# Quantitative assessment of human whole blood RNA as a potential biomarker for infectious disease<sup>†</sup>

Claire L. Smith,<sup>*ab*</sup> Paul Dickinson,<sup>*a*</sup> Thorsten Forster,<sup>*a*</sup> Mizanur Khondoker,<sup>*a*</sup> Marie Craigon,<sup>*a*</sup> Alan Ross,<sup>*a*</sup> Petter Storm,<sup>*a*</sup> Stewart Burgess,<sup>*a*</sup> Paul Lacaze,<sup>*a*</sup> Benjamin J. Stenson<sup>*b*</sup> and Peter Ghazal<sup>\**a*</sup>

Received 10th May 2007, Accepted 15th October 2007 First published as an Advance Article on the web 31st October 2007 DOI: 10.1039/b707122c

Infection remains a significant cause of morbidity and mortality especially in newborn infants. Analytical methods for diagnosing infection are severely limited in terms of sensitivity and specificity and require relatively large samples. It is proposed that stringent regulation of the human transcriptome affords a new molecular diagnostic approach based on measuring a highly specific systemic inflammatory response to infection, detectable at the RNA level. This proposition raises a number of as yet poorly characterised technical and biological variation issues that urgently need to be addressed. Here we report a quantitative assessment of methodological approaches for processing and extraction of RNA from small samples of infant whole blood and applying analysis of variation from biochip measurements. On the basis of testing and selection from a battery of assays we show that sufficient high quality RNA for analysis using multiplex array technology can be obtained from small neonatal samples. These findings formed the basis of implementing a set of robust clinical and experimental standard operating procedures for whole blood RNA samples from 58 infants. Modelling and analysis of variation between samples revealed significant sources of variation from the point of sample collection to processing and signal generation. These experiments further permitted power calculations to be run indicating the tractability and requirements of using changes in RNA expression profiles to detect different states between patient groups. Overall the results of our investigation provide an essential first step toward facilitating an alternative way for diagnosing infection from very small neonatal blood samples, providing methods and requirements for future chip-based studies.

# Introduction

Infection is an important source of morbidity and mortality in neonates and infants. In the developed world, 65% of extremely low birth weight infants develop presumed sepsis in the neonatal period.<sup>1</sup> With mortality rates of 10-50%, a four-fold increase in cerebral palsy and increased risk of hearing, growth and neuro-developmental impairments, the costs are great.<sup>1</sup> On a global perspective, infection accounts for more than half of all deaths worldwide of children younger than age  $5.^2$  During the first year of life, the developing human immune system encounters many challenges from both infections and vaccinations. Systemic deficiencies of both innate and adaptive immunity are thought to contribute to impaired neonatal host defences while protection through maternal antibodies, which is deficient in preterm babies, wanes after approximately six months in term infants. Early diagnosis of infection is key to providing timely and appropriate treatment. Blood is the primary source of clinical

material available and when one considers that some patients may have no more than 40 ml total blood volume and that current procedures often withdraw several millilitres for various blood tests then this sets a stringent ethical and research challenge of working with extremely small quantities of blood. It is noteworthy that the standard diagnostic tool for infection is the blood culture, but this does not give a rapid result (up to 48 h), has poor sensitivity (50–80% at best but often considerably less<sup>3</sup>) and requires blood volumes that represent a significant proportion of an infant's circulating blood volume.

Many signaling molecules and concurrent biological pathways responsible for the initiation and propagation of an inflammatory response to infection have been identified in circulating serum – constituting what has been termed 'Systemic Inflammatory Response'. Individual molecular components identified as part of the systemic inflammatory response have been helpful toward understanding the underlying physiology of inflammation and have also shown potential diagnostic and therapeutic value, including C-reactive protein (CRP) and tumor necrosis factor- $\alpha$ (TNF- $\alpha$ ).<sup>4</sup> From early developmental stages onward, the various activities of circulating immune cells contribute to local as well as systemic levels of cytokines and inflammatory molecules. In this context, blood serves as an integrative tissue whereby its cells and associated signaling and cytokine

<sup>&</sup>lt;sup>a</sup>Division of Pathway Medicine, The University of Edinburgh, The Chancellor's Building, 49 Little France Crescent, Edinburgh, Midlothian, UK EH16 4SB. E-mail: p.ghazal@ed.ac.uk <sup>b</sup>Neonatal Unit, Simpson Centre for Reproductive Health, The Royal Infirmary of Edinburgh, 51 Little France Crescent, Edinburgh, Midlothian, UK EH16 4SA

<sup>†</sup> This paper was published as part of the special issue on the Rapid Diagnostic Testing of Infection and Disease States.

networks relay or enhance the contribution made by sites of infection or tissue damage, to effect protection or cell injury repair responses.

Infections not only elicit but also modify, in a pathogenspecific manner, immune inflammatory responses. This occurs at two levels, the infected cell and the systemic host response levels. At the cell level, analysis of a wide range of studies looking at infection in tissue-culture experiments has defined differential and common host transcriptional responses.<sup>5</sup> This is due in part to the recognition of pathogen-associated molecular patterns by an array of cellular receptors to specific pathogen products. This is especially true for professional antigen presenting cells which orchestrate the selective and appropriate protective immunity. Accordingly, the nature of the systemic inflammatory response generated in vivo to an infective agent will also vary depending upon the specific pathogen resulting in both qualitative and quantitative aspects of the immune response being markedly influenced by the various countermeasures enacted by the pathogen. The systemic responses can be seen in altered cytokine levels, specific lymphocyte responses and can also be detected by alterations in the host RNA phenotype in response to infection.<sup>6-10</sup> Microarrays have proven themselves to be useful means for global analysis of gene or protein content and expression. Studies of variation in gene expression among individuals has revealed a surprising consistency, but also evidence of distinct patterns of inter-individual and temporal variation.<sup>11</sup> Microarray technology has shown many potential clinical applications including classification of cancer patients on the basis of disease outcome and prediction of treatment response.<sup>12–15</sup> In recent months there have been increasing numbers of publications reporting microarray experiments using RNA extracted from clinical whole blood samples in adults and children.<sup>11,16-30</sup> In this report we show for the first time that it is feasible to isolate RNA of sufficient quality and quantity from small volumes of whole blood collected from neonates in order to be able to use RNA as a potential early biomarker for infectious disease.

# Experimental

# Ethical consent

Ethical permission was obtained from the local research ethics committee for this study. Written informed consent was taken from the parent(s) in each case.

# Sample collection

Work to determine optimal RNA extraction from neonatal blood was carried out using umbilical cord blood. Umbilical cord blood sampling took place from the cord segment still attached to the placenta after the cord was cut at delivery. The umbilical cord was cleaned with a sterile swab soaked in phosphate buffered saline and the cord cut using sterile scissors. The umbilical vein was then catheterised with a sterile 5-gauge nasogastric tube and blood aspirated into sterile syringe(s). The samples were then injected immediately into sample collection tubes.

For array analysis neonatal whole blood was used. For these samples neonatal blood sampling was performed by trained members of clinical staff. Gloves were worn during the procedure to avoid contamination. The infant's skin was cleaned with an iodine-based solution and then washed with sterile saline and dried with a sterile swab. The needle or cannula was then inserted and a blood sample of approximately 0.5 ml drawn into a syringe. The sample was injected immediately into a PAXgene<sup>TM</sup> blood RNA tube which was then inverted ten times. Samples were taken from needles, newly inserted venous cannulae or newly inserted arterial cannulae. Samples were not taken from heparinised lines. If samples were to be processed the same day they were transferred to the laboratory and incubated at room temperature for a minimum of 2 h. Otherwise samples were put directly into a -20 °C freezer located within the clinical area until they were transported to the laboratory. In each case data were gathered for each infant including the age of the infant and the mode of sampling (needle or cannula).

# Blood collection tube assessment

Five different blood collection media were investigated. Blood was injected into one each of clinical blood tubes containing EDTA, Lithium Heparin and Sodium Citrate. Blood was also injected directly into a PAXgene<sup>TM</sup> blood RNA tube and into a microcentrifuge tube containing TRIZOL<sup>®</sup> LS reagent. All tubes were transferred to the laboratory on ice with the exception of the PAXgene<sup>TM</sup> tube which was transferred at room temperature. With the exception of the PAXgene<sup>TM</sup> tubes was processed immediately using the TRIZOL<sup>®</sup> LS extraction method followed by an on-column cleanup. The PAXgene<sup>TM</sup> tubes were processed using the PAXgene<sup>TM</sup> blood RNA extraction kits after at least 2 h incubation at room temperature.

# RNA extraction assessment

Five different methods for RNA extraction were performed ranging from organic phase extraction to the use of magnetic bead technology. These were TRIZOL<sup>®</sup> LS (Invitrogen<sup>TM</sup> Corporation), QIAamp RNA Blood Mini Kit (QIAgen Ltd.), TRIZOL<sup>®</sup> LS followed by QIAamp RNA Blood Mini Kit, MagaZorb<sup>®</sup> (CorTex Biochem<sup>TM</sup>, Inc., San Leandro, CA) and PAXgene<sup>TM</sup> (PreAnalytiX GmbH).

At the time of sampling for these experiments, blood was injected into PAXgene<sup>TM</sup> and EDTA tubes. RNA extraction was performed using the PAXgene<sup>TM</sup> system and using 0.5 ml of blood from the EDTA for each of the other methods except MagaZorb<sup>®</sup>. As a separate experiment later, 0.5 ml of blood processed in a PAXgene<sup>TM</sup> tube was compared to 0.2 ml of blood collected in a clinical EDTA tube and processed using the MagaZorb<sup>®</sup> magnetic bead extraction (results shown later for the MagaZorb<sup>®</sup> yield in Table 2 multiplied to give the yield equivalent to 0.5 ml).

# TRIZOL<sup>®</sup> LS extraction

0.5 ml of RNase-free water was added to 0.5 ml of blood, then 3 ml of  $TRIZOL^{(R)}$  LS solution was added and repetitive

pipetting was carried out to lyse the cells. RNA extraction was carried out as per TRIZOL<sup>®</sup> LS instructions with the exception of the initial centrifugation step being carried out at 4000 rpm in an Eppendorf 5810R centrifuge for 1 h.

#### On-column cleanup following TRIZOL® LS RNA extraction

100  $\mu$ l samples obtained from the TRIZOL<sup>®</sup> LS reaction were carried into the first step of the QIAamp RNA Mini Protocol for RNA Cleanup and this protocol was followed to the end. The optional on-column DNase step, second centrifugation step and repeated elution step (to give a final elution volume of 100  $\mu$ l) were incorporated.

# **QIAamp RNA extraction**

The QIAamp RNA Mini Protocol for Isolation of Total Cellular RNA from Whole Human Blood was followed with the following variations: the blood/Buffer EL mix was incubated on ice for 20 min; after centrifugation and removal of the supernatant the pellet was incubated on ice for a further 10 min; and after further addition of Buffer EL, 600  $\mu$ l (rather than 350  $\mu$ l) of buffer RLT was added to the sample. The optional on-column DNase step and second centrifugation step after the addition of buffer RPE were incorporated. The sample was eluted in 100  $\mu$ l of RNase-free water.

# PAXgene<sup>TM</sup> blood RNA extraction

Each of the PAXgene<sup>TM</sup> tubes was processed according to the PAXgene<sup>TM</sup> blood RNA protocol from PreAnalytix dated April 2001. Variations from the protocol were: the incubator steps were carried out in a water bath rather than a shaker–incubator (in step 5 the samples were vortexed once during the incubation), the centrifugation step was increased to 10 min, the optional on-column DNase step and the 1 min drying centrifugation were incorporated.

# MagaZorb<sup>®</sup> extraction

RNA extraction was carried using 200  $\mu$ l of whole umbilical cord blood according to the MagaZorb<sup>®</sup> RNA Purification Protocol (CorTex Biochem<sup>TM</sup>) with Supplementary Protocol B (DNase protocol) incorporated.

#### Quantification and quality assessment of RNA

RNA was quantified and an  $A_{260}$ :  $A_{280}$  ratio calculated for each sample after analysis on a ThermoSpectronic Biomate 5 v1.6 spectrophotometer. RNA quality was assessed running each sample on an Agilent 2100 Bioanalyser using an RNA LabChip kit. RNA quality was assessed qualitatively by looking at the electropherogram of each sample, and quantitatively by means of the RNA Integrity Number (RIN).

#### Microarray processing and analysis

The CodeLink<sup>TM</sup> Human Whole Genome Bioarray was comprised of approximately 55 000 30-mer probes designed to conserve exons across the transcripts of targeted genes. These 55 000 probes represent well-annotated, full length, and partial human gene sequences from major public databases.

The biotin-labelled cRNA target is prepared by a linear amplification method using tailed oligo dT priming of total RNA. After second-strand cDNA synthesis, the cDNA undergoes an in vitro transcription (IVT) reaction to produce the target cRNA. This method produces approximately 1000fold to 5000-fold linear amplification. Various quality control procedures are incorporated. Hybridisation is performed overnight and post-hybridisation processing includes a stringent wash to remove unbound and non-specifically hybridised target molecules and staining with Cy<sup>TM</sup>5-streptavidin conjugate. Several non-stringent washes remove unbound conjugate. The bioarrays are then dried and scanned on the Agilent G2567A scanner at 5 µm resolution. Raw data were obtained from the scanned images using CodeLink<sup>TM</sup> EXPv4.1 (GE Healthcare) feature extraction software. Subsequent data validation comprising data quality control and normalisation involved the use of the statistical software package R (v 2.2.1) and Bioconductor modules for R (v 1.7). Microarray data have been deposited in the GPX MIAME compliant database at http://www.pathwaymedicine.ed.ac.uk/ gpx (Accession number: GPX000071 will be made available upon publication).

#### Statistical methodology and analysis

For comparison of blood collection methods, sample storage and RNA extraction procedures a two-tailed paired Student's *t*-test was employed. For microarray data analysis, a simple (per-gene) analysis of variance model was employed. To define the model, suppose that the median spot intensity  $X_{g(i,j,k,l,m,n)}$ corresponds to the *g*th gene, *i*th operator, *j*th way of taking the sample, *k*th freezing status, *l*th level of time to extraction, *m*th category of age, and *n*th sample. We assume that the median intensity  $X_{g(i,j,k,l,m,n)}$  consists of a systematic component  $S_{g(i,j,k,l,m,n)}$  and a non-systematic (random error) component  $R_{g(i,j,k,l,m,n)}$  and are related by a multiplicative relationship. That is,

$$X_{g(i,j,k,l,m,n)} = S_{g(i,j,k,l,m)} \times R_{g(i,j,k,l,m,n)},$$
(1)

where the random component  $R_{g(i,j,k,l,m,n)}$  is distributed as lognormal with scale parameter 1 [or,  $\log_e(1) = 0$ ] and shape parameter  $\sigma_{g(i,j,k,l,m)}$ . Therefore,

$$R_{g(i,j,k,l,m,n)} \sim LN(0,\sigma^2_{g(i,j,k,l,m)}).$$
 (2)

A random variable *X* has the distribution  $LN(\mu, \sigma^2)$  if  $Y = \log_e(X)$  has the normal distribution  $N(\mu, \sigma^2)$ . Therefore, fitting of a log-normal distribution, in principle, reduces to fitting of a normal distribution. Now with the transformations

$$S_{g(i,j,k,l,m)} = \exp(\eta_{g(i,j,k,l,m)})$$
, and  $R_{g(i,j,k,l,m,n)} = \exp(\varepsilon_{g(i,j,k,l,m,n)})$ ,

model corresponding to eqn (1) for log-intensity  $Y_{g(i,j,k,l,m,n)} = \log_e(X_{g(i,j,k,l,m,n)})$  can be written as

$$Y_{g(i,j,k,l,m,n)} = \eta_{g(i,j,k,l,m)} + \varepsilon_{g(i,j,k,l,m,n)},$$
(3)

where  $\eta_{g(i,j,k,l,m)}$  is the log-systematic component and  $\varepsilon_{g(i,j,k,l,m,n)}$  is the log-error. Assuming that the log-systematic component

 $\eta_{g(i,j,k,l,m)}$  is as an additive function of the effects of the systematic sources of variation,

$$\eta_{g(i,j,k,l,m,n)} = \mu_g + \alpha_{g(i)} + \beta_{g(j)} + \gamma_{g(k)} + \delta_{g(l)} + \tau_{g(m)}, \quad (4)$$

the per-gene analysis of variance (ANOVA) model for logintensity  $Y_{g(i,j,k,l,m,n)}$  is expressed as,

$$Y_{g(i,j,k,l,m,n)} = \mu_g + \alpha_{g(i)} + \beta_{g(j)} + \gamma_{g(k)} + \delta_{g(l)} + \tau_{g(m)} + \varepsilon_{g(i,j,k,l,m,n)},$$
(5)

where  $\mu_g$  is the overall log-expression of gene g,  $\alpha_{g(i)}$  the effect of the *i*th operator,  $\beta_{g(j)}$  the effect of the *j*th way of taking blood,  $\gamma_{g(k)}$  the effect of the *k*th freezing status,  $\delta_{g(l)}$  the effect of the *l*th level of time to extraction,  $\tau_{g(m)}$  the effect of the *m*th category of age, and  $\varepsilon_{g(i,j,k,l,m,n)}$  is the corresponding log-error term distributed as

$$\varepsilon_{g(i,j,k,l,m,n)} \sim N(0,\sigma^2_{g(i,j,k,l,m,n)}).$$
(6)

The parameters of model (5) can be estimated using a leastsquares method by minimising the error sum of squares given by,

$$SSE_{g} = \sum_{i,j,k,l,m,n} \varepsilon_{g(i,j,k,l,m,n)}^{2} = \sum_{i,j,k,l,m,n} (Y_{g(i,j,k,l,m,n)} - \left\{ \mu_{g} + \alpha_{g(i)} + \beta_{g(j)} + \gamma_{g(k)} + \delta_{g(l)} + \tau_{g(m)} \right\} \right)^{2}.$$
(7)

This is done by partially differentiating  $SSE_g$  with respect to each of the parameters, and then solving the resulting equations by setting them equal to zero. Details of the methods have previously been described<sup>31</sup> and are implemented in all standard statistical analysis software. We used the R program aov to fit the model (5). If the normality assumption of the log-errors (6) is true, then the least-squares estimates of the model (5) are equivalent to the maximum likelihood estimates. We employed a factorial design to analyse the data. There are five factors with a number of levels resulting in 144 factor combinations. The experiment was difficult to implement as balanced for such multi-factor analysis and

Table 1 Comparison of blood collection tubes

therefore have cell frequencies ranging from 0 to 5. The reason for this primarily relates to lack of control on the selection of infants, ages and clinical procedures performed. Thus, while interpretation of results obtained from analysis of variance of unbalanced factorials may sometimes be less precise, the current analysis should reflect the overall behaviors of data in terms of the multiple factors, and provide guidelines for future designs and analyses using larger replication studies.

#### **Results and discussion**

#### Sample collection and processing

Sample collection was performed using a closely defined set of standard operating procedures as in the Experimental section. Tables 1 and 2 show results of the investigations into finding the optimal blood collection tube and extraction method (using umbilical cord blood as a surrogate for neonatal blood) comparing five different commercial RNA extraction procedures. The PAXgene<sup>TM</sup> blood RNA system consistently gave the best quality RNA while yielding sufficient quantity for microarray analysis. Using standardised protocols for sample collection and processing following identification of optimal methods means that data are as robust and reproducible as possible. Importantly our subsequent studies using neonatal samples have shown that it is possible to obtain sufficient RNA of consistently high quality from neonatal whole blood samples of 0.5 ml. For the 58 samples examined in this report, the mean yield of RNA was 8.45 µg (range 1.84-43.82), the mean RIN value was 9.1 (range 6.6–10) and the mean  $A_{260}$ /  $A_{280}$  ratio was 1.86 (range 1.24–2.47). We also show in Table 3 that storage of these PAXgene<sup>TM</sup> blood samples at -20 °C for 7 days prior to RNA extraction leads to no loss in quality of RNA.

#### Chip-based measurements – contending with systematic and nonsystematic variation

Next we implemented the defined clinical and RNA extraction standard operating procedures for collection and processing of

Blood tube	п	Mean (range) RNA per 0.5 ml blood/µg	Mean (range) $A_{260 : 280}$	Mean (range) RIN values	RIN <i>p</i> -value <i>cf</i> . PAXgene <sup>TM</sup>
EDTA	8	9.80 (3.51-24.91)	1.73 (1.01-2.10)	5.0 (1.0-7.6)	0.002
Li heparin	7	14.33 (4.35–30.60)	1.99 (1.79–2.17)	7.9 (7.1-8.9)	0.0004
Na citrate	8	10.20 (2.34–20.23)	1.85 (1.62–2.0)	5.5 (1.1-8.3)	0.01
TRIZOL <sup>®</sup> LS	9	18.21 (5.18–39.96)	1.86 (1.63-2.04)	8.2 (6.8–9.2)	0.01
PAXgene <sup>TM</sup>	8	8.16 (3.33–13.17)	2.03 (1.93–2.2)	9.5 (8.3–10)	

 Table 2
 Comparison of RNA extraction methods

RNA extraction method	п	Mean (range) RNA per 0.5 ml blood/µg	Mean (range) $A_{260 : 280}$	Mean (range) RIN values	RIN <i>p</i> -value <i>cf</i> . PAXgene <sup>TM</sup>
MagaZorb <sup>®</sup>	7	15.92 (9.98-21.85)	1.71 (1.53–1.85)	1.9 (1.0-2.9)	0.008
QIAamp	7	2.91 (1.34-4.18)	2.77 (0-7)	4.7 (1.0-8.2)	0.1
TRIZOL <sup>®</sup> LS	7	26.35 (11.88-44.48)	1.80 (1.65–1.92)	2.2(0.0-6.4)	0.00009
TRIZOL <sup>®</sup> LS and QIAamp	7	16.29 (9.03-25.41)	1.91 (1.5–2.02)	6.0 (0.0-7.8)	0.2
PAXgene <sup>TM</sup>	7	9.25 (4.41–16.66)	1.98 (1.95–2.01)	7.5 (5.5–8.9)	

Table 3         Comparison of storage conditions						
Sample storage	n	Mean (range) RNA/µg	Mean (range) A <sub>260 : 280</sub>	Mean (range) RIN values	RIN <i>p</i> -value	
Not frozen Frozen at -20 °C for 7 days	6 6	4.16 (1.72–5.81) 5.61 (3.56–7.92)	1.86 (1.55–2.06) 1.81 (1.64–2.00)	7.8 (1.4–9.6) 9.4 (8.4–9.9)	0.29	

58 neonatal whole blood samples. Whole blood RNA samples are comprised of RNA from a range of blood cells including reticulocytes, and the presence of high levels of globin mRNA from these cells has led to the use of globin reduction protocols for whole blood samples.<sup>29</sup> We evaluated the use of a globin reduction protocol for our samples (data not shown) but observed limited improvement on sensitivity and specificity of chips. For this reason we processed our samples directly without globin reduction. Gene expression data were determined for the 58 neonatal whole blood samples and analysed for the purpose of investigating the sources and magnitude of systematic and non-systematic (random) variation and to explore an appropriate error model for such data. Specifically, five sources of variation were examined: age of patient (<5 days, 5–10 days or >10 days), blood sample collection method (needle/cannula), freezing status of sample (frozen/not frozen), time to extraction (<3 h, 3-4 h or >4 h), and technical operator (one of four operators). The data were comprised of 58 samples, corresponding to RNA extracted from neonatal whole blood of 58 infants and these samples were analysed using CodeLink<sup>TM</sup> Human Whole Genome Bioarrays (GE Healthcare), providing expression profiling of ca. 55 000 human gene targets in a single array. The first step of our analysis was to look for any significant operator variation in the data. We employed a simple (per-gene) analysis of variance model as outlined in the Experimental section [eqn (5)]. Results of applying the model to the first row of data (g = 1) are summarised in Table 4 and show that the factor 'time to extraction' has the highest level of variation, but none of these sources of variation are statistically significant. A summary of these results is graphically presented in Fig. 1. The top panel, Fig. 1A shows the number of genes with a mean squared error (MSE) greater than or equal to a certain level plotted against that level of variation. The bottom panel, Fig. 1B shows the number of genes having a significance level less than or equal to a certain value plotted against the corresponding level of p-values. Mean squared error (MSE) due to all the systematic sources of variation and error, and the *p*-values for the significance of systematic sources of variation, are computed on the basis of the per-gene ANOVA model (5). These analyses show that a significant source of variation can be attributed to the

**Table 4** Results of applying the analysis of variance model to the first row of data (g = 1)

Source of variation	Mean squared error (MSE)	<i>p</i> -Value	
Operator	0.42	0.48	
Blood draw method	0.19	0.55	
Freezing status	0.78	0.22	
Time to extraction	1.21	0.10	
Age	0.21	0.67	
Random error	0.51		

'operator'. Another significant source appears to be the factor 'blood draw method'. That is whether a needle or a cannula was used for taking blood. We see that up to a certain level of variation (MSE = 1.4) the number of genes exceeding a certain threshold of operator variation is higher than that exceeding the same level of variation due to other sources. However, after the level 1.4, 'blood draw method' takes over the 'operator'. Testing if other confounding variables also contribute to these particular variations will require further investigation. Nevertheless, the proportion of genes showing significant variation at p = 0.05 for all five sources of variation considered in this study are shown in Table 5.

# Error models and power calculations required for RNA biomarker identification

We assumed in our basic model (1) that the error associated with the untransformed spot intensity  $X_{g(i,j,k,l,m,n)}$  is  $R_{g(i,j,k,l,m,n)}$ and follows a log-normal distribution,  $R_{g(i,j,k,l,m,n)} \sim$  $LN(0,\sigma^2_{g(i,j,k,l,m,n)})$ . Equivalently, the distribution of log-error, which is the error associated with the log-spot-intensity  $Y_{g(i,j,k,l,m,n)} = \log_e(X_{g(i,j,k,l,m,n)})$  is denoted by  $\varepsilon_{g(i,j,k,l,m,n)}$ , and follows a normal distribution

$$\varepsilon_{g(i,j,k,l,m,n)} \sim N(0,\sigma^2_{g(i,j,k,l,m)}).$$

So, fitting a log-normal distribution to errors  $R_{g(i,j,k,l,m,n)}$  is equivalent to fitting a normal distribution to log-errors  $\varepsilon_{g(i,j,k,l,m,n)}$ . The second option is more commonly used and convenient in many ways. Our per-gene ANOVA model (5) actually fits a normal distribution to the log-errors. Therefore, if log-normal is an appropriate choice for errors  $R_{g(i,j,k,l,m,n)}$ , residuals from the fitted model (5), given by,

$$\begin{split} e_{g(i,j,k,l,m,n)} &= Y_{g(i,j,k,l,m,n)} - \\ \left\{ \hat{\mu}_g + \hat{\alpha}_{g(i)} + \hat{\beta}_{g(j)} + \hat{\gamma}_{g(k)} + \hat{\delta}_{g(l)} + \hat{\tau}_{g(m)} \right\} \end{split}$$

should be approximately normally distributed. Empirical distributions and the corresponding fitted normal distributions to the residuals (log-errors) corresponding to ten randomly selected genes are plotted in Fig. 2. It is seen that except for the minor multimodal features of the empirical distributions, normal distributions provide a reasonable fit to the data. General conclusion from Fig. 2 may be that model (5) with normally distributed log-errors provides a reasonable fit to the data. It would, however, be worth investigating if the multimodality as seen in Fig. 2A may be captured with a more appropriate distribution.

These analyses support the possibility of performing microarray experiments on neonatal whole blood and raise the question of whether or not patient group variance is amenable to profiling RNA biomarkers. Fig. 3A shows the



Fig. 1 Investigation of systematic and non-systematic variation in microarray data generated from neonatal whole blood samples. The number of genes having mean squared error (MSE) greater than or equal to a certain level plotted against the level of variation (A), and number of genes having a significance level less than or equal to a certain value plotted against the level of p-values (B). Mean squared error (MSE) due to the systematic sources and random error, and the p-values for the significance of systematic variations are computed on the basis of the ANOVA model (5).

**Table 5**Proportion of genes having significant variation correspond-<br/>ing to all five systematic sources considered in this study (p = 0.05)

	Proportion of genes showing significant variation at $p = 0.05$						
Sources	Operator	Needle/ cannula	Freezing	Time to extraction	Age		
Proportion	0.89	0.38	0.05	0.01	0.25		

level of variation seen in the expression data generated from the preliminary clinical data. Plots of coefficient of variation of patient samples show acceptable levels of variation. Some exploratory plots of the data are shown in Fig. 3A. If we ignore the outliers in the mean *vs.* CV plot a non-linear trend can be seen in the mean–CV relationship. Determining sample size per experimental condition for a given level of confidence in inferring differential expressions is an important issue, and needs to be decided as an essential first step. The multiple number of RNA markers estimated for each sample makes it difficult to apply traditional sample size calculation techniques and has left most practitioners to rely on rule-of-thumb techniques. A method for computing the sample size for microarray experiments for a given pre-determined level of confidence (power) in inferring differential expressions has been described.<sup>32</sup> The method is based on the assumption that the microarray is set up to compare gene expressions between one treatment group and one experimental group. It is further assumed that the data have been normalised and transformed so that the data for each gene are sufficiently close to a normal distribution so that a standard two-sample pooled variance t-test will reliably detect differentially expressed genes. Here we compute the sample size separately for each gene according to the standard formula for the



Fig. 2 Empirical distributions (A) and the corresponding fitted normal distributions (B) to the residuals (log-errors) corresponding to ten randomly selected genes.

two-sample *t*-test:

$$1 - \beta = 1 - T_{n_1 + n_2 - 2} \left( t_{\alpha_G/2, n_1 + n_2 - 2} | \frac{\Delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) + T_{n_1 + n_2 - 2} \left( -t_{\alpha_G/2, n_1 + n_2 - 2} | \frac{\Delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)$$

where  $T_d(\bullet/\theta)$  is the cumulative distribution function for noncentral *t*-distribution with *d* degrees of freedom and the noncentrality parameter  $\theta$ .

We apply the above method to calculate the sample size to achieve 90% power on the basis of the standard deviations computed from the quantile normalised data of 58 control samples of the neonatal whole blood gene expression study. The results are plotted in Fig. 3B representing the sample size required to achieve 90% power for a given proportion of genes on the arrays. Power calculations and sample size estimates based on n = 58 samples with the same type of sample and on the same microarray platform suggest that 100 samples per group will be required to detect two-fold differential expression with 90% power for at least 90% of the probes on the array, at a significance level of  $\alpha = 0.001$  (corrected for multiple testing by the Bonferroni method<sup>33</sup>). We conclude from these analyses that identifying RNA biomarkers is tractable with a case control group size of 100 patients.

#### Stringency of transcriptome and variation of RNA phenotype

The above studies indicate that while there is a significant level of variation there is nevertheless a relatively stable, wellcorrelated RNA phenotype that can be used to identify a specific set of RNAs as potential biomarkers. Overall, this observation is indicative of homeostatic mechanisms underpinning a stringently regulated genome. In support there are a number of well-documented genetic diseases, mainly of nonprotein coding mutations, that lead to an alteration of very subtle changes (up to two-fold or less) in gene expression in comparison with normal individuals and which result in

CV vs. mean plot 0.6 0.5 Coeff.of variation 0.4 0.3 0.2 0.1 0.0 5 15 Mean(log-expression в Power Calculations (n = 58 arrays) Sample size to obtain 90% power for a given proportion of genes on the array 8 u => 8 Prop. of Genes Needing Sample Size 8 20 8 fold.change=2 alpha=0.001 9 Power=90% 9 10 30 50 55 5 15 20 25 35 40 45 60 65 70 75 80 85 90 95 100 Sample size (n) per group

A Coefficient of variation (CV) plot (n = 58 arrays)

Fig. 3 Scatter plots of inter-patient variance and power calculations.

marked clinical phenotypes.<sup>34–37</sup> These studies indicate that gene expression is tightly regulated and relatively intolerant of dramatic variation. In support, microarray studies of variation in expression among individuals have revealed a surprising level of overall consistency but also evidence of distinct patterns of inter-individual and temporal variation.<sup>11</sup> A recent chip study shows the possibility of even detecting slight alterations in gene expression due to allelic variation.<sup>38</sup> Our recent experience in performing a range of clinical molecular profiling studies from intestinal, mammary, adult blood, endometrial, ovarian and testicular biopsies supports the view for a remarkable stringency in the inter-individual regulation of the transcriptome (see ref. 39 and unpublished observations). To date, neonatal whole blood shows the highest level of variability in our experience. This could be due to greater levels of complexity of procedural, sample handling and age differences as well as a degree of biological variation. Nevertheless, even with these samples our microarray profiling observations show an excellent correlation between individuals, further indicating the human transcriptome to be under stringent homeostatic regulation. Accordingly, even without standardisation, early microarray studies have shown potential

clinical applications including classification of cancer patients on the basis of disease outcome and prediction of treatment response.<sup>12–15</sup>

#### RNA phenotype of systemic host response

One particularly exciting application of microarrays could be investigating infection by detecting alteration in host RNA phenotype in response to infection.<sup>6-9</sup> This would be particularly useful if unique host signatures could identify individual pathogens.<sup>10</sup> Recently, there have been a few small/ pilot studies looking at gene expression profiling in response to infection.<sup>17,19,23,27</sup> Also, there have been an increasing number of publications reporting microarray experiments using RNA from clinical whole blood samples in adults and children.<sup>11,16-30,40</sup> These reports are at this time limited not only in terms of statistical power but also to the gene-analytic approach applied. However, one study has attempted to explore networks of inter-related genes.<sup>41</sup> In this connection it is worth noting that biological pathways provide a central level of physiological organisation and, to date, a pathway-centric approach is markedly absent.

#### Perspective: bio-chip platforms for point-of-care

Biochip platforms based on pathogen detection both at the nucleic acid<sup>42</sup> and protein<sup>43</sup> level are seen as key in the accurate diagnosis of infection. However, the multi-parameter testing of changes in whole blood RNA expression has the potential to use extremely small quantities of material which does not require the presence of the infective agent in the sample. It is possible to envision that micro-devices would have the capacity not only to process and extract RNA but also to detect directly the presence of specific host pathway responses to infection. It is also likely that these pathway responses may also be detected at the protein level based on predictions from the RNA phenotype. Biochips incorporating a combination of both pathogen detection and host response would give the widest possible coverage to detect signatures diagnostic of infection. Nevertheless, regardless of the platform technology used to detect such signatures it is of fundamental importance that a clear understanding of the levels and contributions of systematic and non-systematic variation are fully appreciated. In this report we have shown that the operator and the point of collection can provide a significant source of changes in gene expression. Therefore if RNA is to fulfil a future role as a potential biomarker it is essential that appropriately powered studies are performed which account for known systematic variation and error models developed to account for nonsystematic variation.

# Conclusion

In conclusion we show that multi-parameter testing of changes in RNA expression offers innovative potential and an amenable means for measuring an RNA phenotype using relatively small quantities of whole blood. We provide optimal methods and procedures and attendant clinical and experimental SOPs for applying a rigorous, chip-based investigation. Even with such methods, significant challenges and limitations remain, related to both systematic and non-systematic variations. From an analysis of the RNA phenotypes from 58 samples of neonatal whole blood considered in this study we identify among the systematic sources of variation as potentially contributing variability: laboratory operator, the way the blood is drawn (by needle/cannula), and the age of the infants. Further case studies are required to validate these findings. We show that a linear additive analysis of variance model for log-transformed data with Gaussian-distributed logerrors seems reasonable to adjust the data for systematic variation. Except for the minor multimodal features of the empirical distributions, log-normal distributions for errors or, equivalently, normal distributions for log-errors provide a reasonable fit to the data accounting for non-systematic variation. A sample size of about 100 per-group seems reasonable to achieve 90% power for 90% of markers on the chip.

# Acknowledgements

We would like to acknowledge funding by the INFOBIOMED Network of excellence (European Commission FP6 Programme: contract number IST-2002-507585), Wellcome Trust, Scottish Enterprise, and Chiesi Pharmaceuticals, and help from all of our colleagues in the Division of Pathway Medicine and clinical staff.

# References

- 1 B. J. Stoll, N. I. Hansen, I. Adams-Chapman, A. A. Fanaroff, S. R. Hintz, B. Vohr and R. D. Higgins, *J. Am. Med. Assoc.*, 2004, 292, 2357–2365.
- 2 J. Bryce, C. Boschi-Pinto, K. Shibuya and R. E. Black, *Lancet*, 2005, **365**, 1147–1152.
- 3 J. S. Gerdes, Pediatr. Clin. North Am., 2004, 51(4), 939-959.
- 4 R. Berner, C. M. Niemeyer, J. U. Leititis, A. Funke, C. Schwab, U. Rau, K. Richter, M. S. Tawfeek, A. Clad and M. Brandis, *Pediatr. Res.*, 1998, 44, 469–477.
- 5 R. G. Jenner and R. A. Young, Nat. Rev. Microbiol., 2005, 3, 281–294.
- 6 J. C. Boldrick, A. A. Alizadeh, M. Diehn, S. Dudoit, C. L. Liu, C. E. Belcher, D. Botstein, L. M. Staudt, P. O. Brown and D. A. Relman, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, 99, 972–977.
- 7 C. J. Campbell and P. Ghazal, J. Appl. Microbiol., 2004, 96, 18–23.
- 8 C. A. Cummings and D. A. Relman, *Emerg. Infect. Dis.*, 2000, 6, 513–525.
- 9 D. Yowe, W. J. Cook and J. C. Gutierrez-Ramos, *Microbes Infect.*, 2001, **3**, 813–821.
- 10 Q. Huang, D. Liu, P. Majewski, L. C. Schulte, J. M. Korn, R. A. Young, E. S. Lander and N. Hacohen, *Science*, 2001, 294, 870–875.
- 11 A. R. Whitney, M. Diehn, S. J. Popper, A. A. Alizadeh, J. C. Boldrick, D. A. Relman and P. O. Brown, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 1896–1901.
- A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt, *Nature*, 2000, **403**, 503–511.
- 13 M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts and V. Sondak, *Nature*, 2000, **406**, 536–540.
- 14 T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Science*, 1999, **286**, 531–537.
- 15 C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown and D. Botstein, *Nature*, 2000, **406**, 747–752.
- 16 F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. D. Rosas, S. M. Hersch, P. Hogarth, B. Bouzou, R. V. Jensen and D. Krainc, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 11023–11028.
- 17 R. J. Feezor, C. Oberholzer, H. V. Baker, D. Novick, M. Rubinstein, L. L. Moldawer, J. Pribble, S. Souza, C. A. Dinarello, W. Ertel and A. Oberholzer, *Infect. Immun.*, 2003, **71**, 5803–5813.
- 18 M. Ghielmetti, M. Bellis, M. O. Spycher, S. Miescher and G. Vergeres, *Mol. Immunol.*, 2006, 43, 939–949.
- 19 M. J. Griffiths, M. J. Shafi, S. J. Popper, C. A. Hemingway, M. M. Kortok, A. Wathen, K. A. Rockett, R. Mott, M. Levin, C. R. Newton, K. Marsh, D. A. Relman and D. P. Kwiatkowski, *J. Infect. Dis.*, 2005, **191**, 1599–1611.
- 20 A. D. Hershey, Y. Tang, S. W. Powers, M. A. Kabbouche, D. L. Gilbert, T. A. Glauser and F. R. Sharp, *Headache*, 2004, 44, 994–1004.
- 21 P. A. Horwitz, E. J. Tsai, M. E. Putt, J. M. Gilmore, J. J. Lepore, M. S. Parmacek, A. C. Kao, S. S. Desai, L. R. Goldberg, S. C. Brozena, M. L. Jessup, J. A. Epstein and T. P. Cappola, *Circulation*, 2004, **110**, 3815–3821.
- 22 C. C. Liew, J. Ma, H. C. Tang, R. Zheng and A. A. Dempsey, J. Lab. Clin. Med., 2006, 147, 126–132.

- 23 M. Prucha, A. Ruryk, H. Boriss, E. Moller, R. Zazula, I. Herold, R. A. Claus, K. A. Reinhart, P. Deigner and S. Russwurm, *Shock*, 2004, **22**, 29–33.
- 24 Q. Tan, K. Christensen, L. Christiansen, H. Frederiksen, L. Bathum, J. Dahlgaard and T. A. Kruse, *Hum. Genet.*, 2005, 117, 267–274.
- 25 Y. Tang, T. A. Glauser, D. L. Gilbert, A. D. Hershey, M. D. Privitera, D. M. Ficker, J. P. Szaflarski and F. R. Sharp, *Acta Neurol. Scand.*, 2004, **109**, 159–168.
- 26 Y. Tang, M. B. Schapiro, D. N. Franz, B. J. Patterson, F. J. Hickey, E. K. Schorry, R. J. Hopkin, M. Wylie, T. Narayan, T. A. Glauser, D. L. Gilbert, A. D. Hershey and F. R. Sharp, *Ann. Neurol.*, 2004, 56, 808–814.
- 27 D. C. Thach, B. K. Agan, C. Olsen, J. Diao, B. Lin, J. Gomez, M. Jesse, M. Jenkins, R. Rowley, E. Hanson, C. Tibbetts, D. A. Stenger and E. Walter, *Genes Immun.*, 2005, 6, 588–595.
- 28 Z. Wang, D. Neuburg, C. Li, L. Su, J. Y. Kim, J. C. Chen and D. C. Christiani, *Environ. Health Perspect.*, 2005, 113, 233–241.
- 29 R. J. Feezor, H. V. Baker, M. Mindrinos, D. Hayden, C. L. Tannahill, B. H. Brownstein, A. Fay, S. MacMillan, J. Laramie, W. Xiao, L. L. Moldawer, J. P. Cobb, K. Laudanski, C. L. Miller-Graziano, R. V. Maier, D. Schoenfeld, R. W. Davis and R. G. Tompkins, *Physiol. Genomics*, 2004, **19**, 247–254.
- 30 H. R. Wong, T. P. Shanley, B. Sakthivel, N. Cvijanovich, R. Lin, G. L. Allen, N. J. Thomas, A. Doctor, M. Kalyanaraman, N. M. Tofil, S. Penfil, M. Monaco, M. A. Tagavilla, K. Odoms, K. Dunsmore, M. Barnes and B. J. Aronow, *Physiol. Genomics*, 2007, **30**, 146–155.
- 31 D. C. Montgomery, *Design and Analysis of Experiments*, John Wiley & Sons Inc., New York, 6th edn, 2004.

- 32 G. R. Warnes, Technical Report, University of Rochester, 2006 (http://www.urmc.rochester.edu/smd/biostat/people/faculty/manuscripts/ 0606\_warnes\_liu.pdf).
- 33 C. E. Bonferroni, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936, 8, 3–62.
- 34 D. N. Cooper, Ann. Med., 1992, 24, 427-437.
- 35 E. Di Pierro, M. D. Cappellini, R. Mazzucchelli, V. Moriondo, D. Mologni, B. Zanone Poma and A. Riva, *Exp. Hematol. (N. Y.)*, 2005, **33**, 584–591.
- 36 X. M. Sun, C. Neuwirth, D. P. Wade, B. L. Knight and A. K. Soutar, *Hum. Mol. Genet.*, 1995, **4**, 2125–2129.
- 37 F. Tassone, R. J. Hagerman, W. D. Chamberlain and P. J. Hagerman, Am. J. Med. Genet., 2000, 97, 195-203.
- 38 P. V. Pant, H. Tao, E. J. Beilharz, D. G. Ballinger, D. R. Cox and K. A. Frazer, *Genome Res.*, 2006, 16, 331–339.
- 39 H. O. Critchley, K. A. Robertson, T. Forster, T. A. Henderson, A. R. Williams and P. Ghazal, *Am. J. Obstet. Gynecol.*, 2006, 195(406), e401–416.
- 40 O. Ramilo, W. Allman, W. Chung, A. Mejias, M. Ardura, C. Glaser, K. M. Wittkowski, B. Piqueras, J. Banchereau, A. K. Palucka and D. Chaussabel, *Blood*, 2007, **109**, 2066–2077.
- 41 S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins and S. F. Lowry, *Nature*, 2005, **437**, 1032–1037.
- 42 S. W. Yeung, T. M. Lee, H. Cai and I. M. Hsing, *Nucleic Acids Res.*, 2006, **34**, e118.
- 43 X. Duburcq, C. Olivier, F. Malingue, R. Desmet, A. Bouzidi, F. Zhou, C. Auriault, H. Gras-Masse and O. Melnyk, *Bioconjugate Chem.*, 2004, 15, 307–316.