

## FULL PAPER

# Peripheral blood gene expression profiling in rheumatoid arthritis

FM Batliwalla<sup>1,2</sup>, EC Baechler<sup>3</sup>, X Xiao<sup>1,2</sup>, W Li<sup>1,2</sup>, S Balasubramanian<sup>3</sup>, H Khalili<sup>1,2</sup>, A Damle<sup>1,2</sup>, WA Ortmann<sup>3</sup>, A Perrone<sup>4</sup>, AB Kantor<sup>4</sup>, PS Gulko<sup>1,2</sup>, M Kern<sup>1,2</sup>, R Furie<sup>2</sup>, TW Behrens<sup>3</sup> and PK Gregersen<sup>1,2</sup>

<sup>1</sup>Robert S Boas Center for Genomics and Human Genetics, North Shore-Long Island Jewish Research Institute, Manhasset, NY, USA;

<sup>2</sup>Department of Medicine, NSUH, Manhasset, NY, USA; <sup>3</sup>Department of Medicine, University of Minnesota, Minneapolis, MN, USA;

<sup>4</sup>SurroMed Inc., Menlo Park, CA, USA

We carried out gene expression profiling of peripheral blood mononuclear cells (PBMCs) in 29 patients with active rheumatoid arthritis (RA) and 21 control subjects using Affymetrix U95Av2 arrays. Using cluster analysis, we observed a significant alteration in the expression pattern of 81 genes ( $P < 0.001$ ) in the PBMCs of RA patients compared with controls. Many of these genes correlated with differences in monocyte counts between the two study populations, and we show that a large fraction of these genes are specifically expressed at high levels in monocytes. In addition, a logistic regression analysis was performed to identify genes that performed best in the categorization of RA and control samples. Glutamyl cyclase, IL1RA, S100A12 (also known as calgranulin or EN-RAGE) and Grb2-associated binding protein (GAB2) were among the top discriminators. Along with previous data, the overexpression of S100A12 in RA patients emphasizes the likely importance of RAGE pathways in disease pathogenesis. The altered expression of GAB2, an intracellular adaptor molecule involved in regulating phosphatase function, is of particular interest given the recent identification of the intracellular phosphatase PTPN22 as a risk gene for RA. These data suggest that a detailed study of gene expression patterns in peripheral blood can provide insight into disease pathogenesis. However, it is also clear that substantially larger sample sizes will be required in order to evaluate fully gene expression profiling as a means of identifying disease subsets, or defining biomarkers of outcome and response to therapy in RA.

Genes and Immunity (2005) 6, 388–397. doi:10.1038/sj.gene.6364209; published online 23 June 2004

**Keywords:** gene expression; rheumatoid arthritis; peripheral blood mononuclear cells; monocytes; microarray

## Introduction

Rheumatoid arthritis (RA) is a heterogeneous disorder characterized primarily by chronic inflammation and destruction of bone and cartilage in diarthrodial joints. The cause of RA is unknown, with both genetic and environmental factors contributing to disease susceptibility.<sup>1</sup> Inflammatory rheumatoid synovium is characterized by intimal lining layer hyperplasia and angiogenesis coupled with infiltration by a complex mixture of T and B lymphocytes, mast cells, macrophages and dendritic cells.<sup>2</sup> The relative importance of these cell types for disease initiation and perpetuation remains uncertain. Activated monocytes and T cells may be found in peripheral blood as well as synovial tissues, and these cells are a potential source of proinflammatory cytokines

such as TNF- $\alpha$  that play a key role in disease pathogenesis. Current biologic therapies that neutralize TNF have shown a significant clinical improvement in RA patients. However, only a small percentage of patients achieve a dramatic response (ACR70) and a significant proportion of patients do not respond at all to TNF blockade,<sup>3,4</sup> consistent with the heterogeneous nature of the RA phenotype.

Genome-wide gene expression profiling has been used to better classify many cancers<sup>5</sup> and to understand the molecular pathways involved in several disease processes. Recently, we have successfully used peripheral blood cells to obtain a gene expression profile of patients with systemic lupus erythematosus (SLE).<sup>6</sup> This work has shown that a subset of SLE patients exhibits a 'signature' of interferon (IFN)-inducible genes. In this report, we use a similar strategy to identify gene expression profiles that distinguish RA patients from healthy control individuals. The results clearly reflect the increased percentage of activated monocytes in the peripheral blood in RA and hint at a role for several different inflammatory and signaling pathways.

Correspondence: Professor PK Gregersen, Robert S Boas Center for Genomics and Human Genetics, North Shore-Long Island Jewish Research Institute, 350 Community Drive, Manhasset, NY 11030, USA.

E-mail: peterg@nshs.edu

Received 3 January 2005; revised 15 February 2005; accepted 16 February 2005; published online 23 June 2004

## Results

### Gene expression profiles of peripheral blood mononuclear cells from RA patients compared with normal controls

We analyzed the gene expression profile in the peripheral blood mononuclear cell (PBMC) samples obtained from 29 RA patients and 21 normal control individuals. Out of 4500 genes that were expressed by PBMCs and used for this data analysis, we identified 81 genes with significantly different expression values between the two groups (see Supplementary Table 1 for complete data set). These 81 genes met all three of the following criteria: (1) significant difference between the groups by unpaired *t*-test,  $P < 0.001$ ; (2) a 1.4-fold difference in the mean expression value between groups; (3) a difference of at least 100 average difference (AD) units between the mean expression values in each group. These data were clustered and visualized as shown in Figure 1. All 29 RA patients were clustered together (maroon), and three control individuals (blue) clustered with the patient samples. Of the 81 genes that differentiated RA from controls, 29 genes were downregulated in the RA group, and 52 genes were upregulated in the patients compared to controls. In order to confirm the significance of these results, we performed a permutation analysis of the entire data set of 4500 genes, with 50 repetitions. This analysis yielded 4.5 as the average number of expected false positives using our filter criteria. Therefore, the 81 genes that are differentially expressed in RA patients compared to controls are far greater than what would be expected by chance alone ( $\sim 4.5$  false positives predicted at  $P < 0.001$ ).

### Genes overexpressed in PBMCs of RA patients are enriched for monocyte-specific transcripts

A total of 52 genes were expressed at increased levels in RA patients compared with controls (see Supplementary Table 1). A number of these genes are known to be expressed in monocytes, including CD14 antigen, CD163, CD13, S100 calcium binding protein A12 (calgranulin C), chemokine (C-C motif) receptor 1 and interleukin 1 receptor antagonist (IL-1Ra). In order to develop a more complete list of genes that are preferentially expressed in monocytes, fluorescence-activated cell sorter (FACS)-sorted monocytes ( $n = 3$ ) and T cells ( $n = 2$ ) from normal individuals were processed and hybridized to Affymetrix U133A microarrays. Genes with greater than 10-fold overexpression (mean fold change between 10 and 560) in monocytes compared with T cells were provisionally designated as 'monocyte enriched' (see Supplementary Table 2). Additionally, genes with a greater than 10-fold overexpression in T cells compared with monocytes were provisionally designated as 'T-cell-enriched' genes. Of the 52 genes that are overexpressed in RA PBMCs, 21 (40%) were in this 'monocyte-enriched' group. Thus, the presence of these genes may reflect in part the increased numbers<sup>7,8</sup> or activation of monocytes in RA PBMCs compared with controls (see below).

### Genes underexpressed in PBMCs of RA patients

A total of 22 genes had lowered expression in patients compared with controls (see Supplementary Table 1). Several genes in this list appeared to be lymphocyte-

specific genes, such as CD72 and CD79b. Additionally, molecules associated with signal transduction in lymphoid cells, such as lymphocyte-specific tyrosine kinase, protein kinase C theta, death-associated transcription factor 1, granzyme A, had lower transcript levels in RA patients compared with healthy individuals. Of the 29 genes that are underexpressed in PBMCs of RA patients, 10 genes (34%) were in a group of 'T-cell-enriched' genes (data not shown) and may reflect the decreased number of T cells in the PBMCs of RA patients.

### Pattern of cell subset expression in RA patients compared with controls

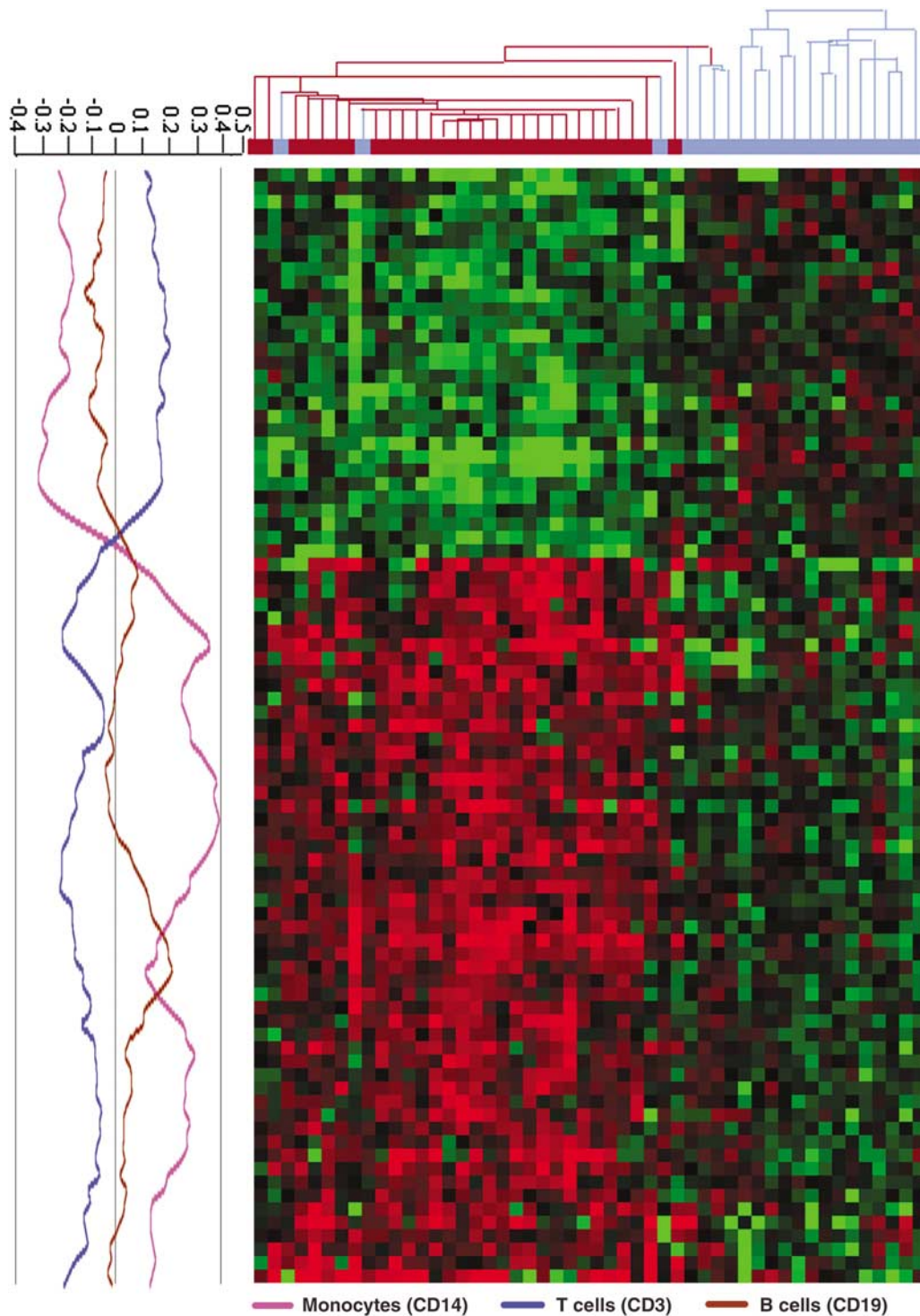
Since several genes that were either overexpressed or underexpressed in the RA patients are associated with T cells, B cells or monocytes, we examined the cell subtype distribution in our samples. Immunophenotyping was carried out on PBMCs from patients and control individuals using microvolume laser scanning cytometry (MLSC) to determine the expression of cell surface parameters. For this analysis, we had access to the frozen PBMCs from a subset of patients ( $n = 24$ ) and controls ( $n = 8$ ). Previous experiments showed that Surroscan analysis of fresh and frozen PBMCs yielded highly comparable results for the assays reported herein. Within PBMCs, a significant increase in the percentage of monocytes was seen in the RA patients (24.64% + 9.8 in patients, 16.15% + 5.3 in controls,  $P = 0.0052$ ).

### Correlation of gene expression profile with subsets of PBMCs

Correlation coefficients were calculated between the percent of each cell subset in PBMCs and the relative expression of genes that distinguished RA patients from healthy controls. Initially, the correlation values were calculated as moving averages across 11 genes, and the results are shown in Figure 1, to the left. A positive correlation was seen between CD3-positive T cells (blue line) and genes that were underexpressed in RA patients. In contrast, the percentages of monocytes (pink line) and B cells (brown line) were positively correlated with genes that were overexpressed in RA patients relative to controls. These observations suggested that at least some of the differences in gene expression between cases and controls were related to differences in cell counts in the PBMCs used as a source of RNA.

Since the monocyte differentiation antigen CD14 is preferentially expressed on the surface of mature cells of the monocytic lineage, we plotted CD14 gene expression levels vs CD14 cell surface expression as shown in Figure 2. The two parameters correlated significantly ( $r = 0.6$ ,  $P = 0.00026$ ). Similarly, correlation coefficients were calculated for each of the 81 genes and the percent expression of monocytes, T-lymphocyte and B-lymphocyte levels. Figure 3 shows genes that exhibited significant positive correlations with monocyte CD14 counts (15 genes,  $r > 0.4$ ,  $P < 0.01$ ).

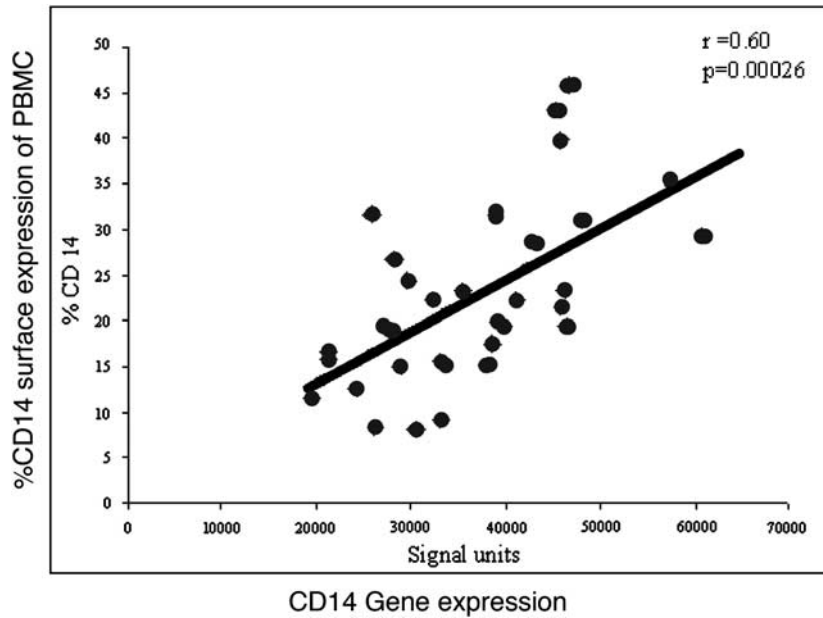
Of the genes that positively correlated with monocyte counts, several are specifically expressed by monocytes, such as CD14 and CD163. Most of the genes shown in Figure 3 are reported to be expressed by monocytes, although the expression is not restricted to monocytes alone; they may also be expressed in multiple tissues ranging from the liver, prostate, cardiac muscle and other



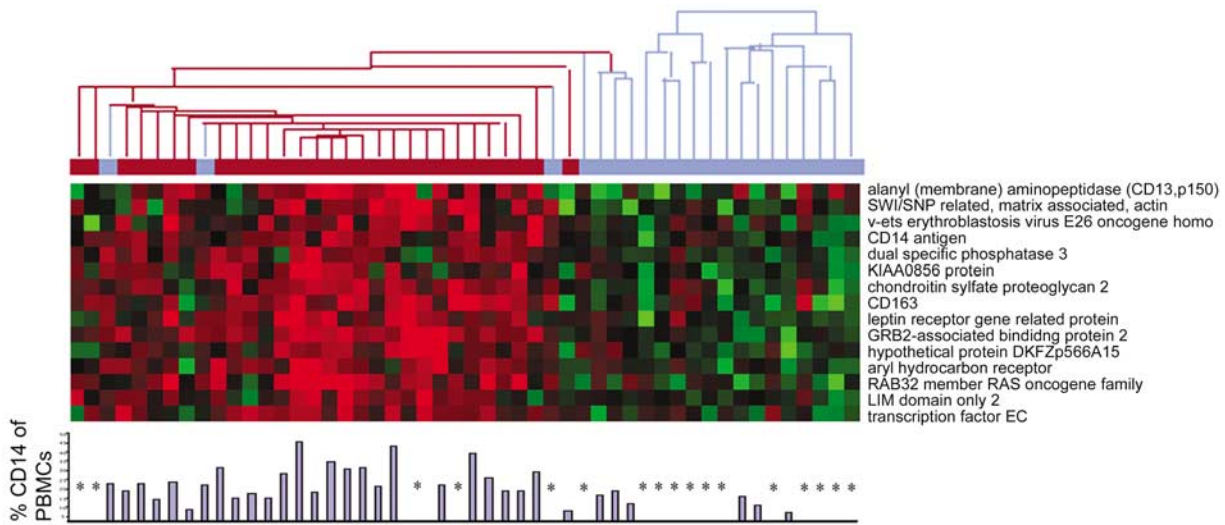
**Figure 1** Gene expression profiles of PBMCs from 21 control individuals and 29 RA patients. Hierarchical clustering of 81 genes that distinguish RA patients from healthy controls is shown. Each row represents a gene; each column shows the expression of 81 genes expressed by each individual in the study. Red indicates genes that are expressed at higher levels compared with the control mean. Green indicates genes that are expressed at lower levels relative to the control mean (see Supplementary Table for the raw gene expression values). Correlation coefficients with cell counts for T lymphocytes, B lymphocytes and monocytes are shown on the left. These correlation values were calculated as moving averages of 11 genes along the vertical axis shown to the left of the figure. The dotted lines show the level of positive and negative correlation ( $r=0.42$ ,  $P=0.01$ ).

blood cells (Novartis atlas, <http://symatlas.gnf.org/SymAtlas/>). Nevertheless, within the total population of PBMCs, we see a direct correlation between all of these genes and the percent of circulating monocytes. Nine of

the 17 genes (Figure 3) that correlate with CD14 expression were also found to be present in our group of 'monocyte-enriched' transcripts (see Supplementary Table 2).



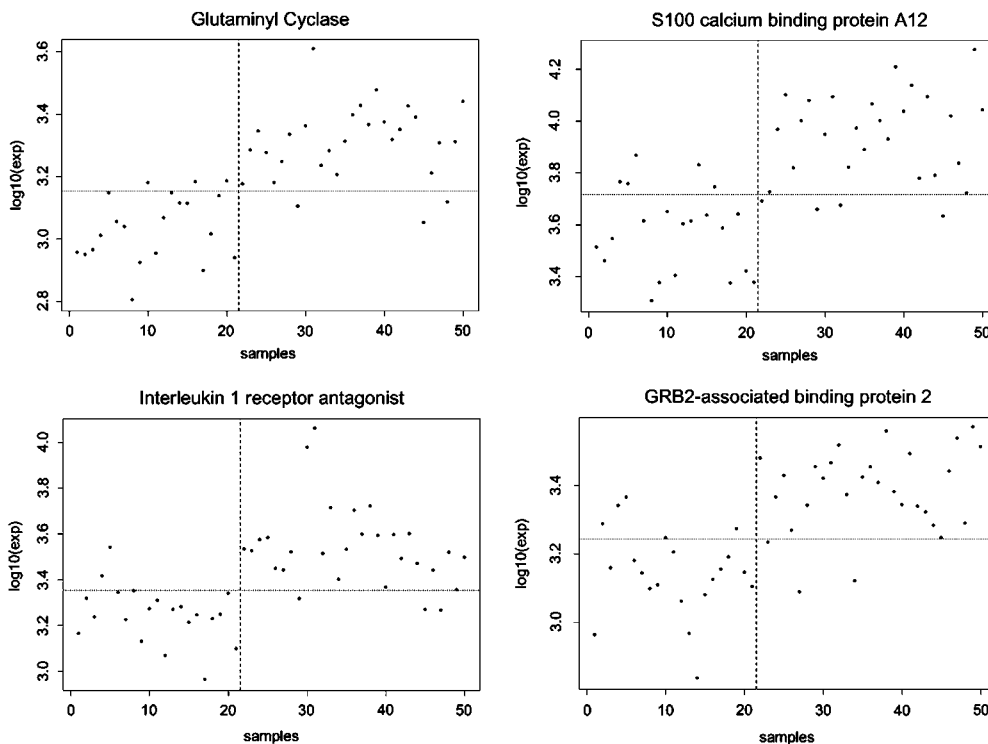
**Figure 2** Correlation between gene expression and cell surface expression for CD14. A significant correlation between the two parameters was detected ( $r=0.60$ ,  $P=0.00026$ ).



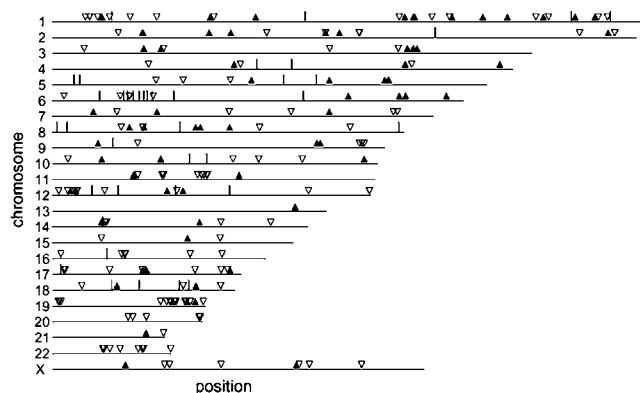
**Figure 3** Differential gene expression between RA patients and control PBMCs. Correlation coefficients were calculated individually for each of the 81 genes with CD14 cell counts. Genes that correlated with monocyte (CD14) counts ( $r>0.4$ ) are shown from the data shown in Figure 1. The bar graph below the gene expression profile depicts the percent of CD14 expressed in the PBMCs of each sample. \* indicates that the CD14% was not obtained for that sample.

Logistic regression analysis can be used to predict the group membership (RA *vs* control) of a sample based on the expression value of a gene. The lower the number of misclassifications, the higher the ranking of a gene (see Patients and methods). Logistic regression analysis was applied to our data set to identify genes whose expression profiles best classify RA patients *vs* controls. Supplementary Table 3 lists 200 genes with the highest logistic regression ranking. Figure 4 shows the results for four genes that had the highest rank by logistic regression—glutamyl cyclase, S100 calcium binding protein A12, IL-1Ra and Grb2-associated binding protein

2 (GAB2). Perhaps not surprisingly, these genes show enriched expression in monocytes (10- to 53-fold increased expression relative to T cells; see Supplementary Table 2); however, these genes are also expressed in a number of other tissues including the brain and other hematopoietic cell subsets (Novartis atlas, <http://symatlas.gnf.org/SymAtlas/>). Furthermore, an analysis with all 12600 probe sets (including *ex vivo* stress response genes; see Patients and methods) showed that the genes discussed above remain in the top seven discriminators for RA patient samples compared with control subjects (the other genes being BCL2-related



**Figure 4** Logistic regression analysis of RA patients and control individuals. Four genes with the highest ranking by logistic regression are glutaminyl cyclase, s100A12, IL1-Ra and GAB2. The plots depict the level of gene expression (log10 transformed value for fluorescence intensity) for each of these four genes for each sample. The 21 controls (samples 1–21) are shown to the left of the vertical dashed line, and the 29 RA patients (samples 22–50) are shown to the right of the vertical dashed line in each plot. The horizontal dashed line gives the threshold value ( $-\hat{a}/\hat{c}$ ) for each analysis.



**Figure 5** Correlation of logistic regression analysis with genome-wide linkage peaks. The chromosomal locations of the top 200 genes identified by logistic regression analysis were compared with linkage peaks from previous studies (Supplementary Table 3). Vertical lines indicate areas of linkage identified from previous studies, filled triangles indicate genes with increased expression levels in RA patients relative to controls and open triangles indicate genes with lower expression levels in RA patients compared with control individuals.

protein A1, adrenomedullin and CD63 antigen). The classification rate for logistic regression analysis in this data set is about 80%.

Finally, we wished to determine whether any of the 200 genes differentially expressed in RA (by logistic regression) are clustered near to the linkage peaks that

we have previously observed in RA affected sib-pair studies.<sup>9,10</sup> Figure 5 shows the chromosomal location of these genes (taken from Supplementary Table 3) and the chromosomal location of the linkage peaks that we previously reported. We did not observe any striking aggregation of differentially expressed genes within the various regions of linkage. The few specific genes that are found near (within 5 cM) to particular linkage peaks (indicated in Supplementary Table 3) may be appropriate targets for further genetic association studies.

## Discussion

In this study, we have used gene expression profiling of PBMCs in an attempt to define transcriptional patterns that differentiate RA patients from control subjects. Strikingly, we observed that a number of genes that are monocyte specific are overexpressed in the RA patients (Figures 1 and 3). For example, CD14 and CD163 expression is highly correlated with the increased percentage of monocytes in the peripheral blood of our RA patient population, compared to controls (Figures 2 and 3). Monocytosis is a feature of RA, although primary data to support this are rather sparse in the literature, and recent textbooks often fail to emphasize this point.<sup>11,12</sup> Our data are based on a relatively small sample of controls. However, we have replicated this finding in an independent set of RA patients ( $n = 76$ ) and controls ( $n = 48$ ) using SurroScan technology, and these studies confirmed an approximately 35% increase in the

number of circulating CD14+ monocytes in RA patients compared with controls (data not shown). A very recent study of gene expression profiling in PBMCs also reports an increase in the number of monocytes in a group of RA patients.<sup>13</sup>

A number of studies on the monocyte lineage in RA have focused on the activation of these cells and the appearance of particular subsets in the periphery. The CD14+ CD16+ monocyte subset appears to be increased in RA patients,<sup>14</sup> although the degree of increase is modest and not all studies agree with this finding.<sup>15</sup> CD14+ CD16+ cells have features that are similar to activated tissue macrophages,<sup>14</sup> express altered levels of chemokine receptors and adhesion molecules, and have an enhanced capacity for transendothelial migration. Both subsets (CD16+ and CD16-) of monocytes are able to differentiate into dendritic cells *in vitro*.<sup>16</sup> The binding of small IgG-RF immune complexes to the CD16 monocytes could enhance TNF- $\alpha$  induction in tissue macrophage.<sup>17</sup> We did not specifically analyze for the presence of CD14+ CD16+ cells by flow cytometry. The CD16 transcript was not significantly elevated in the RA samples; this is not surprising, since CD16 is also expressed by several other cell types (natural killer (NK) cells, T cells), and the overall percentage increase in CD14+ CD16+ cells is probably not large. Thus, the expression profile we observed in RA patients reflects, in part, an increase of peripheral blood monocytes and/or their activation state<sup>18</sup> and subset distribution.

The fact that this monocyte-related gene expression profile is so prominent in our data set may be due to the fact that the RA patients we studied had relatively active disease, and were studied just before starting therapy with a new DMARD (either methotrexate or an anti-TNF agent). Some patients were taking methotrexate at the time of blood collection. Methotrexate has been reported to suppress the spontaneous generation of cells that express CD14 from CD14-negative bone marrow progenitor cells in RA patients.<sup>19</sup> Therefore, it is unlikely that monocytosis is related to methotrexate treatment, and in any case, such an association was not observed. In addition, the 14 patients who were taking methotrexate at the time of blood draw did not all cluster together, indicating that methotrexate does not account for the difference seen between RA and control subjects.

Early studies of gene expression in RA have mainly focused on the characterization of gene expression in synovial tissues, leading to the suggestion that disease subgroups can be characterized using this technology.<sup>20</sup> Several groups have also examined gene expression profiles of PBMCs in a variety of autoimmune disorders, including SLE,<sup>21,22</sup> RA<sup>20,23</sup> and multiple sclerosis.<sup>24</sup> With the exception for SLE,<sup>6</sup> these studies have not revealed clear-cut expression 'signatures' of autoimmune disease. Bovin *et al*<sup>13</sup> have recently examined peripheral blood gene expression profiles from a data set of 14 RA patients and seven healthy controls with the aim of identifying differences between rheumatoid factor (RF)-negative and RF-positive patients. No significant differences between these two groups were observed, although 25 genes were reported to distinguish RA patients from controls. Two of these genes, CD14 and S100A12, are present in our set of genes. Despite the likelihood that filtering criteria are quite different in the two studies (details are not given in the Bovin *et al*<sup>13</sup> study), our core data set includes 15 out

of the 25 genes reported by Bovin *et al*. Of these, nine exhibit a nominally significant difference from controls ( $P < 0.05$ ) using our method of analysis. In addition to CD14 and S100A12, these genes include Ribonuclease RNaseA2, Guanine nucleotide binding protein 10, fatty-acid-Coenzyme A ligase, vanin 2, aquaporin 9, transaldolase 1 and leukotriene A4 hydrolase.

In addition to the cluster analysis, we also carried out a logistic regression analysis of our data in order to rank genes according to their ability to classify samples (case *vs* control). In our data set, the top four genes ranked by logistic regression are glutaminyl cyclase, S100 calcium binding protein A12, IL-1Ra and GAB2. These genes are all expressed in monocytes, as well as other cell types, and at least two of them, S100A12 and IL1-RA, have a prominent role in the inflammatory response.

S100A12, also known as calgranulin C or EN-RAGE, is one of a group of S100 calcium binding proteins found predominantly in monocytes and neutrophils. S100A12 is involved in an inflammatory pathway that utilizes an immunoglobulin family receptor, RAGE. Binding of S100A12 to RAGE induces a variety of proinflammatory changes, including cytokine release and enhanced endothelial cell adhesion by neutrophils and monocytes, mediated in part through activation of NF- $\kappa$ B.<sup>25</sup> In addition, blockade of this pathway significantly reduces inflammation in the collagen-induced arthritis (CIA) disease model.<sup>26</sup> Interestingly, a polymorphism (G82S) in the RAGE molecule appears to enhance RAGE signaling by S100A12, and the high-activity (82S) allele is found in strong linkage disequilibrium with DRB1\*0401.<sup>25</sup> It is currently unclear whether the 82S allele confers risk for RA independent of DRB1. Nevertheless, these data emphasize the likely importance of this pathway in the pathogenesis of RA. In addition, soluble S100A12 can be detected in the serum and synovial fluid, and correlates with disease activity.<sup>27</sup> The utility of this protein as a clinical biomarker has not been thoroughly evaluated, although it appears to be useful in following treatment responses in inflammatory bowel disease.<sup>28</sup>

The presence of glutaminyl cyclase and GAB2 in the top set of genes in the logistic regression analysis is unexpected. Glutaminyl cyclase is a metalloenzyme<sup>29</sup> that maps to chromosome 2p22.2 and catalyzes the addition of a glutaminyl residue to pituitary neuroendocrine peptides. In addition to the pituitary, this enzyme is expressed in adrenal gland, pancreatic islet cells and in monocytes (Novartis atlas, <http://symatlas.gnf.org/SymAtlas/>). Its role in monocytes is unknown.

GAB2 is an adaptor molecule that is involved in signaling for cytokines and antigen receptors.<sup>30</sup> Based on mouse knockout experiments, its most important role appears to be in mast cells and allergic responses.<sup>31</sup> Biochemical studies in a number of cell types have implicated GAB2 in signaling pathways involving the intracellular phosphatase SHP-2, and particularly in signaling through Fc receptors and IL-2-dependent responses in T cells.<sup>32</sup> GAB2 is expressed widely in both hematopoietic cells and brain tissue (Novartis atlas, <http://symatlas.gnf.org/SymAtlas/>). Given our recent identification of the intracellular phosphatase PTPN22 as a risk factor for RA,<sup>33,34</sup> it is interesting to speculate whether GAB2 may have some role in regulating phosphatase activities, other than SHP-2, that may be relevant to RA pathogenesis. PTPN22 has been reported

to bind a related adaptor molecule, Grb2, and GAB2 itself contains SH2 and SH3 binding sites that might be involved in these interactions. These observations indicate the need for further investigation of the role of these pathways in monocytes and as well as other cell types.

We also examined our list of differentially expressed genes to determine whether any of them are located in regions of genetic linkage identified in our previous studies of RA affected sibling pairs.<sup>9,10</sup> As shown in Figure 5, there is no clear aggregation of these genes around regions of linkage, and very few of them lie within 5 cM of a linkage peak. Indeed the only examples of this are the genes for integrin family members alpha 7 (ITGA7) and beta 7 (ITGB7) that fall near to a linkage peak on chromosome 12q13. These genes are therefore candidates for future association studies to determine whether polymorphisms in these genes, possibly regulatory, may explain the linkage evidence.

Finally, in addition to providing insights into pathogenesis, gene expression profiling of peripheral blood cells may be useful for the future development of clinically useful biomarkers. The diagnosis of RA is currently based on characteristic clinical manifestations, in the absence of other identifiable causes of inflammatory arthritis, and it requires disease persistence for the diagnosis to be established. Thus, early diagnosis is often provisional. It may be that gene expression studies can assist with early diagnosis, and permit earlier DMARD therapy for some patients. However, some of the genes overexpressed in RA are also likely to exhibit changes in other inflammatory disorders such as SLE<sup>6</sup> or psoriatic arthritis (unpublished data). Thus, it will require a much more comprehensive study of RA and other autoimmune disorders before particular sets of genes can be identified as a useful diagnostic panel. In many cases, it may be more practical to utilize protein assays based on the gene expression data: S100A12 may be an example of such a biomarker. The studies reported here have been carried out as part of the Autoimmune Biomarkers Collaborative Network (ABCOn)—a collaborative research effort to identify biomarkers that are predictive of response to anti-TNF therapy in RA, as well as disease outcome prediction in other disorders, such as SLE (see [www.boasgeneticscenter.org/genetics/Abcon/Abcon.aspx](http://www.boasgeneticscenter.org/genetics/Abcon/Abcon.aspx)). By combining these studies across multiple disease subsets, we hope to identify combinations of biomarkers that can serve as robust indicators of disease subsets, outcome and response to therapy. However, much larger sample sizes will be required before this goal can be achieved.

## Patients and methods

### Study participants

Informed consent was obtained from patients who provided a blood sample. The study included 29 patients with RA, all of whom had active disease and were beginning a new medication at the time of blood draw (seven patients starting methotrexate, 22 patients starting anti-TNF agents). Patient demographic and clinical information is listed in Table 1. Blood was also obtained from a group of age-matched normal control individuals. All of the blood samples were obtained locally and were

**Table 1** RA patient demographics

Gender	
Female	<i>n</i> = 21
Male	<i>n</i> = 8
Mean age	54 years ± 12.6 (range = 22–80 years)
Mean disease duration	12 years ± 12.66 (range 0–43 years)
Joint count	
Tender joints	13.3 ± 7.1
Swollen joints	13.4 ± 5.3
Medications at blood draw	
NSAIDs	<i>n</i> = 17
Prednisone (<10 mg/day)	<i>n</i> = 19
Methotrexate	<i>n</i> = 14

not subjected to the problem of overnight shipment that we have previously shown alters the gene expression profiles in PBMCs.<sup>35</sup> However, some samples were processed a few hours after blood draw. To rule out the possibility of obtaining false positive genes, we deleted the stress response genes that were altered during *ex vivo* handling of samples.<sup>35</sup>

### Sample processing and microarray hybridization

**RA patients and controls.** PBMCs were isolated from whole blood drawn into EDTA or CPT tubes. RNeasy<sup>®</sup> (Ambion Inc., Austin, TX, USA) was added to the PBMCs to stabilize the RNA. RNA was isolated using RNeasy (Qiagen, Valencia, CA, USA). A 5 µg portion of total RNA was used to synthesize cRNA using the Affymetrix expression protocol (Expression analysis technical manual, Affymetrix Inc., Santa Clara, CA, USA). A 10 µg portion of labeled, fragmented cRNA was hybridized to a U95Av2 chip and then stained and scanned.

**Cell subset-specific gene expression.** T cells, neutrophils, NK cells and monocytes were isolated from whole blood by FACS sorting to obtain cell subset-specific gene expression profiles. Blood was collected into three ACD tubes from a healthy donor. Total white cells were separated from most of the RBC using a density gradient Lympholyte-Poly column (Cedarlane Laboratories, Canada), and residual RBCs were removed with RBC lysis buffer (Roche), according to the manufacturer's instructions. This heterogeneous population consists of PBMCs (including T and B lymphocytes and NK cells), granulocytes and monocytes. After blocking with 10% human serum, the cells were stained for 15 min, on ice, with CD3-APC (T cells), CD66B-FITC (neutrophils), CD64-CyC (monocytes) and CD-56 PE (NK cells). Excess stain was washed with cold PBS + 2% FBS and cells were sorted out into four pure populations of T cells, neutrophils, monocytes and NK cells. Four color, four-way sort was carried out using FACS Vantage SE Turbo with FACS Diva option (BD Biosciences). A post-sort purity analysis was carried out. RNA was isolated from the four cell types using the Qiagen RNeasy kit. Labeled cRNA probes were made according to the protocol from Affymetrix and these probes were used to obtain gene expression profiles using Affymetrix Microarray technology.

### Data acquisition and cluster analysis

Affymetrix microarray suite (MAS) 5.0 software was used to obtain gene expression (signal) values for each gene. To permit accurate comparison between chips and to correct for minor variations in the overall intensity of hybridization, each chip was scaled to an intensity of 1500. The Affymetrix U95Av2 arrays have a total of 10 260 genes represented by 12 626 probe sets. A group of 4173 genes are not expressed in PBMCs leaving 6249 genes expressed in PBMCs. A group of 2076 genes that were found to have a high degree of variability (the *ex vivo* stress response genes) were excluded from this data analysis,<sup>35</sup> leaving 4500 genes for the current data analysis.

Three criteria were used to generate lists of genes that were differentially expressed between RA patients and normal control individuals: (i)  $P < 0.001$  by an unpaired Student's *t*-test; (ii) difference in expression of 100 signal units or greater when comparing the means of the two groups; and (iii) a greater than 1.4-fold change in the mean gene expression between the two groups. The expression value for each gene was converted to 'fold differences' by dividing each signal value by the mean signal value of that gene in the control group. The ratios were then  $\log_2$  transformed and hierarchically clustered using the program CLUSTER and visualized with the TREEVIEW software.<sup>36</sup>

### Significance testing

The number of genes ( $N'$ ) out of total  $N$  genes that exceeds a certain significance level  $p$  by chance alone can be derived by the formula  $N' = Np$ . In order to confirm the validity of the formula, a permutation analysis was carried out, in which the sample labels were reshuffled and the gene selection by the same significance level threshold,  $p$ , was repeated. The number of genes that exceeded the  $p$  threshold (ie the number of false positives) in these permutations was then averaged over the number of permutations.

### Logistic regression analysis

Logistic regression is a statistical discriminant model, in which the probability for a sample to be in one of the two classes is given by an equation ('model') that contains adjustable parameters. We utilized the equation<sup>37</sup>

$$\text{Prob}(\text{sample label} = \text{RA}) = \frac{1}{1 + e^{-a - (cx)}}$$

where  $a$  and  $c$  are the two adjustable parameters in the model and  $x$  is a quantity related to the mRNA expression level (in this case, the logarithm of the fluorescence intensity).

After the training, the two parameters choose the value  $a = \hat{a}$  and  $c = \hat{c}$ . In order to measure the classification performance of the fitted statistical model, we determined the 'deviance', defined as the sum of squares of residuals (difference between the fitted-model-predicted and the true sample label value). The smaller the deviance, the better the model fits the data. Since at the expression level  $x = -\hat{a}/\hat{c}$ , the probability of the sample label to be both RA and control, according to the model, is  $1/2$ ,  $-\hat{a}/\hat{c}$  is the threshold value. If the model classifies the data perfectly, the expression levels of RA samples and control samples would be on two sides of the threshold

value, respectively, without exception. Logistic regression has been applied previously to microarray data analysis.<sup>38–40</sup>

### Cellular assays

Immunophenotyping to obtain percentages of cellular subsets in PBMCs was completed by MLSC on the SurroScan™ platform (SurroMed, Menlo Park, CA, USA).<sup>41–43</sup> Monoclonal antibodies and fluorescent tags were obtained from commercial vendors (BD Biosciences, including BD PharMingen, San Jose, CA, USA; Beckman Coulter, Miami, FL, USA; Serrotec, Raleigh, NC, USA; eBiosciences, San Diego, CA, USA; Amersham Biosciences, Piscataway, NJ, USA). Three different fluorophores, Cy5, Cy5.5 and the tandem dye Cy7-APC<sup>44–46</sup> were used as direct conjugates to monoclonal antibodies specific for different cellular antigens in each assay. The antibody-dye reagents were combined into pre-made cocktails. PBMC samples from patients ( $n = 24$ ) and controls ( $n = 8$ ) previously frozen in FBS with 10% DMSO were thawed and checked for cell viability ( $> 95\%$ ) with Trypan blue. Three million cells were suspended in 1 ml of buffer, and 20  $\mu$ l of the cell suspension per well added to 96-well microtiter plates containing the appropriate reagent cocktails for 32 separate assays and incubated in the dark at room temperature for 20 min. All assays were homogeneous without removal of unreacted antibody reagents. After the incubation, samples were diluted with sample diluent buffer, loaded into Flex32™ capillary arrays (SurroMed) and analyzed with SurroScan™ (SurroMed). Images were converted to flow cytometry standard format with in-house software<sup>41</sup> and analyzed with FlowJo™ cytometry analysis (Tree Star Inc., Ashland, OR, USA). Fluorescence intensities were compensated for spectral overlap of the dyes, so values would be proportional to antigen density. Percentage of cells expressing leukocytes, monocytes, B-cell and T-cell subset markers was determined.

### Acknowledgements

We thank Leslie Goodwin and Sarah Lombardi for assistance with the microarray assays, and Jen Deng, Harini Govindarajan, V Kakkanaiah and Chris Todd for support with SurroScan assays. We also thank Robert Lundsten, Jubal Dais and Ismael Rodriguez for database support by the NSLIJHS Biorepository Informatics Group. We are grateful to Cerdi Beltre for assistance with patient recruitment, and to all the RA patients who agreed to participate in this study. This work was supported by a NIAMS contract NO1-AR-1-2256 (PKG).

### References

- 1 Gregersen PK. Teasing apart the complex genetics of human autoimmunity: lessons from rheumatoid arthritis. *Clin Immunol* 2003; **107**: 1–9.
- 2 Gulko PS, Winchester RJ. Rheumatoid arthritis. In: Frank Austen K, Frank MM, Atkinson JP, Cantor H (eds). *Samter's Immunologic Diseases*. Lippincott Williams and Wilkins: Philadelphia, PA, 1995, pp 427–463.

- 3 Weinblatt ME, Kremer JM, Bankhurst AD *et al*. A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. *N Engl J Med* 1999; **340**: 253–259.
- 4 Lipsky PE, van der Heijde DM, St Clair EW *et al*. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. *N Engl J Med* 2000; **343**: 1594–1602.
- 5 Staudt LM. Gene expression profiling of lymphoid malignancies. *Annu Rev Med* 2002; **53**: 303–318.
- 6 Baechler EC, Batliwalla FM, Karypis G *et al*. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc Natl Acad Sci USA* 2003; **100**: 2610–2615.
- 7 Finnin M, Hamilton JA, Moss ST. Characterization of a CSF-induced proliferating subpopulation of human peripheral blood monocytes by surface marker expression and cytokine production. *J Leukoc Biol* 1999; **66**: 953–960.
- 8 Lohn M, Mueller C, Langner J. Cell cycle retardation in monocytoid cells induced by aminopeptidase N (CD13). *Leuk Lymphoma* 2002; **43**: 407–413.
- 9 Jawaheer D, Seldin MF, Amos CI *et al*. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am J Hum Genet* 2001; **68**: 927–936.
- 10 Jawaheer D, Seldin MF, Amos CI, *et al*, North American Rheumatoid Arthritis Consortium. Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis Rheum* 2003; **48**: 906–916.
- 11 Harris ED. Clinical features of rheumatoid arthritis. In: Kelley WN, Ruddy S, Harris Jr ED, Sledge CB (eds). *Textbook of Rheumatology*, 5th edn. WB Saunders Company: Philadelphia, 1997, pp 898–932.
- 12 Fuchs HA, Sargent JS. Rheumatoid arthritis: the clinical picture. In: Koopman WJ (ed). *Arthritis and Allied Conditions*, 13th edn. Williams and Wilkins: Baltimore, 1997, pp 1041–1070.
- 13 Bovin LF, Rieneck K, Workman C *et al*. Blood cell gene expression profiling in rheumatoid arthritis. Discriminative genes and effect of rheumatoid factor. *Immunol Lett* 2004; **93**: 217–226.
- 14 Kawanaka N, Yamamura M, Aita T *et al*. CD14+, CD16+ blood monocytes and joint inflammation in rheumatoid arthritis. *Arthritis Rheum* 2002; **46**: 2578–2586.
- 15 Cairns AP, Crockard AD, Bell AL. The CD14+ CD16+ monocyte subset in rheumatoid arthritis and systemic lupus erythematosus. *Rheumatol Int* 2002; **21**: 189–192.
- 16 Geissmann F, Jung S, Littman DR. Blood monocytes consist of two principal subsets with distinct migratory properties. *Immunity* 2003; **19**: 71–82.
- 17 Abrahams VM, Cambridge G, Lydyard PM, Edwards JC. Induction of tumor necrosis factor alpha production by adhered human monocytes: a key role for Fc gamma receptor type IIIa in rheumatoid arthritis. *Arthritis Rheum* 2000; **43**: 608–616.
- 18 Stuhlmuller B, Ungethum U, Scholze S *et al*. Identification of known and novel genes in activated monocytes from patients with rheumatoid arthritis. *Arthritis Rheum* 2000; **43**: 775–790.
- 19 Hirohata S, Yanagida T, Hashimoto H, Tomita T, Ochi T. Suppressive influences of methotrexate on the generation of CD14(+) monocyte-lineage cells from bone marrow of patients with rheumatoid arthritis. *Clin Immunol* 1999; **91**: 84–89.
- 20 van der Pouw Kraan TC, van Gaalen FA, Huizinga TW, Pieterman E, Breedveld FC, Verweij CL. Discovery of distinctive gene expression profiles in rheumatoid synovium using cDNA microarray technology: evidence for the existence of multiple pathways of tissue destruction and repair. *Genes Immun* 2003; **4**: 187–196.
- 21 Bennett L, Palucka AK, Arce E *et al*. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J Exp Med* 2003; **97**: 711–723.
- 22 Crow MK, Wohlgemuth J. Microarray analysis of gene expression in lupus. *Arthritis Res Ther* 2003; **5**: 279–287.
- 23 Maas K, Chan S, Parker J *et al*. Cutting edge: molecular portrait of human autoimmune disease. *J Immunol* 2002; **169**: 5–9.
- 24 Bompreszi R, Ringner M, Kim S *et al*. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Hum Mol Genet* 2003; **12**: 2191–2199.
- 25 Foell D, Roth J. Proinflammatory S100 proteins in arthritis and autoimmune disease. *Arthritis Rheum* 2004; **50**: 3762–3771.
- 26 Hofmann MA, Drury S, Hudson BI *et al*. RAGE and arthritis: the G82S polymorphism amplifies the inflammatory response. *Genes Immun* 2002; **3**: 123–135.
- 27 Rouleau P, Vandal K, Ryckman C *et al*. The calcium-binding protein S100A12 induces neutrophil adhesion, migration, and release from bone marrow in mouse at concentrations similar to those found in human inflammatory arthritis. *Clin Immunol* 2003; **107**: 46–54.
- 28 Foell D, Kucharzik T, Kraft M *et al*. Neutrophil derived human S100A12 (EN-RAGE) is strongly expressed during chronic active inflammatory bowel disease. *Gut* 2003; **52**: 847–853.
- 29 Schilling S, Niestroj AJ, Rahfeld JU *et al*. Identification of human glutamyl cyclase as a metalloenzyme. Potent inhibition by imidazole derivatives and heterocyclic chelators. *J Biol Chem* 2003; **278**: 49773–49779.
- 30 Nishida K, Hirano T. The role of Gab family scaffolding adapter proteins in the signal transduction of cytokine and growth factor receptors. *Cancer Sci* 2003; **94**: 1029–1033.
- 31 Nishida K, Wang L, Morii E *et al*. Requirement of Gab2 for mast cell development and KitL/c-Kit signaling. *Blood* 2002; **99**: 1866–1869.
- 32 Arnaud M, Crouin C, Deon C, Loyaux D, Bertoglio J. Phosphorylation of Grb2-associated binder 2 on serine 623 by ERK MAPK regulates its association with the phosphatase SHP-2 and decreases STAT5 activation. *J Immunol* 2004; **173**: 3962–3971.
- 33 Begovich AB, Carlton VE, Honigberg LA *et al*. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004; **75**: 330–337.
- 34 Lee AT, Li W, Liew A *et al*. The PTPN22 R620W polymorphism associates with RF positive rheumatoid arthritis in a dose-dependent manner but not with HLA-SE status. *Genes Immun* 2005; **6**: 129–133.
- 35 Baechler EC, Batliwalla FM, Karypis G *et al*. Expression levels for many genes in human peripheral blood cells are highly sensitive to *ex vivo* incubation. *Genes Immun* 2004; **5**: 347–353.
- 36 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
- 37 Dyke GV, Patterson HD. Analysis of factorial arrangement when the data are proportional. *Biometrics* 1952; **8**: 1–12.
- 38 Eilers PH, Boer JM, van Ommen GJ, van Houwelingen HC. Classification of microarray data with penalized logistic regression. *Proc SPIE* 2001; **4266**: 187–198.
- 39 Li W, Yang Y. How many genes are needed for a discriminant microarray data analysis. In: Lin SM, Johnson KF (eds). *Methods of Microarray Data Analysis*. Kluwer Academic: Boston/Dordrecht/London, 2002, pp 137–149.
- 40 van't Veer LJ, Dai H, van de Vijver MJ *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**: 530–536.

- 41 Walton ID, Dietz LJ, Frenzel G *et al*. Microvolume laser scanning cytometry platform for biological marker discovery. *Proc SPIE Int Soc Opt Eng IBOS Soc Photo-Opt Instrum Eng* 2000; **3926**: 192–201.
- 42 Kantor AB, Alters SE, Cheal K, Dietz LJ. Immune systems biology: immunoprofiling of cells and molecules. *BioTechniques* 2004; **36**: 520–524.
- 43 Kantor AB, Wang W, Lin H *et al*. Biomarker discovery by comprehensive phenotyping for 2 autoimmune diseases. *Clin Immunol* 2004; **111**: 186–195.
- 44 Mujumdar RB, Ernst LA, Mujumdar SR, Lewis CJ, Waggoner AS. Cyanine dye labeling reagents: sulfoindocyanine succinimidyl esters. *Bioconjug Chem* 1993; **4**: 105–111.
- 45 Roederer M, Kantor AB, Parks DR, Herzenberg LA. Cy7PE and Cy7APC: bright new probes for immunofluorescence. *Cytometry* 1996; **24**: 191–197.
- 46 Beavis AJ, Pennline KJ. Allo-7: a new fluorescent tandem dye for use in flow cytometry. *Cytometry* 1996; **24**: 390–395.

Supplementary information accompanies the paper on Genes and Immunity website (<http://www.nature.com/gene>).