Microarray Analysis of Autoimmune Diseases by Machine Learning Procedures

Rubén Armañanzas, Borja Calvo, Iñaki Inza, Marcos López-Hoyos, Víctor Martínez-Taboada, Eduardo Ucar, Irantzu Bernales, Asier Fullaondo, Pedro Larrañaga, and Ana M. Zubiaga

***Abstract*—Microarray-based global gene expression profiling, with the use of sophisticated statistical algorithms is providing new insights into the pathogenesis of autoimmune diseases. We have applied a novel statistical technique for gene selection based on machine learning approaches to analyze microarray expression data gathered from patients with systemic lupus erythematosus (SLE) and primary antiphospholipid syndrome (PAPS), two autoimmune diseases of unknown genetic origin that share many common features. The methodology included a combination of three data discretization policies, a consensus gene selection method, and a multivariate correlation measurement. A set of 150 genes was found to discriminate SLE and PAPS patients from healthy individuals. Statistical validations demonstrate the relevance of this gene set from an univariate and multivariate perspective. Moreover, functional characterization of these genes identified an interferon-regulated gene signature, consistent with previous reports. It also revealed the existence of other regulatory pathways, including those regulated by PTEN, TNF, and BCL-2, which are altered in SLE and PAPS. Remarkably, a significant number of these genes carry E2F binding motifs in their promoters, projecting a role for E2F in the regulation of autoimmunity.**

***Index Terms*—Antiphospholipid syndrome, DNA microarrays, gene profiling, machine learning, systemic lupus erythematosus.**

I. Introduction

DNA MICROARRAY technology [1] offers the possibility to simultaneously analyze the expression of hundreds to thousands of genes [2]. This approach has been applied successfully to better classify many cancers and to understand the molecular pathways involved in several pathologies [3]. Genome-wide gene expression profiles of autoimmune diseases, such as systemic lupus erythematosus, rheumatoid arthritis, or Sjogren's syndrome have also been obtained [4]. These studies have identified genes with a dysregulated expression in autoimmune diseases. Further application of microarray analyses should facilitate the identification of pathways that are common in autoimmunity, but more importantly, genes and pathways that uniquely define patients with a particular disease phenotype, which could be useful for the development of specific treatments [5].

Despite the demonstrated power of high-throughput gene expression profiling approaches, some important limitations have been noted. DNA microarray analyses are typically hypothesis-driven, in the sense that the experiments are designed to address a scientific question [6], an approach that could lead to a biased interpretation of the results. Additionally, because microarrays are inherently noisy, they impact on data quality [7]. Moreover, present microarray studies usually include a very low number of samples under study. In this context, the reliability of a single data mining technique is no guarantee at all. Clear evidence of these effects is the differences found within the results of data analysis techniques for the same microarray data [8].

The discipline of machine learning, in combination with data mining techniques has been very useful in diverse fields of research, including the bioinformatics discipline, to overcome the technology-intrinsic data noise and to obtain relevant knowledge out of a high volume of data [9]. An important feature of this method relies on the fact that no prior knowledge of the system under study is necessary to run the analysis, thus constituting a blind process for which the final results are only based on the characteristics of the raw data. Due to this blindness, a strict validation of the results needs to be tackled: statistical relevance, laboratory qPCR validation, bibliographic revision, regulatory activity evaluation, and dysregulation of transcription factors, among others.

We have now applied machine learning procedures to DNA microarray data derived from samples of patients suffering from systemic lupus erythematosus (SLE) and primary antiphospholipid syndrome (PAPS), two autoimmune conditions with overlapping immunological features, in order to obtain an unbiased identification of genes that could be relevant to the pathogenesis of these diseases. We propose that the computational tools employed in this paper add robustness to the microarray-based identification of biomarkers.

Manuscript received March 15, 2007; revised December 4, 2007. Current version published May 6, 2009. This work was supported in part by Basque Government under Grant Etortek-IE019, Grant Saiotek-SA-2005/00093, and Research Groups 2007–2012 Program. The work of P. Larrañaga and A. M. Zubiaga was supported in part by the Spanish Ministry of Science and Innovation under the TIN2008-06815-C02-01 Research Project. The work of V. Martínez-Taboada was supported in part by the Grant PI050475 from the Spanish Ministry of Health, and API 06/05 from the Fundación Marqués de Valdecilla and from the Fundación Mutua Madrileña. The work of R. Armañanzas was supported by Basque Government Fellowship for graduate studies (BFI05.430).

R. Armañanzas, B. Calvo, and I. Inza are with the Department of Computer Science and Artificial Intelligence, University of the Basque Country, 20080 San Sebastian, Spain (e-mail: ruben@si.ehu.es; borxa@si.ehu.es; inza@si.ehu.es).

M. López-Hoyos and V. Martínez-Taboada are with Marqués de Valdecilla Hospital, 39008 Santander, Spain (e-mail: inmlhm@humv.es; vmartinezt@medynet.com).

E. Ucar is with Basurto Hospital, 48013 Bilbao, Spain (e-mail: educar@hbas.osakidetza.net).

I. Bernales, A. Fullaondo, and A. M. Zubiaga are with the Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, 48080 Bilbao, Spain (e-mail: ggpbepui@lg.ehu.es; ggpfuela@lg.ehu.es; ggpzuela@lg.ehu.es).

P. Larrañaga is with the Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, 28660 Madrid, Spain (e-mail: pedro.larranaga@fi.upm.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2008.2011984

The paper is divided into four main sections. Section II presents the materials and methods used in this research giving as much technical information as possible for each tackled stage. The results and their validations are gathered and discussed in Section III. This section is divided into two sections (Section III-A and III-B), each of which focus on the statistical and the biological validation, respectively. Brief conclusions are finally provided in Section IV.

II. Materials and Methods

A. Study Participants

After informed consent, patients and controls provided a peripheral blood sample, and peripheral blood mononuclear cell (PBMC) were isolated from whole blood by ficoll gradient purification (GE Helathcare Bio-Sciences, Piscataway, NJ, USA). All patients were Caucasian women, and had physician-verified SLE or PAPS. Data on age, clinical characteristics, disease activity, and current medication are summarized in the online *supplementary content.*[1] Disease activity in SLE patients was determined using systemic lupus erythematosus disease activity index (SLEDAI) score [10].

B. Sample Processing and Chip Hybridization

For microarray experiments, four patients with SLE, two patients with primary APS, and five healthy individuals were used (see supplementary content). RNA was extracted from PBMC using triZOL (Invitrogen, Carlsbad, CA, USA) followed by RNeasy cleanup (Qiagen, Hilden, Germany). The isolated RNA was amplified and labeled as described in the *GeneChip Expression Analysis Technical Manual* (Affymetrix, CA, USA), and subsequently hybridized to HG-U133A Genechip microarrays (Affymetrix, CA, USA) and scanned according to the manufacturer's recommendations. The labeling, hybridization and scanning procedures were carried out in Progenika (Derio, Spain).

C. Data Acquisition and Preprocessing

After scanning the arrays, Affymetrix microarray suite (MAS) 5.0 software was used for the compilation and initial analysis of the raw datasets. Four different quality values were measured in order to evaluate the reliability of the microarrays that were used in the experiment: spike control BioB, housekeeping control GAPDH, P-call percentage, and array outlier percentage. Only the arrays that complied with the reliability criteria [11] were considered for further analysis.

Affymetrix is a one-channel technology [1] that only includes one sample on each microarray. To study the intensity change shown by each probe between test and reference samples, comparisons were made between the reference and the test microarray. On the basis of these comparisons, a second filter level was constituted: all probes showing an absent detection value in both microarrays were discarded from the analysis. Microarray internal control sequences were also removed, and the final amount of probes under study was set at 8,808. The known logRatio between the two channel values was computed for each of the 8,808 probes, producing a total of 40 instances or comparisons (healthy versus healthy; SLE versus healthy; PAPS versus healthy).

[1]Supplementary data and results for this study is available at: http://www.sc.ehu.es/ccwbayes/members/ruben/sle/

D. Relevant Gene List Identification

From the statistics and data mining fields, the expression level of each gene is represented as a *random variable* of a probabilistic process. A great number of machine learning methods are designed to deal only with discrete data, so, it becomes necessary to translate the microarray data from continuous to discrete value domains. However, this process can bias the original data and degrade their original quality. In order to amplify the robustness of the knowledge discovery process, and to overcome this possible bias, the use of a set of different discretization techniques is suggested [12]. Thus, beginning with a microarray dataset discretized in different ways, we search for a consensus result with larger reliability and robustness than usual single-discretization modeling. Due to the small sample sizes of gene expression databases, the community of researchers was rapidly aware about the potentialities of model consensus approaches to improve the robustness and stability of final discriminative rules [13], [14].

The consensus procedure that was put forward comprises the best techniques used in various data analysis fields. First, a set of discretization policies, namely: equal width [15], equal frequency [16], and entropy [17]. Given a number of bins, b, equal width simply sorts the values a feature can take and divides the observed range into b equally sized intervals. Equal frequency divides the range into b bins that gather the same number of occurrences. As a usual criterion in this field [18], [19], our assumption is that a gene could be in three possible states, using the idea of over-, under- or baseline activity, so the number of bins is set to three. For its part, entropy discretization makes use of the phenotype distribution over the data, in conjunction with a minimum description length-based algorithm [20]. For each feature independently, this technique finds the appropriate cutoff points in such a way that the phenotype entropy within each resulting interval is minimum, while balancing this by introducing as few cutoff points as possible.

Second, a filter-like selection procedure to detect differentially expressed genes among the studied phenotypes (correlation feature selection) [21]. CFS belongs to the filter optimal subset selection techniques [22], it addresses two fundamental issues: avoid both redundancy and irrelevancy in the selected subset of genes. Making use of the uncertainty coefficient [21] and a classical forward greedy hill-climbing search strategy, CFS is able to identify the genes most correlated with the phenotype distribution keeping the redundancy among them minimum. Bear in mind that there can be genes with a high phenotype correlation coefficient that are not included in the subset finally selected, so a third stage to amplify the relevant gene set is mandatory.

And third, a statistical coexpression measure (a classical probability theory metric—mutual information—[23]). At this last

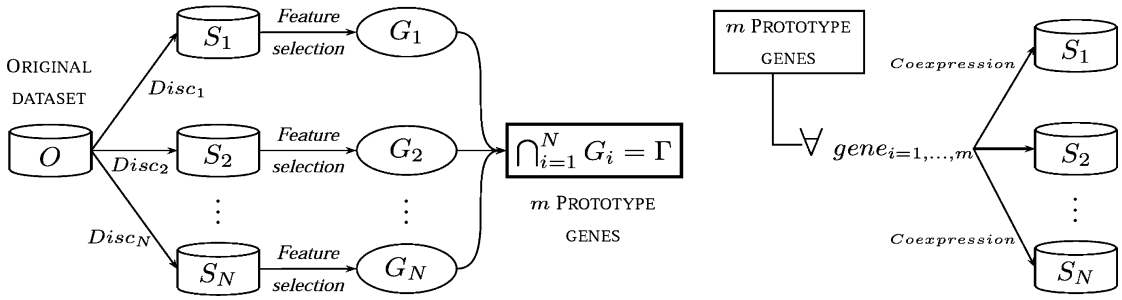

Fig. 1. Machine learning data flow to identify the relevant gene set: identifying the prototype genes and the genes mostly correlated with them.

stage, we look for those possible genes that, having a biologically relevant role in the problem, have not enough statistical power to be considered as a separate prototype. Notice that, using the mutual information metric, and due to the fact that it has no sign consideration, the relationships found could be direct and inverse in the gene profiling. Thus, it can cover a key biological process: positive or negative transcription regulation.

Formally, let the discrete datasets $S_1, \ldots, S_N$ be the result of different discretization policies of the original O microarray dataset. N different feature subset selections are performed on the basis of these S_i discrete datasets, producing the following subsets of genes: $G_1, \ldots, G_N$. The consensus gene subset Γ (hereafter statistical prototypes) will be the intersection between all of them, that is $\Gamma = \bigcap_{i=1}^{N} G_i$, with $\mid \Gamma \mid = m \leq \min_{i=1,\ldots,N} \mid G_i \mid$. In order to amplify the final output gene set, for each statistical prototype gene, its q most univariately correlated genes are also selected. Fig. 1 graphically exemplifies the whole information flow.

The objective of the overall process was to enhance the robustness of the final solution. The use of different discretization procedures adds independence from a specific discretization task, and obtains a compact and biologically meaningful gene set $\Gamma = \bigcap_{i=1}^{N} G_i$. The posterior enlargement of this statistical outcome can add information contained in each of the starting gene sets. The complete formulation of this novel combination of techniques and their comparison with other state-of-the-art approaches can be widely reviewed in [12].

E. Quantitative PCR

DNA purified from six healthy donors, three SLE patients, and five PAPS patients (see supplementary content), all of them different from those used in microarray experiments, was reversed transcribed into cDNA.

Quantitative TaqMan PCR analyses were performed for the following genes: H1F0, PPIA, GNLY, SSB, and SP100. In addition, qPCR of TBP was performed on all samples, which served as internal control. The primers and TaqMan probes for all the genes were obtained from Applied Biosystems (ABI, Foster City, CA, USA). The reactions were run by triplicate on an ABI 7900HT fast real-time PCR system from Applied Biosystems at the Genomics Facility of the University of the Basque Country, using standard cycling conditions. Results were analyzed with the sequence detection system (SDS) Software v2.0 (Applied Biosystems) to obtain the Ct values for each sample.

A DCt value was calculated reflecting the difference between the average Ct of the replicate samples obtained for the control gene (TBP) and the average Ct of the replicate samples obtained for the test gene to be validated. Using these DCt values as the raw expression value in the qPCR experiment, we first determined the median DCt for all the healthy control samples. Next, we calculated the difference between the DCt of each test sample and the DCt values of the healthy controls, thus obtaining a set of DDiff values for each phenotype and gene.

III. RESULTS AND DISCUSSION

SLE is an autoimmune disease characterized by the production of autoantibodies with specificity for a wide range of self-antigens, resulting in injury to various organ systems, including skin, joints, kidney, and central nervous system [24]. PAPS is a related autoimmune disease characterized by recurrent thrombosis and miscarriages, associated with antiphospholipid autoantibodies (aPL) directed against phospholipid-binding plasma proteins as well as with other autoantibodies shared with SLE [25].

The antiphospholipid syndrome can manifest on its own, in the absence of lupus symptoms (primary APS, PAPS), or can develop secondarily in a subset of lupus patients, implying that some pathogenic pathways are common to both autoimmune diseases. To identify genes that discriminate SLE or PAPS patients from healthy controls, a novel set of computational tools was used based on a machine learning approach using the logRatio data from DNA microarrays, as indicated in Section II. A total of 150 probes (hereafter genes) out of an original set containing more than 22,000 genes were detected as the relevant genes whose differential expression confidently discriminates among SLE patients, PAPS patients, and healthy controls.

The complete list of relevant genes, including the Affymetrix probe ID, the gene symbol, their locus, and their relative gene expression in SLE or PAPS patients, and a short description are available online through the Supplementary content page. The significantly different expression profile exhibited by these genes in patients samples relative to samples of healthy controls could contribute to the pathogenesis of the autoimmune conditions analyzed in this paper.

The sample labeled as LB10 corresponds to an SLE patient who developed secondary APS. We wished to determine whether the gene expression profile of this individual was more similar to primary APS or whether it maintained an SLE pattern of gene expression. To this end, hierarchical clustering

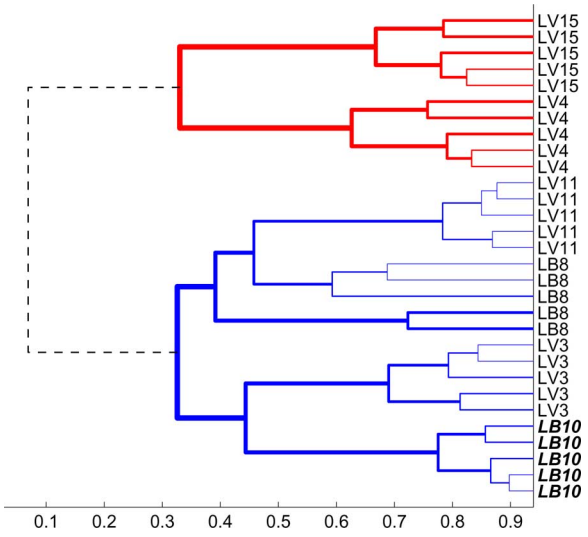

Fig. 2. Clustering of the illness instances (SLE and PAPS). Branches colored in blue belong to SLE-class-labeled instances, while branches in red belong to PAPS instances. The expression profiles of each test sample were compared separately with the expression profiles of control samples, resulting in five comparisons for each SLE or PAPS sample. The five instances generated by sample LB10 flawlessly behave as the rest of SLE instances, being all clustered in the same group.

analysis was carried out considering the most common clustering parameters: Pearson correlation and average linkage. This analysis clearly showed that LB10 is clustered with the rest of SLE patients and not with PAPS patients (see Fig. 2). Furthermore, clustering of all SLE patients was very similar and homogeneous, implying that a secondary acquisition of an antiphospholipid syndrome does not modify significantly the transcriptional profile that characterizes SLE.

Many of the genes identified in our study have not been previously implicated in SLE or PAPS, and represent new biomarkers of these diseases. Interestingly, a link with autoimmunity, and in particular to lupus, has already been established for a significant number of the genes included in this group of dysregulated genes. Such is the case of the genes SSB, SP100, H1F0, all of which are lupus autoantigens [26]–[30] or TAP-1, a transporter gene with polymorphisms showing genetic association with SLE [31] and [32], among others.

A. *Statistical Analysis*

From the 150 total genes returned by the machine learning stage, there were a total of eight statistical prototypes (CPSF1, SLC25A12, UQCRB, NADK, MICAL2, KIAA0776, PARL, and CECR1). It is mandatory to note that prototype genes are the result of a statistical process and their aim is to comprise the statistical axes of the problem. Although a direct biological translation could be made, it would have no biological link to the statistical and biological interpretations: the prototype genes could not have any special biological contribution in the domain of the diseases under study.

The statistical analysis of the selected genes was performed in two ways: by measuring the relevance of the selected genes, and the estimated prediction accuracy in a supervised class prediction procedure. For these two validations, the supervised target dataset was comprised of three different classes or phenotypes and 40 instances: 10 PAPS instances, 20 SLE instances, and 10 control instances (see Section II-C). *A priori*, the selected genes should be relevant to the problem, showing a high correlation degree with the phenotype distribution.

TABLE I
POSITIONS OF THE STATISTICAL PROTOTYPES OVER THE CONSENSUS RANKINGS FOR THE THREE DISCRETE SETS (EF, EW, AND ENTROPY)

Gene	Ranking EF	Ranking EW	Ranking Entropy
CPSF1	22	93	46
SLC25A12	13	26	32
UQCRB	33	33	51
NADK	106	106	142
MICAL2	11	14	20
KIAA0776	31	35	30
PARL	1	3	12
CECR1	19	38	68

Using the Elvira platform [33], and on the basis of the three discrete datasets, seven different univariate filter rankings [34] were calculated. Each of these seven metrics computes the correlation coefficient of each gene with respect to the class variable. In order to obtain an average ranking for each dataset, each gene was weighted with a coefficient proportional to the relative positions shown in each ranking. The consensus rankings are accessible through the online supplementary content page, presenting a list of the 8,808 (924 in the case of entropy discretization) genes ordered by their correlation level with respect to the problem class label.

Table I shows the positions of each statistical prototype in the consensus rankings of each discrete dataset. It is easy to check that the selected genes appear in the top positions of the rankings, with average positions of 29.5, 43.5, and 50.1 for equal frequency (EF), equal width (EW), and entropy discretization, respectively. Such average positions significatively differ from the ones obtained if a random selection is made: 4004.5, 4004.5, and 462.5, respectively.

The second aspect of the statistical analysis comprised the estimation of the prototypes' strength when classifying a new instance not included in the original set. This strength was evaluated on the basis of different classifier performance tests. Due to the great difference between the number of genes (predictive variables) and the instances of each experiment, many distinct and equally effective classifiers may exist for the same training set [34]–[37]. This fact led us to consider four different classification models instead of only one. Furthermore, and trying to cover a wide range of classical paradigms, the four models chosen come from different classification families and are commonly used in DNA microarray class prediction studies [36].

1) *Logistic Regression*—Logistic regression [38] has become a very widely used classification paradigm in life sciences because its parameters can be interpreted as risk factors. Logistic regression is based on the *logistic function* and it allows an interpretation in probability terms. A set of parameters has to be estimated from the problem data, usually known as *regression coefficients*. Usually, regression

coefficients are estimated using the maximum likelihood estimation method, but there are adaptations that penalize this maximum likelihood with other factors. The logistic regression model used in this paper penalizes the likelihood estimation with an estimator known as the *ridge* estimator [39].

2) *k-Nearest Neighbor*—The k-NN algorithm [40] proceeds with the classification task in terms of similarity: unlabeled examples are classified based on their distance to the examples in the training set. k-NN is a classification paradigm with no explicit classification model. In other words, there is no learning stage at which a mathematical model is induced, and from which the categorization stage is tackled. It finds the k closest features in the data and assigns them to the class that most frequently appears within the k-subset. In this work, k-NN is computed with Euclidean distance and a k value of three.
3) *Naïve Bayes*—Continuous naïve Bayes [41] belongs to the Bayesian classifier family. This family is based on the Bayes formulation of conditional dependencies [42]. Bayesian classifiers need to specify a structure, and then, a series of *a priori* and conditional probabilities, or model parameters. The simplest structure is the naïve Bayes structure based on the assumption of the conditional independence between the predictor variables given the class. The model parameters are estimated with a factorization based on the normal distribution assumption for each variable.
4) *Random Forest*—This classification paradigm belongs to the tree-like classification family. Random forest [43] builds a forest composed of t random trees. When building these trees, a random variable selection is performed. The random tree set is learnt using a bootstrap instance selection, and the built trees are not pruned. For our paper, no variable selection is configured at the induction step, because a feature selection has already been performed. Thus, using all the predictive variables provided, ten random trees are built for each forest.

In order to obtain a fair estimation of each classifier performance, a crucial task arose: the choice of the most suitable accuracy estimation method in the context of the microarrays. Classical methods such as hold-out, simple, or leaving-one-out cross validations [44] have been demonstrated not to fit on the intrinsic microarray dimensionality problem [45], [46]. Nowadays, there are two main approaches commonly accepted as the best estimation techniques: the *corrected bootstrap estimator* [47] and *nested stratified cross validation* [48].

For the present study, we chose the use of a nested stratified cross-validation scheme as the accuracy estimation method. This method comprises the performance of two different stages: one internal (or *inner*) loop in which the parameters of the classification methods are estimated; and an external (or *outer*) loop in which the classifiers are induced and validated against previously unseen instances. In our specific case, the feature selection methods were run throughout the inner loop and the different classifiers were induced on the basis of these selected features.

TABLE II
ESTIMATED ACCURACIES OBTAINED FOR THE TEN TIMES FOLD CROSS-VALIDATION ON EACH CLASSIFICATION PARADIGM

	$\Gamma = \bigcap_3 G_i$	$G_{\text{Eq.Width}}$	$G_{\text{Eq.Freq.}}$	G_{Entropy}
Set size	6.17±0.72	25.57±3.19	38.83±2.01	41.42±3.02
Log. reg.	86.75±3.72	95.25±3.05	95.25±2.61	95.00±3.35
Naïve Bayes	86.75±5.01	97.00±1.00	96.25±2.02	96.75±2.25▲
k-NN	88.50±4.50	100.0±0.00▲	100.0±0.00▲	98.75±2.02▲
Rand. forest	80.25±8.98	95.75±2.75	93.00±4.44	90.25±4.39

This classifier is tested over instances not previously seen on the induction stage.

The next parameter to adjust was the number of times that all this process is done for each classifier and for each feature selection method. Taking advantage of previous studies, it is proven that the ten times repetition of ten of these stratified cross validations obtains suited accuracy estimations [46], [49]. This validation scheme is usually known as ten times tenfold cross validation.

As for the study of the prototype's classification strength, we performed the validation over four different gene sets: the intermediate genes selected by the correlation feature selection over the three different discretized data sets, and the consensus prototype genes. The number of selected genes and estimated accuracies are gathered in Table II. All the induction and validation processes were computed using the *WEKA* framework [50].

To assess the significance and reliability of the consensus genes in comparison with each one of the intermediate gene sets, a *corrected repeated k-fold cv test* [49] was performed. This statistical test has been proven as one with the most suited relation between the type I and II errors [49], [51] and a high replicability degree [49]. The test compares the differences between two different classification algorithms by a special corrected t-test. The null hypothesis is that both algorithms have the same classification behavior; the alternative hypothesis states that one algorithm outperforms the classification degree of the other.

For each base classifier, the accuracy of each single discretization policy was compared with respect to the consensus approach. For all the 12 comparisons, only four of them (values with a ▲ symbol on Table II) rejected the null hypothesis for an $\alpha = 0.05$ significance level. For a 0.01 level, none of them showed statistical differences—the null hypothesis was not rejected in any of them. These results let us state that, while suffering a decrease on the classification accuracies, the differences between the use of the consensus prototypes and the intermediate selected gene sets are not statistically significant in many comparisons.

B. Biological Analysis

1) Verification of Microarray Hybridization by Quantitative RT-PCR Analysis: For verification of hybridization signals, quantitative TaqMan PCR analysis was performed for five selected genes. Two of these genes (H1F0 and SP100) are IFN-regulated genes previously associated with SLE [29], [30]. Our microarray data showed that both genes were upregulated in

TABLE III
qPCR OUTPUT VALUES AND EXPECTED ACTIVITY FOR FIVE GENES FROM THE RELEVANT GENELIST

Symbol	Phenotype	Median	1st-quartil	3rd-quartil	Expected	p-value
H1F0	Control	0.0	-0.99	0.99	–	–
	SLE	1.66	1.0	2.16	UP	< 0.0001
	PAPS	2.08	1.01	3.14	UP	0.00264
PPIA	Control	-0.32	-1.37	1.04	–	–
	SLE	0.8	0.4	1.56	BASELINE	0.12506
	PAPS	-1.39	-1.99	-0.7	DOWN	0.06598
GNLY	Control	0.0	-1.34	1.26	–	–
	SLE	-0.1	-0.79	0.58	BASELINE	0.95635
	PAPS	-2.65	-3.26	-1.93	DOWN	< 0.0001
SSB	Control	0.0	-1.51	1.51	–	–
	SLE	0.8	-0.28	1.66	BASELINE	0.20110
	PAPS	-3.17	-3.56	-1.41	DOWN	< 0.0001
SP100	Control	-0.16	-1.09	0.97	–	–
	SLE	3.19	2.26	3.33	UP	0.00076
	PAPS	-1.22	-2.03	-0.89	DOWN	0.07220

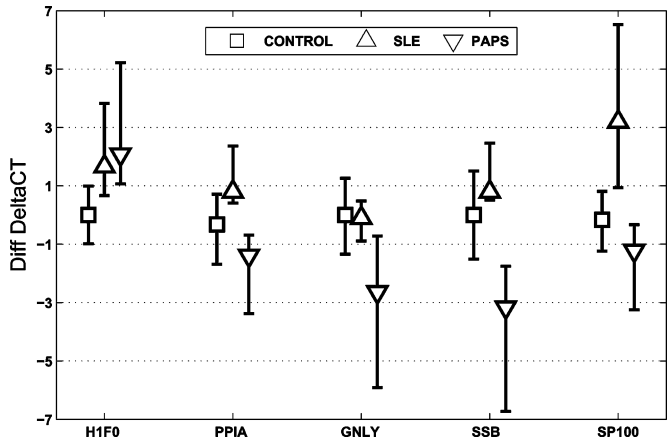

Fig. 3. qPCR validation summary for five genes from the relevant genelist.

SLE patients, but only H1F0 expression was increased in PAPS, whereas SP100 expression was downregulated.

SSB, also called La autoantigen, is a ribonucleoprotein involved in chromatin metabolism, and is known to elicit autoantibody responses in SLE [28]. Its expression in SLE samples was similar to controls, but was downregulated in PAPS. The remaining two genes (GNLY and PPIA) that were selected for PCR analysis are known to have a role in the execution of immune functions [52]–[54] and were found to be downregulated in PAPS samples.

In order to perform a rigorous validation, we used a second cohort of SLE and PAPS samples, a total of 14 new blood samples not used in the microarray analyses. The qPCR sample distribution was as follows: six healthy controls, three SLE samples, and five PAPS samples. Detailed information of each sample is gathered online.

For each gene and phenotype, DDiff values were calculated (see Section II), and the median values of these DDiff, together with the expected gene expression activities are shown in Table III. As a dispersion measure of the results, the values for the first and third quartile of each group of values are also shown. Fig. 3 graphically summarizes the results obtained for each gene within the qPCR validation.

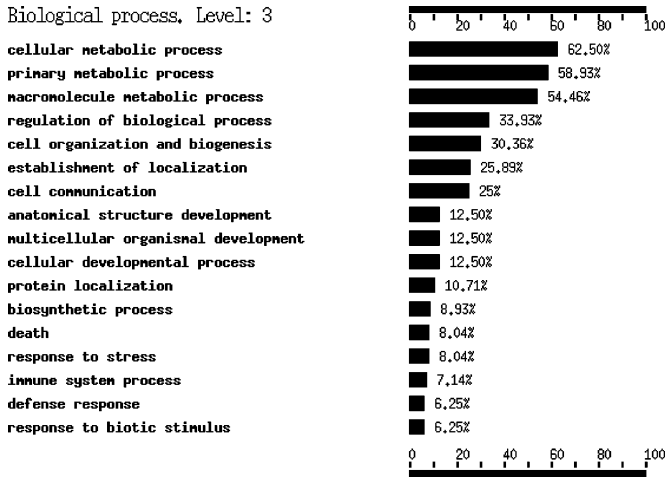

Fig. 4. Gene ontology biological process (level 3) annotations that are significantly overrepresented in the list of dysregulated genes. Annotations with an incidence level lower than 5% are not shown.

As a criterion, a median DDiff value between -1 and 1 was considered a baseline activity, that is, unchanged with respect to healthy controls. Median values higher than 1 reflect an upregulated activity, while values lower than -1 reflect a downregulated activity. Clearly, the expected gene expression profiling as measured by microarray quantitation is fully validated by the qPCR experiment.

As a final validation criterion for the expected gene expression activities, a statistical test was performed comparing the DDiff expression values between the control samples and either SLE or PAPS samples. Column p-value in Table III gathers the output of a nonparametric Mann–Whitney hypothesis test [55], showing that all the values of over or underexpression are statistically significant for a 90%; confidence level. In addition, the p-values for the baseline activities show no statistically significant differences between these cases and the control expression. Thus, all these results are consistent with the expected expression activity for each gene.

2) Functional Characterization of the Relevant Genes in SLE and PAPS: To check whether the results made sense from a biological point of view, we have analyzed the list of dysregulated genes with the FatiGO+ tool [56]. FatiGO+ can be used to search for the GO annotations[2] that are overrepresented in a list of genes. The significance of the overrepresentation is assessed by means of a Fisher exact test. From the 150 dysregulated genes identified, only 112 have GO annotations. Fig. 4 shows the results obtained with this tool (for terms in the level 3 of GO biological process annotations). As we can see in the figure, immune-system-related annotations, such as *defense response* or *immune system process*, are overrepresented in the list of dysregulated genes.

A comparison with previous studies on microarray analyses revealed a notable similarity in the functional categories of the genes found to be dysregulated in our analysis [57], confirming their importance in the pathogenesis of the disease IV. Such

[2]Gene Ontology Consortium http://www.geneontology.org

TABLE IV
SOME OF THE GO FUNCTIONAL GROUPS IDENTIFIED FROM THE RELEVANT GENELIST AND THE GENES THAT BELONG TO EACH GROUP

Functional group	Genes
Cellular metabolic process	GAS7, DDX5, RRAGD, MBD4, CECR1, CUTL1, PTMA
Primary metabolic process	SNRPD2, CPSF1, SSB, SFRS2IP, PPIG, NADK, DNPEP, HSD17B4, NAGPA, ZNF202, ABCA1, COMT
Macromolecule metabolic process	EIF4A1, EIF3S8, RPL18A, MRPL9, HSF2, HSPA6, PFDN4, HSPA8, HSP90B1
Regulation of biological process	KRAS, TMEM97, GPS1, UQCRB, CYB561, TNKS2, ZNF83, ZNF587, POLR2K, GMEB2, SP100
Cell organization and biogenesis	HIST1H1C, TSPYL4, H1F0
Establishment of localization	RIN3, SLC25A12, UQCRB, CYB561, SYPL1, GOSR1, KDELR2, C14orf108, GOLGA4, KPNB1, SLC25A5, PEX1
Cell communication	GNB2L1, CDC42EP3, IQGAP1, PKN1, RAB2, TAX1BP3, ARF3, ANK3
Defense response	WAS, TAP1, GNLY, NMI, CD160
Immune system processes	ISG15, IGHM, HLA-DQB1, GPSM3, MX1
Response to chemical stimulus	HSPA6, HSPA8
Death	BAG5, TNFRSF10B, CASP1, SIRT1

is the case of the categories defense response, immune system process, death, cell communication, or response to chemical stimulus. In addition, our analysis revealed other relevant functions, including metabolism, establishment of localization, or regulation of biological processes that have not been previously associated with SLE or PAPS, and that may provide important clues about the pathogenesis of these diseases.

3) Regulatory Pathways Dysregulated in SLE and PAPS: It is believed that mutations in susceptibility genes that contribute to the pathogenesis of a given disease result in an altered expression and/or activity of genes regulated by them, thus revealing a particular molecular signature of the disease [58]. Taking this idea into account, we searched for factors that could be regulating the genes whose expression is altered in SLE and/or PAPS. For each of the 150 genes identified in the original set, and using the ingenuity pathways analysis tool,[3] the factors that could be involved in the regulation of their expression or activity were identified. Only those genes with known regulators were considered for subsequent analyses. Furthermore, genes regulated by other genes included in the original set were discarded because it is not possible to know whether these genes are dysregulated due to a mutation in the genes that regulate their expression or because their regulators are dysregulated themselves.

The resulting filtered set includes a total of 129 genes (out of 150), and the gene set known to regulate them contains a total of 299 genes. Only the genes regulating three or more target genes were considered, which resulted in a final number of 17 regulatory genes controlling the expression of a total of 45 dysregulated genes (see Table V). Finally, their location within the genome was sought. Remarkably, nearly half of them (8 out of 17) were found to be located in chromosome regions previously reported as susceptibility regions for SLE [59]–[66].

[3]For a detailed description of ingenuity pathways analysis, visit: http://www.ingenuity.com

TABLE V
LOCATION OF THE DETECTED REGULATOR GENES

Regulator	Regulated genes	Locus	Ref.
IL15	BTG1, GNLY, PFDN4, POLR2K, SELPLG, SLC25A5, SP100	4q31	
IFNG	ABCA1, CYB561, HSPA8, MX1, PSMB4, TNFRSF10B	12q14	
MYC	ARF3, DDX5, PPIA, SLC25A5, TNFRSF10B	8q24.12-13	[59], [60]
TP53	COMT, DLG1, ISG15, THRAP2, TNFRSF10B	17p13.1	[59], [60]
EGF	GNB2L1, HK1, MVP, WAS	4q25	
HGF	ANK3, ISG15, HK1, TMEM97	7q21.1	[61]
TNF	ANXA2, BTG1, ISG15, MVP	6p31.3	[63]–[65]
PTEN	BTG1, CYB561, ISG15, RPL36A	10q23.31	
TGFB1	ANXA2, KDELR2, KPNB1, TAX1BP3	19q13.1	[59], [60]
IFNA1	CYB561, ISG15, MX1, NMI	9p22	
IFNA2	MX1, SP100, TNFRSF10B	9p22	
CDC42	IQGAP1, PKN1, WAS	1p36.1	[60], [62]
BCL2	CASP1, IGHG1, KRAS	18q21.3	[59], [62]
FOS	CPSF1, SNRPD2, HSP90B1	14q24.3	
TNFSF10	ISG15, SP100 , TNFRSF10B	3q26	[59]
MYCN	EIF4A1, RPL37A, RPS25	2q24.1	
SRC	ACP1 , ANXA2 , GNB2L1	20q12-13	[60], [61]

Recent microarray reports have suggested that the interferon regulatory pathway could be an important contributor of SLE, based on the dysregulated expression of numerous interferon-inducible genes in lupus samples [67]–[69]. Remarkably, our analyses revealed that three of the regulatory genes are interferon proteins (IFNG, IFNA1 and IFNA2), regulating the expression and/or activity of nine genes identified in the microarray experiments. Seven of these nine genes were overexpressed in SLE patients, consistent with previous findings. Moreover, three more regulator proteins (IL15, MYC, TNFSF10) are also known to be regulated by IFNs. These results indicate that the IFN pathway regulates nearly half of the regulatory genes identified in our analysis, either directly or indirectly, corroborating the importance of the interferon signature in SLE, and suggesting an important role for this pathway also in PAPS pathogenesis.

Other regulatory genes with a known or suspected role in autoimmunity were also present in the search. Such is the case of PTEN, a phosphatase involved in the regulation of the PI3K pathway [70], TNF, tumor necrosis factor [71], or the antiapoptotic protein Bcl-2 [72]. It will be worth examining these regulatory networks in more detail, to determine their contribution to the pathogenesis SLE or PAPS, as well as their usefulness as markers of these diseases.

4) Analysis of Transcription Factor Binding Sites in the Promoters of Genes Relevant for SLE and PAPS Identification:

TABLE VI
DEREGULATED TRANSCRIPTION FACTORS FOUND IN BASIS OF THE IDENTIFIED RELEVANT GENE SET

	IRF2	IRF1	PAX2	SP1	MEF2	P300	E2F	CDXA	CREBP1CJUN
Frequency									
z-value	2.81	2.05	2.71	2.23	2.15	2.14	2.10	-2.33	-2.08
sample mean	0.061	0.078	0.026	2.278	0.035	2.148	0.157	2.470	0.009
population mean	0.021	0.039	0.006	1.824	0.013	1.857	0.095	3.234	0.055
ratio	2.9	2.0	4.2	1.2	2.8	1.2	1.7	0.8	0.2
p-value	0.0049	0.0408	0.0067	0.0256	0.0318	0.0323	0.0353	0.0200	0.0377
Incidence									
number	5	7	3	78	4	105	17	81	1
sample mean	4.35%	6.09%	2.61%	67.83%	3.48%	91.30%	14.78%	70.43%	0.87%
population mean	2.02%	3.79%	0.62%	68.09%	1.25%	82.54%	8.86%	79.47%	5.28%
ratio	2.2	1.6	4.2	1.0	2.8	1.1	1.7	0.9	0.2
p-value	0.0840	0.1474	0.0355	0.4043	0.0572	0.0060	0.0251	0.0059	0.0132

Genes that participate in a particular pathway often share a common transcription factor binding site. We next explored the possibility that the dysregulated genes in SLE and PAPS could be regulated by common transcription factors involved in the development of autoimmunity. We reasoned that if a significant number of dysregulated genes in SLE and PAPS were regulated by a common transcriptional factor, then this factor could somehow be associated with the disease.

We made use of the transcription element listening system [73], also known as TELiS, to identify transcription factor-binding motifs (TFBM) that are overrepresented in the promoters of a given gene set. This analysis considers two variables for any binding motif: 1) the frequency of this motif per gene: a comparison is made between the average frequency within the whole microarray and the frequency in the genes of the relevant list; 2) the number of genes exhibiting this motif in the relevant list compared to the whole microarray. Using these values, it is possible to compute a z-value statistic [74] and to perform a two-tailed hypothesis test based on a Bernouilli-set trial that examine these TFBMs that are overrepresented in the test list with respect to the expected occurrence computed from the original microarray list.

In our case, the parameters for the genome scan were as follows: the promoter search interval was fixed between -1000 and $+200$ bp, and the stringency of the test is fixed to a 0.9 value. TELiS analysis identified a total of 115 genes from the total of 141 mapped genes in the relevant gene set. Within these parameters, TELiS reported a total number of seven overrepresented transcription factor binding motifs (see Table VI). Importantly, two interferon response elements (IRF1 and IRF2) appeared as overrepresented, again revealing the importance of the IFN-regulated pathway in these autoimmune diseases. The binding sites for SP1 and P300 were also significantly overrepresented, particularly with regard to the frequency of binding sites per gene. However, the increase in frequency as well as in incidence were minimal with respect to the reference control, and it is unlikely to be biologically meaningful.

PAX2 and MEF2 are transcription factors that are known to be involved in cellular differentiation and organ development [75]. The finding that their binding sites are overrepresented in our analysis suggests that factors regulating differentiation also play a role in autoimmunity. Remarkably, E2F binding sites were found overrepresented in this analysis. E2F constitutes a family of transcription factors involved in the transcriptional regulation of genes necessary for cell cycle control [76]. Recently, functional inactivation of E2F2, a member of this family, has been found to promote a lupus-like autoimmune disease in a mouse model, linking cell cycle regulation to autoimmunity [77]. Additionally, reduced expression of E2F2 has been reported in SLE patients [68]. These findings project a role for E2F in the regulation of autoimmunity, and suggest that modulation of E2F levels could be beneficial in these diseases.

IV. CONCLUSION

Consensus approaches are alternative techniques that try to overcome the technology-intrinsic data noise in microarray experiments. In the present paper, we applied a supervised consensus gene selection method, aiming to add robustness to the biomarker identification procedures by means of DNA microarrays.

Microarray studies must deal with "the curse of dimensionality" and "the curse of sparsity" [78]: in a problem with a huge number of variables (features or genes), there are only a small number of instances (cases or samples) whereas there are several thousand variables. Therefore, their results must be strictly proven to assess reliability over the given statements. Throughout this paper, in-depth statistical and biological validations have been successfully carried out.

Readers should note the importance of being conservative when dealing with findings coming from a low number of samples. Within some rare diseases, such as SLE and PAPS, it is very difficult for physicians and clinics to find samples cohorts. Therefore, studies in these fields must be able to deal with these adversities while they bring some light into the present genomic research. Throughout this paper, from the starting feature selection to the final biological validations, we have exposed a battery of techniques, both from statistics and biology, to add reliability and proofs to the results of such researches. Authors consider of special importance the posterior validation of the findings by means of qPCR analysis with outer samples not used in the previous statistical stages.

Among these findings, the statistical techniques applied have corroborated the importance of the IFN pathway in SLE and PAPS, and have also revealed the existence of other gene

signatures that could be playing an important role in the pathogenesis of these diseases. Future clinical and/or biological tests over the presented results could throw light on the molecular basis of SLE and PAPS diseases.

ACKNOWLEDGMENT

The authors are grateful to all the patients for their participation. They would like to thank N. Zorrilla and N. Liaño for their technical support.

REFERENCES

[1] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat. Biotechnol.*, vol. 14, pp. 1675–1680, 1996.

[2] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, pp. 467–470, 1995.

[3] E. C. Baechler, F. M. Batliwalla, A. M. Reed, E. J. Peterson, P. M. Gaffney, K. L. Moser, P. K. Gregersen, and T. W. Behrens, "Gene expression profiling in human autoimmunity," *Immunol. Rev.*, vol. 210, pp. 120–137, 2006.

[4] D. Alarcón-Segovia, "Shared autoimmunity: The time has come," *Curr. Rheumatol. Rep.*, vol. 6, pp. 171–174, 2004.

[5] P. K. Gregersen and T. W. Behrens, "Genetics of autoimmune diseases-disorders of immune homeostasis," *Nat. Rev. Genet.*, vol. 7, no. 12, pp. 917–928, 2006.

[6] C. G. Fathman, L. Soares, S. M. Chan, and P. J. Utz, "An array of possibilities for the study of autoimmunity," *Nature*, vol. 435, no. 7042, pp. 605–611, 2005.

[7] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman, and J. Quackenbush, "Within the fold: Assessing differential expression measures and reproducibility in microarray assays," *Genome Biol.*, vol. 3, no. 11, pp. research0062.1–research-0062.12, 2002.

[8] L. Li, L. G. Pedersen, T. A. Darden, and C. R. Weinberg, "Computational analysis of leukemia microarray expression data using the GA/KNN method," in *Proc. Methods Microarray Data Anal.: Papers CAMDA'00*, 2001, pp. 81–95.

[9] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Brief. Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[10] C. Bombardier, D. D. Gladman, M. B. Urowitz, D. Caron, and C. H. Chang, "Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE," *Arthritis Rheum.*, vol. 35, no. 6, pp. 630–640, 1992.

[11] Affymetrix, "Manufacturing quality control and validation studies of genechip arrays," Affymetrix, Inc., Santa Clara, CA, Tech. Note 701309 Rev. 2, 2002.

[12] R. Armañanzas, "Solving bioinformatics problems by means of Bayesian classifiers and feature selection," University of the Basque Country, Leioa Telefono, Spain, Tech. Rep. EHU-KZAA IK-2/06, 2006.

[13] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Mach. Learn.*, vol. 52, no. 1/-2, pp. 91–118, Jul. 2003.

[14] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam, "Consensus clustering and functional interpretation of gene-expression data," *Genome Biol.*, vol. 5, no. 11, pp. R94.1–R94.16, 2004.

[15] R. Kerber, "Chimerge: Discretization for numeric attributes," in *Proc. Nat. Conf. Artif. Intell.*, 1992, pp. 123–128.

[16] J. Catlett, "On changing continuous attributes into ordered discrete attributes," in *Proc. Eur. Working Session Learn.*, 1991, pp. 164–178.

[17] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 1993, pp. 1022–1027.

[18] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Comput. Biol.*, vol. 7, pp. 601–620, 2000.

[19] H. C. Causton, J. Quackenbush, and A. Brazma, *Microarray Gene Expression Data Analysis: A Beginner's Guide. Malden* MA: Blackwell, May 2003.

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[21] M. A. Hall and L. A. Smith, "Feature subset selection: A correlation based filter approach," in *Proc. 4th Int. Conf. Neural Inf. Process. Intell. Inf. Syst.*, 1997, pp. 855–858.

[22] M. Ben-Bassat, "Use of distance measures, information measures and error bounds in feature evaluation," in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds.. vol. 2. Amsterdam, The Netherlands: North-Holland, 1982, pp. 773–791.

[23] T. M. Cover and J. A. Thomas, *Elements Of Information Theory*. New York: Wiley Interscience, 1991.

[24] D. J. Wallace and B. H. Hahn, *Dubois' Lupus Erythematosus*: Baltimore, MD: Williams & Wilkins, 2002.

[25] M. L. Bertolaccini, M. A. Khamashta, and G. R. Hughes, "Diagnosis of antiphospholipid syndrome," *Nat. Clin. Pract. Rheumatol.*, vol. 1, no. 1, pp. 40–46, 2005.

[26] R. H. Scofield, "Genetic knock out of 60 kD Ro (or SSA), a common lupus autoantigen, induces lupus," *Trends Immunol.*, vol. 25, no. 1, pp. 1–3, 2003.

[27] M. Srivastava, A. Rencic, G. Diglio, H. Santana, P. Bonitz, R. Watson, E. Ha, G. Anhalt, T. Provost, and C. Nousari, "Drug-induced, Ro/SSA-positive cutaneous lupus erythematosus," *Arch. Dermatol.*, vol. 139, no. 1, pp. 45–49, 2003.

[28] P. J. Maddison, D. A. Isenberg, N. J. Goulding, J. Leddy, and R. P. Skinner, "Anti La (SSB) identifies a distinctive subgroup of systemic lupus erythematosus," *Br. J. Rheumatol.*, vol. 27, no. 1, pp. 27–31, 1988.

[29] I. Wichmann, M. A. Montes-Cano, N. Respaldiza, A. Alvarez, K. Walter, E. Franco, J. Sanchez-Roman, and A. Nunez-Roldan, "Clinical significance of anti-multiple nuclear dots/Sp100 autoantibodies," *Scand. J. Gastroenterol.*, vol. 38, no. 9, pp. 996–999, 2003.

[30] J. A. Hardin and J. O. Thomas, "Antibodies to histones in systemic lupus erythematosus: Localization of prominent autoantigens on histones h1 and h2b," *Proc. Nat. Acad. Sci. USA*, vol. 80, no. 24, pp. 7410–7414, 1983.

[31] P. Correa, J. Molina, L. Pinto, M. Arcos-Burgos, M. Herrera, and J. Anaya, "TAP1 and TAP2 polymorphisms analysis in Northwestern Colombian patients with systemic lupus erythematosus," *Ann. Rheum. Dis.*, vol. 62, no. 4, pp. 363–365, 2003.

[32] C. Hualupusng, L. Hang, C. Chen, J. Wu, and F. Tsai, "Polymorphisms of TAP1 transporter genes in Chinese patients with systemic lupus erythematosus in Taiwan," *Rheumatol. Int.*, vol. 24, no. 3, pp. 130–132, 2004.

[33] Elvira Consortium, "Elvira: An environment for probabilistic graphical models," presented at the Electron. Proc. 1st Eur. Workshop Probabilistic Graph. Models, J. A. Gámez and A. Salmerón, Eds., Cuenca, Spain, Nov. 2002.

[34] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artif. Intell. Med.*, vol. 31, no. 2, pp. 91–103, Jun. 2004.

[35] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statist. Sin.*, vol. 12, pp. 111–139, 2002.

[36] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarrays data," *Comput. Statist. Data Anal.*, vol. 48, pp. 869–885, 2005.

[37] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: A multiple random validation strategy," *Lancet*, vol. 365, pp. 488–92, 2005.

[38] D. G. Kleinbaum, *Logistic Regression. Statistics in Health Sciences*. New York: Springer-Verlag, 1994.

[39] S. L. Cessie and J. C. V. Houwelingen, "Ridge estimators in logistic regression," *Appl. Statist.*, vol. 41, no. 1, pp. 191–201, 1992.

[40] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, pp. 37–66, Jan. 1991.

[41] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. 11th Conf. Uncertainty Artif. Intell*. San Mateo, CA: Morgan Kaufmann, 1995, pp. 338–345.

[42] T. Bayes, "Essay towards solving a problem in the doctrine of chances," *Philos. Trans. Roy. Soc. London*, vol. 53, pp. 370–418, 1764.

[43] L. Breiman, "Random forests," *Mach. Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[44] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Statist. Soc. B*, vol. 36, no. 2, pp. 111–147, 1974.

[45] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
[46] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.
[47] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. Amer. Statist. Assoc.*, vol. 78, pp. 316–331, 1983.
[48] S. M. Weiss and C. A. Kulikowski, *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufmann, 1991.
[49] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining—PAKDD*, Sydney, Australia, 2004, pp. 3–12.
[50] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
[51] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.
[52] A. Deng, S. Chen, Q. Li, S. C. Lyu, C. Clayberger, and A. M. Krensky, "Granulysin, a cytolytic molecule, is also a chemoattractant and proinflammatory activator," *J. Immunol.*, vol. 174, pp. 5243–5248, 2005.
[53] Z. G. Jin, A. O. Lungu, L. Xie, M. Wang, and C. W. B. C. Berk, "Cyclophilin A is a proinflammatory cytokine that activates endothelial cells," *Arterioscler. Thromb. Vasc. Biol.*, vol. 24, no. 7, pp. 1186–1191, 2004.
[54] K. Zander, M. Sherman, U. Tessmer, K. Bruns, V. Wray, A. T. Prechtel, E. Schubert, P. Henklein, J. Luban, J. Neidleman, W. C. Greene, and U. Schubert, "Cyclophilin A interacts with HIV-1 Vpr and is required for its functional expression," *J. Biol. Chem.*, vol. 278, no. 44, pp. 43 202–43 213, 2003.
[55] H. B. Mann and D. R. Whitney, "On a test of whether one of 2 random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, pp. 50–60, 1947.
[56] F. Al Shahrour, P. Mínguez, J. Tárraga, D. Montaner, E. Alloza, J. M. Vaquerizas, L. Conde, C. Blaschke, J. Vera, and J. Dopazo, "BABELOMICS: A systems biology perspective in the functional annotation of genome-scale experiments," *Nucleic Acids Res.*, vol. 34, pp. w472–w476, 2006.
[57] M. Mandel, M. Gurevich, R. Pauzner, N. Kaminski, and A. Achiron, "Autoimmunity gene expression portrait: Specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erithematosus," *Clin. Exp. Immunol.*, vol. 138, pp. 164–170, 2004.
[58] G. R. Burmester and T. Haupl, "Strategies using functional genomics in rheumatic diseases," *Autoimmun. Rev.*, vol. 3, no. 7/8, pp. 541–549, 2004.
[59] C. Johansson, "Exploring the genetics of SLE with linkage and association Analysis" Ph.D. dissertation, Uppsala Univ., Uppsala, Sweden, 2004.
[60] S. Koskenmies, "Mapping of susceptibility genes for systemic lupus erythematosus (SLE)," Ph.D. dissertation, Univ. Helsinki, Helsinki, Finland, Mar. 2004.
[61] B. P. Tsao, "The genetics of human systemic lupus erythematosus," *Trends Immunol.*, vol. 24, no. 11, pp. 595–602, Nov. 2003.
[62] R. Shai, F. P. Q. Jr, L. Li, O. Kwon, J. Morrison, D. J. Wallace, C. M. Neuwelt, C. Brautbar, W. J. Gauderman, and C. O. Jacob, "Genome-wide screen for systemic lupus erythematosus susceptibility genes in multiplex families," *Hum. Mol. Genet.*, vol. 8, no. 4, pp. 639–644, 1999.
[63] M. Aringer and J. S. Smolen, "Tumour necrosis factor and other proinflammatory cytokines in systemic lupus erythematosus: A rationale for therapeutic intervention," *Lupus*, vol. 13, no. 5, pp. 344–347, 2004.
[64] T. Horiuchi, C. Morita, H. Tsukamoto, H. Mitoma, T. Sawabe, S. Harashima, Y. Kashiwagi, and S. Okamura, "Increased expression of membrane TNF-alpha on activated peripheral CD8+T cells in systemic lupus erythematosus," *Int. J. Mol. Med.*, vol. 17, no. 5, pp. 875–879, 2006.
[65] Y. H. Lee, J. B. Harley, and S. K. Nath, "Meta-analysis of TNF-alpha promoter −308 A/G polymorphism and SLE susceptibility," *Eur. J. Hum. Genet.*, vol. 14, no. 3, pp. 364–371, 2006.
[66] J. A. Croker and R. P. Kimberly, "Genetics of susceptibility and severity in systemic lupus erythematosus," *Curr. Opin. Rheumatol.*, vol. 17, no. 5, pp. 529–537, 2005.
[67] S. J. Rozzo, J. D. Allard, D. Choubey, T. J. Vyse, S. Izui, G. Peltz, and B. L. Kotzin, "Evidence for an interferon-inducible gene, Ifi202, in the susceptibility to systemic lupus," *Immunity*, vol. 15, pp. 435–443, 2001.
[68] E. C. Baechler, F. M. Batliwalla, G. Karypis, P. M. Gaffney, W. A. Ortmann, K. J. Espe, K. B. Shark, W. J. Grande, K. M. Hughes, V. Kapur, P. K. Gregersen, and T. W. Behrens, "Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 5, pp. 2610–2615, 2003.
[69] L. Bennett, A. Palucka, E. Arce, V. Cantrell, J. Borvak, J. Banchereau, and V. Pascual, "Interferon and granulopoiesis signatures in systemic lupus erythematosus blood," *J. Exp. Med.*, vol. 197, no. 6, pp. 711–723, 2003.
[70] R. K. Patel and C. Mohan, "PI3K/AKT signaling and systemic autoimmunity," *Immunol. Res.*, vol. 31, no. 1, pp. 47–55, 2005.
[71] H. C. Hsu, Y. Wu, and J. D. Mountz, "Tumor necrosis factor ligand-receptor superfamily and arthritis," *Curr. Dir. Autoimmun.*, vol. 9, pp. 37–54, 2006.
[72] P. B. Deming and J. C. Rathmell, "Mitochondria, cell death, and b cell tolerance," *Curr. Dir. Autoimmun.*, vol. 9, pp. 95–119, 2006.
[73] S. W. Cole, W. Yan, Z. Galic, J. Arevalo, and J. A. Zack, "Expression-based monitoring of transcription factor activity: The TELiS database," *Bioinformatics*, vol. 21, no. 6, pp. 803–810, 2005.
[74] G. K. Kanji, *100 Statistical Tests*. London, U.K.: Sage, 2006.
[75] C. A. Berkes and S. J. Tapscott, "MyoD and the transcriptional control of myogenesis," *Semin. Cell Dev. Biol.*, vol. 16, no. 4/5, pp. 585–595, 2005.
[76] J. R. Nevins, "The Rb/E2F pathway and cancer," *Hum. Mol. Genet.*, vol. 10, no. 7, pp. 699–703, 2001.
[77] M. Murga, O. Fernández-Capetillo, S. J. Field, B. Moreno, L. R. Borlado, Y. Fujiwara, D. Balomenos, A. Vicario, A. C. Carrera, S. H. Orkin, M. E. Greenberg, and A. M. Zubiaga, "Mutation of E2F2 in mice causes enhanced t lymphocyte proliferation, leading to the development of autoimmunity," *Immunity*, vol. 15, no. 6, pp. 959–970, 2001.
[78] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003.

Authors' photographs and biographies not available at the time of publication.