AD_____

Award Number: W81XWH-04-1-0669

TITLE: Epidemic Outbreak Surveillance (EOS)

PRINCIPAL INVESTIGATOR: Thomas C. Scofield
Dr. Elizabeth Walter
LTC Samuel J.P. Livingstone

CONTRACTING ORGANIZATION: The Henry M. Jackson Foundation for the Advancement of Military Medicine, Incorporated
Rockville, MD 20852

REPORT DATE: July 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-07-2006	Final	1 Jul 2004 – 30 Jun 2006

4. TITLE AND SUBTITLE

Epidemic Outbreak Surveillance (EOS)

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0669

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Thomas C. Scofield
Dr. Elizabeth Walter
LTC Samuel J.P. Livingstone

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Henry M. Jackson Foundation for the
Advancement of Military Medicine, Incorporated
Rockville, MD 20852

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT

This funding is established to support Operations and Management. The goal of the project is to develop and test new technologies for the diagnosis and surveillance of respiratory tract pathogens. This funding supported purchase of supplies and shipping services necessary to carry out protocols to standardize methods of specimen collection and to optimize processing of these specimens. After these processes were established, initial supplies were purchased in order to begin enrollment of healthy, ill, and recovered Basic Military Trainees (BMTs) in order to compare methods of detection of respirator and host response. This funding enabled successful initiation of the project.

15. SUBJECT TERMS
Epidemic Outbreak surveillance, pathogens, respiratory disease, Advanced Medical Testing Demonstrator (AMTD)

16. SECURITY CLASSIFICATION OF:				18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	17	

Table of Contents

Introduction

This funding was grated to The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., (HJF) to provide the administrative, management, logistical and programmatic services in collaboration with HQ, USAF/SGR in accordance with the statement of work (SOW) and with tasks developed by the research. This has occurred through daily interactions of HJF resource personnel assigned to the EOS project with the USAF Principal Investigator for the project and with guidance from HQ, USAF/SGR.

Body

Statement of Work

Rapid diagnosis and surveillance of respiratory pathogens.

Key Research Accomplishments

This funding was established to support Operations and Management. The goal of the project was to develop and test new technologies for the diagnosis and surveillance of respiratory tract pathogens. This funding supported hiring of technical and administrative personnel necessary to carry out protocols to standardize methods of specimen collection and to optimize processing of these specimens. After these processes were established, initial enrolment of healthy, ill, and recovered Basic Military Trainees (BMTs) begin to compare methods of detection of respiratory pathogens and host response.

Following is a list of the most significant activities and platforms tested during the past two years.

- **Nucleic Acid Extraction**
  Laboratory technicians were trained in the Nucleic Acid Extraction protocol established for the Affymetrix Pathogen Identification platform. The project was designed to build consistency in total nucleic acid extraction procedures using the MasterPure Complete DNA & RNA Purification Kit (Epicentre). Training continued with the Affymetrix Respiratory Pathogen Microarray (RPM) protocol using the nucleic acid extraction protocol on normal nasal wash samples spiked with titered Influenza A virus from Virapur, LLC.

- **FluMist Vaccine Testing on Affymetrix Respiratory Pathogen MicroArray**
  Training was conducted using the Random Amplification (Modified DeRisi-RNA) protocol on FluMist Vaccine, Lot #500339 (extracted RNA 03-10-2005). Total Nucleic Acid from FluMist Vaccine was tested on the Respiratory Pathogen Microarray to determine specificity between FluMist and wild type influenza. The testing concluded that FluMist was undistinguishable from wild type influenza on the microarray due to the use of highly conserved genetic probes on the Affymetrix chip. Samples would have to be sequenced to determine FluMist from wild type influenza in patients exhibiting influenza symptoms that had received the FluMist Vaccine.

- **Gene Expression Automated Processing**
  Training was conducted on the Affymetrix Array Station, an automated method to prepare extracted nucleic acid from PAXgene ® (Qiagen) blood tubes for Affymetrix Gene Expression Microarrays. Robotic training consisted of equipment familiarization, set up, Fragmentation and Labeling, Random Amplification, Target Hybridization, Washing and

Staining for the 24 PEG Array platform. Training also included the Affymetrix Gene chip Hybridization Oven and the Gene Chip HT Scanner. This training was designed as a precursor for up coming gene expression studies.

- **Manual Processing of Affymetrix Gene Expression Protocol**
  Manual processing was completed on RNA extracted from normal, healthy blood samples to get a better understanding of the Affymetrix process before proceeding to the automated process. The Affymetrix protocol optimized by the Naval Research Lab was utilized for these experiments.

- **RT-PCR Experiment to Test Integrity of 2003 Nasal Wash and Throat Swab Samples Stored at -80°C. September 2005 (Sue Worthy/Luke Daum)**
  Seven matched pairs of Nasal Wash and Throat Swab samples, previously tested in September 2004 for Influenza A, were pulled from the -80°C storage freezer for re-testing to make sure the virus was still detectable. All samples were treated in the same manner as previously tested. The nucleic acid was extracted using the Qiagen QiAmp Viral RNA Mini Kit. Isolated nucleic acid was set up for RT-PCR using the Invitrogen One-step RT-PCR Kit with SuperScript III and Luke Daum's (AFIOH) Flu A Matrix Gene Universal Primer/Probe Set and Flu B Matrix Gene Primer/Probe Set. All samples tested positive for Influenza A as confirmed by real time RT-PCR on the Roche LightCycler. Additionally, the samples were run using the One-Step RT-PCR Kit and specific primer/probe sets for H1 gene and H3 gene to subtype the Influenza A group. All samples tested positive for H3 Influenza A. Conclusion was that samples are still viable for virus isolation and testing after 2½ years storage at -80°C.

- **Qiagen BioRobot Training**
  The Qiagen Bio Robotic 8000 Universal System is an automated process to extract and purify RNA from 24-96 PAXgene blood tubes. During training 24 of the 48 PAXgene blood samples drawn from six healthy participants were processed on the robot. The Qiagen protocol was followed without modification. After software updates and protocol modifications, a second stage of training was conducted by Qiagen on the remaining 24 blood samples. Each of the two runs yielded different results based on software changes and temperature variances. All 48 samples were run on the 2100 BioAnalyzer (Agilent) and NanoDrop to check for sample quality and quantity. Extracted RNA was stored at -80°C for future studies.

- **Implementation of Globin Clear Protocol**
  October through November 2005 involved addition of the GeneChip Globin-Reduction Kit (Affymetrix) to clear globin transcripts from the extracted RNA from PAXgene blood samples. Globin is an interfering factor leading to reduced sensitivity of gene expression analysis on the GeneChip arrays.

- **Total Inventory of All Stored Samples**
  December 2005 - January 2006 a detailed process was completed to inventory and organize the sample storage freezers for all Nasal wash, Throat swab and PAXgene blood tubes. All samples were inventoried and checked for location against a master database list. Any discrepancies were recorded and the database was corrected.

- **Applied Biosystems (AB) Training on Infectious Disease Surveillance System (IDSS) Protocol**

Training was conducted at AFIOH, Brooks City-Base for the AB protocol that included extraction of total nucleic acid from nasal wash samples spiked with 4 bacterial and 6 viral pathogens, instruction on the AB 6100 Nucleic Acid Prep Station, preparation and setup of samples on TaqMan Low Density Arrays (TLDA) (384-well microfluidics cards), and operation of the AB 7900HT Real-Time PCR instrument.

- **Small Preliminary Validation Study for real clinical samples using the AB Infectious Disease Information System (IDIS) Protocol (TLDA Platform)**
  The ADL was supplied with 20 TLDA cards (Eagle version) and reagents to perform a pre-validation study; testing the specificity and the efficiency of the TLDA cards using true clinical nasal wash samples. A summary of the cards and results is attached.

- **Small Gene Expression Study Conducted**
  The Small Gene Expression Study was conducted comparing gene expression profiles between 3 different patient populations with confirmed infections of influenza A, adenovirus, or Group A strep. A control group consisting of patients with febrile respiratory illness (FRI) that were culture negative for influenza A, adenovirus or Group A strep will also be included in the study. The goal of this study is to use the gene expression profiles of the 4 different patient populations to design Applied Biosystems TaqMan Low-Density Array cards with unique gene expression profiles for each patient population.

Protocol Procedure: In this study, 3 target preparation runs of 24 samples each were loaded on Affymetrix Array Station Robot for a total of 72 samples. 48 samples consisted of the study groups and the rest (24 samples) consisted of positive controls (Jurkat RNA, Hela RNA, Lymph Node RNA). The 24-peg array plates were scanned on the GeneChip HT Scanner.

Preliminary data analysis indicates that introduction of technical variables into the study were minimal. The high percentage of present calls calculated for all the positive controls (lower graph) are indications of excellent results. The low standard deviations found in all the groups demonstrate that the 3 separate robotic runs were consistent. Currently, extensive data analysis is being conducted by Luis Perez with the help of Chris Myers and the bio-statistician at the Naval Health Research Center in San Diego, CA to determine the extent of different gene expression profiles among the four different population groups.

- Personnel Hired for the Project:
  Research Technician
  Clinical Laboratory Technician (2)
  Clinical Research Nurse
  Data Base Administrator
  Administrative Assistant/Data Entry Technician

- Study Form Development: Since this was a new project, data collection forms had to be designed prior to any samples being collected and tested. EOS personnel generated the first draft/template as well as all changes and revisions to the actual documents. Four forms were created:

BMT Study: Four forms were created (Initial, Sick, and Follow-Up) and the Presymptomatic Survey form
Hospital/Clinic Study: Encounter Form

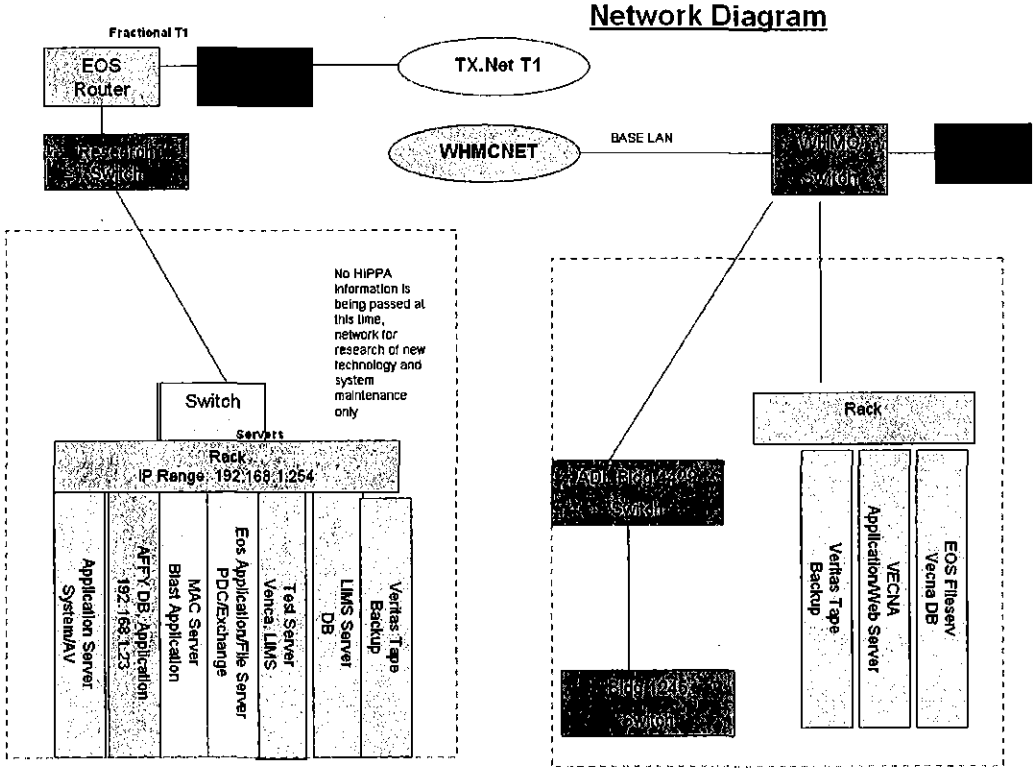
Hospital/Clinic Exempt Study: Sample Handling Form
An electronic "blank" form of each type is available if needed.

- Data Entry – HJF EOS Personnel developed a system for entry/capture of all the data collected on study forms. As mandated by the Principal Investigaor, system requires dual-entry of all paper forms and automatically checks the entries for a match before "marking" the data as valid. With coordination from the clinical team, the database administrator helped to accomplish this task. The administrative staff was responsible for entering the data

- Current Computer/Network Infrastructure – The "Research Network" was brought to life to support all the laboratory and administrative needs of the program. Everything from the purchase of all equipment, coordination of wiring contractors and base resources, and server hook-up and configuration of all servers, printers and workstations (including laboratory equipment) was completed by the HJF EOS IT staff located at Lackland AFB

[redacted]

Epidemic Outbreak Surveillance: Current Network Diagram

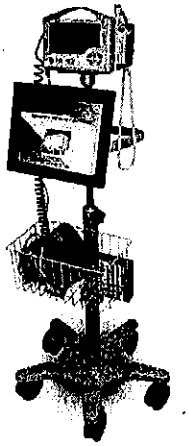


[redacted]

- Patient Kiosk Prototype/Proof of Concept – HJF EOS IT staff developed a working model/prototype to demonstrate that all the input/knowledge gathered from the end users *(clinics/battalion aid stations) was achievable. The prototype included scanning of ID* cards, electronic capture of vital signs, and printing of a custom SF600 based on chief complaint/reason for visit. This prototype lead to acceptance and funding for 15 kiosks currently being used at Lackland in three locations (Number of kiosks and locations still

expanding at Lackland) and the same equipment will soon be used at multiple sites worldwide as part of the "Silent Guardian II" Avian Flu surveillance endeavor. Maintenance and upkeep of this equipment/system falls on the entire EOS IT staff.

Vitals Kiosk



- Laboratory Information Management System (LIMS) – The purchased LIMS solution is currently being configured and customized for EOS. Vendor: LabWare Inc., Product: LIMS v5.X Ongoing project that has the basic laboratory sample receiving portion ready for actual use, key instrument interfaces in testing. Coding is done in proprietary LIMS programming language. While the LIMS would not be considered part of the EOS solution package for roll-out to all Medical Treatment Facilities (MTFs), it must be evaluated for its continued use and value in a testbed executing numerous human research protocols.

Reportable Outcomes

In accordance with the SOW, HJF successfully established administrative and programmatic support services to support the operations and management of the EOS project. HJF effectively and productively provided other study-related tasks as deemed necessary for execution of the EOS project as requested by HQ, USAF/SGR

Conclusion

During the term of this grant, HJF efficiently provided the necessary administrative, management and programmatic support services as deemed necessary for implementation and as requested by HQ, USAF/SGR for the development of an integrated health surveillance venue, focused upon epidemic outbreaks of acute respiratory disease (ARD) and other endemic and seasonal respiratory infections.

References

Not applicable.

Appendices

The following information is a full paper entitled Surveillance of Transcriptomes in Basic Military Trainees with Normal, Febrile Respiratory Illness, and Convlescent Phenotypes (8 pages).

Genes and Immunity (2005) 6, 588–595

www.nature.com/gene

FULL PAPER

Surveillance of transcriptomes in basic military trainees with normal, febrile respiratory illness, and convalescent phenotypes

DC Thach[1], BK Agan[2], C Olsen[3], J Diao[4], B Lin[1], J Gomez[3], M Jesse[3], M Jenkins[2], R Rowley[5], E Hanson[5], C Tibbetts[5,6], DA Stenger[1], E Walter[5,7] and EOS[8]

[1]*Center for Biomolecular Science and Engineering, Naval Research Laboratory, Washington, DC, USA;* [2]*Department of Infectious Diseases, Wilford Hall USAF Medical Center, LAFB, TX, USA;* [3]*Henry M Jackson Foundation, San Antonio, TX, USA;* [4]*Spin Systems, Inc., Sterling, VA, USA;* [5]*HQ USAF/SGR, Falls Church, VA, USA;* [6]*The George Washington University, Washington, DC, USA;* [7]*Texas A&M University Systems, College Station, TX, USA*

Gene expression profiles permit analysis of host immune response at the transcriptome level. We used the Pax gene Blood RNA (PAX) System and Affymetrix microarrays (HG-U133A&B) to survey profiles in basic military trainees and to classify them as healthy, febrile respiratory illness (FRI) without adenovirus, FRI with adenovirus, and convalescent from FRI with adenovirus. We assessed quality metrics of RNA processing for microarrays. Class prediction analysis discovered nested sets of transcripts that could categorize the phenotypes with optimized accuracy of 99% (nonfebrile vs febrile, $P<0.0005$), 87% (healthy vs convalescent, $P=0.001$), and 91% (febrile without vs with adenovirus, $P<0.0005$). The discovered set for classification of nonfebrile vs febrile patients consisted of 40 transcripts with functions related to interferon induced genes, complement cascades, and TNF and IL1 signaling. The set of seven transcripts for distinguishing healthy vs convalescent individuals included those associated with ribosomal structure, humoral immunity, and cell adhesion. The set of 10 transcripts for distinguishing FRI without vs with adenovirus had functions related to interferon induced genes, IL1 receptor accessory protein, and cell interactions. These results are the first in vivo demonstration of classification of infectious diseases via host signature transcripts and move us towards using the transcriptome in biosurveillance.
Genes and Immunity (2005) **6**, 588–595. doi:10.1038/sj.gene.6364244; published online 21 July 2005

Keywords: *transcriptome; surveillance; infection phenotypes*

Introduction

Gene expression analysis using microarray technology holds the potential to advance our understanding of disease pathogenesis and to improve diagnosis, prognosis, and targeted therapeutics. Cancer research[1,2] demonstrates that gene expression profiles can classify malignancies to a level of detail not previously possible, stratifying for prognosis and therapeutic response. Recently, profiles of blood cells from experimental smallpox infection of non-human primates were assessed to determine global immune responses to a pathogen.[3] We further study the use of microarray technology to survey gene expression profiles from blood samples of a human population to determine whether specific immunologic signatures exist for a particular pathogen and to simultaneously elucidate pathways of pathogenesis at the transcriptome level.

Correspondence: Dr D Thach, Center for Biomolecular Science and Engineering, Naval Research Laboratory, 4555 Overlook Ave., SW Bldg 30, Washington, DC 20375, USA. E-mail: dthach@cbmse.nrl.navy.mil
[8]Epidemic Outbreak Surveillance (EOS) members listed at the end of paper.
Received 27 April 2005; revised 3 June 2005; accepted 6 June 2005; published online 21 July 2005

The setting for this study is Lackland Air Force Base (LAFB) in San Antonio, Texas, where Basic Military Trainees (BMTs) enter a 6-week training course and approximately 20% develop febrile respiratory illness (FRI). Since the discontinuation of the adenoviral vaccine in the late 1990s, approximately 60% of FRI is due to adenovirus serotype 4.[4,5] We undertook a study to compare the gene expression profiles of healthy, FRI without adenovirus, FRI with adenovirus and those who had convalesced from FRI with adenovirus in this BMT population. Previously, we have shown methods for handling and processing blood specimens in which RNA integrity was maintained for transcriptome measurements.[6] Now, we extend this experience to survey the gene expression profiles of this cohort to determine whether a set of transcripts exists that differentiate healthy from febrile and, specifically differentiate adenoviral from nonadenoviral causes of FRI.

Results

Clinical phenotypes

A total of 30 healthy, 19 with FRI and negative by culture for adenovirus, 30 with FRI and positive by culture for

adenovirus, and 30 convalescing from adenovirus-positive FRI were enrolled in this study. Enrollees in these four infection status phenotypes were matched for age ±3 years and race/ethnicity. Only male BMTs were enrolled. After selection of samples meeting standards for gene expression analysis, 17 FRI without adenovirus had been ill for 5±3 days (median±s.d.), whereas 26 FRI with adenovirus had been ill for 8±4 days ($P=0.006$, Wilcoxon). The incidence of symptoms over all the groups was sore throat (95.3%), cough (93%), sinus congestion (90.7%), headache (88%), chills (84%), rhinorrhea (81%), body aches (65%), malaise (63%), nausea (54%), diarrhea (14%), pleuritic chest pain (14%), vomiting (14%), and rash (0%), with no significant differences between the FRI groups. There was also no significant difference in allergies, recent injuries, and smoking history among the infection status phenotypes.

Quality and variations of RNA derived from PAX system from the BMT population

In order to identify clinically relevant gene expression profile differences for phenotypes in a population, it is essential that the RNA sample applied to the microarray is representative of the amount of transcripts *in vivo*. The PAX system was used to minimize handling of blood cells postcollection and to immediately stabilize RNA and halt transcription. We previously have shown two methods using this PAX system that provide stable RNA for microarray analysis.[6]

To assess RNA quality on each of the 95 microarrays analyzed in this study, recently published metrics derived from electropherograms of the RNA were used.[7] Assessment of the degradation factor, which is the ratio of the average intensity of bands of lesser molecular weight than the 18S ribosomal peak to the 18S band intensity multiplied by 100, demonstrated minimal degradation of RNA (Figure 1). This degradation factor for the samples correlated with *gapdh* 3′/5′ on the A arrays (Figure 1c; $r=0.3$, $P=0.008$, ANOVA) and *actin* 3′/5′ on the B arrays ($r=0.2$; $P<0.05$, ANOVA), the internal measurements for assessment of RNA quality on the microarray. There was no significant correlation between 28S/18S *vs* degradation factor, *gapdh* 3′/5′, and *actin* 3′/5′, suggesting that the degradation factor is a superior method for assessing RNA quality for microarray analysis. No significant difference in degradation factor was seen among the phenotype groups.

Assessment of the apoptosis factor, which is the ratio of the *height* of the 28S to 18S peak,[7] suggested that a high percentage of blood cells underwent apoptotic cell death. The distribution of the degradation factor, apoptosis factor, 28S/18S, and yields of total RNA are shown in Figure 1b. No significant difference in apoptosis factor was seen among the phenotype groups. There was no significant correlation between duration of freezing and degradation factor (Figure 1d), nor was there correlation with apoptosis factor, RNA yield, 28S/18S, or *gapdh* and *actin* 3′/5′.

We determined if blood cell type heterogeneity affected the sensitivity of transcript detection. Assessment of complete blood count (CBC) variables that affect the number of present calls on the microarray demonstrated a linear correlation between number of probesets called Present and Mean Corpuscular Hemoglobin (MCH). A significant effect was detected ($r=0.272$; $P=0.008$, ANOVA) for the B arrays only (Figure 1e). The equation of the regression line suggested that for every picogram increase in hemoglobin, there is a loss in present detection calls of 100 probesets or 2% of the average number of present called probesets on the B arrays. There was no difference in MCH among the infection status phenotypes.

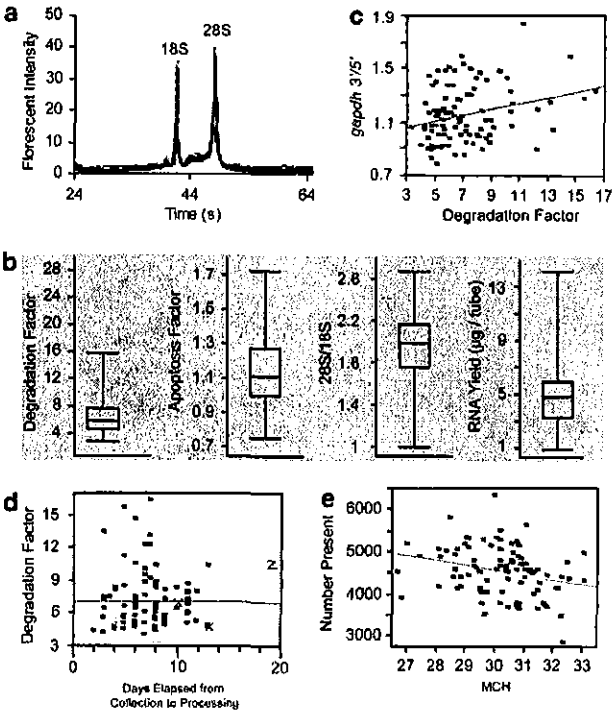


Figure 1 Quality of RNA derived from the PAX system of samples from the BMT population. (a) Overlay of electropherograms from BMTs with various phenotypes and handling conditions. The 18S and 28S ribosomal peaks are indicated. (b) Box plots of quality metrics calculated from the electropherograms. (c) Correlation between *gapdh* 3′/5′ values on the A arrays *vs* degradation factor ($r=0.3$, $P=0.008$, ANOVA). (d) Lack of RNA degradation over days elapsed from blood collection to processing. Samples marked by '+', 'x', or 'z' had an additional freeze thaw cycle before final thawed for RNA isolation. (e) Correlation between the mean corpuscular hemoglobin (MCH) and number of probesets called Present in the B arrays ($r=-0.272$; $P=0.008$, ANOVA). Line shown is from equation: number Present = 8108–117 MCH.

Quality of microarray measurements of PAX system-derived RNA from the BMT population

Individual control charts *vs* the date of microarray scanning were plotted to establish stability of quality metrics over time, determine outliers, and compare with values proposed by the array manufacturer. The percentage of transcripts present was 32±10 (average±3s.d.) for A arrays and 21±6 for B arrays. The *gapdh* and *actin* 3′/5′ values were less than three, the upper-limit proposed by Affymetrix.[8] Noise was 3.6±1.3 for A arrays and 2.9±0.8 for B arrays. Average Background was 100±48 for A arrays and 78±33 for B arrays. After exclusions of array sets that were known to have been processed differently or erroneously, a total of 95 A and B array sets with stable quality metrics remained. These 95 sets were processed in batches with nearly equal

representation of the four infection status phenotypes. Therefore, comparisons among these four groups should detect biological differences as these groups have similar variations due to processing.

Gene expression profiles

The gene expression profiles were displayed on a heat-map with hierarchical clustering of transcripts to characterize and visualize patterns in the profiles of our cohort (Figure 2). Initial examination revealed a large number of transcripts with high expression levels (Figure 2, orange bar) and a smaller number of transcripts with low expression levels (Figure 2, purple bar) in the febrile group compared to the nonfebrile healthy and convalescent patients. There were also transcripts that showed differences between healthy and convalescent patients (Figure 2, gray bar), while there was no obvious group of transcripts that showed differences between febrile without adenovirus *vs* febrile with adenovirus from this visual inspection. Within each group, interindividual variation was observed, suggesting diverse immune responses in this population.

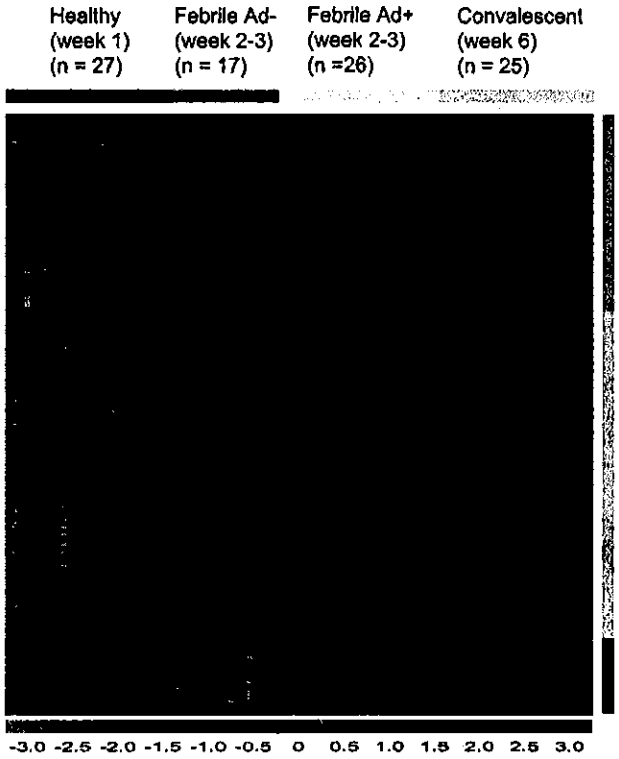


Figure 2 Gene expression profiles of the BMTs. To remove undetected transcripts, those with >80% absent calls across samples were filtered resulting in 15721 from 44928 probesets. To remove uninformative transcripts, probesets in which less than 20% had a 1.5-fold or greater change from the probeset's median value were removed, resulting in 7682 probesets. To focus on transcripts with differences in expression among the four infection status phenotypes, those probesets with $P>0.01$ by ANOVA were excluded, resulting in 4414 probesets. The heat-map shows the transcript abundance (green to red intensities) detected by these 4414 probesets (rows) in each blood sample (column). The rows were hierarchically clustered with 1-correlation distance and average linkage, while the columns were sorted into the infection status phenotypes. Top blue, brown, yellow, and light blue bars denote samples from healthy, febrile without and with adenovirus, and convalescent patients, respectively. Bottom scale denotes standardized values for the green to red intensities in the heat-map. Side gray, orange, and purple bars denote clusters of transcripts that differ among the phenotypes.

Class prediction of infection status phenotype

The pattern recognition above suggested that there were transcripts with differences in expression levels among healthy, febrile, and recovered patients. Therefore, class prediction was performed, to find sets of transcripts that best classify the four infection status phenotypes. Probesets with >80% absent calls across samples were filtered resulting in 15721 probesets for further analysis. For supervised class prediction, the class labels for the febrile group were determine from respiratory viral culture results identifying presence or absence of adenovirus.

Figure 2 suggested that the fever status of individuals was the predominant source of variation in gene expression profiles among samples and this was confirmed by unsupervised clustering of samples. Thus, supervised class prediction analysis was used to find sets of transcripts that classified nonfebrile *vs* febrile patients first (node 1), then of the nonfebrile patients, further classified to healthy or convalescent (node 2), and among the febrile patients, further classified to without or with adenovirus infection (node 3). The segregation of the samples via this nodal scheme was confirmed via binary tree class prediction analysis.

Unlike data from cancer studies,[2,9] there are no reported transcript selection methods or class prediction algorithms that are optimal for classification of infectious diseases. Therefore, we determined the transcript selection method and classification algorithm that would result in the highest percent correct classification during leave-one-out cross-validation. To estimate the optimal transcript selection parameters for classification in each node, the cutoff level of the univariate *P*-value was varied, selecting for probesets that showed statistically significant differences between the two groups at a *P*-value that was equaled to or smaller than a set cutoff level. As the *P*-value cutoffs became more stringent, the number of probesets selected decreased. For each *P*-value cutoff level, the selected probesets were subsequently used to classify the samples using various algorithms along with cross-validation analysis. For classification of node 1, 2 and 3, an optimal *P*-value cutoff level of 10^{-2}, 10^{-3}, 10^{-5} (Figure 3a–c, lower-left corner) was chosen, respectively.

Once an optimal *P*-value cutoff level was estimated and held constant, the additional criterion of fold-change cutoff threshold was varied (Figure 3a–c, *x*-axis) for each node. Figure 3 shows the percent-correct traces for the six algorithms tested tracking closely as fold-change cutoff level increases, but can differ by as much as 10–20% between methods. The black arrows in Figure 3 indicate an optimal percent-correct classification at the specific *P*-value and fold change cutoff. For nonfebrile *vs* febrile, a percent correct call of 99% was achieved using the support vector machines algorithm at a *P*-value cutoff level of 10^{-2} and a fold-change threshold of >5 which selected for 47 probesets to be in the classifier (Figure 3a). For classification of healthy *vs* convalescent patients, an optimal percent correct of 87% using the diagonal linear

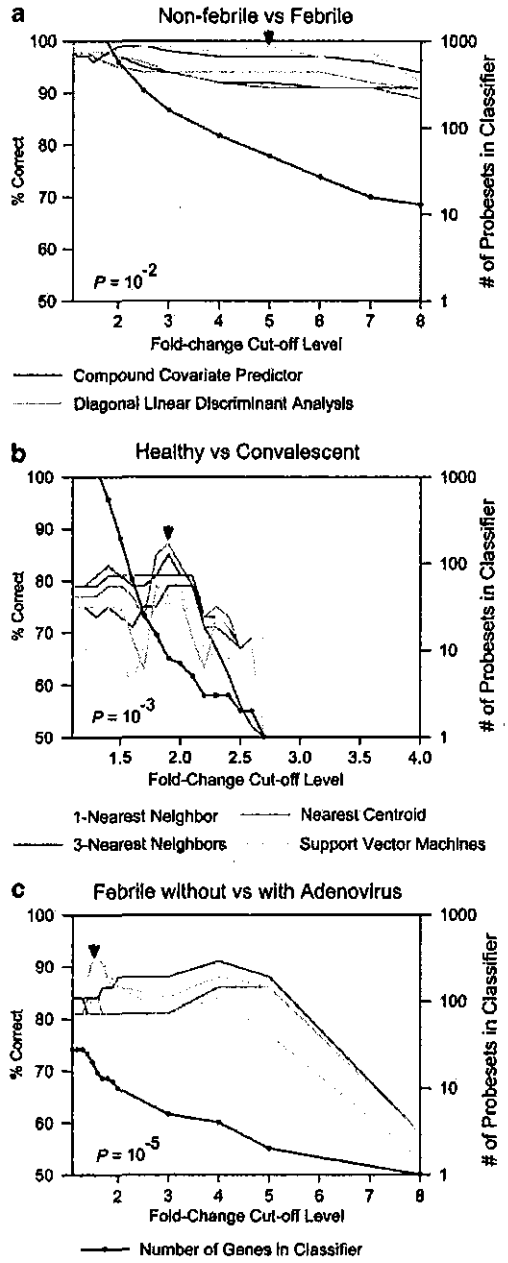


Figure 3 Optimization of class prediction for nonfebrile *vs* febrile (a), healthy *vs* convalescent (b), and febrile without adenovirus *vs* febrile with adenovirus infection (c) phenotypes. Shown in the lower left corners of the three panels are the estimated optimal P-value cutoff levels for each of the three classifications. Classifier transcripts were further filtered by fold change level (*x*-axis), with resulting percent correct classification (left *y*-axis) for various algorithms (color traces), and the number of probesets in the classifier (right *y*-axis, beaded black trace); arrows indicate fold change level that resulted in a highest percent correct classification.

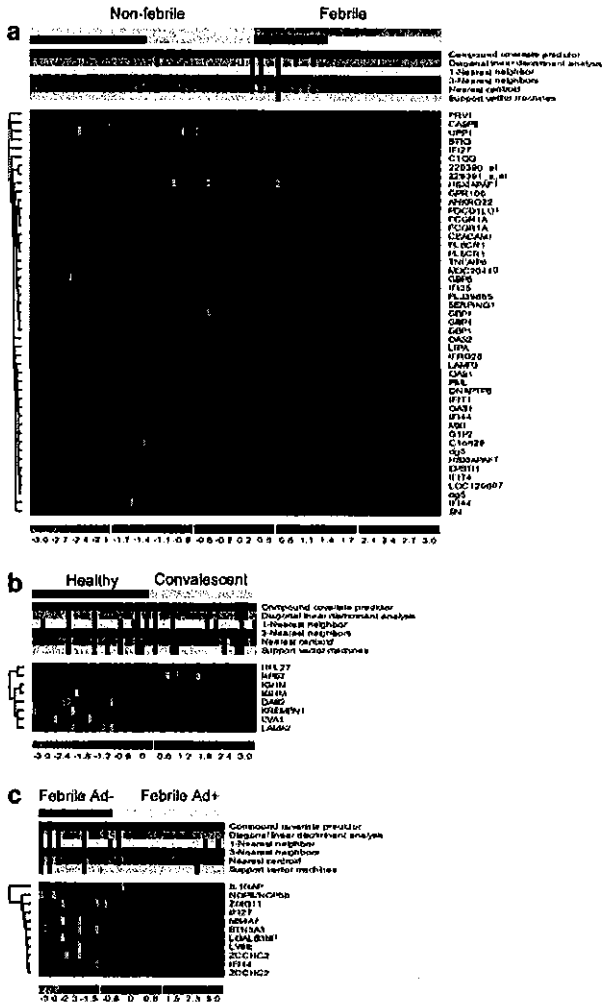


Figure 4 Identities and expression of genes in classifiers found from class prediction analysis. In each panel, top bar indicates the classification phenotypes of the samples (columns). Panel (a) has a second bar that further indicates healthy, convalescent, febrile without and with adenovirus samples as blue, light blue, brown, and yellow, respectively. The middle set of color bars in each panel mark samples that were misclassified (black) by various algorithms. The heat-maps indicate relative expression levels of genes (green to red intensities) identified by gene symbols on the right; for cDNA clones without gene symbols, probeset identifiers are displayed instead. Dendrograms are from clustering of standardized transcript levels (rows) using 1-correlation distance and average linkage. Bottom scale denotes standardized values for the green to red intensities in the heat-map. The transcript sets in panels (a)–(c) gave results marked by arrows in Figure 3a–c, respectively.

discriminant analysis algorithm at a *P*-value cutoff level of 10^{-3} and a fold-change threshold of >1.9, which selected for eight probesets to be in the classifier was obtained (Figure 3b). For classification of febrile patients without- *vs* with adenovirus infection, an optimal percent-correct of 91% using the support vector machine algorithm at a *P*-value cutoff level of 10^{-5} and a fold-change threshold of >1.7, which selected for 11 probesets to be in the classifier was obtained (Figure 3c).

The samples that were misclassified by various algorithms and the associated gene expression profiles for the selected transcript set are shown in Figure 4. For node 1, no individuals were misclassified in the febrile with adenovirus group and misclassified samples tended to belong to the febrile without adenovirus or the convalescent group. For node 2, the misclassified samples seemed to be equally distributed between

healthy and convalescent, while for node 3, the misclassified samples tended to be in the febrile without adenovirus group. One observes that some samples were misclassified regardless of algorithm.

The estimated optimal percent-correct classification of nonfebrile *vs* febrile, healthy *vs* convalescents, and febrile without *vs* with adenovirus infection patients were 99, 87, and 91%, respectively. To determine the reliability of these percentages, the permutation test was performed with 2000 permutations. This resulted in *P*-values of <0.0005, 0.001, and <0.0005, respectively.

Functions of genes in the classifier sets

The identifiers of the discovered transcript sets for the class prediction results are shown in Figure 4. The 47 probesets used to classify fever status (Figure 4a) represent 40 transcripts. These included many that are induced by interferon, including: *IFI27*, *IFI44*, *IFI35*, *IFRG28*, *IFIT1*, *IFIT4*, *OAS1*, *OAS2*, *GBP1*, *CASP5*, *MX1*, and *G1P2*. Furthermore, *OAS1* and *OAS2* catalyze 2′,5′ oligomers of adenosine to activate RNaseL and inhibit cellular protein synthesis, while *MX1* is a member of the GTPase family. *OAS1*, *OAS2*, and *MX1* have been shown to have antiviral functions, and interestingly, have also been found to be activated shortly after infection of nonhuman primates with high titers of smallpox.[3] Transcripts involved in the complement cascade, *C1QG* which is downstream of antibody/antigen complexes and *SERPING1* which inhibits activation of the first component of complement, were associated with fever. The *TNF*-alpha and *IL-1* induced gene, *TNFAIP6*, which is a secretory protein involved in extracellular matrix stability and cell migration, and *STK3* and *CASP5*, which are involved in the *MAPK* signaling pathway and are downstream of the *TNF* and *IL1* receptors were identified as class predictors. *FCGR1A*, which functions in the adaptive immune response and binds IgG, was part of the classifier. Other transcripts with associated known functions less clearly related to FRI or with unknown functions were also identified. Some gene ontology descriptions and, in parenthesis, their ratios of observed to expected number of occurrences were as follows: GTP binding (6), guanyl nucleotide binding (6), response to virus (32), immune response (8), defense response (7), response to pest/pathogen/parasite (6), and response to stress (3).

The eight probeset classifier for distinguishing healthy *vs* convalescent patients mapped to seven transcripts, including *RPl27* and *RPS7* associated with ribosomal structure; *IGHM*, the immunoglobulin heavy constant mu transcript; *LAMA2*, which is involved with cell adhesion, migration, and tissue remodeling; and transcripts related to other functions such as *DAB2*, *KREMEN1*, and *EVA1*.

The 10 transcript classifier for distinguishing febrile without adenovirus *vs* with adenovirus infection included the interleukin-1 receptor accessory protein, *IL1RAP*; two interferon induced genes, *IFI27* and *IFI44*, which were also in the classifier for fever status; and *LGALS3BP*, which is involved in cell–cell and cell–matrix interactions and has been found elevated in individuals infected with the human immunodeficiency virus. Other transcripts with known functions less clearly related to adenoviral FRI or with unknown functions included *ZCCHC2*, *ZSIG11*, *NOP5/NOP58*, *MS4A7*, *LY6E*, and *BTN3A3*.

Discussion

After having rigorously assessed the RNA quality of samples processed with PAX tubes in a relatively large sample of humans with differing infection status phenotypes, we characterized and compared the transcriptomes from whole-blood samples of healthy, FRI without and with adenovirus infection, and convalescent individuals, evaluated class prediction methodologies, discovered nested sets of transcripts that could optimally classify the infection status phenotypes and have begun to implicate pathways and gene functions involved in FRI.

We applied a previously reported quality control metric called the degradation factor[7] to our RNA samples and determined that this factor correlates with quality control metrics (*gapdh* 3′/5′ and *actin* 3′/5′) present on the microarray. This degradation factor can easily be applied to microarray studies on large populations by assessing electropherogram data that is available from a bioanalyzer prior to processing microarrays and an indicator can be set to flag poor quality samples. We find that quality metrics typically used, such as the 28S/18S ratio have high variability outside the traditional standard range of 1.8–2.1 and poorly correlate with the quality control metrics present on the microarray.

When assessing signal-to-noise quality metrics, we discovered that MCH significantly affects number of present calls on the B array only, likely due to detection of low expression transcripts on the B array compared to the A array.[10] At the time of probe design, the probes on the A chip were associated with more annotation than those on the B chip. The MCH is a measure of picograms of hemoglobin per red blood cell and likely is directly related to amounts of globin mRNA in whole-blood samples; prior studies have demonstrated that spiking of increasing amounts of globin mRNA transcripts into total RNA from a cell line decreases the percent present calls linearly.[11] This factor would need to be controlled in future microarray studies or globin mRNA would need to be reduced. In the present study, there was no difference of MCH among the infection status phenotypes.

During supervised analysis, we varied the fold-change cutoff threshold in addition to the *P*-value cutoff to optimize percent correct classification. These combined criteria select for transcripts that not only are statistically different between two groups, but also vary above a specific fold-change threshold, reducing transcripts that may represent noise. The accuracy of classification seemed to be resistant to transcript selection parameters and algorithms when the gene-expression profiles showed large consistent differences, such as between nonfebrile *vs* febrile patients; stricter *P*-value and fold change cutoff levels were needed to select informative transcripts that classify the healthy and convalescent or the febrile patients to an accuracy of 87 and 91%, respectively.

Misclassified samples tended to belong to groups more likely to be heterogeneous, suggesting that the misclassification may be due to the lack of specificity of the class

labels. In future studies of larger size, the convalescent group might be further subclassified based on duration of recovery and the febrile without adenovirus group subclassified based on specific pathogen identified. The majority of transcripts in the classifiers shown in Figure 4 remained in the classifier 100% of the time during leave-one-out cross-validation (100% CV support). Thus, these transcripts in the classifiers are consistently different between individuals of two clinical phenotypes at the time when they present for study, as exemplified in Figure 4a. Individuals in the FRI with adenovirus group tend to present later in illness than those without potentially accounting for gene expression differences in the two groups. The correlation of changes in expression of these genes with infection status may also suggest that these genes are involved in the human host fever and immune responses to adenovirus infection *in vivo*. These transcripts consistently showed the largest fold changes between groups, suggesting that the changes in expression were at the pathway level and were unlikely to be accounted for by differences in cell concentration alone. Furthermore, there were no significant differences in cell-type concentration between the febrile without- *vs* with adenovirus groups. This correlation of transcripts to fever and immune responses was derived from *in vivo* natural infections of humans, suggesting the important role of these genes in the host response at the population level. Nested sets of transcripts resulted in similar percent-correct classifications, likely due to the fact that the expression of each transcript is not independent but correlated with other transcripts in related pathways. The discovery of transcripts with functions unrelated to immune response or with unknown functions implies that these should be further studied in infection phenotype model systems to elucidate mechanistic functions.

Our demonstration that one can predict the class of a patient with FRI due to adenovirus infection from background cases of FRI due to other etiologies support the possibility of using gene-expression in biosurveillance and pathogenesis. To our knowledge, this is the first *in vivo* demonstration of classification of infectious diseases via transcriptional signatures of the host. We intend to extend these findings to other respiratory pathogens, both viral and bacterial and to women, to further determine the capability of applying this technology to biodefense and infectious disease surveillance.

Materials and methods

Entry criteria and sample collection

LAFB is the location of Basic Military Training for all recruits to the United States Air Force. The BMTs are organized into flights of 50–60 individuals that eat, sleep, and train in close quarters. As many as 40–50 BMTs/week present with FRI and 50–70% are due to adenovirus. With approval of the LAFB IRB and after informed consent, approximately 15 ml of blood, filling four to five PAX tubes, were drawn from each volunteer. On days 1–3 of training, blood was drawn from healthy BMTs into PAX tubes by standard protocol,[12] but no nasal wash was collected for this group. During training, BMTs who presented with a temperature of 38.1°C or greater and FRI provided a nasal wash and blood draw. These individuals were categorized into either the FRI without adenovirus or with adenovirus group. Approximately 3 weeks after sample collection from the FRI volunteers with adenovirus, additional blood and nasal wash were collected to constitute samples for the convalescent group. All PAX tubes were maintained at room temperature for 2 h, then frozen at −20°C and shipped on dry-ice to the Navy Research Laboratory in Washington, DC for processing. Nasal washes were performed using a standard protocol with 5 ml of normal saline lavage of the nasopharynx followed by collection of the eluent in a sterile container. Nasal wash eluent was stored at 4°C for 1–24 h before being aliquotted and sent for adenoviral culture. All BMTs underwent standardized questionnaires before each sample collection. Healthy individuals were screened again acute medical illness within 4 weeks of arriving at basic training. BMTs were screened for race/ethnicity, allergies, recent injuries, and smoking history to assess confounding variables for gene expression. The duration and type of respiratory symptoms to include sore throat, sinus congestion, cough, fever, chills, nausea, vomiting, diarrhea, fatigue, body aches, runny nose, headache, chest pain and rash were recorded. A physical examination was recorded.

Sample processing

Blood collection and RNA isolation was performed using the PAX System, which consists of an evacuated tube (PAX tube) for blood collection and a processing kit (PAX kit) for isolation of total RNA from whole blood.[13,14] The isolated RNA was amplified, labeled, and interrogated on the HG-U133A and HG-U133B Genechip® microarrays (Affymetrix), noted here as A and B arrays, respectively.

Total RNA isolation from blood

Frozen PAX tubes were thawed at room temperature for 2 h followed by total RNA isolation as described in the PAX kit handbook,[15] but modified to aid in tight pellet formation by increasing proteinase K from 40 to 80 µl (>600 mAU/ml) per sample, extending the 55°C incubation time from 10 to 30 min, and the centrifugation time to 30 min or more. The optional on-column DNase digestion was not carried out. Purified total RNA was stored at −80°C.

Target preparation

For more complete removal of DNA from purified RNA, duplicate RNA samples were pooled, followed by in-solution DNase treatment using the DNA-free™ kit (Ambion). However, to facilitate removal of the DNase inactivating beads, the completed reaction was spun through a spin column (Qiagen, Cat#79523), rather than attempting to pipette off the supernatant without disturbing the bead pellet. Subsequently, 1 µl from each sample was run on the bioanalyzer (Agilent) for assessment of RNA quality and quantity. The usage of the bioanalyzer was analogous to capillary gel electrophoresis. This resulted in electropherograms displaying florescent intensity *vs* time (Figure 1a), which correlates with the amount of RNA *vs* the size of RNA, respectively. Next, 5 µg of RNA were concentrated via ethanol precipitation as previously described.[6] All subsequent steps were as described in the GeneChip Expression Analysis Technical Manual version 701021 Rev. 3.

Database integration

The database consisted of clinical data such as information transcribed from standardized questionnaires, the CBC, and the handling of blood samples. Laboratory data contained information about the processing of samples, from blood in PAX tubes to RNA extraction, as well as subsequent bioanalyzer measurements. Electropherograms were analyzed by the Biosizing (Agilent) software to output 28S/18S intensity ratios and RNA yields, and by the Degradometer 1.1[7] software to consolidate, scale, and calculate degradation and apoptosis factors. Report files summarizing the quality of target detection for an array were generated by GeneChip® Operating Software 1.1 (Affymetrix). JMP (SAS) was used to join these various data tables together into a metadata table with more than a thousand columns. For gene-expression data, signal values were calculated using the Microarray Suite 5.0 algorithm with no scaling or normalization. This allows for subsequent testing of various scaling and normalization methods.

Statistical analysis

Statistical quality control and relations among metadata variables were analyzed in JMP. ANOVAs and class prediction of phenotypes using gene-expression data were performed in Arraytools 3.2.0 Beta developed by Richard Simon and Amy Lam (http://linus.nci.nih.gov/BRB-ArrayTools.html). Heat-maps and dendrograms were graphed using dChip.[16,17] Analysis of gene functions was aided by Arraytools and EASE.[18] Data analysis was performed primarily by DT.

Scaling was carried out for gene-expression data. For each blood sample, the same hybridization cocktail went onto the A and then the B array, allowing concatenation of the data from the two arrays to form a virtual array. This bypassed issues with analyzing the two data sets separately. The 100 control probesets common between the A and B arrays were selected based on stability in expression from a large study of various tissue types.[10] Thus, all array data were scaled to a target value of 500 using the trimmed mean of the 100 control probesets. This resulted in stable Scale Factors (SF) over time and no differences in SF among the infection status phenotypes (ANOVA, $P = 0.1047$ A arrays, $P = 0.1782$ B arrays). This scaling method allowed for the concatenation of corresponding A and B arrays and should also remove variations that are not gene-specific.

Acknowledgements

We thank the study participants and F Ligler and J Golden for reviewing the manuscript. This work was supported in part by the Defense Threat Reduction Agency, HQ USAF Surgeon General's Office, Office of Naval Research, and the Naval Research Laboratory. The opinions and assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the Department of Defense. The EOS Consortium is an Air Force Medical Service initiative comprised of: *Sponsorship:* P Demitry[1], T Difato[1]; *Executive Board:* E Hanson[4], R Holliday[2], R Rowley[4], C Tibbetts[4]; *Operational Board:* D Stenger[10], E Walter[5], J Diao[2]; *Technical Advisors & Collaborators:* R Kruzelock[6], B Agan[10], L Daum[11], D Metzgar[12], D Niemeyer[11], K Russell[12]; *Research & Clinical Staff:* M Archer[9], R Bravo[3], N Freed[12], J Fuller[12], J Gomez[3], K Gratwick[12], M Jenkins[10], M Jesse[3], B Johnson[3], E Lawrence[3], B Lin[8], C Meador[9], H Melgarejo[3], K Mueller[9], C Olsen[2], D Pearson[3], A Purkayastha[2], J Santiago[3], D Seto[7], F Stotler[3], D Thach[8], J Thornton[9], Z Wang[8], D Watson[3], S Worthy[3], G Vora[8]; *Operations Support Staff:* K Grant[2], C James[2]. Affiliations: Dept. of [1]USAF/SGR, [2]USAF/SGR (Ctr), [3]Lackland AFB, [4]George Washington University, [5]Texas A&M University Systems, [6]Virginia Tech, [7]George Mason University, [8]Naval Research Laboratory, [9]NOVA Research Incorporated, [10]Wilford Hall Medical Ctr, [11]Air Force Institute for Operational Health, [12]Navy Health Research Ctr.

References

1 Bullinger L, Dohner K, Bair E *et al.* Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med* 2004; **350**: 1605–1616.
2 Valk PJ, Verhaak RG, Beijen MA *et al.* Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med* 2004; **350**: 1617–1628.
3 Rubins KH, Hensley LE, Jahrling PB *et al.* The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model. *Proc Natl Acad Sci USA* 2004; **101**: 15190–15195.
4 Barraza EM, Ludwig SL, Gaydos JC, Brundage JF. Reemergence of adenovirus type 4 acute respiratory disease in military trainees: report of an outbreak during a lapse in vaccination. *J Infect Dis* 1999; **179**: 1531–1533.
5 Gray GC, Goswami PR, Malasig MD *et al.* Adult adenovirus infections: loss of orphaned vaccines precipitates military respiratory disease epidemics. For the Adenovirus Surveillance Group. *Clin Infect Dis* 2000; **31**: 663–670.
6 Thach DC, Lin B, Walter E *et al.* Assessment of two methods for handling blood in collection tubes with RNA stabilizing agent for surveillance of gene expression profiles with high density microarrays. *J Immunol Methods* 2003; **283**: 269–279.
7 Auer H, Lyianarachchi S, Newsom D *et al.* Chipping away at the chip bias: RNA degradation in microarray analysis. *Nat Genet* 2003; **35**: 292–293.
8 Affymetrix. GeneChip® expression analysis data analysis fundamentals. Part No. 701190 Rev. 4. p 39. Available at http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf. Accessed Sept. 2004, 2004.
9 Golub TR. Toward a functional taxonomy of cancer. *Cancer Cell* 2004; **6**: 107–108.
10 Affymetrix. Performance and validation of the GeneChip® human genome U133 set. Available at http://www.affymetrix.com/support/technical/technotes/hgu133_performance_technote.pdf. Accessed March 2005, 2002.
11 Affymetrix. Globin reduction protocol: a method for processing whole blood rna samples for improved array results. Available at http://www.affymetrix.com/support/technical/technotes/blood2_technote.pdf. Accessed March 2005, 2003.
12 Preanalytix. Product circular. PAXgene Blood RNA tube. Available at http://www.preanalytix.com/pdf/prodcir.pdf. Accessed Sept. 2004.
13 Jurgensen S, Schram J, Herdman C, Rainen L, Wyrich R, Oelmueller U. Effect of blood collection and storage conditions on gene expression analysis. Available at http://www.preanalytix.com/pdf/SJAMPposterNov01.pdf. Accessed April 2003.
14 Jurgensen S, Schram J, Herdman C, Rainen L, Wyrich R, Oelmueller U. New technology to stabilize cellular RNA in blood. Available at http://www.preanalytix.com/pdf/AMP2000PosterSJ.pdf. Accessed April 2003.

15 Preanalytix. PAXgene blood RNA kit handbook. Available at http://www.preanalytix.com/pdf/RNA_handbook.pdf. Accessed Sept 2004.
16 Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2001; **2** RESEARCH0032.
17 Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 2001; **98**: 31–36.
18 Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003; **4**: R70.