Human Molecular Genetics, 2003, Vol. 12, No. 17 **2191–2199**
DOI: 10.1093/hmg/ddg221

Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease

Roberto Bomprezzi[1,*], Markus Ringnér[1,2], Seungchan Kim[1,3], Michael L. Bittner[1,3], Javed Khan[4], Yidong Chen[1], Abdel Elkahloun[1], Aimee Yu[5], Bibiana Bielekova[5], Paul S. Meltzer[1], Roland Martin[5], Henry F. McFarland[5] and Jeffrey M. Trent[3]

[1]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA, [2]Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden, [3]Translational Genomics Research Institute, Phoenix, AZ 85004, USA, [4]Advanced Technology Center, National Cancer Institute, National Institutes of Health, Gaithersburg, MD 20877, USA and [5]Neuroimmunology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA

Received May 3, 2003; Revised June 12, 2003; Accepted June 26, 2003

Multiple sclerosis (MS) and other T cell-mediated autoimmune diseases develop in individuals carrying a complex susceptibility trait, probably following exposure to various environmental triggers. Owing to the presumed weak influence of single genes on disease predisposition and the recognized genetic heterogeneity of autoimmune disorders in humans, candidate gene searches in MS have been difficult. In an attempt to identify molecular markers indicative of disease status rather than susceptibility genes for MS, we show that gene expression profiling of peripheral blood mononuclear cells by cDNA microarrays can distinguish MS patients from healthy controls. Our findings support the concept that the activation of autoreactive T cells is of primary importance for this complex organ-specific disorder and prompt further investigations on gene expression in peripheral blood cells aimed at characterizing disease phenotypes.

INTRODUCTION

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disorder of the central nervous system (CNS) with various degrees of axonal damage. MS affects mainly young adults with predominance for females and a prevalence in the USA and northern Europe of ~100 per 100 000 people (1). Since it often leads to substantial disability, MS has major socio-economic impact.

Intensive investigations over the last two to three decades have indicated that MS is pathogenetically a T cell-mediated autoimmune disease (2). Its etiology remains elusive, but the interplay between environment and genetic factors is ultimately thought to be critical to the development of the disease. Epidemiological observations at the population level as well as familial aggregation studies (3–6) have prompted, in different MS populations, linkage studies and candidate gene searches in the attempt to dissect genetic components of the susceptibility to MS. The results were similar to other autoimmune diseases, suggesting that MS is genetically complex and there may be no single genes of major attributable risk. Instead MS may result from the contribution of several genes exhibiting low or moderate effect (7,8). While none of the candidate genes suggested to date has been shown to be essential or sufficient for disease development, the most robust association has clearly been established with alleles of the major histocompatibility complex (MHC; HLA in humans, on chromosome 6p21) and in particular with the haplotypes HLA-DQB1*0602, -DQA*0102, -DRB1*1501, -DRB5*0101 (9,10). In summary, etiologic and genetic heterogeneity, epistatic gene interactions and epigenetic modifications are only some of the possible confounding factors to consider in interpreting the results of studies on genetic susceptibility to MS (11).

Evidence of phenotypic heterogeneity in MS was recently reported by Lucchinetti *et al.* (12) who described, in a large series of actively demyelinating lesions from MS patients, different pathology patterns suggestive of distinct pathogenic mechanisms (12).

*To whom correspondence should be addressed at: Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Bldg 50 Room 5150, Bethesda, MD 20892-8000, USA. Tel: +1 3014964655; Fax: +1 3014023241; Email: rbomprez@nhgri.nih.gov

The method of global gene expression analysis using cDNA microarrays has proven to be a sensitive method to develop and refine the molecular taxonomy of cancer and other disorders, as well as a tool to investigate molecular heterogeneity (13). Here, using this technology, complemented with a novel, powerful, analytical method, we compared the gene expression profiles of peripheral blood mononuclear cells (PBMC) from a series of MS patients with those from healthy donors. The specific aim of the study was to identify a panel of molecular markers indicative of disease status and not necessarily the identification of a gene(s) causative for MS. Nonetheless, identifying genes capable of molecular classification of MS from healthy controls using PBMC would be a significant advance, also providing insight into the pathogenetic mechanisms of MS. Results from previous microarray-based studies sustain the rationale for employing microarrays to test PBMC in autoimmune disorders (14–16). In accordance with the reports by Ramanathan *et al.* (14) and Maas *et al.* (15), we found that the differences existing between MS patients and control individuals comprise immune related as well as cell cycle related genes, supporting the idea that, besides the lymphocyte activation and cell–cell interaction, other cellular events play relevant roles during the course of autoimmune disease.

RESULTS

The method of analysis we employed (see Methods) presents two major advantages: (a) extract data from strong feature sets (17) to separate samples having the smallest variance within the group and being least susceptible to both biological and technical noise; (b) test the combined information in sets of genes (pairs or triplets), so that if multiple causes are acting in the sample set, such as gene interaction, the discriminative power can contribute to the overall separation. The disadvantage is the very intense computation requirements. We identified more than a thousand pairs of genes that could distinguish MS samples from controls in our initial 'training set' (termed set 1). The pairs of genes separating MS from healthy controls were strongly dominated by cDNA clones for the heat shock protein-70 followed by *CKS2*. For example, *HSPA1L* together with *H1F2* and *CKS2* together with *PAFAH1B1* by themselves discriminated well between MS and controls (Fig. 1A and B). The results demonstrated a σ-error of less than 5% ($\sigma_n = 0.6$) for the best candidate pairs. Using these pairs, we predicted whether a sample in an independent test cohort (termed set 2), could be molecularly 'diagnosed' with MS, and achieved reasonably good prediction results (Fig. 1C and D). For both pairs, only four individuals were misclassified corresponding to 80% correct predictions. This was also the case for most of the highest ranked pairs. They typically resulted in an 80% classification performance, generally with the same four healthy individuals misclassified as for *HSPA1L* and *H1F2* (Fig. 1C). Even for normalized profiles, the range of values is slightly different between sets 1 and 2 (Fig. 1), either reflecting differences in microarray batches (shift of ranges) or sample preparation (fresh versus frozen, enlarged or reduced variation). Nevertheless, the classifiers identified in set 1 showed very good predictive capability for set 2.

These results indicate that a set of genes differentially expressed in MS patients can be used to predict the disease state in an independent test set. Hence, we proceeded to find candidate discriminatory genes using all individuals from both sample sets. Using minimal criteria requiring each pair to have a 1.4-fold difference in average expression between the two groups, we identified 303 pairs having σ-error less than 14% ($\sigma_n = 0.6$). From this group, 112 represented unique genes that could discriminate MS patients from controls in both sets (for a complete list see Supplementary Material). Of the 112 genes, 53 appeared in at least two pairs (Fig. 2). Of interest, when we used strong feature sets based on gene triplets (rather than pairs) we did not gain significant improvement in misclassification error. For this study, we therefore limited our analysis to pairs. It is worth mentioning that, by generating the final discriminatory gene list from the combination of both sets, we intended to minimize potential experimental artifacts, such as low number of samples. Obviously, differences exist between the two sets, and we are aware of the gender mismatch of the control group of set 1. However, we feel that the validation by a properly matched group (set 2) strengthens the findings. Furthermore, even though the classification error obtained from combining both sets of samples increased when compared with the <5% error achieved in set 1 only, we believe the grouping of both sets may approximate more closely the real biological variation. Being outside the aim of the study, we did not attempt any sub-grouping of the patient samples. We rather see the heterogeneity existing in all the samples (affected and control) and the sample processing itself (fresh and frozen) as 'noise', from which we expect the real signal to emerge.

In a complex disease like MS, we recognize that both multiple interactions of different components of the immune system *in vivo*, and the complexity of the intracellular pathways must be considered in the interpretation of microarray experiments (18). While it is currently impossible to understand the biological outcome of these interaction patterns, we will below highlight some of the genes involved in specific pathways that are of potential interest for MS. These informative genes are divided into two groups, based on their up- or down-regulation in MS versus healthy donors (Table 1).

Genes with comparatively high expression in MS

The *PAFAH1B1* gene encodes for the non-catalytic alpha subunit of the intracellular Ib isoform of platelet-activating factor acetylhydrolase, which is highly specific for platelet-activating factor (PAF). A mutation of this gene is responsible for the human Miller–Dieker lissencephaly, suggesting its importance in controlling PAF levels during brain development (19). Moreover, PAF has a key role as a chemo-attractant, mediating inflammation and allergy.

We found increased expression of the TNF receptor (CD27) transcript in MS patients. CD27 is a disulfide-linked homodimer that serves as a costimulatory molecule during T cell activation. Its production is exclusive to cells of lymphoid lineage (NK, T and B cells) and is crucial for the development and maintenance of immune responses (20).

Unlike those from healthy donors, MS samples showed increased expression of genes related to T cell function and activation, like the transcript for the T cell receptor alpha locus

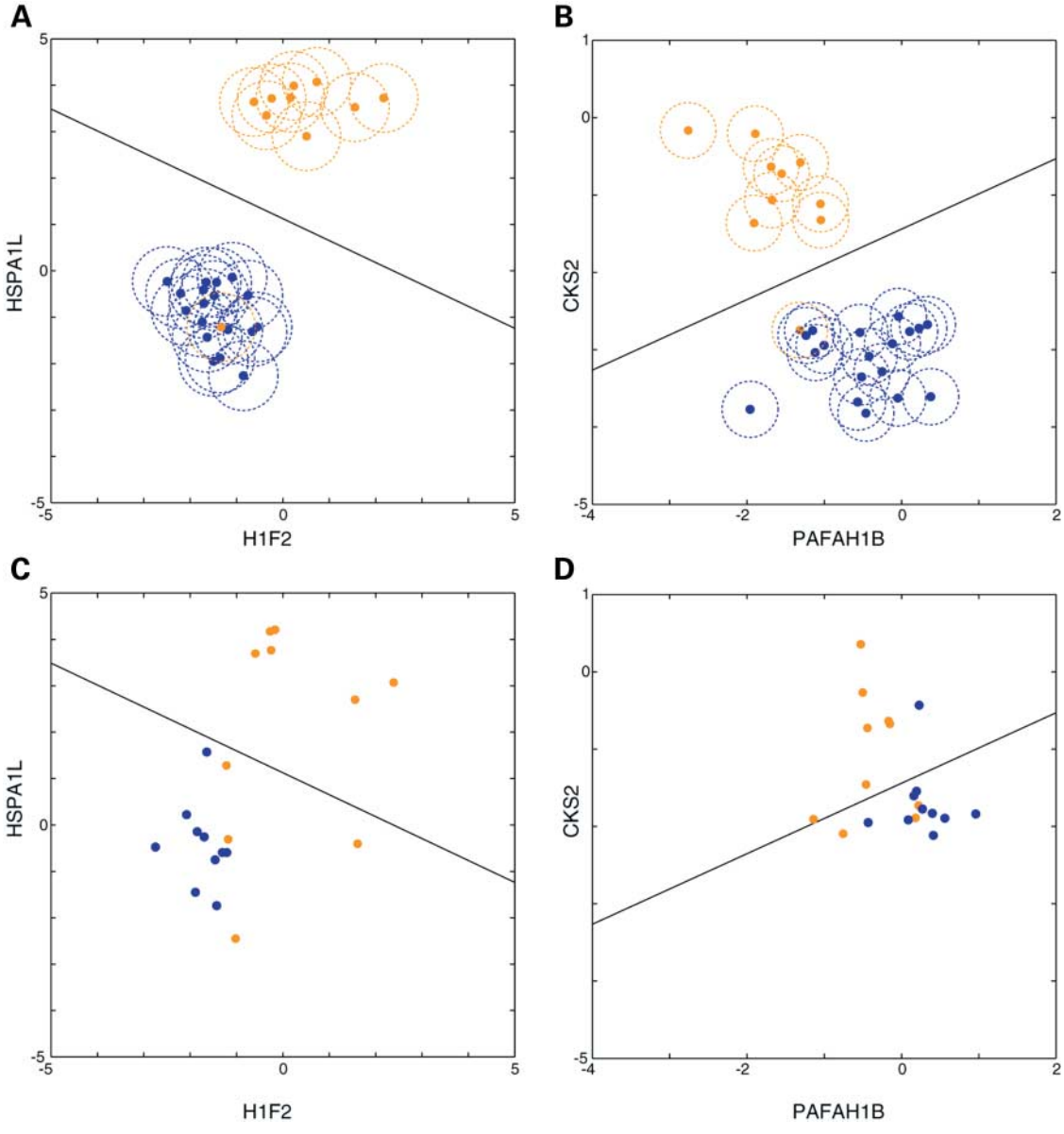

Figure 1. Pairs of genes separating MS patients from healthy controls. The expression levels of *HSPA1L* and *H1F2* (**A**) and *CKS2* and *PAFAH1B1* (**B**) are shown for the MS (blue) and healthy (yellow) samples in set 1. Based on these samples and a σ-error as indicated by the dotted circles around each sample, linear classifiers were designed (solid lines). These classifiers were then used to predict the samples in the test set 2 (**C** and **D**). Each gene pair misclassified four samples, giving a correct prediction rate of 80%.

and its zeta-chain associated protein kinase (ZAP70). T cell receptors are essential for any specific T cell-mediated immune response and have been implicated in MS susceptibility by genetic studies (21). ZAP70 is an important link between T cell receptor engagement and intracellular signaling steps during T cell activation (22). Also, the zinc finger protein 148 (ZNF148), which is known to bind the CACCC box promoter of both the human and mouse TCR genes to activate transcription (23), and the transcription factor-7, a T cell-specific down-stream enhancer element that controls *CD3E* gene expression (24), were up-regulated in the MS samples.

Another gene of interest is the interleukin 7 receptor gene (*IL7R*). IL7R mediates the effects of IL7 during lymphopoiesis, and its function is required for B and T cells, and in particular for gamma/delta T cell development. Of note, increased expression of *IL7R* was observed in a microarray study in PBMC from MS patients (14). Furthermore, an observation on the regulation of this gene and its relationship to T cell activation events has recently been reported (25).

Genes with comparatively low expression in MS

Statistically most significant in our study was the relative underexpression in the MS samples of the transcript for the heat shock protein 70 kDa (HSP70). This ubiquitous, highly conserved protein, previously suggested as a potential

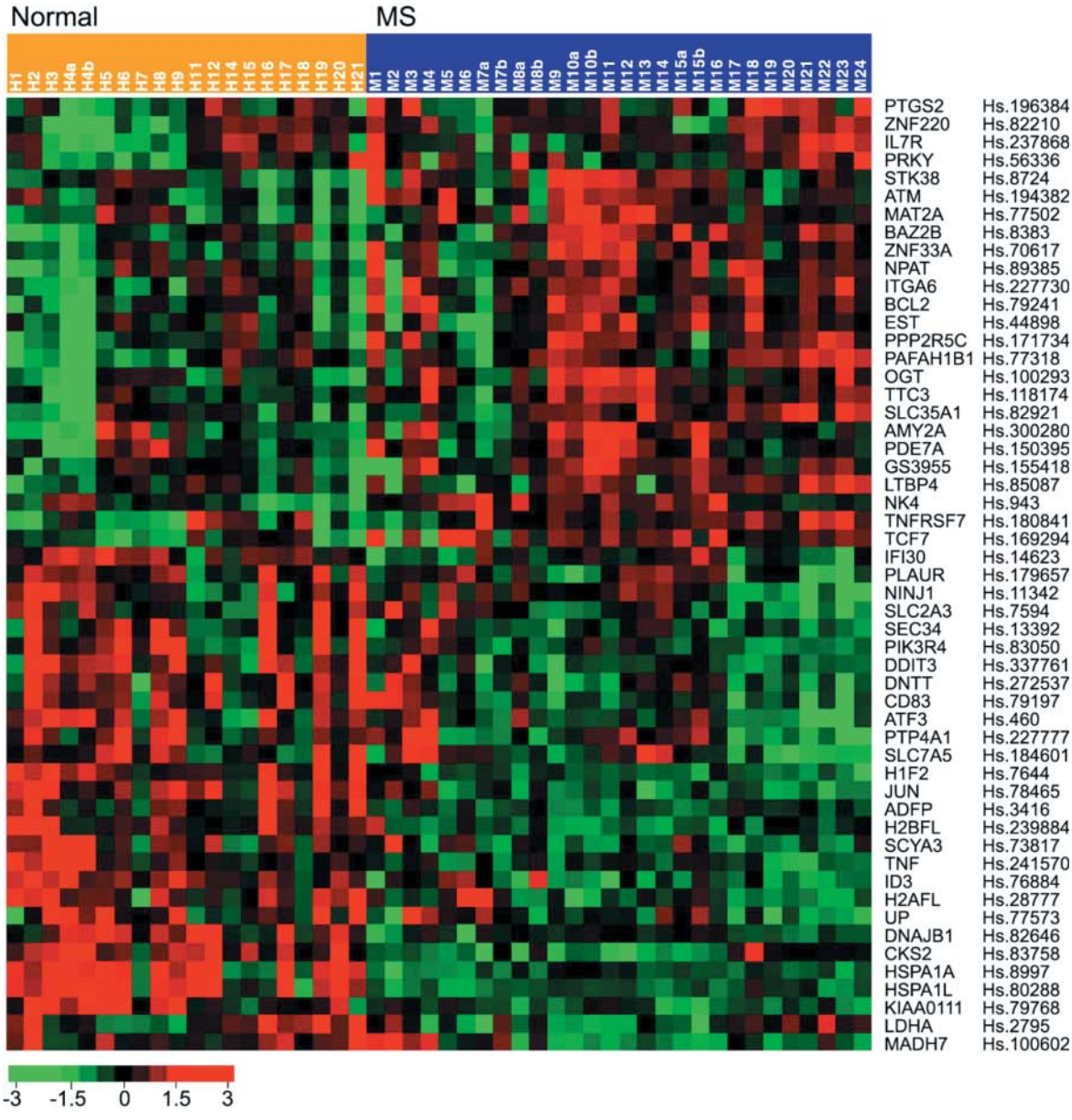

Figure 2. Expression of the 53 discriminatory genes. Each row represents one of the 53 discriminatory genes found in our analysis. Each column represents a sample ordered such that the normal samples are on the left- and the MS samples on the right-hand side. For each gene, red indicates a high level of expression relative to the mean; green indicates a low level of expression relative to the mean. The scale below indicates the number of standard deviations from the mean.

autoantigen in MS (26), maps on chromosome 6p21.3, the MHC/HLA-region. However, HSP70 may not only be of relevance as a potential autoantigen, but may also be particularly important for its involvement in the mRNA degradation of the cytokine in the ubiquitin–proteasome pathway (27) and its chaperone function (28).

Next to HSP70, messengers for histone proteins were detected as differentially regulated. The histone family genes map within a cluster of 35 histone genes to 6p21.3, again in the MHC/HLA region. Lever *et al.* (29) demonstrated that histone H1 is subject to rapid exchange *in vivo* and is modulated by phosphorylation (29). Not surprisingly, we also found that the expression of *CKS2* (CDC28 protein kinase-2 gene) was lower in MS than in healthy donors, and it proved to be a good discriminator of MS versus controls in our study. This latter gene encodes for a protein kinase that is activated during cell cycle, pushes cells into mitosis, and may have histone H1 as a substrate for phosphorylation (30). Moreover, histones are functionally regulated upon acetylation, known to be a mechanism in gene activation, chromatin structure and gene silencing.

Other genes of interest underexpressed in the MS samples included *TNF* and the oncogene *JUN*. The tissue inhibitor of metalloproteinase 1 (*TIMP1*) is also down-regulated in MS, as is *SERPINE1*, which is known to contain several regulatory sequences including heat-shock- and retinoic acid-responsive elements (31).

Table 1. Genes with differential expression in MS and controls

Gene name	Hs cluster	Relative expression[a]	Gene description	Cytogenetic position
Proteins involved in signal transduction and cell–cell interaction				
IKKE	Hs.321045	+	IKK-related kinase epsilon; inducible IkappaB kinase	1q32.1
MAL	Hs.80395	+	Mal, T-cell differentiation protein	2q21.1
ZAP70	Hs.234569	+	Zeta-chain (TCR) associated protein kinase (70 kDa)	2q11.2
DPP4	Hs.44926	+	Dipeptidylpeptidase IV (CD26, adenosine deaminase complexing protein 2)	2q24.3
ITGA6	Hs.227730	+	Integrin, alpha 6	2q31.1
NKTR	Hs.241493	+	Natural killer-tumor recognition sequence	3p22.1
SCYE1	Hs.333513	+	Small inducible cytokine subfamily E, member 1 (endothelial monocyte-activating)	4q25
IL7R	Hs.237868	+	Interleukin 7 receptor	5p13.3
HLA-DRA	Hs.76807	−	Major histocompatibility complex, class II, DR alpha	6p21.1
CD83	Hs.79197	−	CD83 antigen (activated B lymphocytes, immunoglobulin superfamily)	6p23
PTP4A1	Hs.227777	−	Protein tyrosine phosphatase type IVA, member 1	6q12
PDE7A	Hs.150395	+	Phosphodiesterase 7A	8q12.3
ATM	Hs.194382	+	Ataxia telangiectasia mutated (includes complementation groups A, C and D)	11q22.3
TNFRSF7	Hs.180841	+	Tumor necrosis factor receptor superfamily, member 7	12p13.31
DGKA	Hs.172690	+	Diacylglycerol kinase, alpha (80 kDa)	12q13.2
TRA@	Hs.74647	+	T cell receptor alpha locus	14q11.2
NK4	Hs.943	+	Natural killer cell transcript 4	16p13.3
PAFAH1B1	Hs.77318	+	Platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit (45 kDa)	17p13.3
SCYA3	Hs.73817	−	Small inducible cytokine A3 (homologous to mouse Mip-1a)	17q21.1
CCR7	Hs.1652	+	Chemokine (C-C motif) receptor 7	17q21.2
MADH7	Hs.100602	−	MAD (mothers against decapentaplegic, *Drosophila*) homolog 7	18q21.1
TIMP1	Hs.5831	−	Tissue inhibitor of metalloproteinase 1	Xp11.23
Structural proteins, enzymes of cell metabolism and proteins of the intracellular trafficking				
SPTBN1	Hs.107164	+	Spectrin, beta, non-erythrocytic 1	2p16.1
GOLGA4	Hs.183773	+	Golgi autoantigen, golgin subfamily a, 4	3p22.2
PIK3R4	Hs.83050	−	Phosphoinositide-3-kinase, regulatory subunit 4, p150	3q21.3
HSPA1A	Hs.8997	−	Heat shock 70 kDa protein 1A	6p21.1
SLC35A1	Hs.82921	+	Solute carrier family 35 (CMP-sialic acid transporter), member 1	6q15
DNAJA1	Hs.94	−	DnaJ (Hsp40) homolog, subfamily A, member 1	9p21.1
SPTAN1	Hs.77196	+	Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	9q34.11
SERPINH2	Hs.9930	−	Serine (or cysteine) proteinase inhibitor, clade H (heat shock protein 47), member 2	11q14.1
SEC34	Hs.13392	−	Tethering factor SEC34	13q14.13
PPP2R5C	Hs.171734	+	Protein phosphatase 2, regulatory subunit B (B56), gamma isoform	14q32.31
IFI30	Hs.14623	−	Interferon, gamma-inducible protein 30	19p13.11
TTC3	Hs.118174	+	Tetratricopeptide repeat domain 3	21q22.13
Transcription factors, DNA binding and chromatin related proteins				
JUN	Hs.78465	−	v-jun avian sarcoma virus 17 oncogene homolog	1p31.3
BAZ2B	Hs.8383	+	Bromodomain adjacent to zinc finger domain, 2B	2q24.2
XPC	Hs.320	+	Xeroderma pigmentosum, complementation group C	3p25.1
ZNF148	Hs.112180	+	Zinc finger protein 148 (pHZ-52)	3q21.2
TCF7	Hs.169294	+	Transcription factor 7 (T-cell specific, HMG-box)	5q23.3
H1F2	Hs.7644	−	H1 histone family, member 2	6p21.1
CKS2	Hs.83758	−	CDC28 protein kinase 2	9q22.2
DNTT	Hs.272537	−	Terminal deoxynucleotide transferase	10q24.1
BRF1	Hs.85155	−	Zinc finger protein 36, C3H type-like 1	14q24.1
BCL2	Hs.79241	+	B-cell CLL/lymphoma 2	18q21.33
ZNF43	Hs.74107	+	Zinc finger protein 43 (HTF6)	19p13.11

[a]+, Higher average expression in MS; −, higher average expression in controls.

It is worth noting that nine of the over- (2) or under-expressed (7) discriminatory genes map to chromosome 6p21, a region that has reproducibly shown the strongest association with MS and other autoimmune diseases (see Supplementary Material).

DISCUSSION

During recent years, research into the etiology, pathogenesis and clinical presentation of MS has indicated that it probably represents a heterogeneous group of disorders rather than a single disease entity (12), and dissecting the multitude of factors that contribute to the disease development has been difficult and frustrating not only in MS, but also in other autoimmune diseases (32). Within the immune system, many of the cytokines or chemokines and their respective receptors have pleiotropic and redundant functions, and act in complex networks rather than as single factors (33). This complexity results in subtle changes that are difficult if at all possible to detect (34).

The simultaneous analysis of the expression of thousands of genes allows the exploration of interactions within many components of entire networks, becoming elective instrument to recognize pathophysiological pathways (35,36). The results

of our investigation clearly support these points of view. We have found that, even without obtaining perfect separation, subtle but distinguishable gene expression differences between healthy and MS subjects can be identified in peripheral blood cells. In our opinion the misclassification of a few samples of healthy individuals is not surprising and it is conceivable that the set of genes we have identified that best classify MS samples will probably be complemented and change. In fact, consistent with the model of a multifactorial disease, neither single genes nor a fixed group of genes are expected to function as perfect classifiers, and limits are posed by the study design (case–control) with a relatively restricted number of samples (37).

Nevertheless, our findings are in line with previous reports in MS, underscore basic principles and document new concepts. We propose that the process which culminates in an inflammatory cascade in the white matter of the CNS, can be evaluated in the periphery where pathogenetic events take place. However, as also pointed out in a recent microarray study (16), we acknowledge that there is a distinction between the genes that predispose to the disease (susceptibility alleles) and those whose expression is altered in an affected individual as a downstream effect of other genes. Subsets of genes confer 'identity' and account for direct genotype–phenotype correlation, whereas others may function as indicators of the disease status, and by identifying the former we may dissect the interaction of multiple genes each weakly contributing to disease susceptibility. Thus gene expression profiling may serve as an important adjunct in the search for candidate genes in multifactorial disorders (38).

With respect to the identified genes, some of potential interest for MS are presented in the Results section. Pathways may be defined and it is remarkable how the potentially coordinated regulation of genes can be outlined (see Results). For many of the transcripts for which we report a differential expression for the disease samples (e.g. *IL-7R*, *ZAP70*, *TNFRSF7*) (39), the role in autoimmunity is intuitive. However, for other genes it is less so and we gain confidence of their involvement on the solid basis of similar findings in analogous studies (14,15). Among the most informative candidates, we find worth particular attention the reduced transcription of the heat shock protein-70, together with the group of the histones and the kinase-2 of the CDC28 complex (*CKS2*). Although the real link of these candidates to autoimmunity remains speculative, we fully endorse the interpretation provided by Maas and colleagues (15) that these genes as a whole signal for an altered regulation, relative to controls, of the cell cycle in favor of an anti-apoptotic program. The direct correlation between autoimmunity and regulation of apoptosis is a relatively new concept of increasing interest (40), strongly supported by characteristics of the autoimmune lympho-proliferative syndrome (ALPS) in human (41,42), and mouse models of systemic lupus erythemathosus (SLE) (43), and our data suggest that parallels also exist in this T cell-mediated disease, i.e. MS.

When performing a case–control study the information obtained depends on the samples employed. As mentioned above, this may represent an important limit because the samples may not be representative of the general population. It is conceivable, therefore, that two distinct studies may identify different genes. However, we believe that the revealed pathways should be focused on. For instance, *UBE2G*, *UBE2M* and *POH1* encode for proteins functioning in the ubiquitin–proteasome complex and were detected as underexpressed in the autoimmune group of samples in the study by Maas *et al.* (15). We consider the presence in our gene list of *HISPA1A*, also a relatively reduced transcript, as further evidence of the involvement of this pathway in autoimmunity, probably due to its role both in the intracellular processing of cytokine mRNA (27), and the cell cycle (44), or perhaps due to function in protein degradation during antigen presentation by the antigen presenting cells (45), present in lower percentage in PBMC. Similarly, Ramanathan *et al.* (14), reporting the overexpression of transcripts for *LCK*, a protein–tyrosine kinase, a key component of the TCR signal transduction, agrees with our finding on *ZAP70* and transcription factors involved in TCR expression regulation (e.g. *ZNF148*, *TCF7*), in sustaining the relevance of this pathway for autoimmunity. Further examples may be extracted, again by simple means of comparison, such as the recognition of increased levels of matrix metalloproteinase-19 (*MMP-19*) in the previous study (14), paralleled by low levels of the inhibitor of metalloproteinase-1 (*TIMP1*) in our investigation.

In regard to the search for susceptibility loci, it is noteworthy how in this study the chromosome region 6p21 stands out among the others for harboring both the HLA complex, the histone cluster and the heat shock protein-70.

In conclusion, we are providing new evidence that the 'expression signature' of MS as detected in PBMC shows an overall weak signal, suggesting that a fine regulation of a large number of factors has to be kept in place to maintain immunological tolerance and a slight deregulation may account for its unbalance and subsequent pathology. In light of our results, further investigations addressing specific questions on the mechanisms underlying the etiology and pathogenesis of MS will be needed. Such research should encompass longitudinal follow-up of larger patient cohorts, and the information should be complemented by patient stratification based on MS pathology, immunological markers, clinical and magnetic resonance imaging phenotype as well as response to therapeutic intervention. We anticipate that gene expression profiling, besides being utilization for focused research into the role of individual functional pathways, will evolve into an important technique to bridge studies on diseases of complex immunogenetic background, such as MS, to protein expression profiling.

MATERIALS AND METHODS

Patient samples

PBMC samples were either collected fresh or isolated from frozen cells and divided into two sets as they were hybridized onto different batches of slides (see below). Set 1 included freshly collected (14 MS and seven normals) as well as frozen samples (three MS and two normals). Set 2 was composed of 10 MS and 10 healthy donors whose PBMC were stored frozen at the cell bank of the NIB/NIH. All patients enrolled in the study were free of therapy for at least 2 months. Patient and donor characteristics are shown in Table 2. Fresh PBMC were

Table 2. Sample characteristics

Sample code	Age	Gender	Set	Disease type	HLA-DR-B1[a]	Length of disease (years)
Controls						
H1	N.A.	M	1	N.A.	1501	—
H2	37	M	1	N.A.	0302/1501	—
H3[b]	N.A.	M	1	N.A.	N.A.	—
H4a-H4b[b]	N.A.	M	1	N.A.	N.A.	—
H5	34	M	1	N.A.	11/16	—
H6	36	M	1	N.A.	07/1501	—
H7	43	M	1	N.A.	N.A.	—
H8	29	M	1	N.A.	0103/1502	—
H9	36	M	1	N.A.	07/1501	—
H11	31	F	2	N.A.	N.A.	—
H12	29	F	2	N.A.	0101/1501	—
H14	48	F	2	N.A.	07/1501	—
H15	49	M	2	N.A.	0701/1501	—
H16	46	F	2	N.A.	0301/0801	—
H17	39	M	2	N.A.	0302/1501	—
H18	28	F	2	N.A.	0301/1501	—
H19	N.A.	M	2	N.A.	07/15	—
H20	44	M	2	N.A.	0101	—
H21	33	M	2	N.A.	15/16	—
MS patients						
M1	40	M	1	RR	1101/1501	3
M2	48	F	1	RR	07/08	0.3
M3	47	F	1	RR	0701/1401	17
M4	47	F	1	SP	0401/1301	29
M5	31	F	1	RR	01/01	2
M6	33	F	1	RR	1101/1302	5
M7a-M7b	47	F	1; 2	RR	0301/1601	1
M8a-M8b	44	M	1; 2	RR	11/1501	3
M9	51	M	1	RR	01/11	16
M10a-M10b	36	F	1	SP	04/15	13
M11	34	F	1	RR	04/15	0.5
M12	46	F	1	RR	04/15	0.3
M13	36	F	1	RR	0101/1302	5
M14	40	M	1	SP	1101/1401	25
M15a-M15b	44	F	1	SP	01/11	9
M16	41	F	1	RR	4/7	14
M17	51	M	2	RR	11/1501	1
M18	37	F	2	RR	13/1501	0.5
M19	41	M	2	SP	1101/1401	25
M20	49	M	2	RR	1104/1302	5
M21	42	M	2	RR	11/1501	2
M22	42	M	2	SP	14/16	5
M23	36	F	2	RR	0401/1501	15
M24	44	F	2	RR	0404/1501	4

[a]Only the DRB1* allele is shown; if only two digits are listed, serological typing was performed.
[b]RNA pool from multiple controls.
N.A. = not available. 'a' and 'b' designate individuals who donated two samples (fresh and frozen).

isolated from lymphocytopheresis by density gradient centrifugation (Biowitthaker, Walkersville, MD, USA), the cells were lysed by TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and then frozen at −80°C. RNA isolation was performed according to a standard protocol in use at NHGRI (13). RNA from frozen PBMC was prepared as above.

Microarray slides

The experiments were conducted with fluorescent labeling of mRNA samples according to the standard protocol of NHGRI, followed by hybridization onto glass slides (13). Two sets of samples were hybridized onto two sets of slides, with partial overlap of the cDNA clones. Set 1 was hybridized onto slides containing about 6500 clones [4000 obtained under a Cooperative Research and Development Agreement with ResGen (Huntsville, AL) and 2500 from the germinal center library of the 'lymphochip'] (46). For set 2 we used arrays containing 7500 cDNA clones (6500 from ResGen, 1000 from the lymphochip).

Data analysis

We filtered genes by requiring that a gene should have a minimal intensity across all experiments. Each slide was then normalized such that the relative (or normalized) intensity (RI)

for each gene was defined as RI = intensity of that spot/mean intensity of filtered genes (47). The logarithm of RI was used as a measure of the expression levels.

To predict whether samples were from MS patients or healthy donors, we used a recently developed algorithm that finds strong feature sets and design classifiers based on those feature sets (17). The analysis of gene expression data poses a fundamental conceptual problem: there are very large numbers of genes (features) and typically a small number of experiments (samples), rendering it difficult to find the best features from which to construct a classifier (48). Furthermore, once a set of features is selected, a classifier is designed and its error estimated. For small number of samples, an error estimator may still be unbiased, but due to large variance it often yields very optimistic estimates (over-fitting). As a consequence, a large number of feature sets and classifiers may result in low error estimates, while the true error remains high. Conversely, our novel method of analysis mitigates the small-sample problem. (a) It designs classifiers using only a small number of features, and (b) once the features are identified, assesses the probability distribution that results from spreading the mass of the sample points. This makes the classification more difficult, while maintaining sample geometry. By considering increasing amounts of spread, the algorithm finds feature sets whose classification accuracy remains good relative to greater spreading of the sample. The error then gives a measure of the strength of the feature set as a function of the spread.

We then used linear classifiers because of their simplicity and an available analytical representation for both the classifier design and its error, which help exploit many possible feature sets, more than 10 billion in this study. We tested single, two and three genes in a classifier, for four different amounts of spread (σ_n). The misclassification error, defined in terms of σ error as described (17), was estimated for $\sigma_n = 0.6$. A σ error of 5% has been shown for $\sigma_n = 0.6$ to result in robust classifiers with good predictive ability (17). After designing the classifiers using set 1, we applied them to set 2 as an independent set to assess their predictive ability. Lastly, the two sets of samples were combined to generate the final list of marker genes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We acknowledge S. Panavally for her help in the data analysis, D. Leja for help with the illustrations, and A. Glatfelter and J. Leuders for excellent technical support. We also would like to thank S. Leitman and her staff, Department of Transfusion Medicine, Clinical Center, NIH, for providing leukaphereses, T. Simonis for HLA typing, J. Ohayon, J. McCartin and H. Griffith, for patient scheduling and nursing care, and E. Suh and R. Martino for providing computing resources at the Center for Information Technology, NIH. M.R. was supported by a post-doctoral fellowship from the Swedish Research Council.

REFERENCES

1. Noseworthy, J.H., Lucchinetti, C., Rodriguez, M. and Weinshenker, B.G. (2000) Multiple sclerosis. *New Engl. J. Med.*, **343**, 938–952.
2. Martin, R, McFarland, H.F. and McFarlin, D.E. (1992) Immunological aspects of demyelinating diseases. *A. Rev. Immunol.*, **10**, 153–187.
3. Ebers, G.C., Sadovnick, A.D. and Risch, N.J. (1995) A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group. *Nature*, **377**, 150–151.
4. Sadovnick, A.D. (1994) Genetic epidemiology of multiple sclerosis: a survey. *Ann. Neurol.*, **36** (Suppl. 2), S194–203.
5. Sadovnick, A.D., Ebers, G.C., Dyment, D.A. and Risch, N.J. (1996) Evidence for genetic basis of multiple sclerosis. The Canadian Collaborative Study Group. *Lancet*, **347**, 1728–1730.
6. Robertson, N.P., Fraser, M., Deans, J., Clayton, D., Walker, N. and Compston, D.A. (1996) Age-adjusted recurrence risks for relatives of patients with multiple sclerosis. *Brain*, **119**, 449–455.
7. Dyment, D.A., Sadovnick, A.D., Ebers, G.C. and Sadnovich, A.D. (1997) Genetics of multiple sclerosis. *Hum. Mol. Genet.*, **6**, 1693–1698.
8. Sawcer, S., Goodfellow, P.N. and Compston, A. (1997) The genetic analysis of multiple sclerosis. *Trends Genet.*, **13**, 234–239.
9. Compston, A. (1999) The genetic epidemiology of multiple sclerosis. *Phil. Trans. R. Soc. Lond. B Biol. Sci.*, **354**, 1623–1634.
10. Compston, A. and Coles, A. (2002) Multiple sclerosis. *Lancet*, **359**, 1221–1231.
11. Oksenberg, J.R. and Barcellos, L.F. (2000) The complex genetic aetiology of multiple sclerosis. *J. Neurovirol.*, **6** (Suppl. 2), S10–14.
12. Lucchinetti, C., Bruck, W., Parisi, J., Scheithauer, B., Rodriguez, M. and Lassmann, H. (2000) Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. *Ann. Neurol.*, **47**, 707–717.
13. Bittner, M.L., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.
14. Ramanathan, M., Weinstock-Guttman, B., Nguyen, L.T., Badgett, D., Miller, C., Patrick, K., Brownscheidle, C. and Jacobs, L. (2001) *In vivo* gene expression revealed by cDNA arrays: the pattern in relapsing–remitting multiple sclerosis patients compared with normal subjects. *J. Neuroimmunol.*, **116**, 213–219.
15. Maas, K., Chan, S., Parker, J., Slater, A., Moore, J., Olsen, N. and Aune, T.M. (2002) Cutting edge: molecular portrait of human autoimmune disease. *J. Immunol.*, **169**, 5–9.
16. Whitney, A.R., Diehn, M., Popper, S.J., Alizadeh, A.A., Boldrick, J.C., Relman, D.A. and Brown, P.O. (2003) Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA*, **100**, 1896–1901.
17. Kim, S., Dougherty, E.R., Barrera, J., Chen, Y., Bittner, M.L. and Trent, J.M. (2002) Strong feature sets from small samples. *J. Comput. Biol.*, **9**, 127–146.
18. Wandinger, K.P., Sturzebecher, C.S., Bielekova, B., Detore, G., Rosenwald, A., Staudt, L.M., McFarland, H.F. and Martin, R. (2001) Complex immunomodulatory effects of interferon-beta in multiple sclerosis include the upregulation of T helper 1-associated marker genes. *Ann. Neurol.*, **50**, 349–357.
19. Stafforini, D.M., McIntyre, T.M., Zimmerman, G.A. and Prescott, S.M. (1997) Platelet-activating factor acetylhydrolases. *J. Biol. Chem.*, **272**, 17895–17898.
20. Hendriks, J., Gravestein, L.A., Tesselaar, K., van Lier, R.A., Schumacher, T.N. and Borst, J. (2000) CD27 is required for generation and long-term maintenance of T cell immunity. *Nat. Immunol.*, **1**, 433–440.
21. Beall, S.S., Biddison, W.E., McFarlin, D.E., McFarland, H.F. and Hood, L.E. (1993) Susceptibility for multiple sclerosis is determined, in part, by inheritance of a 175-kb region of the TcR V beta chain locus and HLA class II genes. *J. Neuroimmunol.*, **45**, 53–60.
22. Sloan-Lancaster, J., Shaw, A.S., Rothbard, J.B. and Allen, P.M. (1994) Partial T cell signaling: altered phospho-zeta and lack of zap70 recruitment in APL-induced T cell energy. *Cell*, **79**, 913–922.
23. Merchant, J.L., Iyer, G.R., Taylor, B.R., Kitchen, J.R., Mortensen, E.R., Wang, Z., Flintoft, R.J., Michel, J.B. and Bassel-Duby, R. (1996) ZBP–89, a Kruppel-like zinc finger protein, inhibits epidermal growth factor induction of the gastrin promoter. *Mol. Cell. Biol.*, **16**, 6644–6653.

24. van de Wetering, M., Oosterwegel, M., Holstege, F., Dooyes, D., Suijkerbuijk, R., Geurts van Kessel, A. and Clevers, H. (1992) The human T cell transcription factor-1 gene. Structure, localization, and promoter characterization. *J. Biol. Chem.*, **267**, 8530–8536.
25. Xue, H.H., Kovanen, P.E., Pise-Masison, C.A., Berg, M., Radovich, M.F., Brady, J.N. and Leonard, W.J. (2002) IL-2 negatively regulates IL-7 receptor alpha chain expression in activated T lymphocytes. *Proc. Natl Acad. Sci. USA*, **99**, 13759–13764.
26. Salvetti, M., Ristori, G., Buttinelli, C., Fiori, P., Falcone, M., Britton, W., Adams, E., Paone, G., Grasso, M.G. and Pozzilli, C. (1996) The immune response to mycobacterial 70-kDa heat shock proteins frequently involves autoreactive T cells and is quantitatively disregulated in multiple sclerosis. *J. Neuroimmunol.*, **65**, 143–153.
27. Laroia, G., Cuesta, R., Brewer, G. and Schneider, R.J. (1999) Control of mRNA decay by heat shock–ubiquitin–proteasome pathway. *Science*, **284**, 499–502.
28. Young, J.C., Hoogenraad, N.J. and Hartl, F.U. (2003) Molecular chaperones Hsp90 and Hsp70 deliver preproteins to the mitochondrial import receptor Tom70. *Cell*, **112**, 41–50.
29. Lever, M.A., Th'ng, J.P., Sun, X. and Hendzel, M.J. (2000) Rapid exchange of histone H1.1 on chromatin in living human cells. *Nature*, **408**, 873–876.
30. Reed, S.I. and Wittenberg, C. (1990) Mitotic role for the Cdc28 protein kinase of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **87**, 5697–5701.
31. Brenner, D.A., O'Hara, M., Angel, P., Chojkier, M. and Karin, M. (1989) Prolonged activation of jun and collagenase genes by tumour necrosis factor-alpha. *Nature*, **337**, 661–663.
32. Wakeland, E.K., Liu, K., Graham, R.R. and Behrens, T.W. (2001) Delineating the genetic basis of systemic lupus erythematosus. *Immunity*, **15**, 397–408.
33. Kovanen, P.E., Rosenwald, A., Fu, J., Hurt, E.M., Lam, L.T., Giltnane, J.M., Wright, G., Staudt, L.M. and Leonard, W.J. (2003) Analysis of gamma c-family cytokine target genes. Identification of dual-specificity phosphatase 5 (DUSP5) as a regulator of mitogen-activated protein kinase activity in interleukin-2 signaling. *J. Biol. Chem.*, **278**, 5205–5213.
34. Todd, J.A. and Wicker, L.S. (2001) Genetic protection from the inflammatory disease type 1 diabetes in humans and animal models. *Immunity*, **15**, 387–395.
35. Bowcock, A.M., Shannon, W., Du, F., Duncan, J., Cao, K., Aftergut, K., Catier, J., Fernandez-Vina, M.A. and Menter, A. (2001) Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Hum. Mol. Genet.*, **10**, 1793–1805.
36. Zhou, X., Krueger, J.G., Kao, M.C., Lee, E., Du, F., Menter, A., Wong, W.H. and Bowcock, A.M. (2003) Novel mechanisms of T-cell and dendritic cell activation revealed by profiling of psoriasis on the 63,100-element oligonucleotide array. *Physiol. Genom.*, **13**, 69–78.
37. Lander, E.S. and Schork, N.J. (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.
38. Wright, A.F. and Van Heyningen, V. (2001) Medicine: short cut to disease genes. *Nature*, **414**, 705–706.
39. Sharief, M.K. and Hentges, R. (1991) Association between tumor necrosis factor-alpha and disease progression in patients with multiple sclerosis. *New Engl. J. Med.*, **325**, 467–472.
40. Balomenos, D. and Martinez, A.C. (2000) Cell-cycle regulation in immunity, tolerance and autoimmunity. *Immunol Today*, **21**, 551–555.
41. Martin, D.A., Zheng, L., Siegel, R.M., Huang, B., Fisher, G.H., Wang, J., Jackson, C.E., Puck, J.M., Dale, J., Straus, S.E. *et al.* (1999) Defective CD95/APO-1/Fas signal complex formation in the human autoimmune lymphoproliferative syndrome, type Ia. *Proc. Natl Acad. Sci. USA*, **96**, 4552–4557.
42. Wang, J., Zheng, L., Lobito, A., Chan, F.K., Dale, J., Sneller, M., Yao, X., Puck, J.M., Straus, S.E. and Lenardo, M.J. (1999) Inherited human caspase 10 mutations underlie defective lymphocyte and dendritic cell apoptosis in autoimmune lymphoproliferative syndrome type II. *Cell*, **98**, 47–58.
43. Salvador, J.M., Hollander, M.C., Nguyen, A.T., Kopp, J.B., Barisoni, L., Moore, J.K., Ashwell, J.D. and Fornace, A.J. Jr (2002) Mice lacking the p53-effector gene Gadd45a develop a lupus-like syndrome. *Immunity*, **16**, 499–508.
44. Tanahashi-Hori, T., Tanahashi, N., Tanaka, K. and Chiba, T. (2003) Conditional knockdown of proteasomes results in cell-cycle arrest and enhanced expression of molecular chaperones Hsp70 and Hsp40 in chicken DT40 cells. *J. Biol. Chem.*, **278**, 16237.
45. Klein, J. and Sato, A. (2000) The HLA system. First of two parts. *New Engl. J. Med.*, **343**, 702–709.
46. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
47. Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer, P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.
48. Dougherty, E.R. (2001) Small sample issues for microarray-based classification. *Comp. Funct. Genom.*, **2**, 28–34.