Gene-Expression Patterns in Whole Blood Identify Subjects at Risk for Recurrent Tuberculosis

Rohit Mistry,1,3 Jacqueline M. Cliff,2 Christopher L. Clayton,1 Nulda Beyers,4 Yasmin S. Mohamed,1 Paul A. Wilson,1 Hazel M. Dockrell,2 Don M. Wallace,1 Paul D. van Helden,2⁎ Ken Duncan,1,a and Pauline T. Lukey1

1GlaxoSmithKline, Stevenage, and 2Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom; 3Department of Medical Biochemistry and 4Desmond Tutu TB Centre, Faculty of Health Sciences, University of Stellenbosch, Tygerberg, South Africa

Background. The majority of patients with tuberculosis who comply with appropriate treatment are cured. However, ∼5% subsequently have a repeat disease episode, usually within 2 years of successful combination therapy. Presently, there is no way of predicting which patients will experience a relapse.

Methods. We identified 10 subjects who had previously experienced recurrent tuberculosis and carefully matched them to cured subjects who had had only 1 episode of tuberculosis, to patients with active tuberculosis, and to latently infected healthy subjects. We compared their ex vivo whole-blood gene-expression profiles by use of DNA array technology and confirmed the results by use of quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR).

Results. The 4 clinical tuberculosis groups exhibited distinct patterns of gene expression. The gene-transcript profiles of the patients with recurrent tuberculosis were more similar to those of the patients with active tuberculosis than to those of the cured or latently infected subjects. Discriminant analysis of a training data set showed that 9 genes were sufficient to classify the subjects. We confirmed that measurement of the expression of these genes by qRT-PCR can accurately discriminate between subjects in a test set of samples.

Conclusions. A simple test based on gene-expression patterns may be used as a biomarker of cure while identifying patients who are at risk for relapse. This would facilitate the introduction of new tuberculosis drugs.
Biomarkers that could be used to identify individuals at risk for recurrent disease would aid in the clinical management of such patients.

When leukocytes traffic to and from sites of M. tuberculosis infection and interact with mycobacteria, their response includes changes in levels of gene expression. These changes form a gene-expression fingerprint, which might be used to distinguish between latent or active tuberculosis and to predict the risk of recurrent disease after chemotherapy. Using a similar genetic profiling approach, Bleharski et al. [12] were able to discriminate between the tuberculoid and lepromatous forms of leprosy in skin biopsy samples. We aimed to identify novel host-derived biomarkers in blood, a readily obtainable compartment, that would discriminate between subjects who are cured after conventional chemotherapy and those who are susceptible to recurrent disease.

SUBJECTS, MATERIALS, AND METHODS

Subject enrollment and sample collection. Subjects with cured tuberculosis (hereafter, “cured subjects”) had had 1 episode of tuberculosis, were cured for at least 18 months before enrollment into this study, and remained disease free after cure. The definition of cure was in accordance with the guidelines of the World Health Organization and the South African National Tuberculosis Programme—a positive result for a sputum sample smear at diagnosis, a full course of treatment taken, and a negative result for a sputum sample smear during the last month of treatment and at least 1 other negative result for a sputum sample smear. Patients with active disease (hereafter, “active subjects”) had smear-positive tuberculosis at the time of enrollment but had not yet started treatment. Patients with tuberculosis received fixed-dose combination therapy—after the first episode, they received a 6-month regimen containing isoniazid and rifampicin throughout the 6 months as well as ethambutol and pyrazinamide during the first 2 months only. Patients with recurrent disease (hereafter, “recurrent subjects”) had had either 2 \( n = 8 \) or 3 \( n = 2 \) episodes of cured tuberculosis and had received an 8-month regimen for the retreatment episode and were disease free for at least 18 months before enrollment. Subjects with latent M. tuberculosis infection (hereafter, “latent subjects”) had a Mantoux test result \( >15 \) mm and had no record of previously being treated for tuberculosis. All subjects were recruited in the Ravensmead and Uitsig districts of Cape Town, South Africa. In this setting, the high incidence of tuberculosis means that the majority of adults have been exposed to M. tuberculosis and that the latent subjects, therefore, form the healthy control group.

Written, informed consent was obtained from each subject (per the Helsinki declaration and ICH guidelines). At the time of enrollment, each subject was interviewed and examined clinically, a chest radiograph was obtained, sputum samples were collected for smear and culture, and blood samples were collected. At enrollment, no latent, cured, or recurrent subjects had M. tuberculosis cultured from sputum. Detailed information about each subject—including sex, age, number of disease episodes, hemoglobin level, full blood cell counts, and clinical tuberculosis descriptions at the time of recruitment into the study—is given in table 1. All subjects received pre- and posttest counseling before HIV testing—all were HIV negative. Blood samples were collected from each subject within a 2-h period in the morning, to minimize the effects of diurnal variation. The Ethics Committee (Institutional Review Board) of the Faculty of Health Sciences, Stellenbosch University, approved the study.

RNA preparation from whole blood. Seven milliliters of venous blood was mixed immediately with 50 mL of RNA stabilization solution (Roche Diagnostics) and stored frozen at \(-70^\circ\text{C}\) until use. mRNA was isolated using magnetic bead technology in accordance with the manufacturer’s protocol (Roche Diagnostics) and was treated with DNase I (Ambion), and mRNA integrity was confirmed by gel electrophoresis.

DNA arrays. Six matched subjects in each clinical group were assigned to a training set, and the other 4 matched subjects were assigned to a test set. The training set was subjected to DNA array analysis. Matched mRNA samples (e.g., from recurrent subject 1, cured subject 1, latent subject 1, and active subject 1) were treated simultaneously through purification of mRNA, generation of \( \alpha-\)\(^{32}\)P–labeled cDNA probes, DNA array hybridization, and membrane scanning, to minimize the introduction of systematic bias. Two DNA arrays were used [13], one produced from a normalized human leukocyte cDNA library (50,000 clones; Invitrogen), and the other produced from a custom set of sequence-verified cDNA clones representing 8000 genes and expressed sequence tags. Each clone was spotted in duplicate on nylon membrane filters. \( \alpha-\)\(^{32}\)P–labeled cDNA was generated from the mRNA by reverse transcription and was hybridized to the DNA arrays in duplicate, as described elsewhere [14]. Four replicate measurements of expression intensity were obtained for each sample for each clone.

Table 1. Subject characteristics.

<table>
<thead>
<tr>
<th>Subject characteristics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The table is available in its entirety in the online edition of the Journal of Infectious Diseases.</td>
</tr>
</tbody>
</table>

Table 2. Genes significantly differentially expressed between the clinical tuberculosis groups.

The table is available in its entirety in the online edition of the Journal of Infectious Diseases.
Figure 1. Gene signatures in the blood of subjects with active (A), recurrent (R), cured (C), and latent (L) tuberculosis. A hierarchical clustering of the samples and genes is shown. Each row represents the mean signal intensities of each of the 337 differentially expressed genes ($P < 0.05$, analysis of variance), and each column represents a clinical tuberculosis group ($n = 6$ per group). A colored representation of the normalized expression levels is shown, such that red indicates high expression and green indicates low expression (the scale is shown below). On the left is the hierarchical clustering dendrogram of the genes. On the right are selected examples of immune response (black) and discriminant (blue) genes (indicated by National Center for Biotechnology Information reference sequence no. and gene symbol). There are 6 main clusters of genes that show similar expression patterns across the clinical groups. The active and recurrent groups cluster together, and the cured and latent groups cluster together.
Table 3. Stepwise linear discriminant analysis genes and quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR) primer sequences.

The table is available in its entirety in the online edition of the Journal of Infectious Diseases.

Data analysis. Statistical analyses were performed using SAS (version 8; SAS Institute). Clones that were never expressed above background (calculated from the mean intensity of empty spots plus 2 SDs) were removed from the data set, and the remaining expressed clones were analyzed on the log_{10} scale. The data set was normalized by dividing each clone-intensity measurement by the overall median intensity for the corresponding array, so that different arrays and experiments could be compared.

Analysis of variance (ANOVA) was used to assess the variability between replicate spots, duplicate arrays, and samples and to compare intensities between the 4 clinical groups for each clone. Differentially expressed genes with \( P < .05 \) across all 4 groups were identified by sequencing cDNA clones [13]. We performed stepwise linear discriminant analysis analysis with all 337 genes [15] to identify genes that could correctly classify the subjects in the clinical groups, and we performed supervised and unsupervised hierarchical clustering using the Pearson correlation similarity metric in GeneSpring software (Silicon Genetics).

Quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR). mRNA was reverse transcribed using oligo-dT primers and Superscript II (Invitrogen), and the expression levels of 9 discriminant and 3 reference genes (for \( \beta \)-actin, GAPDH, and cyclophilin) were measured in duplicate by real-time qRT-PCR using SybrGreen incorporation (Applied Biosystems) [16]. Transcript copy numbers were calculated from a standard curve generated from sheared genomic DNA [16]. To improve precision and account for sample variability, we performed analysis of covariance on each gene separately, with each of the 3 reference genes and the first principal component scores used as covariates [17]. Primer sequences were designed using the Primer3 program (version 0.2) [18].

RESULTS

Selection of subjects. In 1999, we described 16 patients in South Africa who had had recurrent episodes of tuberculosis after curative treatment [11]. From this group, recurrent subjects were selected for the present study on the basis of the following inclusion criteria: both episodes of tuberculosis had to be cured, both episodes had to be caused by drug-sensitive organisms, subjects had to be treatment compliant throughout both episodes, the last episode of tuberculosis must have been cured at least 18 months before enrollment, and subjects had to be HIV negative and sputum culture negative for \( M. \) tuberculosis at the time of enrollment. Ten individuals fulfilled these enrollment criteria.

These recurrent subjects were matched for age, sex, and residential area (the last to control for environmental factors, including exposure to \( M. \) tuberculosis) to 10 people in each of 3 other clinical groups: cured subjects, active subjects, and latent subjects. In the present setting, most people are vaccinated with bacille Calmette-Guérin shortly after birth, and it was assumed that prior vaccination status would influence all clinical groups to the same extent.

Identification of differentially expressed genes in the training set. We measured full and differential cell counts and hemoglobin levels to rule out the possibility that differences in gene expression were due to variation in the relative abundance of certain cell populations. We found no significant differences between the clinical groups (\( P > .18 \), ANOVA; table 1 and data not shown).

We isolated mRNA from blood samples from 6 matched subjects per clinical group (training set), generated radiolabeled cDNA probes, and hybridized them to 58,000 arrayed cDNA clones. After data reduction to remove nonexpressed clones, 27,566 clones were analyzed, each with quadruplicate data for each subject from each clinical group. ANOVA revealed that 337 genes were differentially expressed (\( P < .05 \) across all 4 groups (table 2). We assigned each gene to a functional group, the largest of which contained genes involved in the immune response (23%), followed by genes of unknown and/or in silico predicted function (17%); metabolic genes (15%); nuclear genes other than transcription factor genes (13%); signal transduction genes (9%); transcription factor genes (6%); structural genes (6%); intracellular trafficking genes (4%); solute transport genes (4%); and proteolysis genes (3%).

To investigate the relatedness of the gene-expression patterns, we subjected the combined expression data for each clinical group to supervised hierarchical cluster analysis [19]. We found that the active and recurrent subjects grouped together, whereas the cured and latent subjects formed a separate cluster (figure 1).

When we then performed unsupervised hierarchical cluster analysis [19], the 337 differentially expressed genes formed into 6 clusters with higher or lower expression across the 4 clinical groups (figure 1). Genes in each cluster belonging to the immune class were examined, because these may be the most relevant to the response to the pathogen (figure 1).
Discriminant analysis of differentially expressed genes in the training set. To identify a subset of the 337 genes that could discriminate between the tuberculosis clinical phenotypes, we subjected the gene-expression data to stepwise linear discriminant analysis [20]. An expression pattern of 22 genes could be used to discriminate between the clinical groups (table 3). Resubstitution and leave-one-out cross-validation determined that the most parsimonious model (table 4) required only 9 of these genes to assign accurately each individual to his or her respective group. These genes are Ras and Rab interactor 3 (RIN3); lymphocyte antigen 6 complex, locus G6D (LY6G6D); testis-expressed gene 264 (TEX264); chromosome 14 open reading frame 2 (C14orf2); suppressor of cytokine signaling 3 (SOCS3); hypothetical protein KIAA2013 (KIAA2013), whose function is unknown; ATP-binding arsenite transporter (ASNA1); mitochondrial ATP synthase (ATP5G1); and nucleolar protein family A, member 3 (NOLA3). The 9 genes that perfectly discriminate between the 4 clinical groups are distributed throughout all 6 gene hierarchical clusters (figure 1), and each gene has a unique expression pattern across the 4 clinical groups (figure 2).

qRT-PCR measurement of 9 discriminant genes. Having identified the set of 9 genes by use of DNA array technology, we moved to a more sensitive, more accurate, and more robust platform—qRT-PCR—for measuring gene expression. We measured transcript levels of each of the discriminant genes in the training set. We compared these data with the expression profiles obtained in the DNA array hybridization experiments and found similar patterns (validating our original observation [figure 2]), and, indeed, the magnitudes of the differences in observed expression levels were enhanced. The housekeeping genes were expressed at the same level in all samples and did not influence or covary with the test genes (data not shown).

We then tested the expression levels of the discriminant genes in the test set samples and found them to be very similar to the training set profiles. The discriminant gene-expression pattern for all 40 subjects is shown in figure 3. The expression patterns were similar for all 10 recurrent subjects, regardless of whether they had had 2 or 3 disease episodes previously.

Development and validation of a discriminant function. We then developed a new discriminant function using the qRT-PCR data from the training set (table 5) and found that only 4 (RIN3, LY6G6D, TEX264, and C14orf2) of the 9 genes were required to correctly assign all subjects to their respective clinical group (figure 4A). We tested the validity of the discriminant function by cross-validation and resubstitution methods, which correctly assigned all 24 subjects to their respective group.

To confirm that this new discriminant function was able to

Figure 2. Expression profile of discriminant genes measured by DNA array and quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR). The expression levels of the discriminant genes (RIN3, LY6G6D, TEX264, C14orf2, SOCS3, KIAA2013, ASNA1, ATP5G1, and NOLA3) as measured by hybridization to cDNA arrays (arbitrary intensity units; shaded bars) are compared with those as measured by qRT-PCR (relative gene copy number detected per nanogram of poly[A]+ RNA; black bars). Data are the mean ± SE levels for the active (A), recurrent (R), cured (C), and latent (L) groups (n = 6 per group).
assign subjects correctly to groups, qRT-PCR data from the test set were analyzed. On basis of the levels of gene expression, the discriminant function assigned 15 of the 16 subjects to the correct groups in a blind analysis (figure 4B). The misclassified subject was clinically diagnosed as having latent tuberculosis at the time of study recruitment but was assigned to the active group by the discriminant function. Clinical reassessment of the subjects in the latent group in 2004 identified active culture-positive pulmonary disease in this subject only.

DISCUSSION

More than 1.8 million deaths were attributed to tuberculosis in 2000 [21], emphasizing the urgent need for new tools to
improve diagnosis, treatment, and prevention. The different clinical outcomes after *M. tuberculosis* infection reflect a combination of factors, including host genetics, previous encounters with other mycobacterial species, coinfection, and immunocompetence.

We identified patients who had experienced recurrent bouts of tuberculosis [14] caused by either reinfection or reactivation of the primary infection, and we hypothesized that all these subjects had an increased susceptibility to developing clinical disease after infection [22]. We then compared them to 2 less-susceptible groups: patients who were successfully cured after 1 disease episode, and subjects who were able to control the infection in a latent form. We identified a panel of biomarkers expressed in blood that could fully discriminate between these 3 groups of subjects, who were all healthy at the time of sampling. Furthermore, these biomarkers were also able to discriminate patients with active tuberculosis. Similar approaches using blood cell transcript profiling have previously been employed to identify biomarkers that can predict clinical outcomes in leukemia [23] and cancers, including renal cell carcinoma [24], and to subclassify subjects with infectious diseases such as HIV-1 [25] and malaria [26].

Measurement of these biomarkers during tuberculosis treatment may be clinically useful to identify patients whose gene-expression profile is typical of a recurrent tuberculosis phenotype, as it may mean that they are at risk for developing tuberculosis again and require close monitoring. This will be tested in future prospective studies. Furthermore, these biomarkers might facilitate clinical trials of new chemotherapeutic regimens or shortened treatment by helping to determine when a patient is cured and by allowing those at risk for relapse to be identified early [27, 28]. Ultimately, these biomarkers could be developed into a test that would generate an image similar to that shown in figure 3.

Most studies examining the effects of mycobacterial infection have used purified cell types, cultured cells, or biopsy material, using processes that may modify gene expression [12, 29]. In the present study, we used whole blood, which may more accurately represent in vivo processes at the time of phlebotomy. Sampling of immune cells trafficking to and from the site of active disease via the systemic circulation is likely to be the mechanism by which gene expression in blood cells reflects clinical status. For any practical use of the markers, it is critical that they can be measured in a sample that is easy to obtain and does not require expensive and time-consuming separation of individual cell types.

Many of the differentially expressed genes we identified here have been implicated in studies of the host immune response to *M. tuberculosis* infection, such as interferon (IFN)–inducible genes, tumor necrosis factor receptor 2, interleukin (IL) 4 receptor, and granulysin [30]. In addition to supporting and enhancing the known immune response to tubercular infection, the gene-expression profiles allow the formulation of new hypotheses to explain further the underlying pathological mechanisms. For example, an overview of the functional classes of genes indicates that a large proportion of the 337 genes encode proteins with roles in inflammation and immunity. Thirty-eight genes were also found to be preferentially expressed in *M. tuberculosis*–activated CD4+ or CD8+ T cells, and genes for signal transducing molecules such as DAP10, which is expressed in activated CD8+ T cells and which was more highly expressed in latent subjects than in active subjects; this may imply a role in protection. The use of a leukocyte array allowed us to identify many genes of unknown function that represent novel genes for future investigation, such as CCDC22, which is up-regulated in cured relative to recurrent subjects and which is expressed preferentially in activated CD8+ T cells.

The 9 genes that discriminated between the 4 clinical groups were distributed throughout all 6 gene clusters and were, thus, representative of the overall gene-expression pattern. Three of the discriminant genes play a role in the immune response. *RIN3* encodes a GTP-Rab5-stabilizing protein [31] and, as such, is involved in the regulation of endocytosis. Rab5 is normally associated with early endosomes and is retained by phagosomes containing *M. tuberculosis* [32]. The increased expression of *RIN3* in infected subjects may reflect increased phagocytosis of *M. tuberculosis*. *LY6G6D* is located in the gene-rich major histocompatibility complex III region (6p21.3), which is associated with susceptibility to many diseases. It encodes a glycosphosphatidyl inositol–linked cell-surface molecule and may play a role in leukocyte activation and differentiation [33]. SOCS3 suppresses cytokine and chemokine signaling via the Janus kinase signal transducer and activation of transcription pathway [34]. It mediates the anti-inflammatory effect caused by the stimulation of Toll-like receptor 2 by heat shock protein 60 [35], and SOCS3 deficiency in mice impairs immune responses to infection via hyperproduction of IL-10 and transforming growth factor β [36]. In our study, the active patients expressed high levels of SOCS3, reflecting an ongoing immune response, whereas the recurrent subjects expressed very low levels, which may indicate impaired immunity. The roles played by the other
6 genes in the immune system are less well defined, and this DNA array screening approach has opened up many new avenues for exploration of immunity to tuberculosis. ASNA1 is involved in heavy metal ion homeostasis and in resistance to arsenic and antimonials [37]; its higher expression in cured patients may reflect restored immune function. ATP5G1 is part of the mitochondrial ATP synthase [38] and is involved in cellular energy production. NOLA3 is involved in pre-rRNA processing [39]. The functions of TEX264 (which is likely to encode a secreted protein), CI4orf2, and KIAA2013 are unknown.

Having identified the set of 9 genes by use of DNA array technology, we moved to a more sensitive, more accurate, and more robust platform—qRT-PCR—for measuring gene expression. A new discriminant function in which the expression pattern of only 4 genes was sufficient to assign the 24 training-set subjects to their respective groups. The validity of this observation was tested on the mRNA isolated from the subjects in the independent test sets and correctly classified 15 of 16 subjects into their respective groups. The misclassified subject had latent tuberculosis at the time of sampling but was classified as an active subject by the discriminant function. Clinical reassessment of the subjects in the latent group in 2004 identified active culture-positive pulmonary tuberculosis in only this subject. When tested originally, this individual may have been progressing from latent to active tuberculosis, and this implies that the pattern of expression may be used to predict progression from latent to active disease before the onset of clinical symptoms. Identification of an active pattern in a skin-test positive but otherwise healthy individual might suggest that chemoprophylaxis is warranted. One of the recurrent subjects had experienced a further episode of tuberculosis at this follow-up point, whereas none of the active or cured subjects had.

In future work, we plan to investigate the expression patterns of these biomarkers in a larger number of subjects with tuberculosis at diagnosis and during therapy. It would be interesting to analyze other groups, such as latenty infected but Mantoux test–negative individuals and nonexposed subjects, to determine whether biomarker expression correlates with known immune function, such as IFN-γ production.

We have shown that whole-blood gene-expression profiling can be used to distinguish clinical tuberculosis groups, and we have identified a panel of 9 host biomarkers that discriminate between active, latent, recurrent, and cured tuberculosis in this setting. Development of such biomarkers might aid the clinical management of tuberculosis by providing a tool that can determine when a patient is cured as well as predict which people are at risk for recurrence. Furthermore, such biomarkers could facilitate the introduction of new tuberculosis chemotherapeutic regimens by shortening the duration and costs of clinical trials.

Acknowledgments

We dedicate this work to Albert Beyers (1954–2000). We gratefully acknowledge the participation of all the subjects and the contribution of the clinical staff. We also thank Dr. N. Carroll and Ms. E. Engelske, for microbiological assessment of sputum samples; Mark Lennon, for advice on data interpretation; and the GlaxoSmithKline Transcriptome Analysis and Sequencing groups, for their assistance with this study.

References

20. Hastie T, Tibshiranri, Friedman J. The elements of statistical learn-


