Genes and Immunity (2013) **14,** 67–82

www.nature.com/gene

ORIGINAL ARTICLE

Common and specific signatures of gene expression and protein–protein interactions in autoimmune diseases

T Tuller, S Atar, E Ruppin, M Gurevich[1] and A Achiron[1]

The aim of this study is to understand intracellular regulatory mechanisms in peripheral blood mononuclear cells (PBMCs), which are either common to many autoimmune diseases or specific to some of them. We incorporated large-scale data such as protein–protein interactions, gene expression and demographical information of hundreds of patients and healthy subjects, related to six autoimmune diseases with available large-scale gene expression measurements: multiple sclerosis (MS), systemic lupus erythematosus (SLE), juvenile rheumatoid arthritis (JRA), Crohn's disease (CD), ulcerative colitis (UC) and type 1 diabetes (T1D). These data were analyzed concurrently by statistical and systems biology approaches tailored for this purpose. We found that chemokines such as *CXCL1-3, 5, 6* and the interleukin (IL) *IL8* tend to be differentially expressed in PBMCs of patients with the analyzed autoimmune diseases. In addition, the anti-apoptotic gene *BCL3*, interferon-γ (*IFNG*), and the vitamin D receptor (*VDR*) gene physically interact with significantly many genes that tend to be differentially expressed in PBMCs of patients with the analyzed autoimmune diseases. In general, similar cellular processes tend to be differentially expressed in PBMC in the analyzed autoimmune diseases. Specifically, the cellular processes related to cell proliferation (for example, epidermal growth factor, platelet-derived growth factor, nuclear factor-κB, Wnt/β-catenin signaling, stress-activated protein kinase c-Jun NH2-terminal kinase), inflammatory response (for example, interleukins *IL2* and *IL6*, the cytokine granulocyte–macrophage colony-stimulating factor and the B-cell receptor), general signaling cascades (for example, mitogen-activated protein kinase, extracellular signal-regulated kinase, p38 and TRK) and apoptosis are activated in most of the analyzed autoimmune diseases. However, our results suggest that in each of the analyzed diseases, apoptosis and chemotaxis are activated via different subsignaling pathways. Analyses of the expression levels of dozens of genes and the protein–protein interactions among them demonstrated that CD and UC have relatively similar gene expression signatures, whereas the gene expression signatures of T1D and JRA relatively differ from the signatures of the other autoimmune diseases. These diseases are the only ones activated via the Fcε pathway. The relevant genes and pathways reported in this study are discussed at length, and may be helpful in the diagnoses and understanding of autoimmunity and/or specific autoimmune diseases.

Genes and Immunity (2013) **14,** 67–82; doi:10.1038/gene.2012.55; published online 29 November 2012

Keywords: autoimmunity; gene expression; protein–protein interactions; apoptosis; signaling pathway; chemotaxis

INTRODUCTION

Autoimmunity is the failure of an organism to recognize its own constituent parts as self, which results in an immune response against its own cells and tissues. Any disease that results from such an aberrant immune response is termed an *autoimmune disease*.[1] Today, more than 80 clinically distinct diseases are classified as autoimmune diseases.[2] They occur in 3–5% of the population,[3] usually as a result of a myriad of genetic and environmental factors, which lead to altered immune reactivity.[4,5] Many of the autoimmune diseases may have life-threatening implications. In addition, as often these diseases lead to substantial disability, autoimmune diseases have major socioeconomic consequences.

Some autoimmune diseases are systemic (for example, systemic lupus erythematosus (SLE)), where various tissues are under attack, whereas in others (for example, multiple sclerosis (MS)), the targets of the autoreactive cells are specific to a single tissue. This research includes six autoimmune diseases, namely, MS, SLE, juvenile rheumatoid arthritis (JRA), Crohn's disease (CD), ulcerative colitis (UC) and type 1 diabetes (T1D), with available gene expression measurements.

T1D is believed to be an autoimmune disease where the immune system attacks pancreatic β-cells in the Islets of Langerhans. This may effectively abolish endogenous insulin production.[6]

SLE is a chronic systemic autoimmune disease. In this disease, the immune system attacks the body's cells and tissues, resulting in inflammation and tissue damage. SLE can affect any part of the body, but most often the tissues under attack are the heart, joints, skin, lungs, blood vessels, liver, kidneys and nervous system. The course of the disease is unpredictable, with periods of illness (called flares) alternating with remissions.[7]

CD is an autoimmune disease that occurs when the immune system attacks the gastrointestinal tract. This autoimmune activity produces inflammation in the gastrointestinal tract.[8]

UC has similarities to CD; the main symptom of the active disease is usually constant diarrhea mixed with blood, of gradual onset. UC is, however, a systemic disease that affects many parts of the body outside the intestine.[9]

JRA is an autoimmune disease that typically appears between the ages of 6 months and 16 years. In this disease the tissues under attack are joint tissues. The signs of this disease include joint pain or swelling and reddened or warm joints.[10]

School of Medicine, Tel Aviv University, Ramat Aviv, Israel. Correspondence: Dr T Tuller, School of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel.
E-mail: tamirtul@post.tau.ac.il.
[1]These authors contributed equally to this work
Received 26 June 2012; revised 24 October 2012; accepted 25 October 2012; published online 29 November 2012

MS is an autoimmune inflammatory T-cell-mediated disease, believed to result from a misdirected immune attack against central nervous system myelin antigens. The damage of myelin in MS leads to neurological dysfunction. MS attacks mainly young adults (usually between ages 20 and 40 years), and is predominant in women.[11–13]

The role of the different autoimmune cells varies among the analyzed diseases. MS is believed to be an inflammatory T-cell-mediated disease.[11–13] Inflammatory T cells are also thought to be central to the pathology of autoimmune arthritis.[14] SLE, on the other hand, is mediated both by T cells and B cells. Specifically, it was shown that there is an increased longevity of autoreactive T cells in patients with SLE; however, B cells also contribute to the production of autoantibodies and thus are central to SLE manifestations.[15–17] A T-cell-mediated immune response has been identified in the mucosa of CD, and is postulated to be the primary precipitating event in CD,[18] whereas in UC there is no strong evidence for T-cell activation, and humoral mechanisms predominate.[18]

Finally, T1D is believed to be T-cell mediated and involves interactions between natural killer cells, different dendritic cell populations and T cells.[19]

Numerous previous studies have demonstrated that transcriptional profiling of peripheral blood mononuclear cell (PBMC) is a useful tool for identifying gene expression signatures that are related to autoimmune diseases[20–29] (and a general review in Chaussabel *et al.*[30]).

A possible explanation for the advantageousness of PMBC in this context is the fact that autoreactive immune cells initiate the autoimmune inflammatory process against the corresponding target organs.[1,6–10,31,32]

The PBMC gene expression analysis of autoimmune diseases is usually performed separately for each illness (see a review in Centola *et al.*[20]): For example, Achiron *et al.*[21] identified a statistically significant transcriptional signature of genes in PBMCs from relapse-remitting MS patients; Jarvis *et al.*[32] identified discrete subsets of functionally related genes relevant to JRA pathophysiology; and Baechler *et al.*[33] showed that patients with SLE exhibit dysregulated expression of genes related to the IFN pathway.

Only a few previous studies compared the gene expression of more than one autoimmune disease.[28,34–36] These studies used basic statistical tests for comparing the sets of genes, which were significantly over/underexpressed in these diseases. For example, in a research that was performed in our lab, Mandel *et al.*[28] compared gene expression of patients with SLE and patients with MS. The study demonstrated that there are genes that are over/underexpressed only in one of the diseases. However, there is a substantial overlap between genes that are over/underexpressed in MS patients and in patients with SLE. Similarly, Bovin *et al.*[36] compared RA, chronic autoimmune thyroiditis and inflammatory bowel disease, and concluded that at least some of the genes activated in RA are predominantly or solely related to general and disease-nonspecific autoimmune processes. Another study[35] focused on the disparate gene expression signatures of a pair of autoimmune diseases with similar symptoms, UC and CD. As these two inflammatory bowel diseases share several demographic and clinical characteristics, such gene expression signatures, which can complement the standard diagnosis of UC and CD, may improve the diagnostic process.

In this study, we generalize the previous studies in two main directions. First, we analyze a larger set of six autoimmune diseases; these diseases were normalized and analyzed simultaneously with computational tools that were tailored specifically for this purpose. Ergo, we aim at understanding the genes' signaling pathways activated in PBMC, in all (or many of) the analyzed autoimmune diseases and in specific diseases.

Secondly, we combine the gene expression measurements and physical interactions (protein–protein interactions (PPIs)) to overcome some of the noise and bias of gene expression measurements,[37] and the fact that there are mechanisms of regulation that are post-transcriptional.[38,39] Such an approach has been successfully demonstrated before for single diseases (that is, cancer[40] and MS[38]).

Thus, in this study we provide the first global systems biology view of gene expression changes in PBMC in autoimmune diseases.

RESULTS

We consolidated a large data set of gene expression measurements of patients with the aforementioned six autoimmune diseases; some of the measurements were downloaded from Gene Expression Omnibus, whereas others were generated in our lab (Figure 1; see details in the Materials and methods section). Each data set included both PBMC gene expression of patients with one of these diseases and gene expression of healthy subjects as a control (see more details regarding the data sets in the Materials and methods section and Table 1). The total number of analyzed samples was 393.

Each data set was normalized separately to minimize batch effects, and for each gene in each data set we computed an analysis of variance (ANOVA) *P*-value (see Materials and methods and Rutherford[41]). This *P*-value is based on the mRNA levels in PBMC of the gene in patients vs healthy subjects in the data set; it was more significant if the expression levels (mRNA levels) of the gene increased or decreased more extremely in the group of patients in comparison to the group of healthy subjects. The ANOVA *P*-value also considers the relevant clinical and demographical characteristics of the patients and healthy subjects, as well as batch effects (Figure 1).

In addition, we used a large-scale data set of human protein–protein interactions (PPI) network, which includes pairs of proteins that have physical interactions with each other (see Materials and methods). These data were used to compute for each gene in each disease, a *P*-value based on the proteins that have PPIs with the protein it encodes and their expression levels (see more details in the Materials and methods section and Figure 1). We named this *P*-value the PPI *P*-value, or in-short PPI *P*-value. Briefly, the PPI *P*-value of a gene is based on the number of ANOVA significant proteins that have PPI with it; however, the PPI *P*-value of a gene is not based on the mRNA levels of the gene itself. Intuitively, the PPI *P*-value is based on the fact that genes that physically interact with relatively many ANOVA significant proteins have a higher probability to undergo regulatory changes themselves (for example, post-transcriptional regulation) and hence are more significant (see more technical details and motivation in the Materials and methods section). Thus, the PPI *P*-values enable us to detect genes that with high probability undergo regulatory changes that are not necessarily observed in their mRNA levels.

One of the aims of the paper is to perform large-scale analysis of gene expression levels in PBMC of many of the diseases concurrently. Specifically, we aimed at finding genes and cellular functions that are differentially expressed in many of the diseases; some of these genes and cellular functions may contribute to the pathogenesis of many of the analyzed diseases.

To this end, we defined two generalized scores related to the ANOVA and PPI *P*-values mentioned above: (1) a gene is *x*-ANOVA significant if it is ANOVA significant in at least *x* diseases and the direction of the gene mRNA fold change in all these diseases is identical (see Figure 1 and Materials and methods); this score is a generalization of the ANOVA *P*-value. (2) A gene is *x*-PPI significant if it is PPI significant based on the *x*-ANOVA significant genes (that is, a gene is PPI significant with respect to a certain disease if it has significantly many PPI with ANOVA significant genes in that disease; *x*-PPI significance is a generalization of the PPI *P*-value;

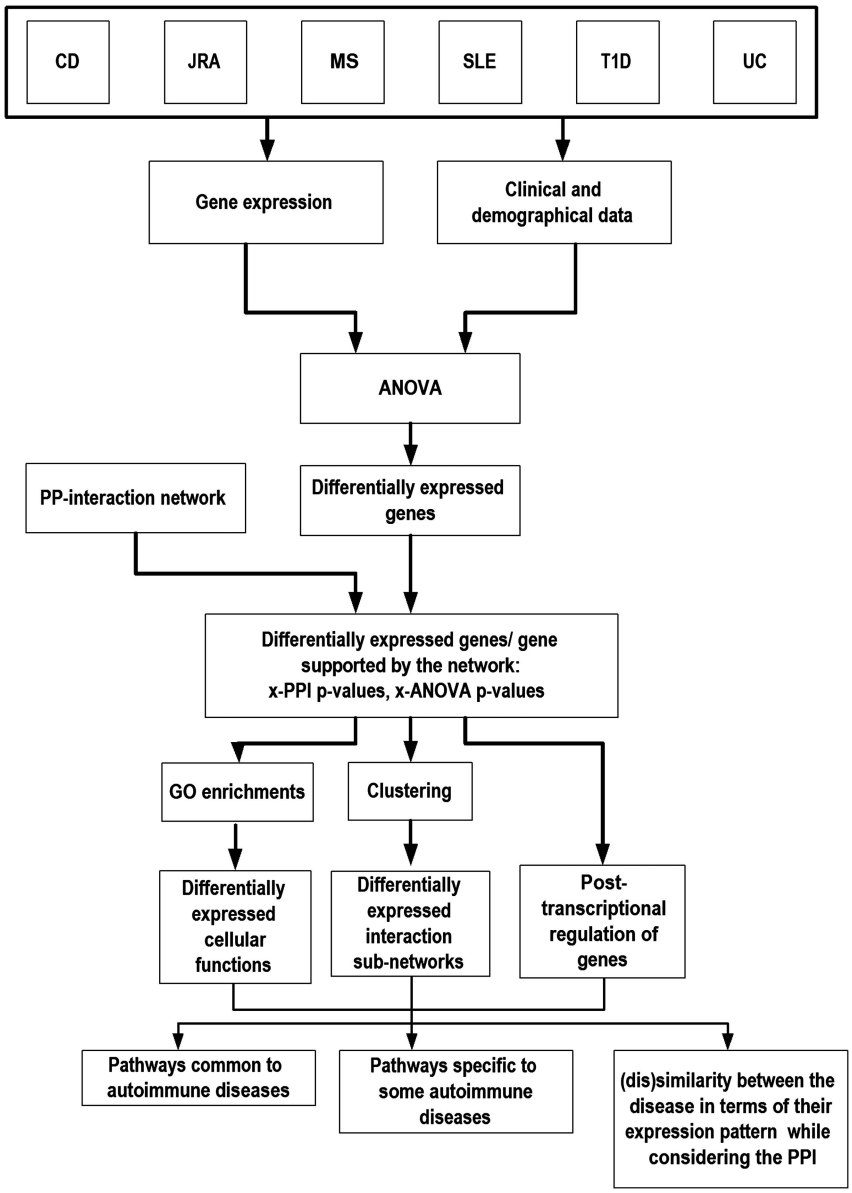

Figure 1. A flow diagram of the analyses performed in this study.

see Figure 1 and Materials and methods). We use a cutoff of $x \geqslant 3$ (50% of the diseases) throughout the paper.

In the following sections, we report a systems biology analysis that is based on the ANOVA and PPI *P*-values or their generalizations (Figure 1). Specifically, based on the *P*-value and scores defined above, we intend to answer the following questions: (1) Which genes are differentially expressed in PBMC in many autoimmune diseases? (2) What cellular processes are differentially expressed in PBMC in many autoimmune diseases, and what are the cellular processes that are specific only to part of the diseases? (3) Is there a significant global gene expression signature of autoimmunity in PBMC in each of the analyzed diseases? Is there a significant global signature in PBMC that is *common* and specific to all the analyzed diseases? (4) Which of the signaling pathways are differentially expressed in PBMC in many autoimmune diseases, and which of the pathways are specific only to part of the diseases? (5) How are the autoimmune diseases clustered based on their gene expression in PBMC?

Genes that are over/underexpressed in many autoimmune diseases

Supplementary Table 1 includes the ANOVA and PPI *P*-values of all the analyzed genes in each of the diseases. Tables 2 and 3 include the genes that have significant 4-ANOVA and 3-PPI *P*-values (see also Supplementary Table 2). The 4-ANOVA significant genes include various genes related to the signaling and triggering of the immune system, including *IFNG* and its receptor interleukin *IL8*, the chemokine *CXCL2* and *TLR6*. These genes are usually related to the upregulation of chemotaxis and proliferation of immune cells, which occur in PBMC of patients with autoimmune diseases.

One prominent gene group with 3-PPI significant *P*-values is the *CXCL* chemokines family (*CXCL1–3, 5, 6* and *CCL4*), supporting the fact that one of the common mechanisms in the analyzed autoimmune diseases is inflammation and chemoattraction, to guide the migration of immunological cells.[42,43]

Another important group that appears in many autoimmune diseases, that is both 4-ANOVA and 3-PPI significant, includes the interleukin *IL8* and its receptor. This IL has an important role in the

Table 1. (A) The gene expression data sets[a] and (B) the clinical and demographical characteristics of the subjects from the MS data set[b]

Disease	*Number of chips (patients/ healthy)*	*Chip type*	*Source/reference*	*Clinical data*
(A)				
MS	123/41	U133A-2	Our lab	See Table 1B
SLE	5/5	HG-U95Av2 HG-U133A_2	Our lab[28]	*Patients:* Age: 42.8 ± 12.6 years F/M = 4/1 *Healthy subjects:* Age: 44.7 ± 7.4 years F/M = 4/1
JRA	15/11	HG_U95Av2	NCBI GEO record GSE1402[27]	*Patients:* Age: 15.9 ± 4.9 years F/M = 8/7 *Healthy subjects:* Age: 14.1 ± 4.6 years F/M = 6/5 Information from Barnes *et al.*[27]
CD	59/42	U133A (we used soft file)	NCBI GEO record GDS1615, GSE3365[35]	*Patients:* Age: 41.3 ± 12.7 years F/M = 38/21 *Healthy subjects:* Age: 44.1 ± 8.8 years F/M = 18/24
UC	26/42	U133A	NCBI GEO record GDS1615, GSE3365[35]	*Patients:* Age: 46.7 ± 13.5 years F/M = 18/8 *Healthy subjects:* Age: 44.1 ± 8.8 years F/M = 18/24
T1D	43/24	U133A, U133B	NCBI GEO record GSE9006[29]	*Patients:* Age: 9.5 ± 3.8 years F/M = 25/18 *Healthy subjects:* Age: 11 ± 4.6 years F/M = 14/10

	Age (years)	*Disease duration*	*Annual relapse rate*	*EDSS*	*F/M*
(B)					
Patients	35.6 ± 10.8	6.6 ± 6.3	1.5 ± 1.6	2.5 ± 1.3	78/45
Healthy subjects	35.1 ± 8.7	N/A	N/A	N/A	21/20

Abbreviations: CD, Crohn's disease; F, female; EDSS, Expanded Disability Status Scale; GEO, Gene Expression Omnibus; JRA, juvenile rheumatoid arthritis; M, male; MS, multiple sclerosis; N/A, not applicable; NCBI, National Center for Biotechnology Information; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis. [a]More details regarding the clinical characteristics of each data set appear in the Materisl and methods section [b]For each variable we report the mean and ± s.d.

inflammatory response that occurs in most of the diseases, and will be discussed at length.[6,7,9,10,38,42]

Interestingly, the gene *VDR* (vitamin D receptor) has a significant 3-PPI *P*-value. This gene encodes the nuclear hormone receptor for vitamin D3. The receptor belongs to the family of *trans*-acting transcriptional regulatory factors, and exhibits sequence similarity to the steroid and thyroid hormone receptors. Downstream targets of this nuclear hormone receptor are principally involved in mineral metabolism, although the receptor regulates a variety of other metabolic pathways, such as those involved in the immune response and cancer. This result supports the belief that vitamin D protects against autoimmune diseases.[44]

Finally, we found that *BCL3* has a significant 3-PPI *P*-value. This protein functions as a transcriptional coactivator through its association with nuclear factor (NF)-κB homodimers.[45] The expression of this gene can be induced by *NF-κB*, which forms a part of the autoregulatory loop that controls the nuclear residence of p50 *NF-κB*. As a result, the proliferation and anti-apoptotic signals are increased.[46]

Gene ontology groups that are enriched in many autoimmune diseases

Next, we performed gene ontology (GO) enrichment analysis based on genes with significant ANOVA and PPI *P*-values in each disease, to uncover general cellular functions that are over-represented in the genes that are differentially expressed in each disease (see technical details in the Materials and methods section). The resultant significant GO functions and the corresponding enrichment *P*-values appear in Supplementary Table 3. Table 4 includes a summary of the GO functions that are enriched in many diseases when considering genes that have either ANOVA or PPI significant *P*-values.

As can be seen, cellular processes, which are related to apoptosis, inflammatory response, regulation of cytokines and regulation of T-cell activation, are over-represented in most of the analyzed diseases. Inappropriate cell death and apoptosis may be one of the causes of autoreactive immune response.[47,48] Thus, in the analyzed diseases there is a proliferation of such autoreactive immune cells in addition to inflammatory response, which is

Table 2. Nineteen genes that are 4-ANOVA significant (see Materials and methods), their fold change and a description of each gene

ANOVA-based P-values			
Gene symbol	*Fold change direction (1, increase; −1, decrease)*	P-value *range with the name of the diseases with the highest lowest* P-values	*Description*
EGR1	1	0.00000119 (CD) $<P<$ 0.040283 (JRA); not significant: SLE	*Early growth response 1*
CXCL2	1	0.00015832 (CD) $<P<$ 0.048917 (UC); not significant: SLE	*Chemokine (C–X–C motif) ligand 2*
PTGS2	1	(T1D) 0.00000266 $<P<$ 0.038208 (UC); not significant: SLE	*Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)*
FOS	1	(CD) 0.0000409 $<P<$ 0.036025 (T1D); not significant: SLE, JRA	*FBJ murine osteosarcoma viral oncogene homolog*
IFNG	−1	(SLE) 0.00011148 $<P<$ 0.035745 (CD); not significant: UC, JRA	*Interferon γ*
IFNGR2	1	(UC) 0.0000366 $<P<$ 0.044235 (SLE); not significant: MS, JRA	*Interferon γ receptor 2 (interferon gamma transducer 1)*
IL8	1	(CD) 2.24E − 07 $<P<$ 0.018736 (UC); not significant: SLE, JRA	*Interleukin 8*
IGH	1	(UC) 1.05E − 08 $<P<$ 0.040442 (SLE); not significant: CD, JRA	*Immunoglobulin heavy locus*
ITGB8	1	(CD) 0.00077538 $<P<$ 0.015449 (UC); not significant: T1D,SLE	*Integrin, β8*
KLRD1	−1	(CD) 0.00000204 $<P<$ 0.033677 (T1D); not significant: JRA, SLE	*Killer cell lectin-like receptor subfamily D, member 1*
IL1R2	1	(UC) 0.00022596 $<P<$ 0.015102 (MS); not significant: JRA, SLE	*Interleukin 1 receptor, type II*
ACOT1	−1	(CD) 0.00366658 $<P<$ 0.035113(SLE); not significant: MS, JRA	*Acyl-CoA thioesterase 1*
CD160	−1	(CD) 3.38E − 09 $<P<$ 0.023071(MS); not significant: SLE, JRA	*CD160 molecule*
SOCS3	1	(CD) 0.00000111 $<P<$ 0.027513 (T1D); not significant:SLE, JRA	*Suppressor of cytokine signaling 3*
ZNF91	−1	(CD) 0.00000881 $<P<$ 0.037897 (MS); not significant: SLE, JRA	*Zinc-finger protein 91*
TLR6	1	(CD) 0.0029267 $<P<$ 0.029474 (JRA); not significant: MS, SLE	*Toll-like receptor 6*
TLE3	1	(CD) 0.001739 $<P<$ 0.039967 (UC); not significant: T1D, SLE	*Transducin-like enhancer of split 3 (E(sp1) homolog, Drosophila)*
ALPL	1	(UC) 0.000165 $<P<$ 0.029585 (T1D); not significant: SLE, JRA	*Alkaline phosphatase, liver/bone/kidney*
RAMP3	1	(JRA) 0.00104094 $<P<$ 0.03312 (UC); not significant: T1D, SLE	*Receptor (G-protein-coupled) activity-modifying protein 3*

Abbreviations: ANOVA, analysis of variance; CD, Crohn's disease; JRA, juvenile rheumatoid arthritis; MS, multiple sclerosis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis.

mediated by secretion of cytokines.[49] We will discuss these processes in greater detail in the following sections.

Global gene expression signature of autoimmunity

We performed several novel tests to demonstrate that there exists a global gene expression signal of autoimmunity. These tests are 'global' as they consider the gene expression of the *entire* gene set, and the PPIs between the products of these genes.

In the first test, we compared the number of genes that are both ANOVA and PPI significant in the real network to randomized ones (see Materials and methods and '# ANOVA and PPI significant' in Table 5). In the second test, in each disease we permuted the genes that were ANOVA significant, and compared the number of genes with PPI significant *P*-values in the randomized data sets to the original one (see Materials and methods and '# PPI significant' in Table 5). In the final test, we compared the mean distance between ANOVA significant genes in the real data to the ones obtained in randomized data sets (see Materials and methods and 'mean distance PPI significant' in Table 5).

As can be seen in Table 4, in each of the six diseases at least one of these global tests was significant (in most of the cases, more than one *P*-value was significant), demonstrating that there exists a significant global gene expression signature in PBMC in all the analyzed diseases. Thus, these data sets can be effectively utilized to study gene expression signatures in autoimmunity.

Next, these global statistic tests were generalized to check the expression levels of many of the diseases simultaneously, based on the 3-ANOVA and 3-PPI significant genes (see Materials and methods). When we permuted the genes with 3-ANOVA significant *P*-values, we found that the number of genes that are both 3-PPI significant and 3-ANOVA significant in the original data set was larger than their number in the randomized data sets (*P*-value $\leqslant$0.05). Furthermore, we found that the mean distance between 3-ANOVA significant genes in the real data is lower than the one obtained in the randomized data sets ($P<0.05$).

In addition, we permuted the *P*-values of the genes in each disease and compared the number of 3-ANOVA significant genes in the original data set and the permuted data sets (see Materials and methods). We found that the number of 3-ANOVA significant genes in the original data set was higher (*P*-value $<$0.05),

Table 3. Genes that are 3-PPI significant (see Materials and methods), their fold change and a description of each gene

Protein-interaction based P-*values*			
Gene symbol	*Fold change direction (1, increase; −1, decrease)*	*PPI* P-*value (3-PPI)*	*Description*
CXCR1/IL8RA	1	9.38E − 05	*Chemokine (C–X–C motif) receptor 1/interleukin 8 receptor*
IL8	1	0.000324	*Interleukin 8*
CXCL2	1	0.000435	*Chemokine (C–X–C motif) ligand 2*
CXCL1	1	0.000435	*Chemokine (C–X–C motif) ligand 1 (melanoma growth-stimulating activity, α)*
CXCL5	N/A	0.000435	*Chemokine (C–X–C motif) ligand 5*
TLR2	N/A	0.000519	*Toll-like receptor 2*
NCOR2	N/A	0.000907	*Nuclear receptor corepressor 2*
CXCR2	1	0.000988	*Chemokine (C–X–C motif) receptor 2*
THRB	N/A	0.001535	*Thyroid hormone receptor, β*
CXCL3	1	0.002343	*Chemokine (C–X–C motif) ligand 3*
CXCL6	N/A	0.002343	*Chemokine (C–X–C motif) ligand 6 (granulocyte chemotactic protein 2)*
RUNX1T1	N/A	0.00273	*Runt-related transcription factor 1; translocated to, 1 (cyclin D-related)*
HLA-E	N/A	0.003412	*Major histocompatibility complex, class I, E*
CCL4	N/A	0.003412	*chemokine (C–C motif) ligand 4*
ELK4	N/A	0.003412	*ELK4, ETS-domain protein (SRF accessory protein 1)*
JUNB	N/A	0.003679	*Jun B proto-oncogene*
CD79A	N/A	0.004827	*CD79a molecule, immunoglobulin-associated α*
DARC	N/A	0.005265	*Duffy blood group, chemokine receptor*
AES	N/A	0.007616	*Amino-terminal enhancer of split*
VDR	N/A	0.007797	*Vitamin D (1,25- dihydroxyvitamin D3) receptor*
HES1	N/A	0.010494	*Hairy and enhancer of split 1, (Drosophila)*
GTF2F2	N/A	0.011765	*General transcription factor IIF, polypeptide 2, 30 kDa*
RALB	N/A	0.013173	*v-Ral simian leukemia viral oncogene homolog B (ras-related; GTP binding protein)*
NFATC3	N/A	0.013173	*Nuclear factor of activated T cells, cytoplasmic, calcineurin-dependent 3*
SYNJ2	N/A	0.013173	*Synaptojanin 2*
RAB11FIP2	N/A	0.013173	*RAB11 family-interacting protein 2 (class I)*
TCEA1	N/A	0.013173	*Transcription elongation factor A (SII), 1*
ELK1	N/A	0.019814	*ELK1, member of ETS oncogene family*
ADM	N/A	0.021255	*Adrenomedullin*
CD40LG	N/A	0.021255	*CD40 ligand*
NR1I3	N/A	0.021255	*Nuclear receptor subfamily 1, group I, member 3*
EIF3C	N/A	0.021255	*Eukaryotic translation initiation factor 3, subunit C*
PRSS2	N/A	0.021255	*Protease, serine, 2 (trypsin 2)*
AGER	N/A	0.021255	*Advanced glycosylation end product-specific receptor*
PSMA6	N/A	0.021255	*Proteasome (prosome, macropain) subunit, α type, 6*
GRINL1A	N/A	0.021255	*Glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A*
PSMC5	N/A	0.022457	*Proteasome (prosome, macropain) 26S subunit, ATPase, 5*
DDIT3	N/A	0.022457	*DNA-damage-inducible transcript 3*
ACVR2A	N/A	0.022457	*Activin A receptor, type IIA*
CCL5	N/A	0.022457	*Chemokine (C–C motif) ligand 5*
GTF2F1	N/A	0.022457	*General transcription factor IIF, polypeptide 1, 74 kDa*
MAP3K14	N/A	0.022667	*Mitogen-activated protein kinase kinase kinase 14*
RRAS2	N/A	0.023096	*Related RAS viral (r-ras) oncogene homolog 2*
BCL3	N/A	0.023096	*B-cell CLL/lymphoma 3*

Abbreviations: N/A, not applicable; PPI, protein–protein interaction. The fold change was determined only if the gene was ANOVA significant in at least three diseases, and had the same direction as the fold change.

suggesting that indeed the same genes tend to be differentially expressed in the various autoimmune diseases.

Altogether, the results reported in this subsection demonstrate that, in comparison to random networks, the protein products of genes that are not only differentially expressed in each disease separately but also that are concurrently differentially expressed in many autoimmune diseases tend to physically interact with each other. These results support the conjecture that the reported signals are biologically meaningful, and that there is a common gene expression signature of autoimmunity in PBMC.

Gene clusters of autoimmunity and disease-specific gene clusters

Next, we concentrated on the set of autoimmune genes—that is, the genes that are significantly expressed (by the ANOVA *P*-value) in at least three diseases, and the direction of their gene expression shift is identical in all these diseases. The projection of these genes on the PPI network appears in Figure 2a. The graph includes 94 genes and is enriched with genes related to inflammatory response ($P = 3.5 \times 10^{-12}$), chemotaxis ($P = 4 \times 10^{-11}$) and cytokine–cytokine receptor interaction (P-value $= 4 \times 10^{-8}$); in addition, it includes process related to small chemokine, IL8-like ($P = 2.95 \times 10^{-6}$), chemokine activity (1.5×10^{-5}), leukocyte activation ($P = 5 \times 10^{-7}$), lymphocyte activation ($P = 6 \times 10^{-6}$) and processes related to regulation of transcription from RNA polymerase II promoter ($P = 3 \times 10^{-4}$; see more results in Supplementary Tables 4 and 5). Specifically, among others, the autoimmune network includes chemokine and chemokine receptors (*CCR4*, *CCL3*, *CCR1*, *CCL7*, *CXCL3*, *CXCL2* and *CXCL1*), IL1 and IL8 and their receptors (*IL8RB*, *IL8*, *IL8RA*, *IL1RAP* and *IL1R2*), IFNs and proteins related to their regulation (*IRF8*, *IRF2*, *IFNG* and *IFNGR2*).

Table 4. Some of the cellular processes that were enriched based on the differentially expressed genes in many of the studied autoimmune diseases (at least three diseases)

Cellular process	*MS*	*SLE*	*JRA*	*CD*	*UC*	*T1D*
Regulation of apoptosis	8.1E − 10	0.000238	2.3E − 06	1.6E − 22	4.2E − 08	0.000038
Inflammatory response	0.002	Not significant	Not significant	0.001	0.0012	9.78E − 06
Regulation of transcription	2.3E − 27	3.3E − 10	Not significant	3E − 07	1.1E − 44	Not significant
Regulation of cell cycle	0.000018	7.29E − 13	Not significant	4.4E − 11	0.00058	Not significant
Regulation of cytokines	Not significant	Not significant	0.0015	2E − 10	9.4E − 07	0.000017
Activation of NF-κB-inducing kinase activity	Not significant	Not significant	1.8E − 09	1.4E − 07	Not significant	0.00049
MAPKKK cascade	Not significant	Not significant	Not significant	4.2E − 10	0.00006230	5.6E − 08
Regulation of T-cell activation	Not significant	Not significant	0.000026	3E − 07	Not significant	1.2E − 08

Abbreviations: CD, Crohn's disease; FDR, false discovery rate; JRA, juvenile rheumatoid arthritis; MS, multiple sclerosis; MAPKKK, mitogen-activated protein kinase kinase kinase; NF-κB, nuclear fator-κB; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis. All the processes passed FDR correction. Full tables appear in Supplementary Table 3.

Table 5. Global *P*-values in each autoimmune disease separately

Disease	*# ANOVA and PPI significant*	*# PPI significant*	*Mean distance PPI significant*
MS	<0.05	<0.05	0.05
SLE	0.05	<0.05	0.05
JRA	0.05	0.05	0.25
CD	<0.05	0.3	0.05
UC	0.05	<0.05	0.05
T1D	0.15	0.1	<0.05

Abbreviations: CD, Crohn's disease; JRA, juvenile rheumatoid arthritis; MS, multiple sclerosis; PPI, protein–protein interaction; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis.

The size of the graph that appears in Figure 2a was significantly larger than the graph sizes obtained for the randomized networks ($P<0.05$; see Materials and methods). In most cases, the size of the graph was also significantly larger when we permutated each of the diseases separately ($P\leqslant0.05$ for T1D, UC, CD and MS; see Materials and methods); these results demonstrate that the graph is the result of a combination of gene expression patterns in most of the diseases, and also due to a small subset of them.

For the most part, in this study we concentrate on the genes and subnetworks that are common to many autoimmune diseases. In the rest of this subsection, we aim to study the subnetworks that are specific to each disease. To this end, we plotted for each disease, the subnetwork composed of the genes that are ANOVA significant only in that disease, which interact with at least one additional gene that is ANOVA significant in the same disease (see technical details in the Materials and methods section). With high reliability, the genes in each subnetwork are differentially expressed only in one disease.

The resultant disease-specific gene clusters appear in Figure 2b; GO enrichment analyses of each cluster appear in Figure 2c (see also Supplementary Tables 6 and 7). Interestingly, most of the disease-specific gene clusters relate to cellular processes, such as 'regulation of apoptosis' or 'chemotaxis'. However, by the definition of these clusters, it seems that these cellular processes are activated via different regulatory processes in different diseases. Specifically, in each disease, cellular processes related to disparate regulatory mechanisms were enriched with 'transcription from RNA polymerase II promoter' in MS; 'chromatin modification' and 'RNA splicing' in SLE; 'regulation of I-κB kinase/NF-κB cascade' in JRA; 'mRNA/rRNA processing and transport' in CD; 'chromatin assembly or disassembly' and 'phosphorylation' in UC; and 'translational initiation' in T1D.

Pathways of autoimmunity

In this section, we perform a comprehensive study of the pathways that are common to many autoimmune diseases. Figure 3 includes pathways that are enriched with significant genes (either ANOVA or PPI significant) in at least one autoimmune disease (see technical explanations in the Materials and methods section). The figure demonstrates that autoimmune diseases tend to have similar enriched pathways. Specifically, a major fraction of the pathways were enriched in more than one disease. Among the top pathways that are enriched in many diseases we found: (1) epidermal growth factor, platelet-derived growth factor, NF-κB, Wnt/β-catenin signaling, stress-activated protein kinase c-Jun NH2-terminal kinase pathways, which are related to proliferation and pathways related to metabolism, such as *PPARa*; (2) general signaling cascades, such as mitogen-activated protein kinase, extracellular signal-regulated kinase, p38 and TRK; and (3) pathways related to the interleukins IL2 and IL6, the cytokine granulocyte–macrophage colony-stimulating factor, inflammatory response via the glucocorticoid pathway and B-cell receptor.

Additional pathways that are enriched in at least three diseases and that have not been mentioned above include: IL4, IL10, IFN, vascular endothelial growth factor, ephrin, tumor growth factor-β, cAMP. The association of some of these pathways with autoimmunity have been previously suggested. For example, the abnormal expression of the IFN pathway in PBMC has been associated with SLE[33,50,51] and MS.[25]

We found only few signaling pathways that are very significant only for a low portion of the diseases (that is, in one to two diseases; Figure 4). One such pathway is ubiqutination, which exhibits a strong signal of regulation only in T1D. Another pathway is the Fcε RI pathway, which is significantly enriched in T1D and JRA.

In the rest of this section, we used the Ingenuity software (http://www.ingenuity.com/) and the literature for generating the pathways that are common to many of the analyzed autoimmune diseases. We analyzed the set of genes that are either 3-ANOVA significant or 3-PPI significant (see Materials and methods)

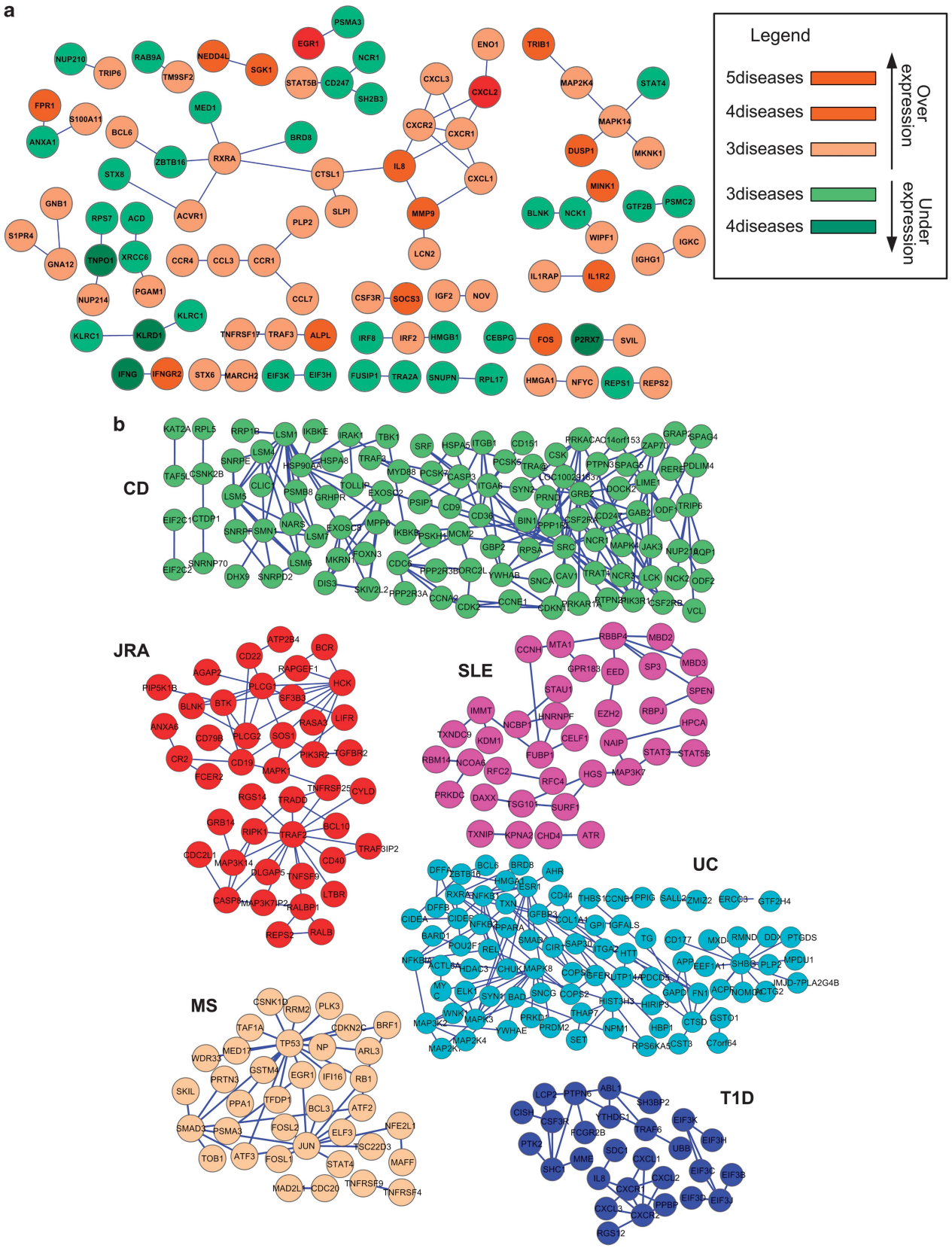

Figure 2. (**a**) Genes that are clustered in the PPI network and that are differentially expressed in at least three autoimmune diseases (the list of genes in this network and additional details about them appear in Supplementary Tables 4 and 5). We considered only genes with at least one neighbor in the projection. (**b**) Clusters of genes specific to each of the diseases (see Supplementary Table 6 and Materials and methods). (**c**). GO functions with top enrichment scores (all of them passed FDR) for each disease-specific cluster; see Supplementary Table 7.

Pathway	3ANOVA/3PPI	MS	CD	UC	T1D	SLE	JRA	
EGF	0.115	0.008	6*E-6	0.005	2*E-4	0.003	0.002	5-6 Diseases
Glucocorticoid	0.001	2*E-14	2*E-5	3*E-6	3*E-5	0.041	0.001	
p38	0.023	1*E-6	0.001	6*E-5	0.040	0.04	0.006	
SAPK JK	0.118	9*E-4	2*E-5	8*E-4	1*E-5	0.017	3*E-4	
B Cell Receptor	0.047	0.017	1*E-6	0.014	3*E-12	0.363	3*E-11	
ERK MAPK	0.463	0.010	4*E-5	0.01	6*E-4	0.427	2*E-4	
GM CSF	0.382	0.003	9*E-5	0.014	2*E-7	0.14	0.004	
EGF1	0.297	0.018	3*E-5	0.019	0.003	0.144	2.*E-5	
IL-2	0.275	0.013	6*E-5	0.196	5*E-6	0.044	1*E-5	
IL-6	0.005	2*E-4	0.003	2*E-4	1*E-5	0.501	0.006	
InsulinR	0.356	0.442	4*E-5	0.005	3*E-4	0.041	0.001	
TRK	0.633	0.130	1*E-4	0.006	0.001	0.032	4*E-4	
NF kappa b	0.726	8*E-5	8*E-8	2*E-4	6*E-4	0.486	1*E-12	
PDGF	0.141	3*E-4	2*E-9	3*4-4	1*E-4	8*E-5	0.05	
PPARa	0.185	0.002	0.003	2*E-4	0.011	0.008	0.477	
Whtb	0.463	0.010	0.003	0.004	0.027	0.024	0.425	
NK	0.001	0.015	5*E-5	0.11	5*E-5	0.84	4*E-4	4 Diseases
TCR	0.340	0.007	5*E-7	0.512	9*E-6	0.073	9*E-6	
Appoptosis	1	7*E-4	0.006	0.118	0.466	0.001	0.035	
Integrin	0.900	0.489	1*E-5	0.024	6*E-6	0.336	1*E-4	
Aryl signaling	0.185	1*E-9	1*E-6	5*E-9	0.111	1*E-6	0.477	
G-protein	1	0.422	1*E-4	0.03	0.004	0.075	3*E-4	
HepaticF	0.047	0.030	0.006	0.194	4*E-6	0.758	0.005	
Huntington's	0.974	0.054	0.031	0.25	0.021	0.043	9*E-4	
IL-10	3*E-4	0.001	0.008	0.004	3*E-4	0.195	0.077	
IL-4	0.654	0.641	0.010	0.197	2*E-9	0.043	1*E-7	
Jak Stat	1	0.331	0.003	0.067	0.007	0.005	0.026	
Neuregulin	0.482	0.857	1*E-4	0.039	0.003	0.158	5*E-5	
cAMP	1	0.043	6*E-4	0.015	0.228	0.188	0.05	
p53	0.894	2*E-6	0.003	0.007	0.82	9*E-5	0.966	
PPAR	0.319	0.201	0.003	0.003	0.015	0.002	0.088	
PTEN	1	0.309	6*E-7	0.117	8*E-4	0.03	0.001	
Cell Cycle G1S CP	1	1*E-6	0.019	0.006	0.505	0.001	0.917	
Toll like	0.002	0.003	4*E-6	1*E-5	6*E-5	0.461	0.10	
VEGF	0.611	0.576	2*E-4	0.041	0.004	0.845	0.009	
Acute	0.430	0.005	0.034	1*E-4	0.006	0.716	0.069	
Cell Cycle G2M CP	1	1*E-6	0.019	0.006	0.505	0.001	0.916	
Interferon	0.045	0.309	0.036	0.365	0.004	0.076	0.001	3 Diseases
Leukocyte	0.947	0.126	2*E-4	0.056	2.*E-5	0.976	0.002	
Chemokine	0.010	0.004	0.036	0.054	0.013	0.313	0.162	
Ephrin	0.894	0.070	3*E-4	0.463	4*E-6	0.459	6*E-7	
Estrogen R	0.554	5*E-6	0.078	5*E-6	0.282	0.001	0.088	
Hypoxia	0.727	5*E-4	0.036	0.065	0.182	0.047	1	
PI3K AKT	1	0.475	3*E-5	1*E-4	0.401	0.007	0.153	
Tgfb	0.260	5*E-4	0.109	0.007	0.644	9*E-4	0.076	
Vdr rxr	9*E-4	3*E-5	0.130	2*E-5	0.400	0.015	0.91	
Fc epsilon RI	1	0.927	0.055	0.361	4*E-8	0.60	1*E-5	1-2 Diseases
ER	1	8*E-4	0.270	0.180	0.64	0.147	0.029	
Axonal signaling	0.923	0.756	0.074	0.574	0.018	0.99	0.0003	
NRF2	0.175	0.078	0.018	1*E-4	0.75	0.297	0.626	
TR RXR	0.029	0.033	0.266	0.02	0.81	0.472	0.962	
Coagulation	1	0.678	0.743	0.198	0.04	0.84	0.978	
Ubiqutination	0.010	0.309	0.437	0.70	2*E-4	0.166	0.064	
Xenobiotic	0.083	0.536	0.051	0.004	0.694	0.70	0.575	
Amytrophic	0.829	0.475	0.038	0.522	0.114	0.716	0.066	
Carmide	0.254	0.880	0.017	0.063	0.922	0.22	0.146	

Figure 3. Pathways enriched in at least one autoimmune disease (see Materials and methods): The significant *P*-values are in red. Pathways that are enriched in many diseases are at the beginning, whereas the disease-specific pathways are at the end. The corresponding *P*-values of each pathway in each disease appear in the figure ($E-x$ denotes 10^{-x}).

together with the pathways that are enriched in the diseases mentioned above. A graph with the network (set of pathways) that includes genes that are significantly expressed in many diseases appears in Figure 4.

As can be seen, the resultant autoimmune pathways include genes related to proliferation and inflammatory responses. One branch that leads to proliferation is initiated with the protein kinases MAP2K4, which has a 3-ANOVA significant score (ANOVA significant in CD, UC and SLE; all *P*-values <0.01). These kinases eventually activate the *FOS* gene (ANOVA significant in CD, UC, MS and T1D; all *P*-value <0.036), which encodes the leucine zipper protein, that can dimerize with proteins of the JUN family,[52]

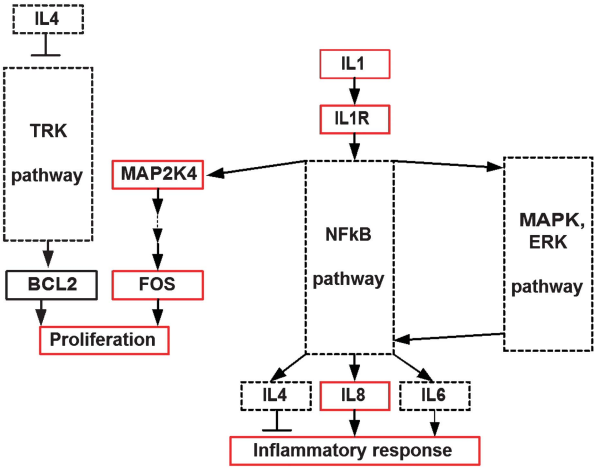

Figure 4. Signaling pathways and sub-PPI networks that are similarly expressed in many autoimmune diseases. Black boxes denote 3-PPI significant genes, dashed boxes denote enriched pathways and red boxes denote overexpressed 3-ANOVA significant genes.

thereby forming the transcription factor complex AP-1. As such, the FOS protein has been implicated as regulators of cell proliferation, differentiation and transformation. In some cases, expression of the *FOS* gene has also been associated with apoptotic cell death.

The branch that leads to inflammatory response initiates with the cytokine *IL1* (*IL1B* significantly upregulated in MS, CD, T1D and JRA; all *P*-value <0.03) and its receptor *IL1R* (IL1R2; ANOVA significant in MS, CD, UC and T1D; all *P*-values <0.01); many other members of the IL1 family were also ANOVA significant in many diseases (for example, the genes *IL1RN*, *IL1RAP*). Il1 is a pleiotropic cytokine involved in various immune responses, inflammatory processes and hematopoiesis. This cytokine is produced by monocytes and macrophages as a proprotein, which is proteolytically processed and released in response to cell injury, and thus induces apoptosis and inflammation. It has been suggested that the polymorphism of these genes is associated with SLE.[53] *IL1* eventually activates the NF-κB pathway (which is enriched in five autoimmune diseases: MS, CD, UC, T1D and JRA; all *P*-values <0.007; Figure 3), which is a transcription regulator that is activated by various intra- and extracellular stimuli, such as cytokines, oxidant-free radicals, ultraviolet irradiation and bacterial or viral products. Activated *NF-κB* translocates into the nucleus and stimulates the expression of genes involved in a wide variety of biological functions. Inappropriate activation of *NF-κB* has been associated with a number of inflammatory diseases,[54–56] whereas persistent inhibition of *NF-κB* leads to inappropriate immune cell development or delayed cell growth.[56] *NF-κB* pathway eventually regulates the *IL8* cytokine (3-PPI *P*-value $=0.0003$; ANOVA significant in MS, CD, UC and T1D; all *P*-values <0.02), which is a member of the CXC chemokine family. The *IL8* activates *IL8R* (ANOVA significant in CD, UC and T1D; all *P*-value <0.009). *IL8* is one of the major mediators of the inflammatory response. It is secreted by several cell types, functions as a chemoattractant with an important role in inflammation (it signals immune cells (neutrophils) to enter the site of inflammation) and is also a potent angiogenic factor. Thus, *IL8* triggers migration and adhesion (see also Tuller *et al.*[38]). Additional cytokines that are regulated by the *NF-κB* pathway are *IL4* (which is enriched in CD, T1D, SLE and JRA; all *P*-values <0.04; Figure 3) and *IL6* (which is enriched in MS, CD, UC, T1D and JRA; all *P*-values <0.006; Figure 3). *IL6* can also act as a proinflammatory (but also anti-inflammatory) cytokine.[57] It was suggested that this gene is relevant to the pathogenesis of diseases such T1D,[58] SLE[59] and RA;[60] the results reported here suggest that it is also relevant to the pathogenesis of additional autoimmune diseases. *IL4* may have an important role in regulating inflammation,[61] usually it has anti-inflammatory effect, and it also regulates the proliferation of B and T cells.[62,63]

Another mechanism that leads to inflammation is via the mitogen-activated protein kinase pathway that is enriched in five autoimmune diseases (MS, CD, UC, T1D and JRA; all *P*-values <0.01; Figure 3), which eventually regulates inflammation via the NF-κB pathway. Finally, our results also suggest the regulation of proliferation via the TRK pathway that is enriched in five autoimmune diseases (SLE, CD, UC, T1D and JRA; all *P*-values <0.03; Figure 3). The activation of this pathway in PBMC has been reported before;[64] here we suggest that it may be relevant to the pathogenesis of the analyzed autoimmune diseases. This pathway starts with the enhancing of TRK with *IL4*,[64] as mentioned above, which enhances the gene *NT-3* (an inducer of TRK). Eventually TRK pathway activated *BCL3* (3-PPI *P*-value <0.02), which, as we mentioned before, induces proliferation.

Distances between autoimmune diseases

The aim of this subsection was to cluster the autoimmune diseases according to their global changes in the PBMC gene expression signature. To this end, we computed a distance between each pair of diseases, which is based on the genes that are ANOVA significant in the diseases, and the *d*istances between them in the PPI network. Thus, the distance between a pair of diseases considers both the ANOVA *P*-values and the PPI network; roughly speaking, if the ANOVA significant genes in a pair of diseases tend to be closer in the PPI network, the diseases are more similar (see details in the Materials and methods section). The distances between all the disease pairs appear in Figure 5a, and the resultant clustering of the diseases (see Materials and methods) appears in Figure 5b.

As can be seen in Figure 5, the two inflammatory bowel diseases, CD and UC, have a relatively similar gene expression signature. Thus, the similar phenotypes of these diseases are manifested also in the changes of the expression levels of patients with the disease in comparison to healthy subjects. Indeed, 85% of the cases with CD and UC appear in the 3-ANOVA significant group (therefore, we also provide a list of 2-ANOVA significant genes in Supplementary Table 2). The systemic autoimmune disease SLE, and also MS, are positioned in the 'middle' of the clustering tree, whereas the gene expression signatures of T1D and JRA are relatively different from the signatures of the other diseases.

DISCUSSION

In this paper, we reported the first large-scale systems biology analysis of autoimmune diseases, by analyzing their gene expression in PBMC, and by employing a novel projection of it on the human PPI network. First, we reported lists of genes with significant *P*-values in many diseases; previous studies analyzed a smaller set of diseases, not necessarily in PBMC, and without estimating post-transcriptional regulation based on the PPI network. At the next step, we performed a GO enrichment analysis of the significant genes in each disease separately, and reported the GO groups that are enriched in many autoimmune diseases. We continued with an analysis of each disease separately, and demonstrated, for the first time, that in each disease there is a significant global gene expression signal in PBMC that is related to the disease; we also found a significant global signal of autoimmunity when we analyzed the genes that are *c*ommon to many autoimmune diseases. Next, we analyzed, for the first time, the projection of the common autoimmune genes and the disease-specific genes on the PPI network, showing that in both cases genes related to apoptosis are

a

	MS	SLE	JRA	CD	UC	T1D
MS	0	3.69	3.68	3.64	3.32	4.16
SLE	3.69	0	4.26	3.16	3.32	4.48
JRA	3.68	4.26	0	2.96	3.7	3.62
CD	3.64	3.16	2.96	0	**2.74**	3.22
UC	3.32	3.32	3.7	2.74	0	3.62
T1D	4.16	4.48	3.62	3.22	3.62	0

b

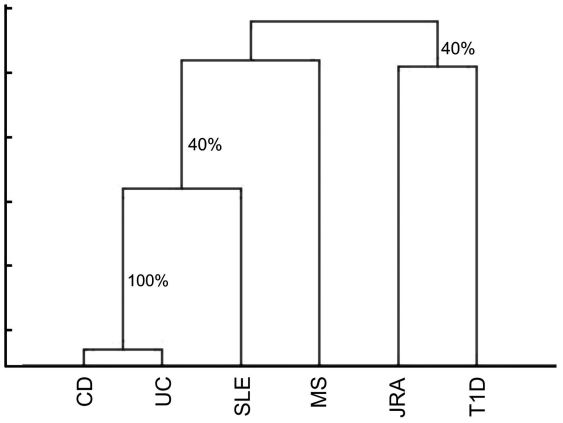

Figure 5. (**a**) Mean distance between differentially expressed genes in each disease based on the protein interaction network (see Materials and methods). (**b**) A hierarchical clustering of the autoimmune diseases based on A (the edges include the Jackknifing results; see Materials and methods). A full colour version of this figure is available at the *Genes and Immunity* journal online.

over-represented, whreas the common autoimmune genes include inflammatory genes. Additional analysis of the common autoimmune genes demonstrates that inflammation and proliferation are the common pathways that are differentially expressed in autoimmune diseases.

Finally, we demonstrated a novel robust approach for comparing gene expression in diseases. We showed for the first time that CD and UC, known to have similar phenotypic characteristics, also exhibit relatively similar gene expression patterns in PBMC. The paper includes discussions about relevant genes and pathways. However, due to lack of space, many additional relevant genes appear in the Supplementary Tables. For example, the protein *IL17* is a highly proinflammatory cytokine that has been reported to be upregulated in transcription in many autoreactive T-cell lines of many autoimmune diseases (see, for example, Nistala and Wedderburn[14] and Mandel *et al.*[65]). Indeed, this gene is ANOVA significant in two diseases (MS and SLE), but it did not pass the threshold to be included in Table 2.

One of the major conclusions derived from this paper is that there is a robust, biologically meaningful, significant common signature, which appears in the analyzed diseases. Thus, although target tissues of the autoreactive cells differ among different autoimmune diseases, there are common 'autoimmune' signaling pathways that are triggered in PBMC in many of the diseases (Figure 2a and Figures 3 and 4). Moreover, cellular pathways that tend to be abnormally expressed in a certain autoimmune disease are expressed, with high probability, in additional diseases (Figure 4). Specifically, we show for the first time that it is relatively difficult to find pathways that are differentially expressed only in a small subset of the diseases (Figure 4). To verify that the set of autoimmune genes is indeed *specific* to autoimmune diseases and *not* differentially expressed in non-autoimmune diseases, we performed a statistical test that compares the set of 3-PPI and 3-ANOVA genes to the set of genes that are differentially expressed in many other non-autoimmune diseases (a total of 367 diseases; details are given in the Materials and methods section). We found that the two sets of genes are indeed not similar ($P = 0.05$). Thus, many of the autoimmune genes and pathways that are reported in this study can be used as a general biomarker of autoimmune disease diagnosis.

On the other hand, we show for the first time that in each disease the autoimmune pathways such as apoptosis and cellular functions are regulated via different regulatory mechanisms. Thus, as we report in this paper, there are dozens of genes, subpathways and subnetworks that are specific to each disease (Figures 2b and c). These genes and pathways can be used as biomarkers for a more refined, disease-specific diagnosis.

The three major pathways that are emphasized in most of the stages of our analysis are the downregulation of apoptosis, and the upregulation of inflammation and cell proliferation (for example, the pathways epidermal growth factor, platelet-derived growth factor, *NF-κB*, Wnt/β-catenin signaling, stress-activated protein kinase c-Jun NH2-terminal kinase pathways, and p53, vascular endothelial growth factor) that are partially mediated via various interleukins and cytokines (for example, IL2 and IL6, granulocyte–macrophage colony-stimulating factor, glucocorticoid pathway, B-cell receptor, IL4, IL10, IFN, tumor growth factor-β), and signaling pathways such as mitogen-activated protein kinase, extracellular signal-regulated kinase, p38, TRK, ephrin and cAMP. Specifically, we found apoptosis to be also enriched when we analyzed genes that are specific to each disease, demonstrating that downregulation of apoptosis is performed by different regulatory mechanisms in each disease. Apoptosis, programmed cell death, is a process that occurs in various cells in multicellular organisms.[39] It can be triggered by extracellular (for example, growth factors and cytokines) or intracellular (for example, damage to the DNA) signals. The intracellular pathways are usually controlled by the Bcl2 family members, whereas the extracellular pathways are usually controlled by death receptors.[39,66]

Apoptosis has an important role in T-cell maturation: thymocytes expressing non-functional or autoreactive TCRs are eliminated by apoptosis during development. Apoptosis also leads to the deletion of expanded effector T cells during immune responses. Similarly, immature B cells are tested for autoreactivity by the immune system before leaving the bone marrow. The immature B cells whose B-cell receptors bind too strongly to self-antigens will not be allowed to mature and may be eliminated by apoptosis.[39]

Thus, the dysregulation of apoptosis in the immune system results in autoimmunity.[66,67] Furthermore, in general, tissue homeostasis is maintained through a balance between cell proliferation and apoptotic cell death. Thus, as we demonstrate here, both the downregulation of apoptosis and the upregulation of cell proliferation contribute to autoimmunity. One contribution of this paper is the detailed lists of genes that are involved in these pathways in autoimmune diseases (see also Supplementary Tables 1, 2, 4 and 6).

A report of the single-nucleotide polymorphisms that are common to many autoimmune diseases, which was published recently, includes a relatively short list of genes that appear in at least three of the analyzed autoimmune diseases:[68] *PTPN22*, *IL23R*, *NRXN1*, *KIAA1109*, *EPHA7*, *TRIM27*, *TNFAIP3*, *TNKS* and *C20orf42*. Interestingly, none of these genes have significant 3-ANOVA *P*-values or 3-PPI *P*-values. However, *EPHA7* is 2-ANOVA significant (in CD and UC); this gene belongs to the ephrin receptor subfamily of the protein–tyrosine kinase family and was mentioned above as a pathway significantly enriched in several diseases.

This result demonstrates again the differences between genes' single-nucleotide polymorphisms analysis and gene expression analysis. It is probable that some of these single-nucleotide polymorphisms are causal and that they eventually affect at least one of the aforementioned pathways (apoptosis, inflammation and proliferation). However, as we previously mentioned, the regulatory changes caused by these genes are not necessarily at the mRNA level, or they may be too weak to be detected by a microarray, or/and by our PPI *P*-value. Our approach may successfully detect regulatory changes 'downstream' to these genes. This fact emphasizes the need for incorporating additional data when analyzing mRNA measurements in diseases, as was done here with PPIs. When an additional qualitative relevant large-scale data (for example, measurements of miRNA and tRNA levels, highly qualitative network of miRNA and their targets, highly qualitative network of transcription factors and their targets) will become available, the approach described here can be further improved.

Finally, this paper includes a novel clustering of autoimmune genes (Figure 2), as well as a disease clustering according to their gene expression signature in PBMC (Figure 5). The results are relatively robust as they are also based on the PPI network. We believe that such an approach can be generalized for the classification of other groups of diseases (for example, cancers, diseases of the nervous system or metabolic diseases), as well as groups of patients within a disease. Such a classification may be useful in personalized medicine, and for evaluating the possibility of employing similar medical treatments in different diseases or patients that have a similar fingerprint. In addition, our approach can be useful for developing an accurate model of the pathogenesis of various diseases, setting the basis for the development of more rational therapeutic approaches.

MATERIALS AND METHODS

The studied populations

Some of the clinical and demographical characteristics and other relevant information of each of the analyzed data sets appear in Table 1.

In all cases we analyzed, the gene expression of the aforementioned diseases in PBMC. As autoreactive PBMCs initiate the autoimmune inflammatory process against the corresponding target organs,[1,6–10,31] transcriptional profiling of PBMCs is a useful tool for identifying disease-related specific and common autoimmune signatures, and has been conventionally used in this setting.[20–23]

Clinical and demographical information regarding the analyzed patients

In this subsection, we provide clinical information about the analyzed data sets.

MS: Clinical information about the MS patients appears in Table 1 (see also Supplementary Information and in Tuller *et al.*[38]). Specifically, the mean age of the patients and the healthy subjects was very similar (around 35 years), and we included more than 50% women in the group of healthy subjects. All the MS patients were diagnosed with definite MS according to the McDonald criteria.[69] MS relapse was defined as the onset of new objective neurological symptoms/signs or the aggravation of existing neurological disability, not accompanied by metabolic changes, fever or other signs of infection, and lasting for a period of at least 48 h accompanied by objective change of at least 0.5 in the Expanded Disability Status Scale score. Confirmed relapses and Expanded Disability Status Scale scores were recorded consecutively.

SLE: The five SLE patients (mean disease duration 8.4 ± 5.6 years) were free of cytotoxic agents or immunomodulatory drugs for at least 30 days before blood was withdrawn. All participants had peripheral blood counts within the normal range.[28] In addition, the male-to-female ratio was identical in both groups and the mean age was very similar (around 43 vs 45 years).

CD and *UC*: In the cases of CD and UC, blood samples for pharmacogenomic analysis were collected at the North American and European clinical sites from a total of 42 apparently healthy individuals, 59 CD patients and 26 UC patients participating in three distinct clinical trials (two CD and one UC trial).[35] The mean age was very similar in the two groups (for example, around 47 vs 44 years). The female to male ratio was different; however, both groups included both females and males and the ANOVA *P*-value controls for this bias.

CD patients had CD activity index scores[70] ranging between 220 and 400, with an abdominal pain rating of $\geqslant 25$ and/or a diarrhea rating of $\geqslant 25$. Diagnosis of CD for at least 6 months was confirmed by radiological studies, endoscopy with histological examination or surgical pathology; patients with a diagnosis of CD were included if the diagnosis was confirmed by a biopsy. UC patients had scores from the Physician's Global Assessment of the Mayo Ulcerative Colitis Scoring System, ranging from mild to moderate (scores of 1 or 2).[71] The diagnosis of left-sided UC was provided by endoscopy with biopsy, in addition to the standard clinical criteria.

Investigation of concomitant medication usage between the two inflammatory bowel disease populations indicated that neither 5-aminosalicylic acid nor any of the other less frequently used drugs reported as concomitant medications confounded the comparisons in this study.[35]

JRA: In the case of JRA, 26 chronic arthritis patients seen at the Cincinnati Children's Hospital Medical Center Rheumatology Clinic were assessed for clinical phenotype by retrospective chart review.[27] The ACR diagnosis and classification criteria[72] are the common system in the United States, and in all instances there was sufficient information in the charts to classify patients by course according to these criteria. Of the 15 patients with polyarticular course, three had pauciarticular onset, nine had polyarticular onset and three had systemic onset. Information about the presence or absence of rheumatoid factor in seven patients was not available. These patients were classified as if they were rheumatoid factor-negative with respect to juvenile idiopathic arthritis[32] classification. The patients and control groups have similar mean age (15.9 vs 14.1 years) and similar female to male ratio (8/7 vs 6/5).

T1D: In the case of T1D, blood samples were obtained from 43 newly diagnosed T1D patients.[29] We did not consider the samples that are 1 and 4 months after diagnosis (to avoid duplicates). Patients with type 2 diabetes were distinguished from T1D on the basis of age, body habitus, presence (11 of 12 patients) of acanthosis nigricans, family history of type 2 diabetes (11 of 12 patients) and absence of autoantibodies to insulin, IA-2 and GAD65. All but two of the T1D patients with positive anti-insulin antibodies were also positive for at least one additional autoantibody. One teenager with putative T1D was excluded from the study because he was negative for all three antibodies. The patients and the control groups have similar mean age (9.5 vs 10.9 years) and close to identical female to male ratio (25/18 vs 14/10).

RNA isolation and microarray expression profiling in MS and SLE

In the case of the MS, which were generated in our lab, PBMCs were separated on Ficol–Hypaque gradient, total RNA was purified, labeled, hybridized to a Gene chip array (HU133A-2 in the case of MS) and scanned (GeneArray scanner G2500A, Hewlett Packard, Palo Alto, CA, USA) according to the manufacturer's protocol (Affymetrix Inc., Santa Clara, CA, USA).

The normalization and analysis of the gene expression data sets

In the case of MS, we used the Sheba MS center-recorded computerized clinical follow-up and blood gene expression measurements data set. This data set includes information about demographical variables and batch effects, such as age, gender and time of blood sampling. In the case of the other diseases, we used the demographical variables and batch effects from the relevant papers and from Gene Expression Omnibus.

The following data analysis was performed by MATLAB (The MathWorks, Inc., Natick, MA, USA). Expression values were computed from raw CEL files by applying the background adjustment and robust multichip average background correction algorithm. Each of the data sets underwent quantile normalization. At this stage, we averaged the expression levels of all the probes of a gene to get an estimation of the expression levels of the gene.

In the next step, in each of the data sets we computed significantly over/underexpressed genes based on ANOVA,[41] considering batch effects that were available in each case (such as scan date, chip type and also the age and gender of each patient). Although it is usually impossible to remove all batch effects, we aim to minimize them.

The following batch effects were considered (these are effects for which the required information was available for all the analyzed patients

and healthy subjects): MS—scan date, serial number, age and gender; T1D—ethnicity, race, age and gender; JRA—scan date; CD—age, gender and race; SLE—scan date and chip type; and UC—age, gender and race.

To avoid the statistical problems corresponding to merging various sources of gene expression, the analysis was performed for the gene expression measurements of each autoimmune disease separately.

In addition, we performed the following procedures:

In the case of T1D:

(1) We normalized each chip (U133A, U133B) separately.
(2) For each gene, we choose the minimal *P*-value (between U133A and U133B).
(3) The fold change of each gene was determined by the chip that gave the more significant *P*-value.

In the case of SLE:

The data set included two types of chips (seven samples from U95 and three samples from U133). We considered only the common genes and performed ANOVA based on all the measurements.

In the case of CD and UC:

We took the relevant files for each data set (excluded UC from CD and vice versa). In this case, there were no CEL files, thus we only performed quantile normalization on the data from the GSE soft file.

In the case of JRA, we considered patients with poly-JRA.

Comparing data from different types of DNA chips

In general, in each disease we translated (based on the Affymetrix annotation tables) the probes to genes (average over all probes), and in the various tests between diseases we consider only the genes that are common to all the diseases (9273 genes).

The *Homo sapiens* PPI network

The human PPI network was gathered from public databases[73,74] and from recently published papers.[75,76] The final reconstructed network included 7915 proteins and 28 972 PPIs (6850 proteins and 25 931 PPIs appear in the analyzed chips).

General notes about the statistical analyses

The paper includes four general types of *P*-values/scores: (1) *P*-values/scores of single genes in a single disease (PPI *P*-value and ANOVA *P*-value); these *P*-values are similar to the ones reported in previous papers. (2) *P*-values/scores of single genes in multiple (at least 50%—three) diseases (3-PPI *P*-value and 3-ANOVA); these *P*-values/scores are generalizations of the *P*-values in (1). (3) Enrichment *P*-values related to pathways or cellular functions based on (1) (specific disease) or (2) (multiple diseases). (4) Global *P*-values related to the entire set of genes and the PPIs between them. These *P*-values can also be related to a specific disease (1) or multiple diseases (2). In some of the cases, false discovery rate (FDR) filtering was not needed (for example, case (4)), but in other cases we perform FDR filtering.

Most of the statistical and computational approaches employed in this study (for example, the PPI *P*-value in (1) and (3), pathway enrichment in multiple diseases in (3), global *P*-values in (4) and clustering of genes on PPI networks) were tailored for this study and cannot be performed by public tools such as the 'gene expression atlas' (http://www.ebi.ac.uk/gxa/).

Genes that are ANOVA significant in many autoimmune diseases

As we defined in the Results section, genes that are ANOVA significant in at least *x* diseases were named *x*-ANOVA significant genes.

Gene-specific *P*-values based on the PPI network and the gene expression of all the diseases

We begin this section with a brief motivation related to the advantages of the PPI *P*-values. Regulation of gene expression is a multistep process that includes, among others, the following steps: transcription of the gene to pre-mRNA, splicing of pre-mRNA, degradation of mRNA, translation of mRNA to proteins and degradation of proteins (see, for example, Alberts *et al.*[39]). Many of these stages are mediated and regulated by proteins via protein–protein and protein–DNA interactions, or by RNA genes such as miRNA via RNA–RNA interactions.

It is important to remember that since proteins are the 'machines' that perform most of the cellular functions, we are interested in detecting changes in the protein levels of genes in the different diseases in comparison to healthy subjects (in the case of this paper, autoimmune diseases). However, researchers in the field usually use mRNA levels (or changes in the mRNA) as a proxy to protein levels (or changes in the protein levels), simply since today it is technically much easier to perform large-scale measurements of mRNA levels than to perform large-scale measurements of protein levels. For example, there are large-scale, relatively 'cheap', technologies such as DNA chips and deep sequencing for measuring mRNA levels, but no similar technologies for measuring protein levels.

Thus, in many cases, because of the post-transcriptional regulatory steps mentioned above, the changes in protein levels cannot be detected at the mRNA levels; similarly, it is possible that we detect changes at the mRNA levels while there are no significant changes in protein levels (for example, because of noise in mRNA measurements and/or regulatory reasons). For example, if in a disease miRNA molecules bind to the mRNA molecule of a gene and decrease its translation rate, the protein levels of the gene will decrease, but its mRNA levels will not change; in this case, if we use microarray technology for measuring mRNA levels, we will not see the decrease in protein levels of the gene in the disease.

Indeed, it was shown that gene mRNA levels explain only around 30% of their protein abundance variance in human and yeast;[77,78] however, there are significant correlations between the various regulatory stages (for example, there is significant correlation between mRNA levels and protein levels[77,78]).

How can we detect post-transcriptional regulatory changes in diseases based on mRNA measurements?

Cellular processes usually involve sets of genes and proteins that interact with each other to perform various functions: for example, the apoptosis process includes the following physical (protein–protein) interactions:

Grb2/SOS → Ras → Raf → Mek-- > MEKK, and so on.

If indeed there are regulatory changes in a pathway, we expect to see changes in the protein levels of all (or most) of the genes in the pathway. Thus, if we observe that many of the proteins that interact with the product of a gene undergo significant changes in their transcription levels (that is, the gene has significant PPI *P*-value), with high probability this gene also undergoes these regulatory changes. It is possible that we do not observe these regulatory changes at the transcription level because of noise or biases in the mRNA measurements, or because of the fact that the regulatory changes in the gene are not at the transcription step but at the other aforementioned regulatory steps (for example, degradation, translation, PPIs, and so on). The PPI *P*-value is based on this idea (that is, 'guilt by association')—if the protein product of a gene has significantly many PPIs with genes that undergo significant changes in their mRNA levels, with high probability it also undergoes regulatory changes (not necessarily at the mRNA levels). For example, in the case of the apoptosis subpathway above, if we observe changes in mRNA levels of the genes encoding the proteins upstream of Raf (the gene *Ras*), and also in the mRNA levels of the genes encoding proteins downstream of Raf (the gene *Mek*), it increases the probability of all the pathways, including Raf that undergo regulatory changes.

To compute a *P*-value related to this idea, we invented the PPI *P*-value approach.

Figure 6 includes an illustration of the PPI *P*-value approach. The left part of the figure includes the PPIs with the product of a target gene (in red). Each node in the graph corresponds to a protein, a node is black/white if the mRNA levels of the protein are significantly different in the disease vs the controls; an edge between two nodes represents PPI. As can be seen,

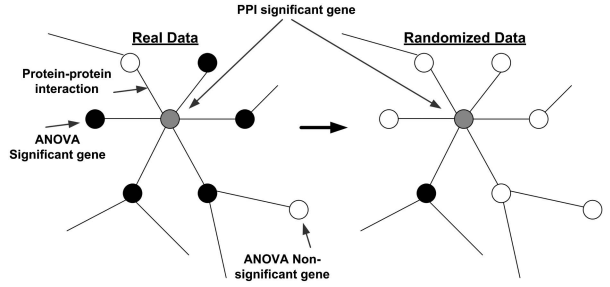

Figure 6. Illustration of the PPI *P*-value computation (see explanations in the main text). A full colour version of this figure is available at the *Genes and Immunity* journal online.

the target gene has PPIs with six proteins, out of which five are ANOVA significant (black).

The right part of the figure represents the same PPI network after randomizing (permutation of the black and white nodes). As can be seen, in this case the target gene has PPI with only one black node (out of six PPIs), which is much lower than in the case of the real data. The PPI *P*-value is based on such a comparison to random networks; specifically, we compute the probability that each of the genes in the network will have similar (or larger) number of PPIs with proteins/nodes that are ANOVA significant in a randomized network.

Indeed, approaches similar but not identical to the PPI *P*-value mentioned in this study have already been successfully employed in cancer and MS.[38,40]

Now we describe the technical details related to computing the PPI *P*-values.

For each gene, we computed a *P*-value (PPI *P*-value) that was based on the protein interaction network in the following manner:

(1) For each protein in the network, we computed a hypergeometric *P*-value that is based on the number of proteins, which interact with it and are differentially expressed in at least *x* diseases (we used $x = 3$; that is, 3-ANOVA significant), the total number of differentially expressed proteins and the topology of the PPI network. Roughly speaking, a larger number of interacting genes that are *x*-ANOVA significant correspond to a more significant *P*-value. We named this *P*-value *x*-PPI-network *P*-value (PPI- network *P*-value) and calculated it as follows: Let n_i denote the number of genes interacting with gene *i*, let m_i be the number of interactions with gene i that are *x*-ANOVA significant, let N be the total number of genes in the network and let M be the total number of *x*-ANOVA significant genes in the network. The *x*-PPI-network *P*-value of gene *i* is:

$$\sum_{j=m}^{n_i} \binom{M}{j} * \binom{N-M}{n_i-j} / \binom{N}{n_i}$$

We also computed an *x*-PPI-network *P*-value for each gene in a specific disease. In this case, the *P*-values were computed as above, but with $x = 1$ (that is, 1-PPI-network related to 1-ANOVA significant genes).

A global *P*-value for the number of genes with significant *x*-PPI-network *P*-values

A global *P*-value, which is related to the number of genes with significant *x*-PPI-network *P*-values, was computed as follows:

Repeat 100 times:

(1) Randomly choose a subset of genes of size *M*.
(2) Assume that the subset of genes that was chosen in 1 includes the *x*-ANOVA significant genes, and compute for all genes the *x*-PPI *P*-value mentioned above.
(3) Compute the empirical probability (frequency) that the number of genes with significant *x*-PPI *P*-values obtained in the random network is larger (or equal) to the number of genes with significant *x*-PPI *P*-values obtained in the original network.

A global *P*-value for the number of genes with significant *x*-PPI-network and *x*-ANOVA *P*-values

This global *P*-value was computed in a manner similar to the *P*-value above but with the modification that we computed the frequency/probability that the number of genes with *both* significant *x*-PPI *P*-values and *x*-ANOVA *P*-values obtained in the random network is larger (or equal) to the number of genes with significant *x*-PPI *P*-values and *x*-ANOVA *P*-values obtained in the original network.

A global *P*-value for the number of genes with significant *x*-ANOVA *P*-values, which interact with *x*-ANOVA significant gene(s)

This *P*-value corresponds to the number of genes in Figure 2a, and was computed in a manner similar to the two previous *P*-values.

A global *P*-value for the distance between *x*-ANOVA significant genes in the PPI network

The aim of the global *P*-value described in this subsection is to demonstrate that genes with *x*-ANOVA significant *P*-values tend to be close to other genes with *x*-ANOVA significant *P*-values in the PPI network. Therefore, the observed changes in the gene expression are not random.

This global *P*-value was computed as follows:

(1) Find for each gene that is *x*-ANOVA significant, the distance to the closest gene in the PPI network that is also *x*-ANOVA significant. Compute the mean distance.
Repeat 100 times.
(2) Randomly select a subset of *M* genes and assign them to the nodes of the *x*-PPI network such that the degree distribution of these nodes will be identical to the degree distribution of *x*-ANOVA significant nodes in the original graph.
(3) Find the mean distance between each of the *M* random *x*-ANOVA significant nodes and its closest neighbor, and compute the mean distance.
(4) Compute the empirical probability (frequency) that the random network has a smaller (or equal) mean distance than the mean distance in the original one.

The effect of a single disease on the global *P*-values

To show that the global *P*-values reported above are not related to a small subset of the diseases, we performed the statistical tests reported above when randomizing *only* one of the diseases at a time.

Specifically, for each disease *i* with M_i ANOVA significant genes, in each of the randomization steps, we randomly selected subsets of M_i genes, assuming that these are the ANOVA significant genes in disease *i*, while keeping fixed the set of ANOVA significant genes in the other diseases.

A *P*-value for comparing the autoimmune genes to general disease genes

To compare general disease genes to the autoimmune genes, we performed the following test:

(1) The autoimmune genes were defined as the set of 3-PPI and 3-ANOVA significant genes.
(2) The set of the general disease genes was based on the 'gene expression atlas' database (http://www.ebi.ac.uk/gxa/). This set included the genes that are significantly expressed in at least 50% out of 367 diseases that appeared in this database.
(3) Let n denote the number of autoimmune genes, let m be the number of autoimmune genes that are also general disease genes, let N be the total number of genes and let M be the total number of general disease genes. The *P*-value related to enrichment of autoimmune genes with general genes is:

$$\sum_{j=m}^{n} \binom{M}{j} * \binom{N-M}{n-j} / \binom{N}{n}$$

A significant *P*-value suggests that autoimmune genes tend to appear in many other diseases. The resultant *P*-value was not significant (0.48), suggesting that the autoimmune genes are not enriched with general disease genes.

GO enrichment

GO enrichment analysis of genes with significant ANOVA *P*-values and/or of genes with significant PPI *P*-values ($P < 0.05$ in both cases) was performed by DAVID[79] (http://david.abcc.ncifcrf.gov/). In all cases, we considered GO groups whose enrichment *P*-values passed FDR as was reported by DAVID (allowing FDR of $< 5\%$). In addition, we considered only GO groups from the biological process ontology.

Pathway enrichment

Pathway enrichments were performed based on the pathways of Ingenuity software (http://www.ingenuity.com/; see Supplementary Table 8)

Let n_i denote the number of genes in pathway *i*, let m_i be the number of genes in pathway *i* that are ANOVA or PPI significant in a certain disease (*y*), let n be the total number of genes, and let M be the total number of ANOVA or PPI significant genes in the network. The enrichment *P*-value of pathway *i* in disease *y* is:

$$\sum_{j=m}^{n_i} \binom{M}{j} * \binom{N-M}{n-j} / \binom{N}{n_i}$$

Controlling for the FDR

We used the MATLAB function, which is the implementation of Storey[80] for controlling the FDR, and we also report the results of the more strict approach of Benjamini and Hochberg.[81] In the case of the GO enrichment analysis that was performed by DAVID, we used the FDR filtering of DAVID.

Distances between the autoimmune diseases and their clustering

Let *D*1 and *D*2 denote two autoimmune diseases.

The distance between *D*1 and *D*2, Dist(*D*1, *D*2), was defined as follows: (1) for each gene, *g1i*, in the data set corresponding to the disease *D*1 that was both ANOVA significant and PPI significant, find the distance to the closest gene in the PPI network that is ANOVA and PPI significant in the data set corresponding to the disease *D*2. Let *d1i* denote this distance. (2) *Dist*(*D*1, *D*2) is the mean *d1i* over all the genes *g1i; Dist*(*D*1, *D*2), = *meani d1i.*

At the next step, we generated a symmetric distance measure between the diseases: SymDist(*D*1, *D*2) = Dist(*D*1, *D*2) + Dist(*D*2, *D*1). We used these distances to generate a clustering of the diseases based on the MATLAB implementation of hierarchical clustering (with default parameters).

To evaluate the robustness of the clustering, Jackknifing (see, for example,Shao and Tu[82]) was performed as described below:

Repeat 100 times:

(1) Randomly select 80% of the genes.
(2) Run the steps described above to generate a hierarchical clustering.
(3) Count for each edge (that is, a partitioning of the leaves) in the hierarchical clustering (Figure 4) the number of times it appears in the resultant random sampling.

Disease-specific clustering

To find clusters of genes that are specific to different diseases (Figure 2b), we performed the following steps. For each disease we considered the graph that is induced by genes with both ANOVA and PPI significant *P*-values (*P*-value $\leqslant 0.05$), and that are *not* significant in any other disease. In addition, both nodes in each edge in the graph should be significant in the same disease.

Fold change

To estimate the fold change in one condition vs another (for example, relapse vs remission), taking into account the batch effects and additional variables, we performed a multivariate linear regression (see, for example, Pedhazur[83]) in which the binary variable (patient with a disease or healthy subject) is the dependent variable and all the variables mentioned in the ANOVA computation are the independent variables (a different set of variables for each disease). The scan date was represented by a set of dummy binary variables, and other variables were either continuous or binary. The sign of the coefficient related to the expression levels determined the fold in one condition vs the other. For example, in the case of the MS database we set the dependent variable to be '1' in the case of a patient with MS (and '0' otherwise—that is, for healthy subjects). Thus, a positive coefficient of the expression levels variable corresponds to increased expression levels of the gene in MS in comparison to the control.

Similar results were obtained when we used partial correlations (Spearman's or Pearson's) between the *MS*/*Healthy* variable and the expression levels given all the other variables.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1 Noel R, Rose IRM. *The Aautoimmune Diseases*. Academic Press: New York, 2006.
2 Karopka T, Fluck J, Mevissen HT, Glass A. The autoimmune disease database: a dynamically compiled literature-derived database. *BMC Bioinform* 2006; **7**: 325.
3 Jacobson DL, Gange SJ, Rose NR, Graham NM. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. *Clin Immunol Immunopathol* 1997; **84**: 223–243.
4 Marrack P, Kappler J, Kotzin BL. Autoimmune disease: why and where it occurs. *Nat Med* 2001; **7**: 899–905.
5 Salaman MR. A two-step hypothesis for the appearance of autoimmune disease. *Autoimmunity* 2003; **36**: 57–61.
6 Eisenbarth GS. *Type 1 Diabetes: Molecular, Cellular and Clinical Immunology*. Springer: Berlin, 2004.
7 Lahita RG. *Lupus: Systemic Erythematosus*. Academic Press: New York, 2003.
8 Prantera C, Korelitz BI. Crohn's disease. *Informa Health Care* 1996; **8**: 198–199.
9 Ó'Moráin CA. *Ulcerative colitis*. CRC Press: Boca Raton, FL, 1991.
10 Earl J, Brewer EHG, Donald A. *Person: Juvenile Rheumatoid Arthritis*. Saunders: Boston, MA, 1982.
11 Waxman SG. Demyelinating diseases—new pathological insights, new therapeutic targets. *N Engl J Med* 1998; **338**: 323–325.
12 Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG. Multiple sclerosis. *N Engl J Med* 2000; **343**: 938–952.
13 Compston A, Coles A. Multiple sclerosis. *Lancet* 2002; **359**: 1221–1231.
14 Nistala K, Wedderburn LR. Th17 and regulatory T cells: rebalancing pro- and anti-inflammatory forces in autoimmune arthritis. *Rheumatology (Oxford)* 2009; **48**: 602–606.
15 Tsokos GC. Systemic lupus erythematosus. *N Engl J Med* 2011; **365**: 2110–2121.
16 Crispin JC, Liossis SN, Kis-Toth K, Lieberman LA, Kyttaris VC, Juang YT *et al.* Pathogenesis of human systemic lupus erythematosus: recent advances. *Trends Mol Med* 2010; **16**: 47–57.
17 Chan OT, Hannum LG, Haberman AM, Madaio MP, Shlomchik MJ. A novel mouse with B cells but lacking serum antibody reveals an antibody-independent role for B cells in murine lupus. *J Exp Med* 1999; **189**: 1639–1648.
18 MacDonald TT, Murch SH. Aetiology and pathogenesis of chronic inflammatory bowel disease. *Baillieres Clin Gastroenterol* 1994; **8**: 1–34.
19 Lehuen A, Diana J, Zaccone P, Cooke A. Immune cell crosstalk in type 1 diabetes. *Nat Rev Immunol* 2010; **10**: 501–513.
20 Centola M, Frank MB, Bolstad AI, Alex P, Szanto A, Zeher M *et al.* Genome-scale assessment of molecular pathology in systemic autoimmune diseases using microarray technology: a potential breakthrough diagnostic and individualized therapy-design tool. *Scand J Immunol* 2006; **64**: 236–242.
21 Achiron A, Gurevich M, Friedman N, Kaminski N, Mandel M. Blood transcriptional signatures of multiple sclerosis: unique gene expression of disease activity. *Ann Neurol* 2004; **55**: 410–417.
22 Achiron A, Gurevich M, Snir Y, Segal E, Mandel M. Zinc-ion binding and cytokine activity regulation pathways predicts outcome in relapsing-remitting multiple sclerosis. *Clin Exp Immunol* 2007; **149**: 235–242.
23 Baranzini SE, Mousavi P, Rio J, Caillier SJ, Stillman A, Villoslada P *et al.* Transcription-based prediction of response to IFNbeta using supervised computational methods. *PLoS Biol* 2005; **3**: e2.
24 Gurevich M, Tuller T, Rubinstein U, Or-Bach R, Achiron A. Prediction of acute multiple sclerosis relapses by transcription levels of peripheral blood cells. *BMC Med Genom* 2009; **2**: 46.
25 Bomprezzi R, Ringner M, Kim S, Bittner ML, Khan J, Chen Y *et al.* Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Hum Mol Genet* 2003; **12**: 2191–2199.
26 Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA *et al.* Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci USA* 2003; **100**: 1896–1901.
27 Barnes MG, Aronow BJ, Luyrink LK, Moroldo MB, Pavlidis P, Passo MH *et al.* Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR + chemokine genes relate to course of arthritis. *Rheumatology (Oxford)* 2004; **43**: 973–979.
28 Mandel M, Gurevich M, Pauzner R, Kaminski N, Achiron A. Autoimmunity gene expression portrait: specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erythematosus. *Clin Exp Immunol* 2004; **138**: 164–170.
29 Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC. Gene expression in peripheral blood mononuclear cells from children with diabetes. *J Clin Endocrinol Metab* 2007; **92**: 3705–3711.
30 Chaussabel D, Pascual V, Banchereau J. Assessing the human immune system through blood transcriptomics. *BMC Biol* 2010; **8**: 84.
31 Compston A, Ebers G, Lassmann H, McDonald I, Matthews B, Wekerle H. *McAlpine's Multiple Sclerosis* 1998.
32 Jarvis JN, Dozmorov I, Jiang K, Frank MB, Szodoray P, Alex P *et al.* Novel approaches to gene expression analysis of active polyarticular juvenile rheumatoid arthritis. *Arthritis Res Ther* 2004; **6**: R15–R32.
33 Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc Natl Acad Sci USA* 2003; **100**: 2610–2615.
34 Aune TM, Maas K, Parker J, Moore JH, Olsen NJ. Profiles of gene expression in human autoimmune disease. *Cell Biochem Biophys* 2004; **40**: 81–96.
35 Burczynski ME, Peterson RL, Twine NC, Zuberek KA, Brodeur BJ, Casciotti L *et al.* Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J Mol Diagn* 2006; **8**: 51–61.

36 Bovin LF, Brynskov J, Hegedus L, Jess T, Nielsen CH, Bendtzen K. Gene expression profiling in autoimmune diseases: chronic inflammation or disease specific patterns? *Autoimmunity* 2007; **40**: 191–201.
37 Scherer A. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley: New York, 2009.
38 Tuller T, Atar S, Ruppin E, Gurevich M, Achiron A. Global map of physical interactions among differentially expressed genes in multiple sclerosis relapses and remissions. *Hum Mol Genet* 2011; **20**: 3606–3619.
39 Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell* New York, 2002.
40 Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007; **3**: 140.
41 Rutherford A. *Introducing to ANOVA and ANCOVA a GLM Approach*. SAGE Publication Inc: Thousand Oaks, CA, 2001.
42 Fernandez EJ, Lolis E. Structure, function, and inhibition of chemokines. *Annu Rev Pharmacol Toxicol* 2002; **42**: 469–499.
43 Rottman JB. Key role of chemokines and chemokine receptors in inflammation, immunity, neoplasia, and infectious disease. *Vet Pathol* 1999; **36**: 357–367.
44 Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A *et al.* A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res* 2010; **20**: 1352–1360.
45 Wulczyn FG, Naumann M, Scheidereit C. Candidate proto-oncogene bcl-3 encodes a subunit-specific inhibitor of transcription factor NF-kappa B. *Nature* 1992; **358**: 597–599.
46 Gilmore TD. Introduction to NF-kappaB: players, pathways, perspectives. *Oncogene* 2006; **25**: 6680–6684.
47 Eggleton P, Harries LW, Alberigo G, Wordsworth P, Viner N, Haigh R *et al.* Changes in apoptotic gene expression in lymphocytes from rheumatoid arthritis and systemic lupus erythematosus patients compared with healthy lymphocytes. *J Clin Immunol* 2010; **30**: 649–658.
48 Achiron A, Feldman A, Mandel M, Gurevich M. Impaired expression of peripheral blood apoptotic-related gene transcripts in acute multiple sclerosis relapse. *Ann N Y Acad Sci* 2007; **1107**: 155–167.
49 Nathan C. Points of control in inflammation. *Nature* 2002; **420**: 846–852.
50 Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N *et al.* A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* 2008; **29**: 150–164.
51 Arasappan D, Tong W, Mummaneni P, Fang H, Amur S. Meta-analysis of microarray data using a pathway-based approach identifies a 37-gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells. *BMC Med* 2011; **9**: 65.
52 Bossis G, Malnou CE, Farras R, Andermarcher E, Hipskind R, Rodriguez M *et al.* Down-regulation of c-Fos/c-Jun AP-1 dimer activity by sumoylation. *Mol Cell Biol* 2005; **25**: 6964–6979.
53 Camargo JF, Correa PA, Castiblanco J, Anaya JM. Interleukin-1beta polymorphisms in Colombian patients with autoimmune rheumatic diseases. *Genes Immun* 2004; **5**: 609–614.
54 Aradhya S, Nelson DL. NF-kappaB signaling and human disease. *Curr Opin Genet Dev* 2001; **11**: 300–306.
55 Pena AS, Penate M. Genetic susceptibility and regulation of inflammation in Crohn's disease. Relationship with the innate immune system. *Rev Esp Enferm Dig* 2002; **94**: 351–360.
56 Yamamoto Y, Gaynor RB. Therapeutic potential of inhibition of the NF-kappaB pathway in the treatment of inflammation and cancer. *J Clin Invest* 2001; **107**: 135–142.
57 Gabay C. Interleukin-6 and chronic inflammation. *Arthritis Res Ther* 2006; **8**(Suppl 2): S3.
58 Kristiansen OP, Mandrup-Poulsen T. Interleukin-6 and diabetes: the good, the bad, or the indifferent? *Diabetes* 2005; **54**(Suppl 2): S114–S124.
59 Tackey E, Lipsky PE, Illei GG. Rationale for interleukin-6 blockade in systemic lupus erythematosus. *Lupus* 2004; **13**: 339–343.
60 Nishimoto N. Interleukin-6 in rheumatoid arthritis. *Curr Opin Rheumatol* 2006; **18**: 277–281.
61 Woods JM, Katschke KJ, Volin MV, Ruth JH, Woodruff DC, Amin MA *et al.* IL-4 adenoviral gene therapy reduces inflammation, proinflammatory cytokines, vascularization, and bony destruction in rat adjuvant-induced arthritis. *J Immunol* 2001; **166**: 1214–1222.
62 Wurster AL, Rodgers VL, White MF, Rothstein TL, Grusby MJ. Interleukin-4-mediated protection of primary B cells from apoptosis through Stat6-dependent up-regulation of Bcl-xL. *J Biol Chem* 2002; **277**: 27169–27175.
63 Or R, Renz H, Terada N, Gelfand EW. IL-4 and IL-2 promote human T-cell proliferation through symmetrical but independent pathways. *Clin Immunol Immunopathol* 1992; **64**: 210–217.
64 Besser M, Wank R. Cutting edge: clonally restricted production of the neurotrophins brain-derived neurotrophic factor and neurotrophin-3 mRNA by human immune cells and Th1/Th2-polarized expression of their receptors. *J Immunol* 1999; **162**: 6303–6306.
65 Mandel M, Achiron A, Tuller T, Barliya T, Rechavi G, Amariglio N *et al.* Clone clusters in autoreactive CD4 T-cell lines from probable multiple sclerosis patients form disease-characteristic signatures. *Immunology* 2009; **128**: 287–300.
66 Zhang N, Hartig H, Dzhagalov I, Draper D, He YW. The role of apoptosis in the development and function of T lymphocytes. *Cell Res* 2005; **15**: 749–769.
67 Eguchi K. Apoptosis in autoimmune diseases. *Intern Med* 2001; **40**: 275–284.
68 Baranzini SE. The genetics of autoimmune diseases: a networked perspective. *Curr Opin Immunol* 2009; **21**: 596–605.
69 McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD *et al.* Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. *Ann Neurol* 2001; **50**: 121–127.
70 Best WR, Becktel JM, Singleton JW, Kern Jr F. Development of a Crohn's disease activity index. National Cooperative Crohn's Disease Study. *Gastroenterology* 1976; **70**: 439–444.
71 Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. *N Engl J Med* 1987; **317**: 1625–1629.
72 Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham III CO *et al.* Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis* 2010; **69**: 1580–1588.
73 Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 2005; **437**: 1173–1178.
74 Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005; **122**: 957–968.
75 Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003; **13**: 2363–2371.
76 Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002; **30**: 303–305.
77 Tuller T, Kupiec M, Ruppin E. Determinants of protein abundance and translation efficiency in *S. cerevisiae*. *PLoS Comput Biol* 2007; **3**: e248.
78 Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* 2010; **6**: 400.
79 Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003; **4**: P3.
80 Storey JD. A direct approach to false discovery rates. *J R Statist Soc* 2002; **64**: 479–498.
81 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc* 1995; **57**: 289–300.
82 Shao J, Tu D. *The Jackknife and Bootstrap*. Springer: Berlin, 1995.
83 Pedhazur EJ. *Multiple Regression in Behavioral Research*. Wadsworth Publishing: Toronto, ON, Canada, 1997.

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