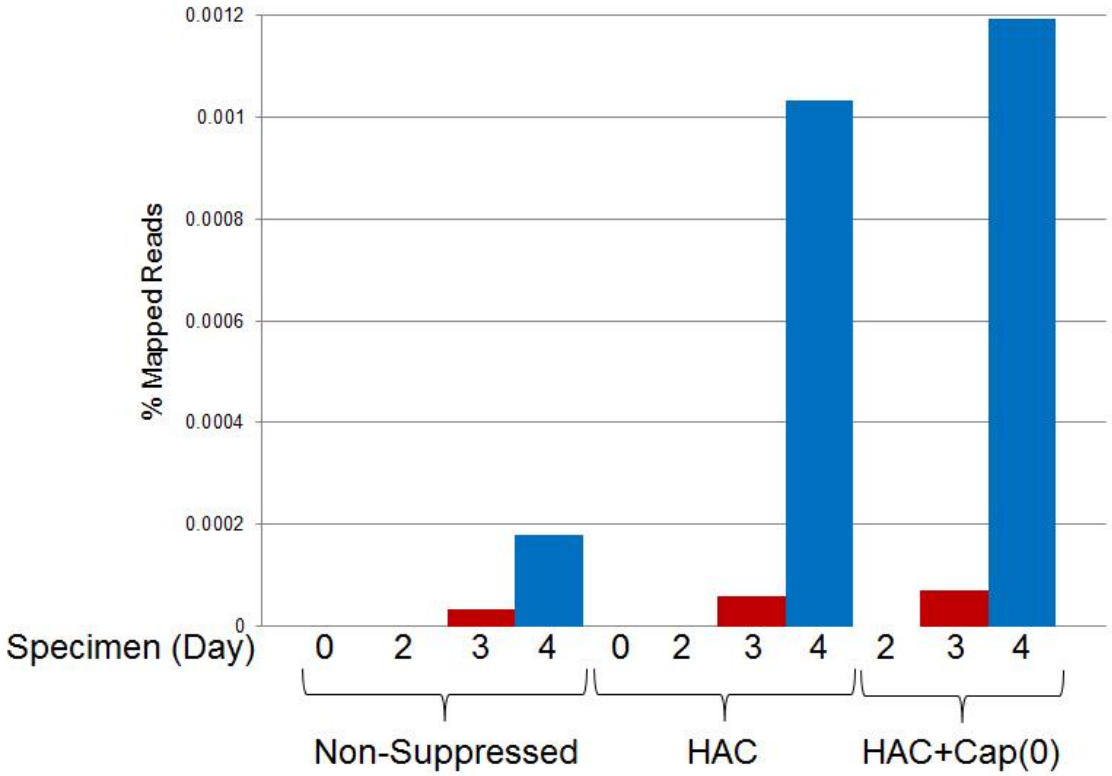

Figure 15. Proportion of Reads Mapping to *Y. pestis* in NHP-A WBC cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Middle) *vs* HAC+Cap(D0) (Right). Y-axis = Percentage of mapped reads.

6.2 Viral Transcripts/Genomes Detected in WBC Fractions

We did not detect transcripts/genomes from endogenous viruses in WBC fractions at appreciable levels. In the only positive case, small numbers of hits to Mason-Pfizer monkey virus were observed (average = 4 hits *per* library; range = 0-15 hits *per* library). Other virus hits appear to derive from environmental or reagent contamination introduced during library preparation. These include small numbers of hits to viruses that were being studied in our laboratory at the time of library preparation [e.g., Dengue (DENV), hepatitis C (HCV), human immunodeficiency virus 1 (HIV1)], and viruses that likely were introduced with contaminated water [*e.g.*, tobacco mosaic virus (TMV), melon necrotic spot virus (MNSV), pepper mild mottle virus (PMMoV), turnip vein-clearing virus (TVCV)]. Virus profiles did not change as a function of time post-infection, and were not affected by suppression.

6.3 Fungal Transcripts Detected in WBC Fractions

We detected fungal transcripts in WBC fractions at appreciable levels only in the cases of *Scheffersomyces stipitis* (average = 277 hits *per* library; range = 32-1285 hits *per* library) and *Pichia pastoris* (average = 32 hits *per* library; range = 2-97 hits *per* library). *S. stipitis* (aka

Pichia stipitis) was being studied in our laboratory at the time of library preparation; these hits likely derived from environmental contamination of the libraries. *P. pastoris* is commonly used for recombinant protein expression; these hits likely derived from contaminants associated with the recombinant enzymes we used to prepare our libraries.

6.4 Non-Host Transcripts Detected in Plasma Fractions

We detected appreciable levels of transcripts from a wide variety of bacterial species in plasma fractions, at ~10-fold higher levels than detected in WBC fractions. The bacterial profiles of the plasma samples were clearly distinct from those of the WBC samples; both were characterized by high levels of *Burkholderia phytofirmans*, but aside from that they held little in common. As with the WBC samples, *Y. pestis* hits were not detected in Day 0 (as expected) or Day 2 plasma samples. Very low levels of *Y. pestis* transcripts (3 hits) were detected in Day 3 plasma samples, and much higher levels (229 hits; 0.02% of mapped reads) in Day 4 plasma samples.

As with the WBC samples, the plasma samples contained very little viral content; in fact, the only virus we detected robustly was DENV, undoubtedly a contaminant. Similarly, we detected few fungal transcripts in the plasma fractions, with those derived from *S. stipitis* and *P. pastoris* the most common.

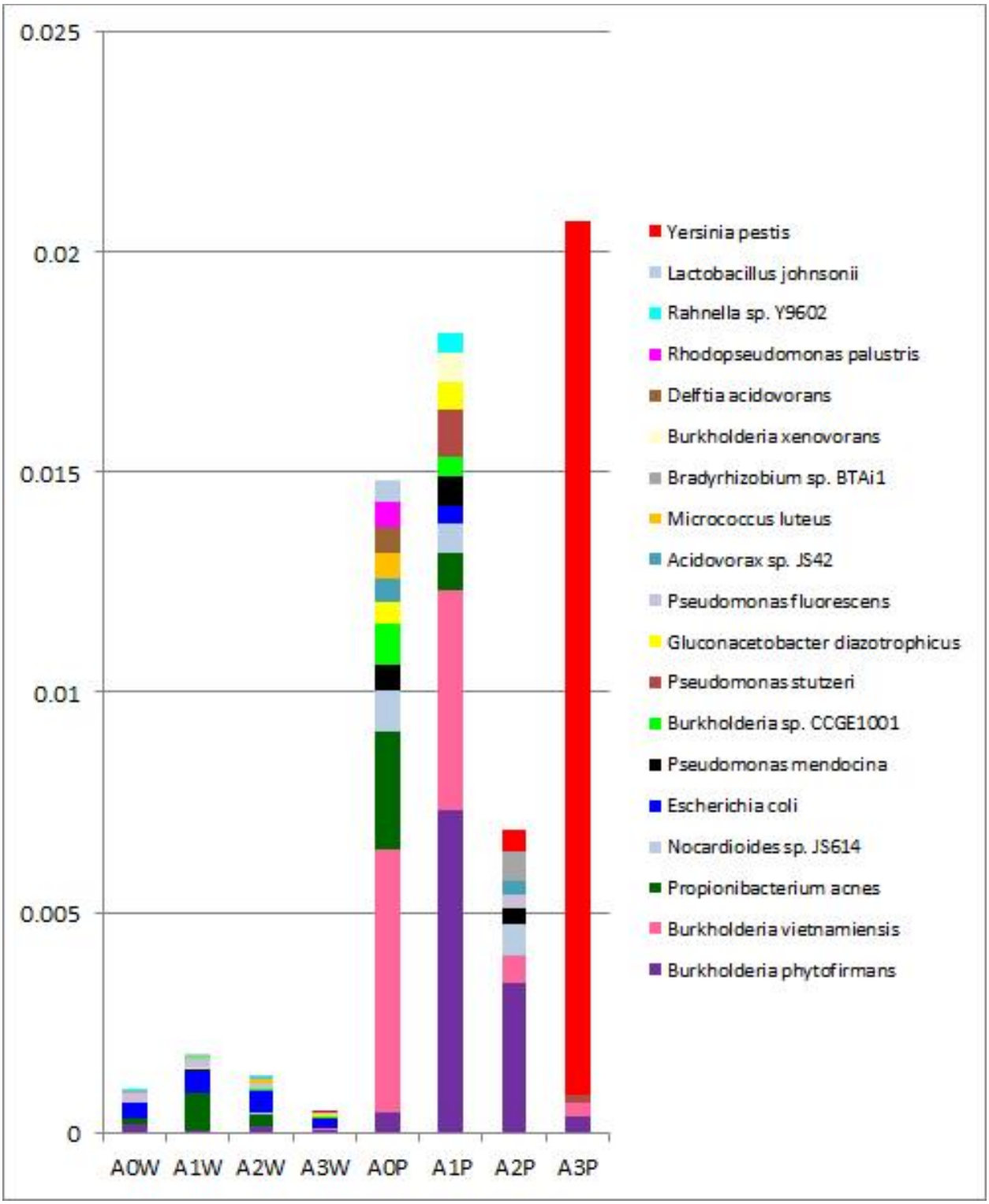

Figure 16. Proportions of Reads Mapping to the Top 19 Most Prevalent Bacterial Species Represented in NHP-A Plasma cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Right), Linear Scale. Y-axis = Percentage of mapped reads.

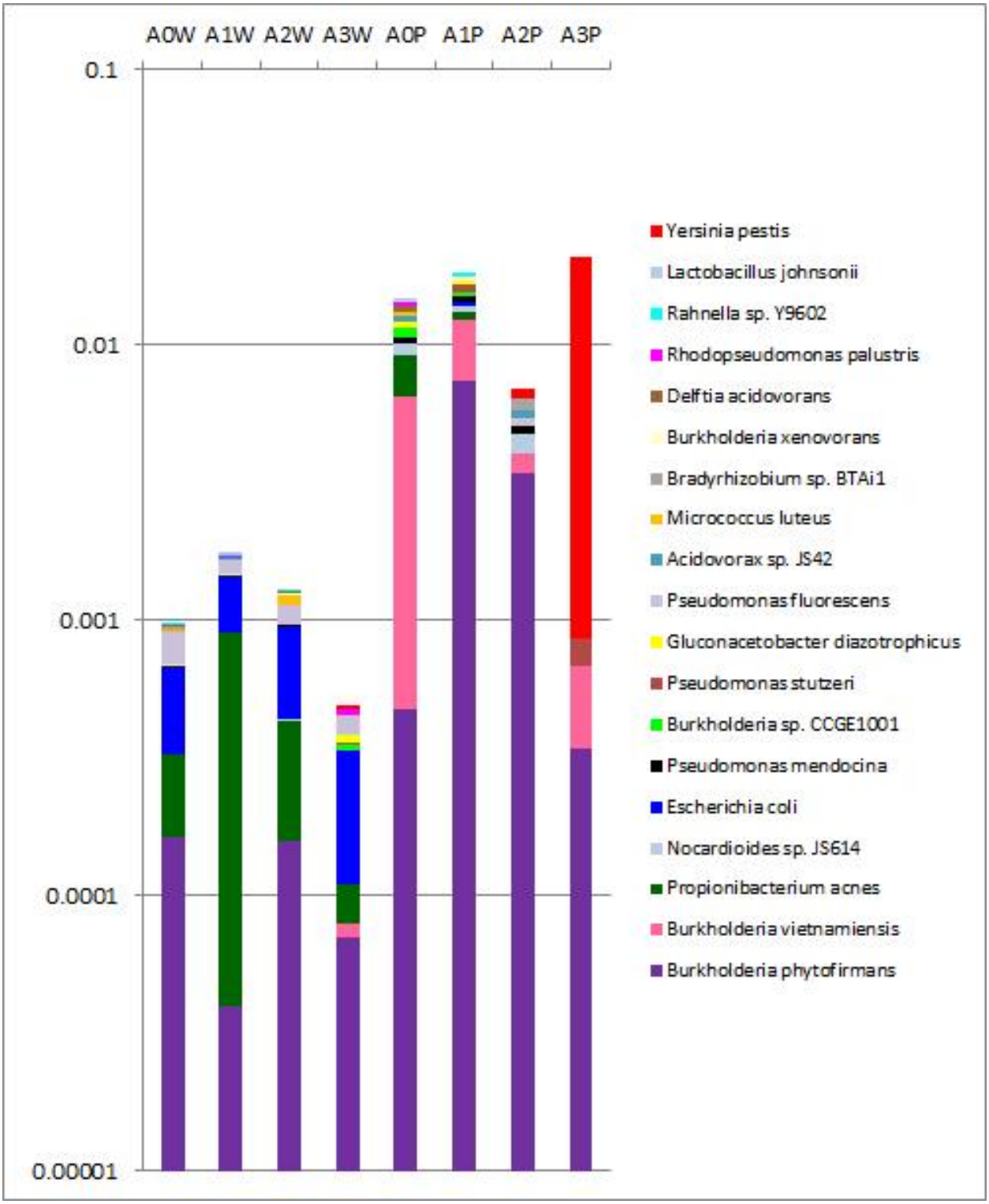

Figure 17. Proportions of Reads Mapping to the Top 19 Most Prevalent Bacterial Species Represented in NHP-A Plasma cDNA Libraries: Non-Suppressed (Left) *vs* HAC-Normalized (Right), Log_{10} Scale. Y-axis = Percentage of mapped reads.

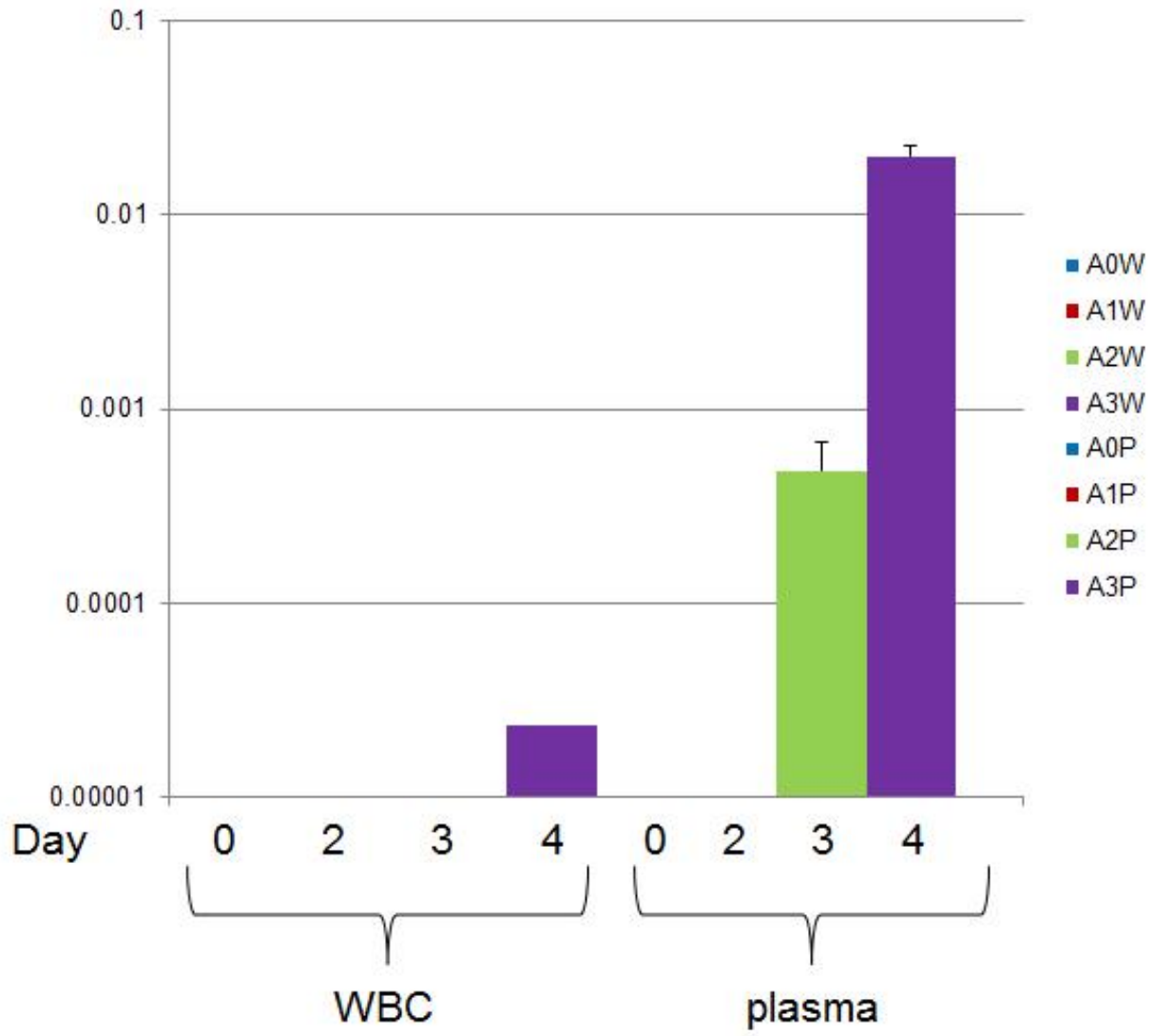

Figure 18. Proportions of Reads Mapping to *Y. pestis* in Non-Suppressed NHP-A WBC *vs* Plasma cDNA Libraries. Y-axis = Percentage of mapped reads.

6.5 Summary

As expected, non-host transcripts, including those derived from *Y. pestis*, were more abundant in plasma cDNA libraries as compared to WBC cDNA libraries. Bacterial transcripts accounted for the vast majority of these "microbiome" RNA species; viral and fungal transcripts were detected in much smaller amounts. Aside from the *Y. pestis* transcripts, there were no obvious indicator species, or sets of species, correlating with infection state, though subtle correlations may be teased out in future analyses. Molecular suppression treatments improved the sensitivity with which *Y. pestis* transcripts could be detected. They may also have improved the sensitivity with which transcripts from microbiome constituents were detected. However, in one notable case HAC-mediated normalization selectively reduced representation of a bacterial species

(*Taylorella equigenitalis*), and increased representation of other microbial species could be due to introduction or amplification of contaminants, rather than enrichment of endogenous (microbiome) species. Future analyses along these lines will require means by which mapped reads can be confidently and accurately assigned to contaminant *vs* endogenous species.

REFERENCES

1. Chaussabel D, Pascual V, Banchereau J (2010) Assessing the human immune system through blood transcriptomics. BMC Biol 8:84.

2. Langevin SA, Bent ZW, Solberg OD, Curtis DJ, Lane PD, Williams KP, Schoeniger JS, Lane TW, Branda SS (2013) Peregrine: A Rapid and Unbiased Method to Produce Strand-Specific RNA-Seq Libraries from Small Quantities of Starting Material. RNA Biology (in press).

3. Vandernoot VA, Langevin SA, Solberg OD, Lane PD, Curtis DJ, Bent ZW, Williams KP, Patel KD, Schoeniger JS, Branda SS, Lane TW (2012) cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. BioTechniques 53:373-80.

4. Bent ZW, Tran-Gyamfi MB, Langevin SA, Brazel DM, Hamblin RY, Branda SS, Patel KD, Lane TW, VanderNoot VA (2013) Enriching pathogen transcripts from infected samples: A capture based approach to enhanced host-pathogen RNA-Seq. Anal Biochem (submitted).

5. Merrick BA, Bruno ME (2004) Genomic and proteomic profiling for biomarkers and signature profiles of toxicity. Curr Opin Mol Ther 6:600-7.

6. Verweij CL (2009) Transcript profiling towards personalised medicine in rheumatoid arthritis. Neth J Med 67:364-71.

7. Bauer JW, Bilgic H, Baechler EC (2009) Gene-expression profiling in rheumatic disease: tools and therapeutic potential. Nat Rev Rheumatol 5:257-65.

8. Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP (2011) Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics 27:i383-91.

9. Van Andel R, Sherwood R, Gennings C, Lyons CR, Hutt J, Gigliotti A, Barr E (2008) Clinical and pathologic features of cynomolgus macaques (Macaca fascicularis) infected with aerosolized Yersinia pestis. Comp Med 58:68-75.

10. Koster F, Perlin DS, Park S, Brasel T, Gigliotti A, Barr E, Myers L, Layton RC, Sherwood R, Lyons CR (2010) Milestones in progression of primary pneumonic plague in cynomolgus macaques. Infect Immun 78:2946-55.

11. Blow JA, Dohm DJ, Negley DL, Mores CN (2004) Virus inactivation by nucleic acid extraction reagents. J Virol Methods 119:195-8.

12. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, Chisholm SW (2010) Unlocking short read sequencing for metagenomics. PLoS ONE 5:e11840.

13. Blencowe BJ, Ahmad S, Lee LJ (2009) Current-generation high-throughput sequencing: Deepening insights into mammalian transcriptomes. Genes Dev 23:1379-86.

14. Costa V, Angelini C, De Feis I, Ciccodicola A (2010) Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol 2010:853916.

15. Ebeling M, Küng E, See A, Broger C, Steiner G, Berrera M, Heckel T, Iniguez L, Albert T, Schmucki R, Biller H, Singer T, Certa U (2011) Genome-based analysis of the nonhuman primate Macaca fascicularis as a model for drug safety assessment. Genome Res 21:1746-56.

16. Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, Du H, Chen J, Chen R, Zhang P, Huang Z, Thompson JR, Meng Y, Bai Y, Wang J, Zhuo M, Wang T, Huang Y, Wei L, Li J, Wang Z, Hu H, Yang P, Le L, Stenson PD, Li B, Liu X, Ball EV, An N, Huang Q, Zhang Y, Fan W, Zhang X, Li Y, Wang W, Katze MG, Su B, Nielsen R, Yang H, Wang J, Wang X, Wang J. (2011) Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol 29:1019-23.

17. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-9.

DISTRIBUTION

Internal (Electronic Copy)

1	MS-0359	LDRD Office, 1911
1	MS–9291	A. Singh, 8620
1	MS-9291	V. VanderNoot, 8621
1	MS-9671	Z. Bent, 8623
1	MS-9671	A. Sinha, 8623
1	MS-0899	Technical Library, 9536

External (Electronic Copy)

None

Sandia National Laboratories