Kidney International, Vol. 65 (2004), pp. 420–430

CELL BIOLOGY – IMMUNOLOGY – PATHOLOGY

Gene expression profiles of circulating leukocytes correlate with renal disease activity in IgA nephropathy

Gloria A. Preston,[1] Iwao Waga,[1] David A. Alcorta, Hitoshi Sasai, William E. Munger, Pamela Sullivan, Brian Phillips, J. Charles Jennette, and Ronald J. Falk

Department of Medicine, Division of Nephrology and Hypertension, Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; Japan Tobacco, Inc., Yokohama, Kanagawa, Japan; and Gene Logic, Inc., Gaithersburg, Maryland

Gene expression profiles of circulating leukocytes correlate with renal disease activity in IgA nephropathy.

Background. The goal of these studies was to explore the possibility of using gene expression profiles of circulating leukocytes as a functional fingerprint of nephritic disease activity.

Methods. This feasibility study utilized IgA nephropathy (IgAN) as a model system. Genes differentially expressed in IgAN patients were identified by Affymetrix GeneChip® microarrays, and compared with gene expression of focal segmental glomerulosclerosis (FSGS), minimal change disease, antineutrophil cytoplasmic antibody (ANCA) glomerulonephritis, and with healthy volunteers. Of the genes identified, 15 transcriptionally up-regulated were validated in a larger cohort of patients using TaqMan® polymerase chain reaction (PCR). To test whether increased expression of these genes correlated with disease activity, cluster analyses were performed utilizing the TaqMan® PCR values. Taking a mathematical approach, we tested whether gene expression values were correlative with kidney function, as reflected by serum creatinine and creatinine clearance values.

Results. We identified 15 genes significantly correlative with disease activity in IgAN. This gene signature of IgAN patients' leukocytes reflected kidney function. This was demonstrated in that mathematically generated theoretical values of serum creatinine and creatinine clearance correlated significantly with actual IgAN patient values of serum creatinine and creatinine clearance. There was no apparent correlation with hematuria and proteinuria. The expression levels of this same gene set in ANCA glomerulonephritis or Lupus nephritis patients were not correlative with serum creatinine or creatinine clearance values.

Conclusion. These data indicate that leukocytes carry informative disease-specific markers of pathogenic changes in renal tissue.

[1]These two authors contributed equally to this work.

Key words: IgA nephropathy, microarrays, serum creatinine, peripheral leukocytes.

Received March 25, 2003
and in revised form June 5, 2003, August 14, 2003, and August 19, 2003
Accepted for publication August 25, 2003

Disease diagnosis and evaluation of disease severity using gene expression patterns are quickly becoming reality rather than conjecture. Indeed, the fields of bioinformatics and molecular epidemiology have emerged as methods to utilize genomics data derived from large-scale sequencing efforts such as the Human Genome Project [1]. Microarrays are now in use to fingerprint biologic or pathologic processes [2, 3]. The value of evaluating gene expression profiles has been documented in animal models of disease. For example, gene expression in leukocytes can distinguish among a variety of experimental cerebral disorders [4]. Despite the enormous promise of this revolutionary technology, its practical application and relevance to the clinical arena has not been widely studied. If patterns of altered gene expression can be established using white blood cells of patients, this could lead to a new approach to diagnostics that is minimally invasive. This prompted us to take the first steps in investigating whether or not gene expression patterns of circulating leukocytes could be used as functional fingerprints of events that occur in the kidney of patients with IgA nephropathy (IgAN).

Here, we test the hypothesis that changes in gene expression patterns in circulating leukocytes of patients with the nephritic syndrome IgAN will correlate with renal disease activity. IgAN, a condition recognized worldwide as one of the most common primary glomerulonephropathies, is characterized by the presence of IgA in the glomerular mesangium [5, 6]. Glomerular filtration and renal blood flow often is progressively impaired by advancing glomerular and tubulointerstitial injury [7]. To screen for genes pertinent to disease activity in IgAN patients, we performed microarray chip profiling on peripheral blood leukocytes and compared the results with those from healthy volunteers and with results from three other glomerular diseases. A subset of the identified genes were verified in a larger cohort of patients and controls using TaqMan® polymerase chain reaction (PCR). Cluster analysis of these data demonstrates a correlation between

gene transcript levels and disease activity. Moreover, we show that gene expression changes reflect serum creatinine and creatinine clearance values in these patients, as determined by multiple regression analyses.

METHODS

Patients

The diagnosis of primary IgAN was based on renal biopsy findings of IgA-dominant or codominant mesangial immune deposits by immunofluorescence microscopy. Measures of disease activity included microscopic hematuria (0 to 4+), 24-hour protein excretion, serum creatinine, and calculated glomerular filtration rate (GFR) (Cockcroft and Gault's equation). The pathology indices included the degree of glomerular mesangial and endocapillary proliferation and the degree of glomerular or interstitial fibrosis, each scored on a scale of 0 (none) to 4+ (severe). An arbitrary classification scheme that closely corresponds with our clinical practice was devised by two of us (R.J.F. and J.C.J.), which incorporates pathology and clinical disease indicators. Classification of degree of clinical disease activity was as follows: Mild disease activity was defined as the presence of hematuria with <1 g proteinuria with normal renal function, with minimal evidence of scaring on biopsy. Moderate disease activity was defined as the presence of hematuria with >1 g proteinuria and stable renal function (i.e., a stable serum creatinine), with some glomerular scarring and interstitial fibrosis. The severe disease category was defined as renal insufficiency with or without hematuria and >1 g of proteinuria, with marked glomerular scarring and interstitial fibrosis, but also included patients with substantial hematuria and acute renal failure. Patients were considered in remission if there was no evidence of hematuria, <1 g of proteinuria and no evidence on renal biopsy of active glomerular inflammation (i.e., hypercellularity, necrosis or cellular crescents). Patients diagnosed with IgAN enrolled in the study ($N = 22$) (22 white; 11 males and 11 females) ranged in age from 7 to 79 years. Of these, five patients donated a second sample later totaling 27 samples analyzed. Eight of the RNA samples were consumed in the microarray chip analyses. The remaining 19 samples were utilized for TaqMan® PCR analyses. Only one sample (IgA 2a) was sufficient in amount for both microarray and TaqMan® PCR analyses. The time lapse between biopsy and leukocyte had a median of 10 months with a mean of $14\frac{1}{2}$ months and a range of 1 week to 42 months. Healthy volunteers ($N = 32$) (32 white; 17 males and 15 females) ranged in age from 18 to 46 years. Patients diagnosed with antineutrophil cytoplasmic antibody (ANCA) glomerulonephritis ($N = 24$) (22 white and two African Americans; 13 males and 11 females), systemic lupus erythematosus (SLE) ($N = 18$) (seven white, nine African American, two Asian; one male and 17 females), focal segmental glomerulosclerosis (FSGS) ($N = 5$) (two white and three African American; three males and two females), and minimal change disease ($N = 6$) (two white, three African American, and one Asian; four males and two females) were included in the study for disease-related comparisons.

Leukocyte isolation

Approximately 20 mL of blood was drawn into four 7 mL ethylenediaminetetraacetic acid (EDTA) vacutainers. Leukocytes were isolated from whole blood by lysis of red blood cells by incubation for 11 minutes in a hypotonic ammonium chloride solution at a 9-to1 ratio. Following a wash with 1× Hank's balanced salt solution (HBSS), the cells were resuspended in RNA Stat 60 at a concentration of $\sim 10^7$ cells/mL (Tel Test, Inc., Friendswood, TX, USA). The RNA Stat 60/cell solution was stored in −70°C for up to 1 week before RNA processing.

RNA isolation procedure

As per protocol from Tel Test, Inc., RNA was isolated from the cell/RNA Stat 60 mixtures by addition of 0.2 mL of chloroform to remove proteins. The aqueous phase containing the RNA was isolated by centrifugation, and the RNA precipitated with 0.5 mL of isopropanol. The RNA pellet was washed once with 75% ethanol and then resuspended in 100 μL of nuclease-free water (Promega, Madison, WI, USA). As per RNeasy protocol (Qiagen, Valencia, CA, USA), RNA solution was applied to column for purification, DNase treated for 15 minutes and eluted in 30 to 50 μL of nuclease-free water. RNA was treated with RNA secure (Ambion, Austin, TX, USA) after both the initial resuspension in nuclease-free water and after the column elution. RNA was quantified and purity determined by obtaining the absorbance at 260 nm and 280 nm using spectrophometric methods. The RNA integrity was determined by visualization of the 28S and 18S RNA bands using 0.5 μg of RNA on a 1% agarose gel stained with Sybr Gold (Molecular Probes, Eugene, OR, USA). RNA was then stored at −70°C.

Microarray data processing

Affymetrix Human 60KMicroarray GeneChip® of genes microarrays were utilized for identification of genes differentially expressed (methodologies described in detail by Affymetrix, Palo Alto, CA, USA). This procedure was performed at Gene Logic, Inc., as previously described [8]. Briefly, hybridization to the Affymetrix GeneChip® HuGeneFL array, and raw data collection was done exactly as described by Tackels-Horne et al [9]. The raw data were analyzed with Affymetrix software, GeneChip® version.3.0 and Experimental data mining tool version1.0. S-Plus was used to perform the analysis

of variance (ANOVA) principal component analysis (PCA) as previously described [8], and hierarchical clustering analyses identified ~341 genes as differentially regulated (153 up and 188 down) in IgAN patients. Arrays were globally scaled to an average intensity of 2500. A value of 200 was assigned to all intensity measurements below 200 before differences in intensities were calculated. Three parameters were used for analysis, average difference intensity change, difference call, and fold change. Gene expression levels that varied less than 1.5-fold relative to the biologic base line or had a difference call of "no change," as determined by the GeneChip® algorithms, were considered unchanged. Further analysis was based on those genes whose expression levels changed between IgA patients and controls. Known constitutively expressed genes were used to normalize the data from different microarray experiments. Briefly, the expression levels of the selected genes were scaled to a "standard experiment" and the geometric mean of the scaling factors was calculated. This value served as the normalization factor for all genes represented on the microarray. Genes were clustered within categories using the Statistica, Gene Cluster, Treeview programs [10].

TaqMan® PCR quality control and analyses

TaqMan® PCR was performed to validate the microarray chip results and to examine a subset of these genes in a larger cohort of patients. TaqMan® primers and probes were designed using Primer Express software (version 1.0) or version 1.5 (Macintosh) (Applied Biosystems, Foster City, CA, USA). TaqMan® PCR reactions were performed in MicroAmp Optical 384-well Reaction Plates. Fluorescence emission was monitored using the ABI Prism 7900 Sequence Detection System, and this information was automatically converted to amplification plots using the ABI Prism 7900 Sequence Detection System software. The log RNA concentration versus the cycle number at specific threshold (Ct) value was plotted. A line was fit to the data and the slope and the linear regression values were determined. Only primer sets that met the following criteria were used for quantitative reverse-transcription (RT)-PCR (TaqMan®) analysis: (*1*) a slope value between −3.0 and −3.7, (*2*) a linear regression value of 0.90 or greater, (*3*) a no template control (NTC) Ct value above 35, and (*4*) an NTC Ct no closer than 3 Cts from the lowest concentration of RNA analyzed. Only sample RNAs with Ct values within the linear range as determined by this qualification process were accepted. Disease specific RNA samples were used for TaqMan® PCR analyses. To circumvent inter-plate variability, we tested one gene per plate using all RNA samples. Fold-change in expression was determined by the $\Delta\Delta$Ct method: Briefly, the ΔCt value was determined by subtracting the Ct value for the housekeeping gene cytochrome C oxidase from the Ct value for the gene of interest. Cytochrome C usage as a normalizing gene was based on comparisons of microarray data analyses of a battery of housekeeping genes, which showed cytochrome C expression to be relatively constant among the samples, whereas glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was not (data not shown). Arbitrarily, a control sample was selected and the remainder of samples was adjusted by subtracting that ΔCt value. Then, the fold-change was calculated using the following equation: fold-change $= 2(-\Delta\Delta Ct)$ and the mean of these calculated values was determined. The differences in expression of genes in patients' samples are calculated by dividing the actual fold-change by the adjusted mean of the normals to give a relative expression level. The Student *t* test was performed on adjusted normal and IgAN-fold expression change values to determine significance ($P < 0.05$). Duplicate patient samples were omitted from statistical analyses.

Patients were divided into two groups, mild/moderate and severe, and expression levels were analyzed to determine statistical difference from normal controls using ANOVA and ranked ANOVA analyses for significant differences ($P < 0.05$). Only genes that were different by either Student *t* test or the ANOVA analyses were selected for further study. All statistical analyses were performed using SAS statistical program (SAS, Durham, NC, USA).

The quantitative values of Q-RT-PCR results were subjected to cluster analysis performed by computational calculations with Statistica (Statsoft, Inc., Tulsa, OK, USA) and Excel (Microsoft, Redmond, WA, USA).

Regression and correlation analyses

For regression modeling, we fit the data to develop a formula that consists of a sum of predictor effects, with each predictor coming from a covariate. We applied statistical methods to find relationships between disease-related gene expression values and clinical parameters using a standard computational spread sheet program [11–14]. To select the disease-related genes as independent predictor variables, we used the *Ru* value:

$$Ru = 1 - (1 - R^2)(n + k + 1)/(n - k - 1)$$

where R is the multiple correlation coefficient, n is fold-change TaqMan® PCR value, and k is the degree of freedom of regression. The multiple regression procedure was used to estimate a linear equation for identification of gene combinations yielding a *Ru* value closest to 1:

$$Y = a + b1 * X1 + b2 * X2 + \cdots + bp * Xp$$

where Y is either serum creatinine or creatinine clearance, $X1, X2 \cdots XP$ are the TaqMan® PCR values, a is the fixed value and $b1, b2 \cdots bp$ are the regression coefficients

for each gene. Transcript levels outside of normal values were identified by comparing patient TaqMan® PCR values to the mean value of 20 healthy volunteers. The efficiency index of a particular gene is the range of calculated fold-change values (max-min value of TaqMan® PCR data) times the coefficient for curve fit from the multiple regression analysis. *Ep* was calculated using as follows:

$$Ep = bp\ x$$

RESULTS

Identification of genes differentially expressed in IgAN

Our experimental strategy was to first identify genes differentially expressed in IgAN using microarray chip technology and then to utilize this information to select a subset of IgAN-up-regulated genes for testing by quantitative TaqMan® PCR in a larger cohort of patients and controls. Patient selection for microarray chip studies included representatives of mild, moderate, and severe (Table 1). To aid in identification of transcripts specifically altered in IgAN ($N = 9$), comparisons were made with healthy volunteers (normals) ($N = 12$), patients with ANCA glomerulonephritis ($N = 5$), FSGS ($N = 5$) and minimal change disease ($N = 6$). We identified 153 transcriptionally up-regulated and 188 down-regulated genes. For example, interleukin-8 (IL-8) was found to be significantly increased 2.42-fold in IgAN (Fig. 1).

From a list of identified, up-regulated genes (fold-increase 1.5- to 3.6-fold) in IgAN patients (Table 2), we selected 14 statistically significant genes by ANOVA and ranked ANOVA analyses ($P < 0.05$) for further study, with addition of galectin 3 because of its carbohydrate-binding capacity. Limiting the number of genes is necessary for additional analyses and for the mathematical formula outputs of this study. Three sets of PCR primers for each gene were developed and characterized for proficiency and specificity. Transcript levels of these 15 genes were analyzed by TaqMan® PCR in 15 new IgAN patients (three of which donated a sample twice within the year) in addition to IgA 2a, whose RNA sample was sufficient to use in both microarray and TaqMan ® PCR analyses ($N = 19$ samples). Table 3 lists the differences in expression of genes in patients' samples, which were calculated by dividing the actual fold-change by the adjusted mean of the normals to give a relative expression level. The results confirmed the microarray results, thus verifying that these particular genes are abnormally expressed under the pathologic conditions of IgAN in some patients, when compared to the mean expression levels of healthy volunteers.

Increased expression clusters with IgAN patients

To address the issue of whether this gene profile correlates with disease activity, we performed cluster analyses, based on TaqMan® PCR fold-change data (Table 3). These hierarchical clustering algorithms provide a general view of the association of the genes with the clinical parameters of IgAN. As depicted in Figure 2, the clustering output highlights two primary clusters, one with the majority of IgAN patients. Lower levels of the dendrogram (the shorter branches denoting higher degrees of similarity) revealed four distinct subgroups**.** Interestingly, clinical parameters of proteinuria and hematuria do not appear to dictate clustering patterns (Fig. 2), nor does age, gender, or race (all subjects in this set of experiments were white). Instead, we observed a general linkage with disease activity. We found that some normal individuals clustered with IgAN patients (normals 1, 2, 3, and 4), and some patients clustered with the normals (IgA 11, IgA 2b, IgA 8, IgA 5b, and IgA 5a). This could not be explained by any of the known variables. However, it is remarkable that replicate samples from the same patient, whose

Table 1. Leukocyte donors for microarray chip analyses

Patient	Age/gender/ race	Serum creatinine	Disease activity
N1679	22/M/W	1.1	NA
N1680	26/F/W	1.1	NA
N1708	44/F/W	1.1	NA
N3538	37/F/W	1.1	NA
N3539	20/M/W	1.1	NA
N3543	20/M/W	nd	NA
N3544	24/M/W	nd	NA
N5094	42/F/W	nd	NA
N5095	31/F/W	nd	NA
N5096	22/M/W	nd	NA
N5097	30/F/W	nd	NA
N5098	32/M/W	nd	NA
IgA 1b (1183)	42/M/W	1.3	Remission
IgA 19 (4205)	33/F/W	1.0	Mild
IgA 21 (5089)	50/F/W	1.1	Mild
IgA 4a (3540)	21/M/W	1.0	Moderate
IgA 20 (5092)	32/M/W	1.9	Moderate
IgA 2a[a](5090)	17/M/W	0.9	Moderate
IgA 24 (4204)	33/F/W	2.6	Severe
IgA 22 (5091)	32/F/W	1.7	Severe
IgA 25 (5093)	20/M/W	2.9	Severe
ANCA 1180	63/F/W	6.4	Severe
ANCA 1648	54/F/W	2.2	Moderate
ANCA 1650	79/M/W	4.6	Severe
ANCA 2165	50/F/W	0.8	Moderate
ANCA 3534	75/M/W	1.8	Moderate
FSGS 1649	44/F/B	2.9	Severe
FSGS 1652	14/M/B	1.0	Moderate
FSGS 2166	28/F/W	0.7	Moderate
FSGS 2269	7/M/B	0.5	Moderate
FSGS 2270	31/M/W	2.3	Moderate
MC 1647	15/F/B	1.4	Severe
MC 1651	38/M/B	1.0	Moderate
MC 1678	25/M/W	0.9	Severe
MC 2267	53/M/W	1.0	Moderate
MC 2268	20/F/B	1.2	Moderate
MC 3535	25/M/A	0.9	Severe

Abbreviations are: ANCA, antineutrophil cytoplasmic antibodies; FSGS, focal segmental glomerulosclerosis; MC, minimal change; NA, not applicable.

[a]This patient sample was used in both microarray and TaqMan® polymerase chain reaction (PCR) analyses. Samples collected from the same patient over the course of disease are indicated (a and b).

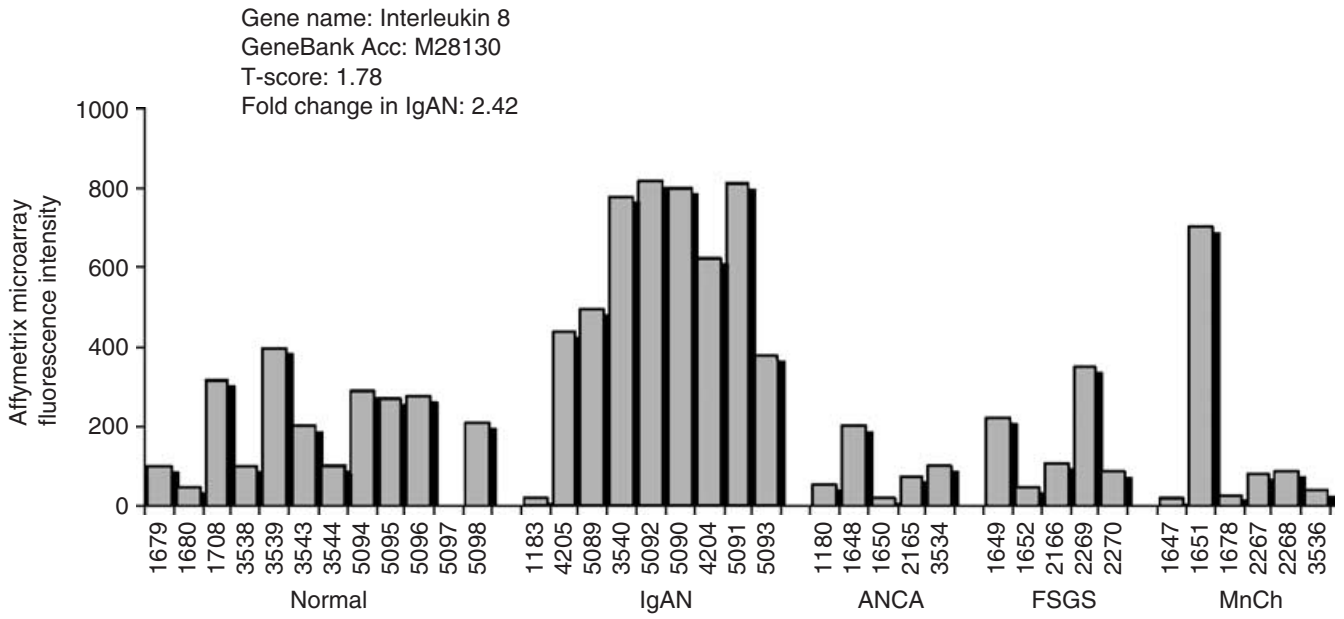

Fig. 1. An example of results from microarray chip analyses indicating that interleukin-8 (IL-8) transcript levels were statistically higher in patients with IgA nephropathy (IgAN), as compared to normals and patients with antineutrophil cytoplasmic antigen (ANCA), focal segmental glomerular sclerosis (FSGS), and minimal change (MnCh) nephritic disorders. Patient characteristics are given in Table 1.

disease status had not changed, clustered tightly together as observed with IgA 16a and b. The same applies for IgA 5a and b in cluster F.

The gene signature of IgAN patients' leukocytes reflects kidney function

We took our analyses a step further and asked if we could correlate the gene expression data with kidney function determined by measurements of serum creatinine concentrations. We first performed multiple regression analyses to generate a mathematical model for approximation of serum creatinine concentration. Analyses of reliability of regression (see **Methods** section) grouped the genes that yielded a *Ru* value of 0.854, indicating that genes *BTG2*, *NCUBE1*, *FLJ2948*, *SRPK1*, *LZS*, *GIG-2*, and *IL-8* correlate mathematically with serum creatinine levels (Fig. 3A). Using these genes a formula was developed: serum creatinine = 1.631676 + (0.198639 × *BTG2*) + (−0.69285 × *NCUBE*) + (0.026568 × *FLJ294*) + (0.41222 × *SRPK1*) + (0.35276 × *LZS*) + (−0.16758 × *GIG-2*) + (−0.14398 × *IL-8*). Substitution of individual TaqMan® PCR values for each of these seven genes, a theoretical serum creatinine level was calculated for each of 18 patients. Graphic representation of the calculated value versus the actual clinical value indicates that expression of these genes is correlative with serum creatinine values (Fig. 3B).

Not all overexpressed genes reflect a detrimental effect. Some overexpressed genes may have a positive effect, when considering the patient's recovery process, or a protective effect when considering disease progression. We calculated the efficiency index of each of the seven genes described above. An efficiency index value of zero would indicate that expression of that gene does not depend on serum creatinine concentration. In contrast, a positive or negative value would indicate that the expression increases or decreases linearly with changes in serum creatinine concentration. Higher levels of *BTG-2* correspond to higher levels of serum creatinine, while *NCUBE1* is inversely proportional (Fig. 3C).

Next, we asked how well the IgAN gene expression data correlate with creatinine clearance. Creatinine clearance takes into account the patient's age, gender, and size. The mathematical iterations described above were performed to select genes from the group of 15 that gave the best *Ru* value, 0.773 (Fig. 4A). The eight genes identified, *PMAIP1*, *B3GNT5*, *SRPK1*, *SSI-3*, *LZS*, *GIG-2*, *AXUD1*, and *PTGS2*, were used to generate a formula that would provide a theoretical creatinine clearance value; creatinine clearance = 25.52501 + (31.07027 × *PMAIP*) + (−35.7845 × *B3GNT*) + (52.24351 × *SRPK1*) + (8.056434 × *SSI-3*) + (33.42925 × *LZS*) + (−26.3692 × *GIG-2*) + (−23.037 × *AXUD1*) + (47.70423 × *PTGS2*). The actual creatinine clearance values (ranging from ~168.8 to 48.0) and the theoretic values were comparative in every case (Fig. 4B), with the largest deviation seen in patient IgA 9 with values of 57.9/actual versus 80.4/theoretic.

Analysis of efficiency index for each gene indicate that *PMAIP1*, *SRPK1*, *SSI-3*, *LZS*, and *PTGS2* are linearly correlated with higher creatinine clearance values, implying that these genes may provide a protective effect, while

Table 2. Top known genes up-regulated in IgA nephropathy (IgAN) patients' leukocytes

Name	Accession	Name	Accession
Transcription factor		Signaling	
►Egr 1 (early growth response)	X52541	M AP KKK3 (kinase)	U78876
Egr 2	AA486027	SGK kinase	R97759
►GOS2 (lymphocyte G_0/G_1 gene)	T52813	GADD34	AA251320
KIAA1100	AA411433	OSR1 (oxidative-stress responsive 1)	AA039663
Transforming protein fos-B	L49169	Ubiquitin enzyme 7 interacting protein 3	AA447671
MYB binding protein (P160) 1a	N49846	HM74 putative chemokine receptor	D10923
		►NCUBE1	AA256528
Proteases and inhibitors		NUP98 (nucleoporin 98 kD)	AA505118
Cathepsin K	T67463	SREBP cleavage-activating protein	D83782
SLPI (antileukoproteinase)	AA026641	SOS1 (son of sevenless homolog 1)	W74256
►LYZ (lysozyme)	M21119	GRN (granulin)	X62320
		►FLJ22948 fis, clone KAT09449	AA447740
Membrane associated		►AXUD1 (AXIN1 up-regulated 1)	T16484
ATP-binding A (ABC1), 7	H45265	► B3GNT5	AA043551
Adaptor-related protein alpha2	H28956	►GIG2 (G-protein-receptor-induced)	AA236455
ATP synthase subunit	T49146	►PMAIP1 (PMA-induced protein 1)	AA262439
Membrane assoc protein 17	AA253473	►SSI-3 (suppressor of cytokine signaling)	R69417
Mitochondrial membrane protein	AA121962	UBE1 (ubiquitin-activating enzyme E1)	M58028
Phospholipid scramblase	AF008445		
		Nucleic acid binding	
Secreted proteins		DDX34 DEAD/H Box 34	R48810
►IL-8	M28130	KIAA1100	AA411433
Inducible cytokine A4	M69203	►SRPK1 (SFRS protein kinase)	R78142
VNN3 Vanin 3	AA461448	BTBD1 [BTB (POZ) domain containing]	R69336
		HDAC5 (histone deacetylase 5)	AA496574
Vasoactive mediators		FHL3 (four and a half LIM domains)	AA460438
►PTGS2 (COX-2)	U04636		
Cell cycle		Miscellaneous	
►BTG family, member 2	Y09943	SLC25A1 (solute carrier family 25)	W86850
BHLHB2	T40999	SLC16A5 (solute carrier family 16)	AA421374
Amplified in ostersarcoma	N25082	Wolf-Hirschhorn syndrome-like	AA286863
PTMS (parathymosin)	W90032	KIAA1536 protein	AA412555
MIR (myosin regulatory interacting protein)	AA129373		
IDN3	AA490868		
NELL2 (NEL-like 2)	H23584		
HLX1 (H2.0-like homeo box 1)	M60721		

►Denotes genes selected for real-time polymerase chain reaction (PCR) analyses.

Table 3. Adjusted TaqMan® polymerase chain reaction (PCR) values (fold-change above the mean of the normals)

	Patient	BTG2	PMAIP	B3GNT	NCUBE	FLJ294	SRPK1	SSI-3	LZS	GIG-2	AXUD1	PTGS2	EGR1	GOS2	Gal3	IL-8
1a	Mild	1.88	2.02	4.15	1.79	1.40	1.36	1.12	1.32	1.53	2.40	2.37	1.80	11.02	0.61	1.55
2a	Mild	1.04	2.08	3.27	2.38	0.68	2.06	2.71	1.98	2.89	1.54	2.05	1.83	4.74	1.08	2.77
3	Mild	0.85	3.74	2.91	1.06	0.41	0.86	1.07	1.11	3.39	1.32	1.63	1.65	4.49	0.46	2.27
4b	Moderate	1.52	2.07	0.95	1.51	1.36	0.97	1.36	1.21	1.90	1.25	1.71	0.91	1.25	0.75	1.65
5a	Moderate	1.09	1.51	1.59	2.05	1.02	1.75	2.02	0.56	2.42	1.38	1.29	1.13	0.80	0.81	0.99
5b	Moderate	1.16	1.27	1.86	2.54	1.60	2.01	1.88	1.03	1.10	1.76	0.56	0.75	0.94	1.03	1.13
6	Moderate	2.69	3.92	5.03	2.24	0.88	1.57	3.84	0.89	2.11	4.79	3.74	1.22	5.83	0.40	2.41
7	Moderate	1.44	1.34	0.95	0.94	0.66	1.49	0.87	1.06	3.67	1.70	1.79	0.81	1.56	0.41	3.08
8	Moderate	0.97	0.74	1.28	1.51	0.88	0.85	0.76	0.74	1.34	1.03	0.78	0.50	0.40	0.58	1.26
9	Moderate	0.72	1.11	1.03	1.27	0.95	0.80	1.58	1.19	1.10	1.42	0.52	0.61	2.41	0.50	1.43
2b	Severe	1.19	1.40	1.04	2.10	1.15	0.65	0.48	2.00	1.71	0.57	0.96	0.99	0.66	1.18	0.30
11	Severe	1.29	0.76	4.66	1.31	65.31	1.72	1.16	2.37	0.82	2.07	0.64	1.39	0.78	1.36	0.89
12	Severe	5.00	4.29	6.16	2.97	1.60	1.23	3.04	2.60	2.87	4.05	3.56	4.86	3.09	0.96	5.76
13	Severe	2.17	1.06	3.50	1.44	0.48	1.29	1.76	1.70	1.69	1.98	1.91	0.97	4.81	0.82	0.97
15	Remission	0.58	1.17	2.11	0.91	1.30	1.04	0.88	0.58	1.31	1.45	0.77	0.98	4.34	0.61	1.61
16a	Remission	2.14	2.65	3.78	2.56	0.42	2.34	4.76	1.92	3.66	3.82	1.63	3.04	4.07	0.59	1.48
16b	Remission	3.13	3.66	8.18	3.35	0.56	3.45	5.70	3.44	4.80	9.19	5.44	11.79	9.19	0.79	4.05
17	Remission	2.89	2.57	1.52	2.02	2.20	1.50	2.23	1.44	1.91	3.88	1.65	1.01	6.43	1.30	3.97

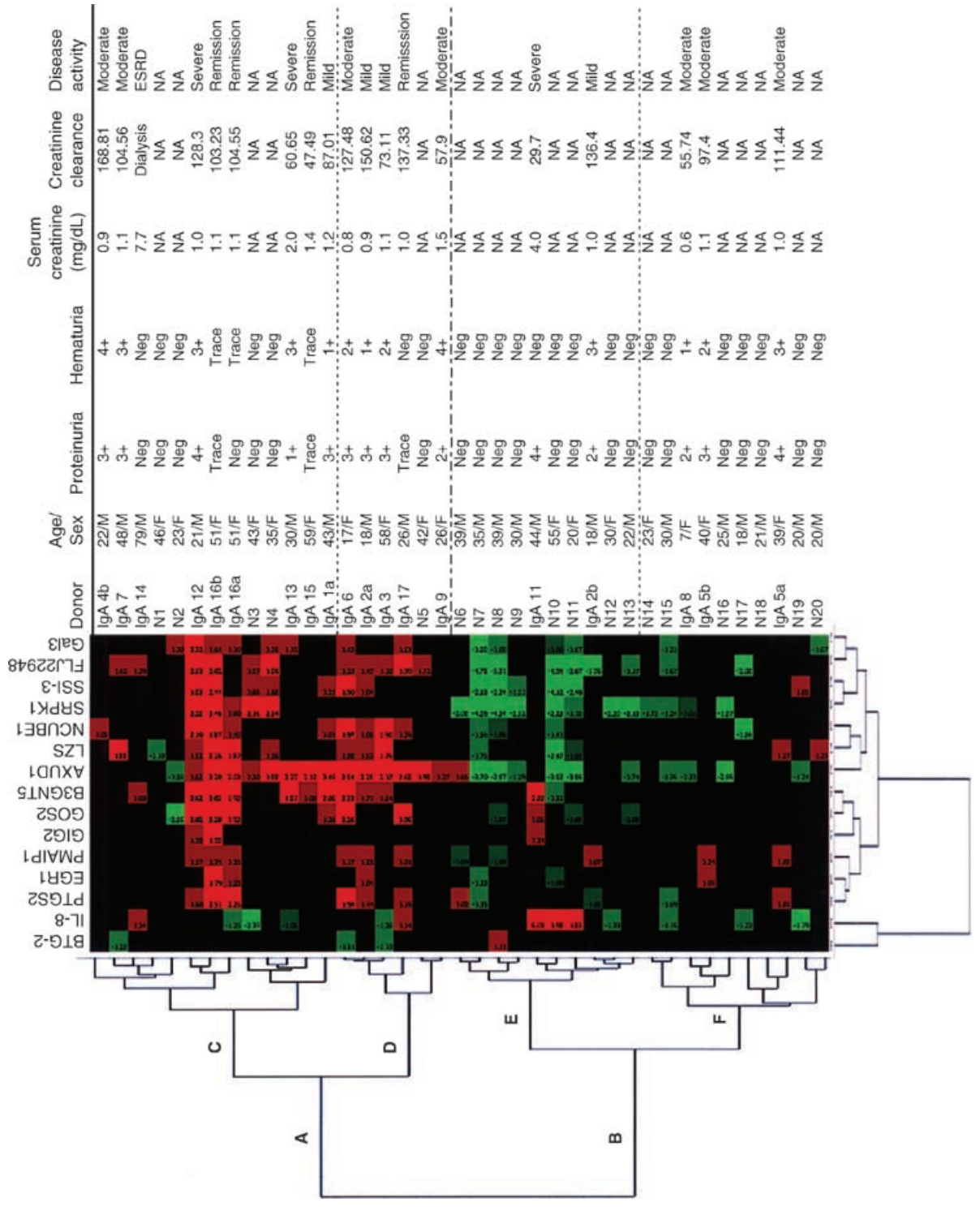

Fig. 2. Hierarchical clustering of genes derived from microarray data, which were confirmed and quantitated by TaqMan® polymerase chain reaction (PCR). The fold-change of the specific genes is plotted on the X-axis (listed in Table 3) and the donors are plotted on the Y-axis [21]. Each column represents the data for a single gene and each row represents an individual. In order to visualize the data, the fold-deviation from the average expression of each gene across the set of samples studied is shown as a colored square ranging from bright green (below average levels of expression for that gene) through black (average expression of that gene) to bright red (above average level of mRNA present for that gene). Genes that show similar expression patterns across different patients cluster together. The hierarchical tree, or dendrogram, is displayed under the clustered genes, and over the clustered subjects, to depict graphically the degrees of relatedness (correlation coefficient) between adjacent subjects and genes; short branches between two samples denote a high degree of similarity, whereas longer branches depict a lesser degree of similarity. Major clusters are denoted as "A" and "B." Subclusters are denoted as "C," "D," "E," and "F."

A

BTG2	0.397	0.296	0.173	0.144	0.132	0.111	0.096	0.082	0.088	0.090	**0.270**	
PMAIP1	0.688	0.518	0.458	**0.485**								
B3GNT5	0.709	0.580	0.384	0.258	0.169	0.199	**0.233**					
NCUBE1	0.169	0.087	0.044	0.026	0.012	0.006	0.003	0.004	0.003	0.008	0.030	0.052
FLJ2948	0.553	0.497	0.440	0.390	0.195	0.041	0.012	0.003	0.001	0.000	0.000	0.000
SRPK1	0.418	0.227	0.149	0.091	0.056	0.054	0.036	0.026	0.024	**0.101**		
SSI-3	**0.797**											
LZS	0.509	0.294	0.193	0.162	0.176	0.160	0.040	0.026	0.024	0.086	0.070	0.058
GIG-2	0.475	0.361	0.281	0.265	0.331	0.400	0.200	0.160	**0.102**			
AXUD1	0.688	0.694	**0.659**									
PTGS2	0.717	0.547	0.475	0.296	0.298	0.219	0.192	**0.498**				
EGR1	0.799	**0.858**										
GOS2	0.721	0.531	0.428	0.371	**0.481**							
Gal 3	0.712	0.707	0.632	0.448	0.427	**0.520**						
IL-8	0.369	0.242	0.164	0.121	0.126	0.108	0.113	0.092	0.082	0.031	0.085	**0.164**
Ru	0.482	0.650	0.742	0.789	0.811	0.828	0.845	0.840	**0.854**	0.831	0.807	0.811

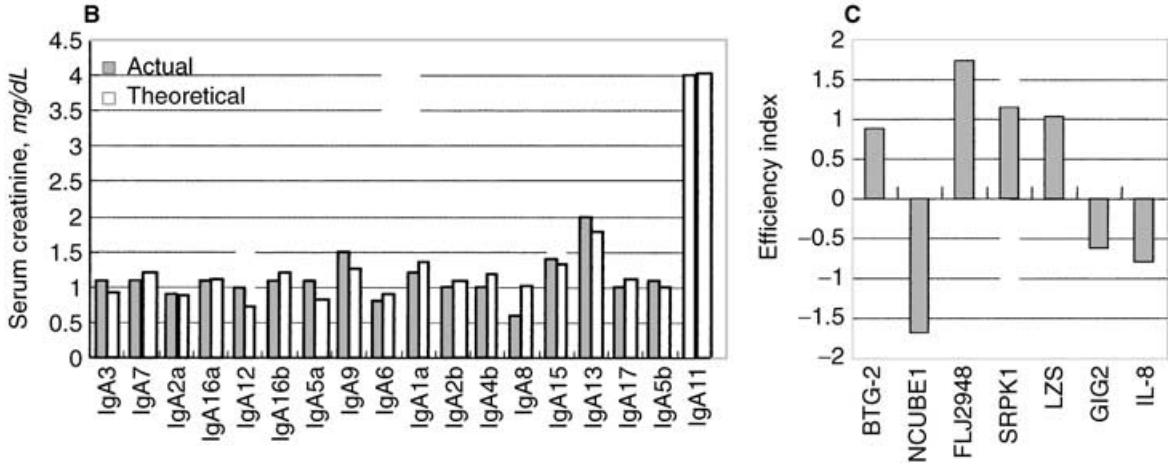

Fig. 3. Theoretical serum creatinine values, generated using a mathematically derived formula using gene expression levels, correlate with actual serum creatinine levels. (*A*) Regression modeling was performed to select disease related genes as independent predictor variables. Seven genes (underlined) gave the best-fit *Ru* value of 0.854, based on *P* values. (*B*) This method was applied to find correlation between disease related gene expression values and clinical serum creatinine values, using a standard computational spread sheet program. (*C*) The efficiency index of a particular gene is the range of calculated fold-change values [Max-min value of TaqMan® polymerase chain reaction (PCR) data] times the coefficient for curve fit from the multiple regression analysis. A positive value implies that a reduction of this transcript may have beneficial effects in reducing serum creatinine levels.

the opposite would be true for *B3GNT5*, *AXUD1*, and *GIG-2* (Fig. 4c).

The gene set is specifically indicative of disease activity in IgAN patients

Is the leukocyte gene expression profile for IgAN specific for this disease or is this particular profile dictated by the general state of renal insufficiency caused by any disease? To test this we used TaqMan® PCR to determine the expression levels of the 15 genes identified as part of the IgAN profile in two other glomerular diseases (i.e., ANCA glomerulonephritis or lupus nephritis). The fold-change values (data not shown) were substituted into the mathematical model developed for IgAN. For theoretic values to be correlative, the actual serum creatinine value minus the theoretical value must be <0.5 or the actual creatinine clearance value minus the theoretical value must be <25. In ANCA patients, the gene expression-generated values were correlative in five of 19, while in lupus patients nine of 19 were correlative, when comparing serum creatinine concentrations (Table 4). Theoretic creatinine clearance values were correlative in four of 19 ANCA patients and one of 19 lupus patients. The data indicate that the mathematical model established for IgAN has little relationship to clinical values in ANCA glomerulonephritis and lupus nephritis patients. These data are compellingly, suggestive that gene expression patterns in leukocytes can serve as specific fingerprints of a particular disease that can be used to distinguish disease activity.

DISCUSSION

This work establishes for the first time that gene expression changes in circulating leukocytes can be useful in assessing in IgAN patients. Using microarray

A

*BTG2	0.919	**0.902**								
PMAIP1	0.512	0.308	0.193	0.056	0.012	0.007	0.004	0.002	0.001	0.026
B3GNT5	0.304	0.108	0.033	0.011	0.002	0.001	0.000	0.000	0.000	0.001
NCUBE1	0.878	0.789	**0.717**							
FLJ2948	0.666	0.457	0.346	0.280	0.215	0.355	**0.381**			
SRPK1	0.676	0.518	0.399	0.126	0.017	0.013	0.010	0.005	0.001	0.051
SSI-3	0.674	0.433	0.307	0.230	0.276	0.361	0.179	**0.253**		
LZS	0.727	0.558	0.433	0.263	0.020	0.011	0.007	0.003	0.003	0.063
GIG-2	0.770	0.582	0.473	0.247	0.032	0.011	0.005	0.003	0.002	**0.083**
AXUD1	0.752	0.225	0.081	0.045	0.024	0.018	0.003	0.003	**0.003**	
PTGS2	0.285	0.087	0.020	0.006	0.002	0.001	0.001	0.000	0.000	0.023
EGR1	**0.944**									
GOS2	0.687	0.533	0.412	0.378	**0.360**					
Gal 3	0.782	0.673	0.603	**0.427**						
IL-8	0.773	0.624	0.543	0.291	0.286	**0.462**				
Ru	-0.209	0.409	0.615	0.706	0.728	0.737	0.760	**0.773**	0.769	0.443

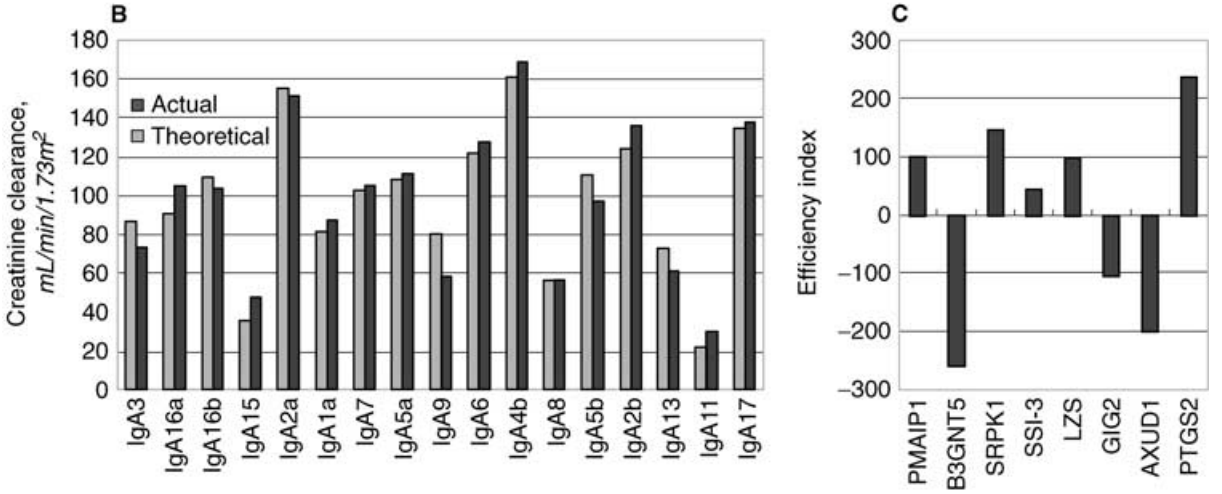

Fig. 4. Theoretic creatinine clearance values, generated using a mathematically derived formula based on gene expression levels, correlate with actual creatinine clearance values. (*A*) Regression modeling was performed to select disease related genes as independent predictor variables. Eight genes (underlined) gave the best-fit *Ru* value of 0.773, based on *P* values. (*B*) This method was applied to find a correlation between disease related gene expression values and creatinine clearance values, using a standard computational spread sheet program. (*C*) The efficiency index of a particular gene is the range of calculated fold-change values [Max-min value of TaqMan® polymerase chain reaction (PCR) data] time the coefficient for curve fit from the multiple regression analysis. A negative value implies that restoration of transcript levels to normal may have positive effects on creatinine clearance.

technology, we identified genes differentially expressed in IgAN that were not up-regulated in ANCA glomerulonephritis, FSGS, or minimal change disease. TaqMan® PCR authenticated the microarray data and computational studies using these PCR values provided a method to generate an accurate and reproducible estimate of clinical parameters and disease activity.

We were astonished to find that calculations based on expression levels of a particular gene group could accurately approximate clinical measures of kidney function in 100% of the IgAN patients. This correlation of the expression of these particular genes with renal impairment appears to be specific for IgAN because it was not observed in patients with lupus nephritis or ANCA glomerulonephritis. Given these results, we are perplexed as to why the cluster analyses using this gene group resulted in several patients who clustered with the healthy volunteers. Looking at the patients' characteristics, there were no apparent differences that would explain these results.

Also of interest, included in our study is a pediatric patient (IgA 8, 7 years old) whose gene expression profile did *not* cluster with the adult IgAN patients. For additional comparisons with the adult values, the Cockcroft-Gault equation was applied (although not conventionally applicable), and we found this pediatric patient's creatinine clearance value correlated with the theoretic value. However, once the researcher realizes that clustering is simply an exploratory data analysis tool and not a typical statistical test, the random patient that falls into the "normal" cluster is understandable. Cluster analysis takes large amounts of information and sorts it into manageable, meaningful piles. For example, initially each object exists in a class by itself. Now imagine that, in very small steps, we "relax" our criterion as to what is and is not unique. Put another way, we lower our threshold

Table 4. IgA nephropathy (IgAN)-related gene expression profiles do not reflect clinical serum creatine levels or creatinine clearance of patients with antineutrophil cyctoplasmic antigen (ANCA) disease or systemic lupus erythematosus (SLE)

Serum creatinine *mg/dL*					
Patient (age/gender/race)	Actual	Theoretic	Patient (age/gender/race)	Actual	Theoretic
ANCA 1 (54/M/W)	2.2	0.47	SLE 1 (63/F/W)	0.86	−0.29
ANCA 2 (15/F/W)	0.8	1.19◀	SLE 2 (60/M/W)	1.6	0.03
ANCA 3 (43/M/W)	0.9	0.57◀	SLE 3 (37/F/A)	0.6	0.10
ANCA 4 (45/M/B)	2.1	0.44	SLE 4 (32/F/W)	1.1	1.04◀
ANCA 5 (52/F/W)	1.7	1.72	SLE 5 (47/F/B)	1.8	0.64
ANCA 6 (55/M/W)	2.6	0.62	SLE 6 (33/F/B)	0.9	−0.91
ANCA 7 (67/M/W)	1.7	1.21	SLE 7 (23/F/W)	1	0.6◀
ANCA 8 (36/M/W)	3.8	1.05	SLE 8 (21/F/B)	1.6	1.12
ANCA 9 (53/M/W)	0.9	1.16◀	SLE 9 (24/F/W)	1	0.58◀
ANCA 10 (43/M/W)	2.7	1.16	SLE 10 (28/F/B)	0.7	1.16◀
ANCA 11 (50/F/B)	9.2	2.06	SLE 11 (28/F/B)	0.9	0.33◀
ANCA 12 (64/M/W)	1.1	2.01	SLE 12 (35/F/W)	0.8	−0.41
ANCA 13 (55/M/W)	2.4	1.29	SLE 13 (22/F/B)	0.8	0.54◀
ANCA 14 (24/F/W)	1	1.53	SLE 14 (44/F/B)	4.4	1.24
ANCA 15 (2/M/W)	2.4	0.77	SLE 15 (57/F/B)	1.1	1.22◀
ANCA 16 (45/F/W)	1.1	1.57◀	SLE 16 (48/F/B)	0.7	0.74◀
ANCA 17 (33/F/W)	1	0.46	SLE 17 (20/F/A)	5.2	1.17
ANCA 18 (36/F/W)	0.7	1.59	SLE 18 (24/F/W)	0.9	1.20◀
ANCA 19 (61/F/W)	4.5	1.66			
Creatinine clearance					
Patient	Actual	Theoretic	Patient	Actual	Theoretic
ANCA 1	42.1	137.24	SLE 1	62.1	10.68
ANCA 2	100.9	150.76	SLE 2	53.9	336.93
ANCA 3	116.2	196.09	SLE 3	110.8	244.29
ANCA 4	48.8	105.36	SLE 4	63.4	130.14
ANCA 5	33.4	308.40	SLE 5	33.4	157.94
ANCA 6	35.2	195.42	SLE 6	76.8	165.21
ANCA 7	46.7	118.32	SLE 7	75.6	141.75
ANCA 8	29.5	224.54	SLE 8	48	92.89
ANCA 9	104.2	102.36◀	SLE 9	74.9	222.30
ANCA 10	38.7	225.45	SLE 10	103.4	211.01
ANCA 11	6.3	234.90	SLE 11	80.3	175.04
ANCA 12	74.5	184.79	SLE 12	84.7	235.62
ANCA 13	38.2	113.64	SLE 13	95.3	144.56
ANCA 14	74.9	85.09◀	SLE 14	14.1	107.43
ANCA 15	53	38.67◀	SLE 15	48.7	197.53
ANCA 16	55.8	61.87◀	SLE 16	84.8	132.82
ANCA 17	69.1	128.40	SLE 17	14.9	229.61
ANCA 18	30	159.60	SLE 18	83.2	98.29◀
ANCA 19	18.4	75.81			

◀Denotes theoretic values that coorelate with actual values; correlative criteria: serum creatinine–actual vs. predictive must be <0.5; creatinine clearance–actual vs. predictive must be <25.

regarding the decision when to declare two or more objects to be members of the same cluster. As a result, we link more and more objects together and aggregate larger and larger clusters of increasingly dissimilar elements. Finally, in the last step, all objects are joined together. Thus, the resulting clusters are by nature not homogeneous. Outliers within the clusters are difficult to interpret. These may stem from the fact that the similarities/dissimilarities between different clusters may pertain to or be caused by somewhat different subsets of variables. Nevertheless, and in support of the efficacy of our studies, in two instances cluster analysis of samples from the same individual, collected as much as a year apart, gave results that fell into the same cluster.

The pathophysiologic basis for the correlation between this gene expression profile in leukocytes and renal function in IgAN is not revealed by our studies, but the data should provide a fertile ground for exploration. It was interesting to make comparisons of our microarray analysis of total leukocytes with a published expression profile of bacterially exposed neutrophils. Newburger, Subrahmanyam, and Weissman [15] utilized a gel-based method to display 3′ end fragments of cDNAs on isolated neutrophils. Comparisons revealed that some gene families are represented in both studies [mitogen-activated protein (MAP) kinase, GADD, and ubiquitin pathway, GOS, Ras-related] and both studies had cyclooxygenase-2 (COX-2) and IL-8. Bacteria are known to starts flares

of active disease in IgAN and this overlap in genes, although small, may reflect bacterially induced changes in our IgAN patients. These similarities imply that the neutrophil population is responsible in part for the altered leukocyte expression profile in our patients. However, further studies are needed for verification.

The lack of correlation between the IgAN expression profile and renal function in ANCA glomerulonephritis and lupus patients indicates that the altered gene expression is not dictated merely by the renal insufficiency. Thus, the expression profile, most likely is a reflection of altered leukocyte function that is more directly related to the pathologic events in the kidney in IgAN. This is in accord with the earlier observations by our research group of an expression profile of leukocyte genes that is characteristic of ANCA glomerulonephritis patients [16]. Medications might, and probably do, affect leukocyte gene expression. Vasculitis and SLE is usually treated with corticosteroids, while most patients with IgAN do not receive steroids. On the other hand, angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor inhibitors are probably more widely used in patients with IgAN. This is an area that remains to be studied.

Although familial links in our IgAN patient population have been ruled out, a familial form of IgAN has been linked to a gene on chromosome 6q22-23 [17]. Interestingly, one of the genes that clustered with IgAN in our studies was *NCUBE1*, which is a gene on chromosome 6q. The product of this gene is a member of a family of ubiquitin-conjugating enzymes. These enzymes selectively target proteins for proteasomal degradation by the covalent attachment of ubiquitin moieties [18]. The efficiency index analyses indicate *NCUBE1* to be inversely related to disease activity (i.e., increased expression was associated with lower creatinine concentrations). The data do not prove causality but suggest there may be a relationship between this gene and the induction or progression of IgAN. Information from mathematical modeling of this type may identify potential drug targets. The information gleaned from this type of analysis is the identification of potential drug targets such as *NCUBE1*, which if increased may have a beneficial effect on serum creatinine levels, at least in IgAN patients.

The era of bioinformatics and computational biology allows inferences about causality of disease at the genetic level. New reports are surfacing daily that describe the efficacy and utility of microarray information [19]. By combining the patients' clinical data with gene cluster bioinformatics, we were able to cluster specific genes with biologic outcomes. These studies were not designed as large population analyses to prove disease specificity or as predictors of renal function. Large population studies will be needed to determine if in fact leukocyte gene expression profiles can be used as bioindicators for diagnostics and for predicting disease responses. As we consider the efficacy versus pitfalls of revolutionary new technologies on the horizon, as stated by King and Sinha, "the potential payoff remains large" [20].

Reprint requests to Dr. Gloria Preston, CB #7155, 346 MacNider Bldg., Division of Nephrology and Hypertension, Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7155.
E-mail: Gloria_Preston@med.unc.edu

REFERENCES

1. Collins FS, Patrinos A, Jordan E, *et al*: New goals for the U.S. Human Genome Project: 1998–2003. *Science* 282:682–689, 1998
2. Lockhart DJ, Dong H, Byrne MC, *et al*: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14:1675–1680, 1996
3. Chee M, Yang R, Hubbell E, *et al*: Accessing genetic information with high-density DNA arrays. *Science* 274:610–614, 1996
4. Tang Y, Lu A, Aronow BJ, Sharp FR: Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: Blood genomic fingerprints of disease. *Ann Neurol* 50:699–707, 2001
5. Ibels LS, Gyory AZ: IgA nephropathy: Analysis of the natural history, important factors in the progression of renal disease, and a review of the literature. *Medicine (Baltimore)* 73:79–102, 1994
6. Novak J, Julian BA, Tomana M, Mesteck J: Progress in molecular and genetic studies of IgA nephropathy. *J Clin Immunol* 21:310–327, 2001
7. Donadio JV, Grande JP: IgA nephropathy. *N Engl J Med* 347:738–748, 2002
8. Prakash K, Pirozzi G, Elashoff M, *et al*: Symptomatic and asymptomatic benign prostatic hyperplasia: Molecular differentiation by using microarrays. *Proc Natl Acad Sci USA* 99:7598–7603, 2002
9. Tackels-Horne D, Goodman MD, Williams AJ, *et al*: Identification of differentially expressed genes in hepatocellular carcinoma and metastatic liver tumors by oligonucleotide expression profiling. *Cancer* 92:395–405, 2001
10. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868, 1998
11. Cook R: Detection of influential observations in linear regression. *Techology* 19:15–18, 1977
12. Cox D: Regression models and life tables. *J Royal Stat Soc* 34:187–220, 1972
13. Ueda T: *Data Mining Practice With Excel*, Tokyo, Doyukan, 2001
14. StatSoft I: *Electronic Statistics Textbood*, Tulsa, *http://www.statsoft.com/textbook/stathome.html*, 2002
15. Newburger PE, Subrahmanyam YV, Weissman SM: Global analysis of neutrophil gene expression. *Curr Opin Hematol* 7:16–20, 2000
16. Yang JJ, Preston GA, Alcorta DA, *et al*: Expression profile of leukocyte genes activated by anti-neutrophil cytoplasmic autoantibodies (ANCA). *Kidney Int* 62:1638–1649, 2002
17. Gharavi AG, Yan Y, Scolari F, *et al*: IgA nephropathy, the most common cause of glomerulonephritis, is linked to 6q22-23. *Nat Genet* 26:354–357, 2000
18. Lester D, Farquharson C, Russell G, Houston B: Identification of a family of noncanonical ubiquitin-conjugating enzymes structurally related to yeast UBC6. *Biochem Biophys Res Commun* 269:474–480, 2000
19. Manger ID, Relman DA: How the host "sees" pathogens: Global gene expression responses to infection. *Curr Opin Immunol* 12:215–218, 2000
20. King HC, Sinha AA: Gene expression profile analysis by DNA microarrays: Promise and pitfalls. *JAMA* 286:2280–2288, 2001
21. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868, 1998