Published online 28 June 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 18* ***7881–7899***
doi:10.1093/nar/gkr503

Internal standard-based analysis of microarray data2—Analysis of functional associations between HVE-genes

Igor M. Dozmorov[1,*], James Jarvis[2], Ricardo Saban[2], Doris M. Benbrook[2], Edward Wakeland[3], Ivona Aksentijevich[4], John Ryan[4], Nicholas Chiorazzi[5,6,7], Joel M. Guthridge[1], Elizabeth Drewe[8], Patrick J. Tighe[8], Michael Centola[1] and Ivan Lefkovits[9]

[1]Oklahoma Medical Research Foundation, [2]Oklahoma University Health Science Center HSC, Oklahoma City, OK 73104, [3]The University of Texas Southwestern Medical Center, Dallas, Texas 75390, [4]National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, Maryland 20892, [5]The Feinstein Institute for Medical Research, [6]The Departments of Medicine and of Cell Biology, North Shore University Hospital, [7]Albert Einstein College of Medicine, Manhasset, NY, USA, [8]University of Nottingham, Nottingham, UK and [9]Department of Biomedicine, University Clinics Basel, Vesalianum, Vesalgasse 1, CH-4051 Basel, Switzerland

Received January 25, 2011; Revised May 12, 2011; Accepted June 1, 2011

ABSTRACT

In this work we apply the Internal Standard-based analytical approach that we described in an earlier communication and here we demonstrate experimental results on functional associations among the hypervariably-expressed genes (HVE-genes). Our working assumption was that those genetic components, which initiate the disease, involve HVE-genes for which the level of expression is undistinguishable among healthy individuals and individuals with pathology. We show that analysis of the functional associations of the HVE-genes is indeed suitable to revealing disease-specific differences. We show also that another possible exploit of HVE-genes for characterization of pathological alterations is by using multivariate classification methods. This in turn offers important clues on naturally occurring dynamic processes in the organism and is further used for dynamic discrimination of groups of compared samples. We conclude that our approach can uncover principally new collective differences that cannot be discerned by individual gene analysis.

INTRODUCTION

The microarray technology has revolutionized the study of biology by allowing for simultaneous examination of thousands of genes—the total genome expression profile. However, the most exciting prospect is to characterize the organism as a whole by defining the functional associations among their genes. It turns out that it is not possible to visualize genetic associations in a steady state. To understand the dynamic features of interest, the underlying system must be stimulated to elucidate the features of the biological regulatory networks. A common practice in experimental biology has been to make single, stepwise changes in one variable at a time and to follow the system's response as it proceeds from an initial steady state to a final steady state.

Although such changes lead to results that are interpretable from a biochemical point of view, step changes do not persistently excite the network since most of the data will be biased because of approaching the new steady state. As a result, many dynamic features remain unidentified, even with extensive prior knowledge. Capturing the multivariate nature of biological regulatory networks requires the introduction of multivariate random perturbations, especially when the underlying data contain high levels of noise. As it was shown earlier (1), random, independent inputs enable better identification of relevant results, and such identification is more robust to noise.

In most biological systems, random stimulations from the environment continue throughout the life span of the organism, and the organism persistently reacts in turn to such random stimulations. Genes participating in this reaction are in dynamic states. Thus, it is possible to reveal genes displaying an extraordinarily high variability of expression, and we call these genes 'hypervariably expressed genes' or *HVE-genes*. It has been shown that

*To whom correspondence should be addressed. Tel: +1 405 271 7052; Fax: +1 405 271 4002; Email: igor-dozmorov@omrf.org

even in genetically identical individuals; tissues display a considerable degree of variation in gene expression (2). There are multiple reasons for the extreme variability of such genes. For example, previously unrecognized heterogeneities could be present in the presumably homogenous group of samples, or there may be genes that are involved throughout different phases of internal dynamic processes.

Genetic diseases are often associated with the manifestation of profound genetic variations. Hence, under such conditions increased variability of some genes will be expected, although the association of these genetic variations with transcriptional changes cannot be directly inferred. Genes that demonstrate variability in expression at the population level could be potential candidates for further studies of the genetic architecture of complex traits associated with pathology, especially if these gens display intra-individual stability. In this context, it is interesting to note that gene expression variability is often increased in autoimmune pathologies and is normalized again after successful treatment [see e.g. (3–5)].

Examples of significant increases of the proportion of HVE-genes in various inflammatory pathologies include lupus, rheumatoid arthritis and TNF Receptor Associated Periodic Syndrome (TRAPS). Because TRAPS is a rare autoinflammatory disorder caused by mutations in the extracellular domain of the TNF receptor superfamily 1A, one does expect to observe differences in gene expression variability when comparing TRAPS patients with healthy donors. Indeed when comparing 14 TRAPS patients with a counterpart of 14 healthy donors, 124 genes displayed increased expression variability in the samples from TRAPS patients (Figure 2A). Many of these genes are members of the TNF receptor pathway and are associated with inflammatory processes (as shown by the Ingenuity Pathway Analysis presented in Supplementary Figure S1). It is of interest that among the outlined entities, Mediterranean fever gene (MEFV) is present—a hallmark of another close to TRAPS pathology—Mediterranean fever (6).

The most prominent problem in studying HVE-genes is the lack of statistical methods to facilitate the selection of HVE-genes from microarray experiments in which sample sizes are too small to use standard statistical techniques. Variable gene expression can be a characteristic feature of pathology, but the lack of adequate methods for multivariate analysis complicates the interpretation of the obtained results, especially regarding the reproducibility and reliability of the established features (7,8). The reasons behind these objections include the instability of existing methods and sample sizes that are too small to support the notion of reliable variability features.

We demonstrated earlier (9), that many problems of genome-scale microarray experiments, which appeared to be consequences of the vast amount of information, were successfully resolved by the use of the Internal Standard strategy. In this method information about nonspecific variations is dissociated from the conventional behavior of genes that share certain features, such as equity in expression, stability and distinctiveness from background noise. Knowledge of the parameters governed by Internal Standards is an added benefit to statistically robust analyses of functional associations by clustering and networking genes.

In this communication, we present the application of the Internal Standard strategy to HVE-gene selection and a functional analysis based on strong statistical criteria. Rather than presenting an orderly, methodological approach, we assembled data obtained throughout several research endeavors, and we present the actual results from applying multivariate procedures to the analysis of HVE-genes in both normal and pathological processes.

Programs created for the selection and analyses of the features of the HVE-genes are implemented in MatLab (Mathworks, MA, USA) and available from authors upon request.

MATERIALS AND METHODS

Gene expression data sets

This work uses a wide spectrum of experimental data. The actual biological portion of the experiments was performed in a collaborative manner separately for each sub-project, and portions of them have already been reported in independent publications or are in preparation for publications. The common denominator of each of these projects is the evaluation procedure. Expression data sets were obtained using various sources of mRNA and several microarray technologies. Fragmented descriptions of the experimental protocols and the microarray experiments are given in Table 1 and in the Supplementary Data. The reason for compiling multiple diverse biological experiments into a single paper is to allow the output microarray data from these experiments to be analyzed using the Internal Standard-based analysis procedure.

Microarray data analysis

The methods used for gene expression analysis are based on the use of Internal Standards, which are constructed by identifying a large family of similarly behaving genes. The application of these Internal Standards to the normalization of microarray data and the differential analysis of gene expression was presented in the first part of this project (9).

The normalization procedure consists of two subsequent steps:

- The first step is the determination of the parameters of the background of the array—the average (Av) and standard deviation (SD) of normally distributed low level expressions in an array with subsequent normalization of all expressions in the array. A normalized score, 'S,' is obtained [$S = (PV - Av)/SD$], where PV is the original pixel value for the spot, and Av and SD are the mean and standard deviation respectively, of the set of background spots. The distribution of S has zero mean and $SD = 1$ over the set of background genes in the normalized array. Only genes expressed

Table 1. Information about projects used in the article

No.	Project	Investigators—primary owners of the data	mRNA source	Microarray platform
1	JRA	J Jarvis, OUHSC, OK	Figure 2D. Peripheral blood of Patients 3–15 years (21 samples) and healthy control donors (19 samples)	Human WG-6 v3.0 beadchip (Illumina, San Diego, CA, USA)
			Figure 8. Peripheral blood of Patients 3–15 years (15 samples) and healthy control donors (12 samples)	Micromax cDNA arrays, Perkin Elmer Life Sci., Boston, MA, USA
2	Chronic Lymphocyte Leukemia (CLL)	N Chiorazzi, Feinstein Inst. Med. Res., NY	Figure 2C. B cells from peripheral blood of CLL patients (20-with mutated IGHV, and 16 with unmutated IGHV) and of 18 healthy control donors	Human WG-6 v3.0 beadchip (Illumina, San Diego, CA, USA)
3	TRAPS	I Aksentijevich, J Ryan, NIAMSD, Bethesda, MD	Figure 2A. Peripheral blood of TRAPS patients (14 samples) and healthy control donors (14 samples)	GeneChip Human Genome U133 Plus 2.0 Array (Affymetrix, Santa Clara, CA, USA).
4	TRAPS	E Drew, PJ Tighe, Univ. Nottingham, UK	Figure 10. Peripheral blood of TRAPS patients (33 samples) and healthy control donors (11 samples)	Human oligonucleotide microarrays (Qiagen #810516, Human Genome Oligo Set V2 Search). Containing 21 329 human genes. List of genes: http://omrf.ouhsc.edu/~frank/human-library.txt
5	SLE	J Guthridge, OMRF, OK	Figure 9. EBV-transformed cell lines form two SLE patients and two healthy control donors. Time course (0, 0.5, 1, 2, 4, 8, 16, 24 h) of the response of B cell lines to stimulation with anti-human IgM $F(ab)'_2$ antibodies. Sixty-four samples altogether including duplicated serum controls (no stimulation).	Human Focus Array, Affymetrix, Santa Clara, CA, USA). The chip contains 8793 genes.
6	Mouse bladder gene regulation	R Saban, OUHSC,OK	Figure 7. Bladder tissues from Neurokinin 1 receptor knockout mice and C57BL/6J mice as controls. Time course: 0, 1, 4, 24 h following stimulation with antigen (DNP_4-human serum albumin) or saline.	Mouse 1.2 Arrays (catalog no. 7853-1; Clontech, Palo Alto, CA, USA) containing 1177 mouse genes. List of genes: http://www.clontech. com/atlas/genelists/index.html.
7	T cells from BALB/c mice	M Centola, OMRF, OK	Spleen T cells from 10 BALB/c female mice	Mouse microarrays were produced at the OMRF core facility using a commercially available library of 70 bp long DNA oligos (70-mers, Qiagen/Operon Technologies). List of genes: (http://www.ncbi.nlm.nih.gov/UniGene/).
8	Endometrial Cancer (EC)	D Benbrook, OUHSC, OK	Figure 2B. Cells for cultures collected from a healthy premenopausal Female. Endometrial organotypic cultures were exposed to DMBA (to induce DNA damage) or solvent control. There were four-replicates in each group	GeneChip Human Genome U133 Plus 2.0 Array (Affymetrix, Santa Clara, CA, USA).
9	SLE- mouse models	E Wakeland, UT Southwestern Med. Center, Dallas, TX	Figure 2S. CD220 B cells and $CD4^+$ T cells from B6, B6.Sle1 and B6.Sle1Sles1 8-week-old mice	Mouse WG-6 v1.1 beadchip (Illumina, San Diego, CA, USA)

above background (>3 SDs) are used for the second step 'adjustment'.
- The second step is the adjustment of the normalized profiles to each other by robust regression analysis of genes expressed above background. This procedure is based on the selection of equally expressed genes as a homogenous family of genes, with normally distributed residuals defined as deviations from the regression line. Outliers are thereafter determined as genes having deviations not associated with this internal standard of equity in expression, which include thousands of members.
- For multi-sample data adjustment an averaged profile is calculated and each sample is adjusted to the averaged profile using the robust regression procedure described above. A new averaged profile is calculated from transformed profiles of the samples and the adjustment procedure is repeated. Several subsequent adjustment may be necessary for the best result, however for the data initially normalized to background two steps of adjustment are usually sufficient.

One of the most important criteria in the selection of HVE-genes and the analysis of their behavior is the choice of the 'Reference Group'—which is an Internal Standard for equity in expression and for stability of the analyzed processes (absence of variability exceeding technological and biological noise).

Procedure for establishing the 'Reference Group'

The Reference Group is constructed by identifying a set of genes expressed above background level with inherently low variability as determined by an *F*-test. The procedure consists of two steps; the first step ensures that an absolute majority of stable genes are identified, while the second step ensures that the outliers are excluded with a simple iterative procedure. At the beginning, all genes are represented by their residuals (relatively averaged profile), which after normalization and log transformation lose their sample-dependent individuality as well as their expression level-dependent individuality (Figure 1A). For the majority of genes, the variation between replicates is relatively small and homogenous and follows the standard *F*-distribution. A small portion of genes that exhibit high variation (statistically distinct from the rest) are the HVE-genes. To obtain the Internal Standard for gene variability, HVE-genes should be excluded by an iterative procedure (9). The *F*-test is used as the criterion for the exclusion of outliers, i.e. genes that exhibit an estimated variability that is considerably higher than that that of the total group. The total group variability is recalculated after each exclusion step, and the procedure is repeated until no additional genes can be excluded by this procedure. The statistical threshold for the exclusion of HVE-genes is chosen such that these exclusions are based on an exceptional *P*-value (usually $P < 0.05$). The completion of all the exclusion process a new Internal Standard called the 'Reference Group', which is composed of genes expressed above the background of control samples with a low variability of expression (as determined by an *F*-test) and whose residuals approximate a normal distribution. Though not all excluded genes are HVE-genes, we can be sure that the majority of them are excluded and will not interfere with the estimation of parameters for the rest of the analysis. The Reference Group is further used for selection of HVE-genes and for analysis of their functional associations in clustering and networking procedures.

List of four résumés of calculations steps

Upon providing in the 'Result' section detailed explanations and arguments about the chosen path of calculations, procedures summarizing the calculation steps are presented in four sequential step-by-step résumés.

Step-by-step Résumé 1: Associative analysis of differences in gene expression variations.
Step-by-step Résumé 2: *F*-means cluster analysis of HVE-genes co-expression.
Step-by-step Résumé 3: Correlation mosaic analysis of HVE-genes co-expression.
Step-by-step Résumé 4: Networking procedure based on the use of partial correlations.

RESULTS

All of the experiments described in this communication were analyzed using the Internal Standard approach, which has been described in our earlier paper (9), in combination with other methods.

Selection of 'hypervariably expressed genes'

Upon establishing the Internal Standard of biological stability (Figure 1A) the selection of HVE-genes was made using strict statistical criteria. HVE genes were identified as those for which the expression level varied significantly ($P < P\text{o}$) when comparing the variability of individual genes to the variability of the 'Reference Group'. The threshold *P*o was chosen either in a restricted manner ($P\text{o} < 1/N$, where N is the number of all genes expressed significantly differently from background noise) or in a moderate manner ($P\text{o} < 0.05$), depending on the purpose of the subsequent analysis. Choosing the threshold as $P\text{o} < 1/N$ (N was often more than half of all genes on the array) can be considered to be a slight modification of the Bonferroni correction for multiple hypothesis tests. Such a choice excludes virtually all false positives, but consequently loses many true positives as well. This choice should be made when selecting HVE-genes that are unique to any given group. In situations in which the traditional $P = 0.05$ is applied, many false positives will be retained. Nevertheless, this choice can be useful when studying HVE-genes that reproducibly appear in several groups, cluster together or reproducibly interconnect in a subsequent networking procedure. All of these subsequent steps refine the list of HVE genes to only those that demonstrate some reproducible features that are probabilistically less likely to be present in false selections.

Hyper-variations appearing from experimental errors (the influence of dirty spots) were statistically filtered

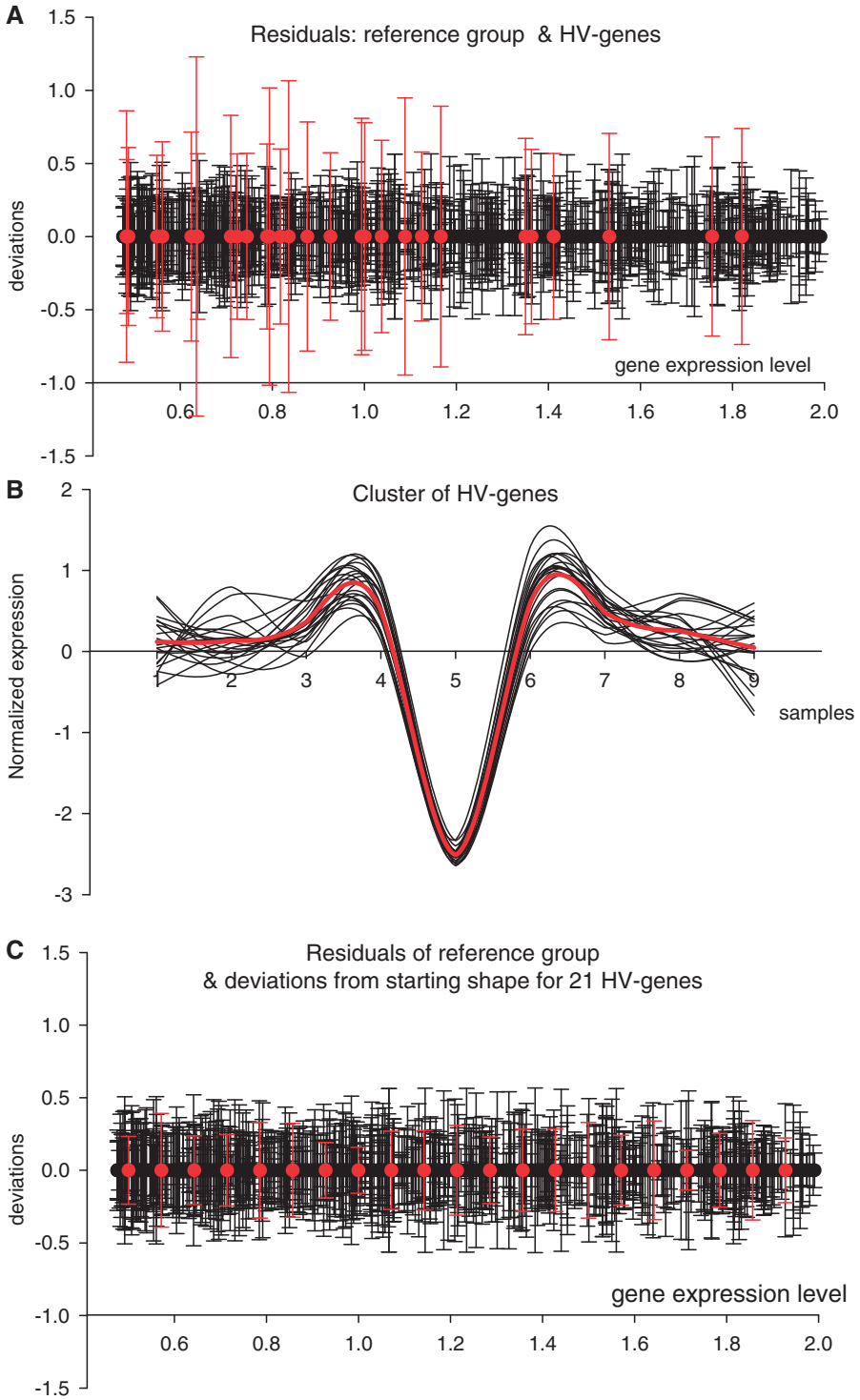

Figure 1. *F*-means clustering procedure. (**A**) The standard deviations of genes from the Reference Group, with HVE-genes (red bars) included. (**B**) Gene content of the cluster with seeding profile shown as a red line. (**C**) Deviations of genes' profiles from the seeding profile (shown as red SD bars) do not exceed the ranges of normal expression noise (gray-Reference Group). Abscissa: (A) and (C). The normalized gene expression level (log10 presentation), (B) The sample numbers. Ordinate: (A) and (B) Gene expression deviations from the equity of expression; (B) Gene expression levels in samples normalized to have zero mean (over all samples) and SD = 1.

from this analysis by comparing the variability of the residuals in a replicated group of samples with the same variability obtained by excluding both the maximum and minimum one at a time. A statistically significant decrease in variability after excluding one replicate provides evidence of a possible error in that particular replicate. Such genes are excluded from the family of HVE-genes as being falsely selected.

Increased gene expression variability associated with pathologies

In replicated microarray experiments, each gene in the array can be characterized by two independent parameters: the level of expression and the variability (except in regions of low-intensity spots that are abundantly contaminated with highly variable background noise). In addition to the conventional comparison of gene expression levels, it is possible to compare their variability using strict statistical criteria. The conventional statistical method for comparison of variability, ANOVA, encounters the same obstacles when applied to the analysis of microarray experiments containing immense amount of information. The conventional low statistical threshold ($P < 0.05$) will produce a large output of false positive selections, whereas any profound adjustments of this threshold will result in the loss of sensitivity of the statistical test. The practice of using the Internal Standard resolves this problem with the same efficiency as was achieved for differential gene expression analysis (9).

Selecting genes with different variabilities relies on the next statistical steps. First, the *F*-test was used to identify HVE-genes in each group of samples. Next, the differences in their variability were determined in a paired comparison.

Résumé 1: Differential analysis of gene expression variability. Two groups are considered: Group 1 has n chips and k genes, while Group 2 has m chips and k genes.

Data is first normalized as described in the 'Materials and Methods' section and presented in log-transformed form, making the variability of the majority of genes independent of the level of their expression.

- Reference groups are created for each group of samples (Groups1 and 2) and HVE-genes are selected in each group as previously described. (Associative *F*-tests, with $m+k-2$ degrees of freedom ($a = \frac{1}{k}$), to establish if the gene associates/belongs to the group of stably expressed genes).
- A paired *F*-test is performed on the genes selected as HVE-genes in both groups (Groups 1 and 2, comparison of the SDs for the same gene in two groups—with $n+m-2$ degrees of freedom and threshold corrected for the multiple hypothesis tests), to determine whether the genes have equal SDs.
- Additional restrictions on the fold change and the minimal average level of expression may applied. The data are grouped into five sets:

  B0: HVE-genes without differences in variability in the case-control comparison
  B1: HVE-genes having significantly higher variation in the Experimental group
  B2: HVE-genes having significantly higher variation in the Control group
  B3: Genes that exhibit the HVE property only in the Experimental group
  B4: Genes that exhibit the HVE property only in the Control group

The ratio of SDs for HVE-genes in groups B1 and B2 was used to exclude changes that are statistically significant but are not biologically significant. The fold change restriction was usually applied as an addition to the statistical analysis to draw attention to the most prominent differences. Upon excluding Bo, all other groups (B1 – B4) contain genes that exhibit some characteristic differences in the variability of expression level when comparing 'experimental versus control'. These genes also establish a pathology-specific fingerprint. Unique variable genes from the B3 group are of special importance in addressing questions about dynamic processes associated with any given pathology.

To understand the mechanisms behind a disease, one should first attempt to establish whether disease-specific differences in gene variability are the consequence or the cause of the pathology. The superfluous variability of normally stable genes as well as the 'freezing' of genes predicted to participate in dynamically adaptive reactions could provide clues towards the understanding of the pathology.

Increased variability can also be of a non-genetic, physiological nature; and one might expect that many pathologies, such as inflammation, that are associated with a burst of dynamic changes are also accompanied with a considerable increase in the portion of genes that display high variability.

Examples of significant increases in the proportion of HVE-genes in various inflammatory pathologies include lupus, rheumatoid arthritis and TRAPS. Because TRAPS is a rare autoinflammatory disorder caused by mutations in the extracellular domain of TNF receptor superfamily 1A, differences in gene expression variability are expected when comparing TRAPS patients with healthy donors. Indeed, when comparing 14 TRAPS patients with a counterpart of 14 healthy donors, 124 genes were found to display increased expression variability in the samples from TRAPS patients (Figure 2A). Many of these genes are members of the TNF receptor pathway and are associated with inflammatory processes (as shown by the Ingenuity Pathway Analysis presented in Supplementary Figure S1). It is of interest that Mediterranean fever gene (MEFV) is present among the outlined entities. This gene is associated with Mediterranean fever, a disease with similar pathology to TRAPS (6).

Increased variability may be associated with the development of pathology. Figure 2B presents the appearance of uniquely variable genes in the course of the transformation of endometrial cells into cancer cells by the action of the carcinogen DMBA (7,12-dimethylbenz[a]anthracene) (10).

Increased variability may also be observed in pathologies that are less dynamic than inflammatory conditions, for example, chronic pathologies that are not associated with a burst of dynamic changes. Figure 2C presents genes that demonstrate stable expression levels in B cells from normal healthy donors and extreme variations in samples from patients with B cell chronic lymphocytic leukemia (non-mutated and mutated subgroups) (11).

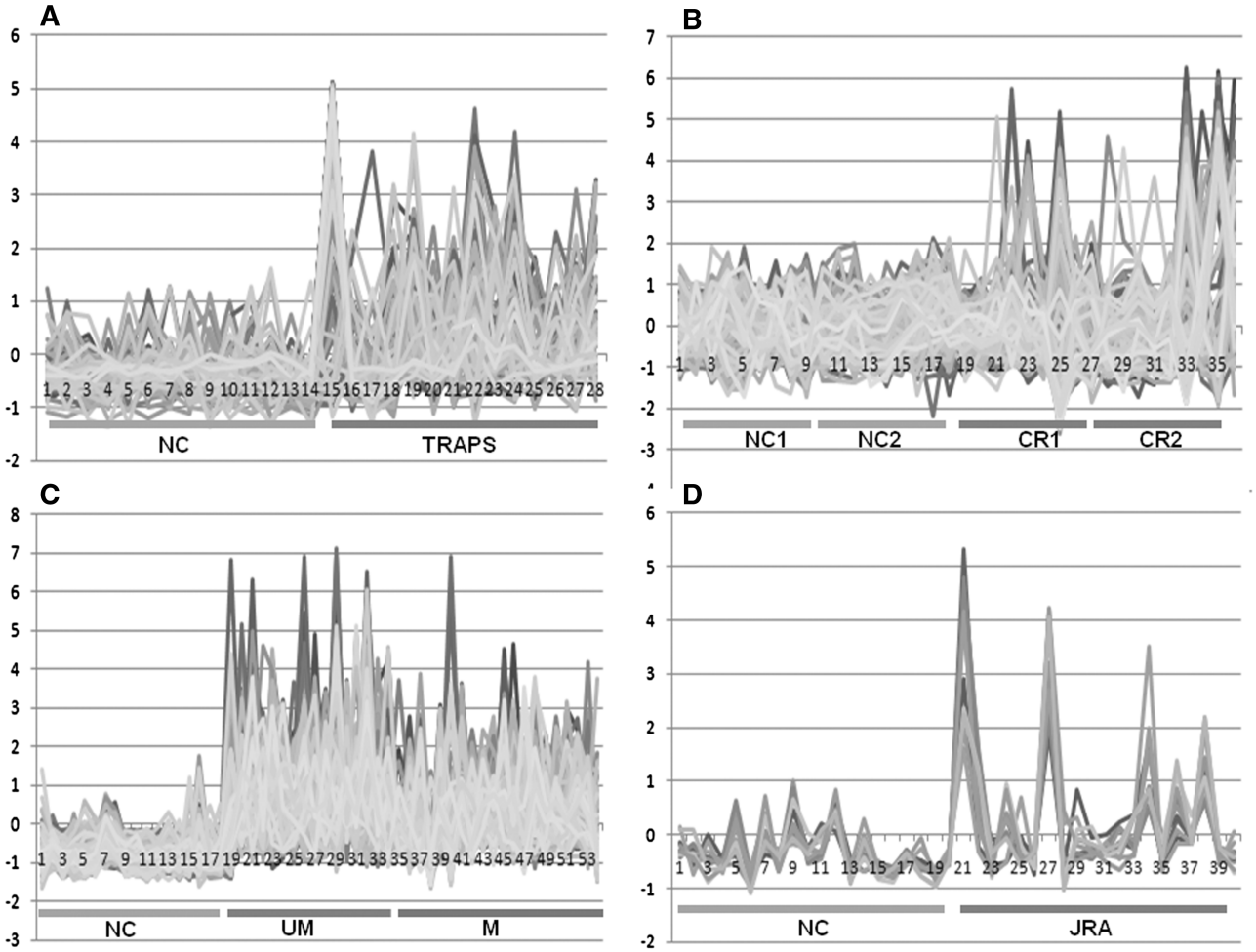

Figure 2. Increase in gene variability associated with different pathologies. Expression data normalized to make the overall Average = 0, SD = 1. Abscissa: the sample numbers. Ordinate: the normalized expression level. mRNA for the transcription study was obtained from various samples: (**A**) Samples from healthy controls (1–14) and TRAPS patients (15–28). (**B**) Endometrial cells: controls (1–9 and 10–18) and cells transformed to cancer cells by DMBA (19–27 and 28–36). The results of two independent experiments are presented. (**C**) Samples from the B cells of healthy donors (1–18), and B cell chronic lymphocytic leukemia patients: (19–34) un-mutated, and (35–54) mutated subgroups. (**D**) Whole blood samples from healthy donors (1–20) and JRA patients (21–40).

The seemingly chaotic behavior of gene expression variation in various pathologies could in fact be a result of the superposition of several co-expressed groups of genes. An example of this phenomenon is presented in Figure 2D, where a group of variable genes in Juvenile Rheumatoid Arthritis (JRA) patients reveal closely related co-expression patterns.

The set of genes that are uniquely expressed in any given pathology is referred to as the 'fingerprint' or 'signature' of the particular pathology (12). We extend this definition to refer to the set of uniquely variable genes and coin the expression 'functional fingerprint'.

An interesting example of a 'functional fingerprint' in autoimmune pathologies was obtained using lupus prone mice. We compared mice with the Sle1 mutation, which makes them susceptible to the development of lupus-like pathology, with mice possessing an additional Sles1 mutation that in turn cancels the effect of the first Sle1 mutation (13–15). We found that in $B220^{+}$ cells, 35 genes that were stable in healthy animals, became variable in B6Sle1 mice and again reverted into stable form in B6Sle1Sles1 mice (Supplementary Figure S2). In $CD4^{+}$ cells, changes in variabilities of 150 genes was associated with the Sle1 mutation.

F-means clustering for inferring functional interconnections

There are diseases in which differences in HVE-genes occur at particular stages of disease manifestation, while no distinctive differences are evident at the onset. The only means of revealing pathology-specific differences is through the analysis of functional associations for such HVE-genes. The most commonly used computational approach to analyzing such functional associations is cluster analysis.

F-means cluster analysis of HVE-genes is an unsupervised method, in which every decision, including the selection of variable genes, the search for the optimal number of clusters, as well as optimization of the distribution of

genes over clusters, is solved using statistical criteria. If we know the precise differences in the gene expression levels among the samples, we would have a 'true' clustering. The residuals from the Reference Group provide an empirical estimate of the error of the distribution, or the 'noise' in the data.

F-means clustering of HVE-genes was initiated by defining a parameter called the connectivity, which is defined as the number of genes that vary in expression in a similar manner as the 'seed' gene. Clusters then were nucleated starting with genes of highest connectivity. Genes of lower connectivity were included in a given cluster if their expression levels deviated from the seeding profile without exceeding the variation of the residuals in the Reference Group based upon an *F*-test (Figure 1B and C). The number of different clusters was determined by the experimental system's ability to distinguish differences exceeding random fluctuations of the normalized residuals in the Reference Group.

Résumé 2: F-*means cluster analysis of the coexpression of HVE-genes*. The clustering procedure consists of the following steps:

- Gene expression normalization, log-transformation and rescaling as noted above.
- Selection of HVE-genes. Exclusion some of them whose extreme variability was produced by the deviation from stable state in only one sample to minimize the influence of technical errors.

Determination of the connectivity, for each of these HVE-genes. Connectivity is defined as the number of genes whose expression patterns does not vary from the expression pattern of a given gene within the ranges derived from the Reference Group (based on the *F*-test). The appropriate correction of threshold for the *F*-test should be used to diminish the proportion of false positive selections ($Po < 1/N$, N- number of HVE-genes).

HVE-genes for each group are sorted by their connectivity and the clustering process begins with the genes exhibiting the highest connectivity. The first cluster contains the gene with the highest connectivity and all genes whose deviations from the expression of this gene in each sample have variabilities that do not exceed the variability of the Reference Group. The next gene of higher connectivity not belonging to the first cluster acts as the starting point for Cluster #2, and other genes are included in this cluster using the same criteria as in the first cluster. This process continues until all genes are analyzed. Genes that appeared in more than one cluster are considered to be likely functional links among these clusters. Genes that have zero connectivity do not belong to any cluster. Additional restrictions on the choice of the thresholds for statistical tests and the minimal cluster content can be elicited from simulation experiments where the gene expression data are replaced with random data having the same characteristic parameters (average and standard deviation). The use of simulated data establishes the minimal cluster content that appears by chance at the chosen statistical thresholds.

Three potentially different results are distinguished:

- functional associations for genes from the B4 set are characteristic of dynamic processes that prevail under normal conditions and are absent in pathology;
- functional associations appear under pathological conditions only for genes from the B3 set, are uniquely variable in the pathological group and are stable in the normal control group
- functional associations for genes from the B0, B1 and B2 sets are significantly modulated in one of the compared groups (normal control or pathology).

Hypervariably expressed genes demonstrate similar patterns of variations

The co-expression of HVE-genes or similarities in their expression profiles are of particular importance to understanding the biological significance of these findings. The idea that co-expression of genes revealed by the clustering procedure implies the participation of these genes in general biological processes was first formulated by the group of Eisen (16). An extension of this idea is that the same should be true for HVE-genes, whose different level of expression can be considered as snapshots of some dynamical process. In contrast to temporal dynamics, the actual shape of the cluster in the case of HVE-genes is of lesser significance as shown in Figure 3. Even if HVE-gene expression in each sample is consistent with some phase of a dynamic process, the absence of information about the real sequence of events makes the shape of the profile useless.

Several practical examples demonstrate the consistent characteristics of the variation in the expression levels of the group of clustered genes. The first example was obtained from analysis of gene expression in T lymphocytes from a homogenous group of mice. Figure 4 demonstrates that dozens of genes with significantly high variations in their expression levels could be gathered in clusters. The very high content of these clusters excludes the possibility of chance variations.

Another example of co-expression of HVE-genes was obtained through analysis of gene expressions in samples from TRAPS patients (Figure 5). The majority of genes in the biggest clusters in samples from two entirely unrelated groups—healthy controls and TRAPS patients—had identical co-expression patterns. The largest clusters in the control group and in the group of TRAPS patients consist of 163 and 51 genes, respectively. We applied the same technique to *F*-means clustering in groups produced from controls and patients by substituting of real data with random values having the same averages and SD for each gene. The largest cluster obtained in this simulation procedure was 10 times smaller than the largest cluster in the actual control group, and no genes were found to cluster in the simulated patient group. Similar results were found when comparing the eight largest clusters obtained from the analysis of real and simulated data (Figure 6).

Another example was created earlier in the course of gene expression analysis in samples of children with

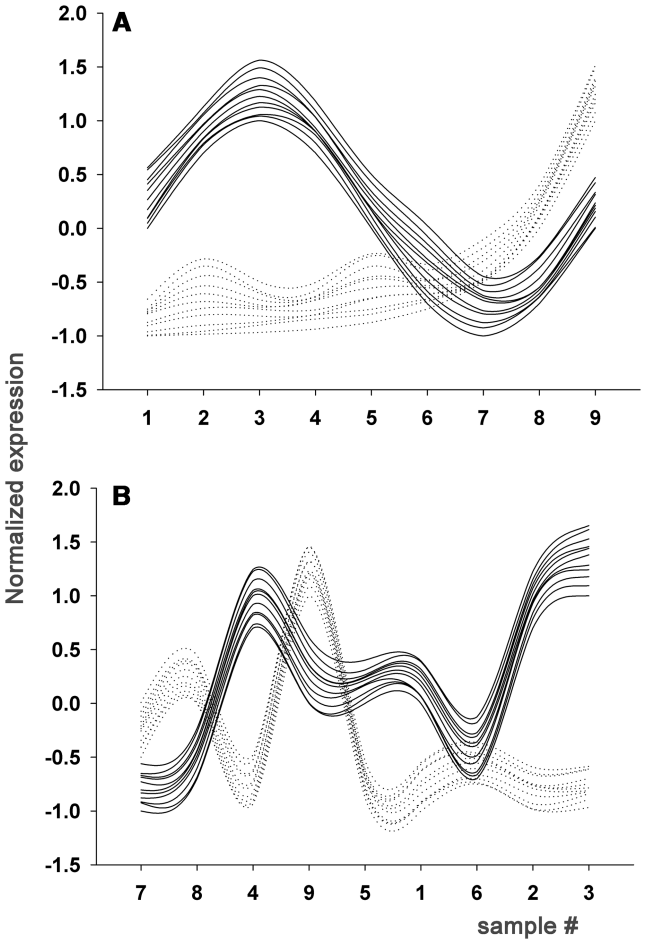

Figure 3. Shapes of the HVE gene expression profiles does not have sense. Diagrams illustrating the formation of the cluster profiles of HVE-genes in a homogeneous group. (**A**) Possible assortment of nine samples representing two dynamical processes with participation of several genes, each of whose profiles are shown in either red or black. (**B**) Variant of A in which the order of the samples is arbitrarily changed. The exact shape of the dynamical process is lost after such rearrangement, but the fact of gene co-expression is still evident.

polyarticular JRA and normal healthy controls (27 samples altogether) (17). In this work the sizes of the HVE-gene clusters also significantly exceeded the sizes of clusters identified in the simulation experiment. Additional validation of the biological meaningfulness of partitioning HVE-genes into clusters was obtained by analyzing of the cluster contents. The two biggest clusters consisted exclusively of genes encoding ribosomal proteins, while others consisted of genes encoding general regulatory proteins, such as insulin and NF-κB, and also of protein involved in mitochondrial protein synthesis, proteasome and mini-chromosome maintenance DNA replication complex. Furthermore, many co-expressed genes shared a common function; for example genes encoding numerous glycolytic enzymes and genes involved in the tricarboxylic acid cycle. (17)

We have reported many other examples of employing *F*-means clustering for the analysis of clinical and experimental data in a series of publications (17–20).

Correlation mosaic analysis to visualize changes in cluster associations

Both the reproducibility and significant differences in the clustering results are usually estimated visually, or qualitatively. Here, we present correlation mosaic based visualization of global patterns in expression data with individually presented interconnections between patterns and genes. This approach can be used as an independent clustering procedure or as an addition to the completed *F*-means clustering results. In this example the clustering procedure is based on the Pearson correlation and consists essentially of the sequence of operations used in *F*-means clustering described above. The primary difference is that instead of using *deviation variability* as a measure of distance, we use a correlation coefficient. The number of clusters and the cluster contents are determined using a threshold that can be established in simulation experiments. The output of this procedure consists of three data sets: first, cluster allocation for all genes in the analysis, second, connectivity parameter for each gene, and third, matrices of correlation coefficients. Matrices of correlation coefficients can be represented in a graphical form known as a correlation mosaic, which is convenient for the visual inspection of the differences in gene associations between cases and controls.

Résumé 3: Correlation mosaic analysis of the co-expression of HVE-genes. The procedure consists of the following steps:

- Normalization of gene expression and identification of HVE-genes is conducted as in Résumé 1. HVE-gene expression data are presented in normalized units.
- A connectivity parameter is defined for each HVE-gene as the number of other genes whose expression profiles correlate with any given gene above the threshold 'tr'. The appropriate choice of threshold is obtained in simulation experiments.
- HVE-genes in each group are sorted by their connectivity, and the clustering process begins with genes of the highest connectivity. The gene with the highest connectivity and all genes that deviate from this gene's expression in each sample with variabilities not higher than the variability of the Reference Group comprise Cluster #1. The next gene not belonging to the first cluster and genes selected as not significantly deviating comprise Cluster #2. The process continues until all genes are analyzed. Genes that have zero connectivity do not belong to any cluster.
- The result is presented as a color-plot with the gene numbers used as the coordinates along the axes, with the same ordering $\boldsymbol{G_1}\ldots\boldsymbol{G_n}$ used along the abscissa and the ordinate).
- When the correlated gene associations are compared between two groups of samples, the order of coordinated genes is the same in both mosaics.

This correlation mosaic method was applied to the analysis of gene expression data and cytokine multiplex data in clinical and experimental samples (17–26). In the

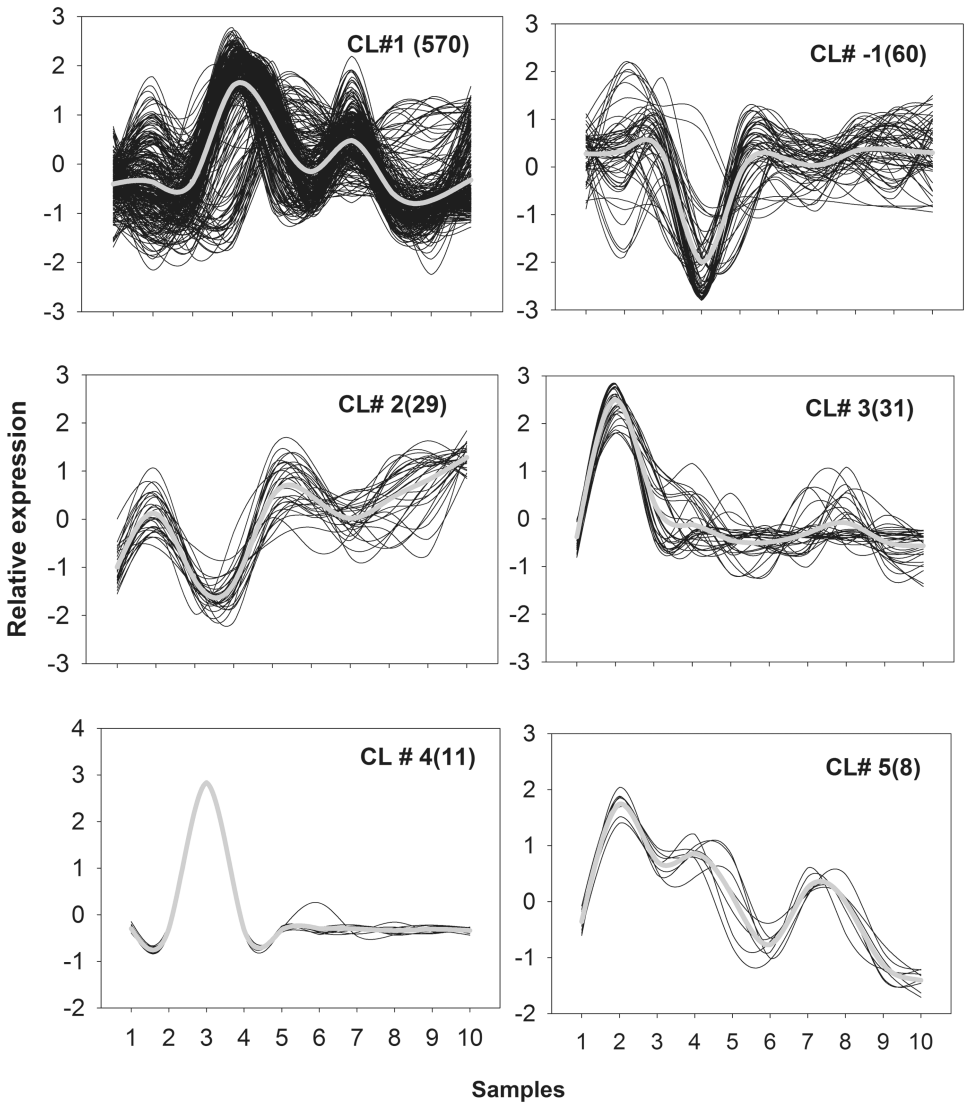

Figure 4. *F*-means clustering of gene expressions in T cells from B6 mice. The six largest clusters are shown. Abscissa: cluster numbers derived from 10 samples from 10 different mice. Ordinate: the normalized expression levels. Figures in brackets: the numbers of genes in each cluster.

very first example a mouse model of bladder inflammation was used to investigate the role of neurokinin 1 receptors (NK1R) and neprilysin (NEP) in neurogenic inflammation. Cystitis was induced in wild-type mice sensitized to human serum albumin after being challenged with the same antigen. Microarray analysis revealed that inflammatory processes in wild mice-type led to a downregulation of neprilysin expression. The most prominent cluster of activator protein 1 (AP-1)-responsive genes included neprilysin (upper portion of Figure 7). In contrast, $NK_1R^{-/-}$ mice failed to mount an inflammatory reaction and the presence of neprilysin negatively correlated with the expression of the same gene(s) in wild-type mice (bottom Figure 7). The switching of NEP correlations from positive in wild-type mice to negative in $NK1R^{-/-}$ mice is very convincing in this presentation. This work (21) provided a suitable model for elucidating the involvement of AP-1 transcription factor in bladder inflammation and suggested a testable hypothesis regarding the role of NK1R and NEP in inflammation.

- The correlation mosaic analysis also was applied to HVE-genes in JRA data as given above. Figure 8 presents an outstanding visualization of the changes in some gene associations with other cluster members during the course of treatment of JRA patients. Analysis of the healthy donor group (HD group) reveals the presence of two highly correlated clusters of genes. The color variation in the mosaic visualizes the differences among the healthy donors (HD), non-treated (AD) and treated partially-responding (PR) patients. On closer inspection, the involvement of genes with altered functional interconnections within each cluster indicates that those genes are directly involved in the pathology (17).
- These examples demonstrate that with the use of color-coded correlation mosaics, complicated

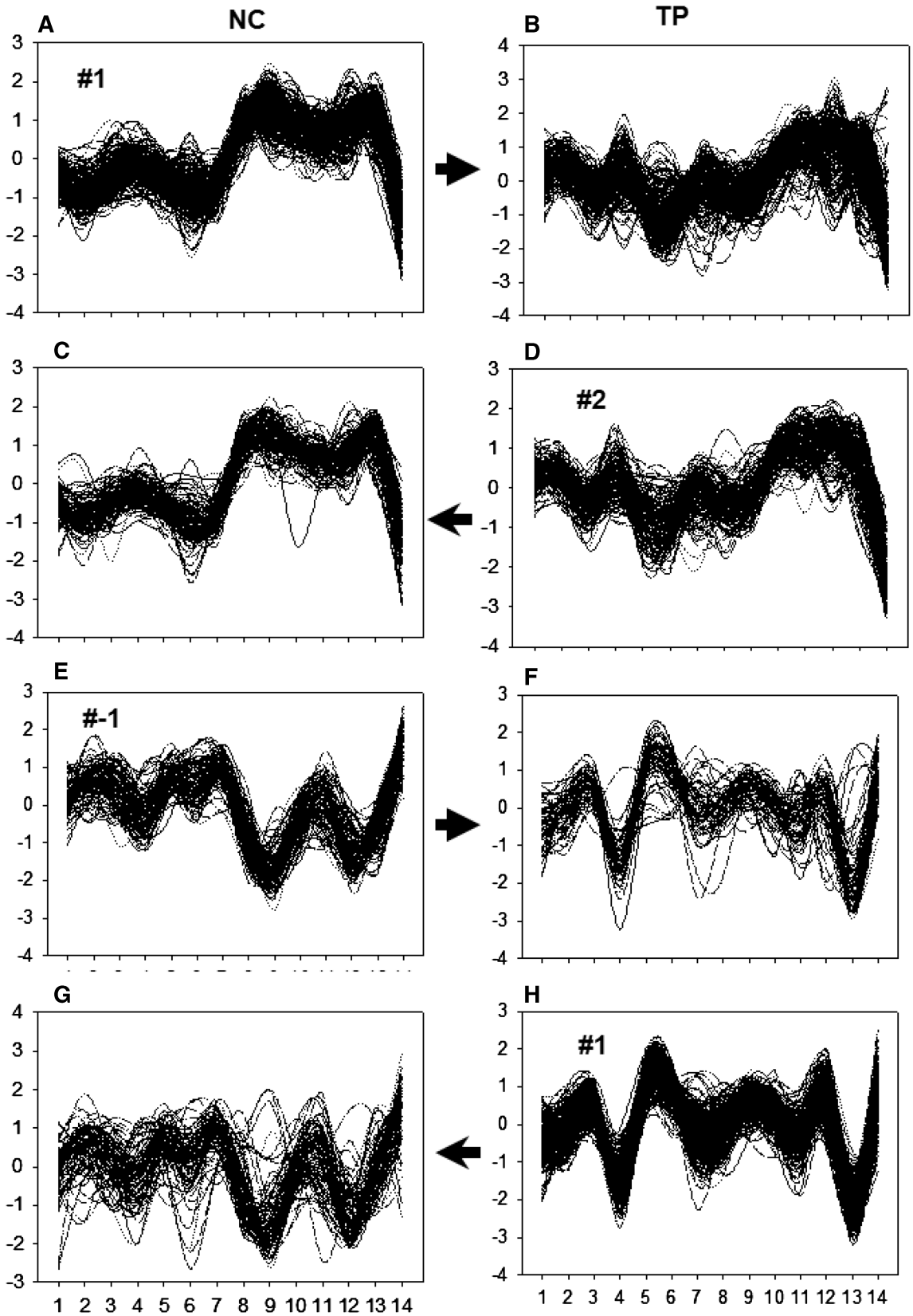

Figure 5. Reproducibility of the HVE gene co-expression in two unrelated sample groups: NC (normal controls) and TP (TRAPS patients). Normalized expression levels (ordinate) are presented against the numbers of samples in each group. Genes in the largest cluster (#1, **A**) in the NC group are also co-expressed in the TP group (**B**). Most of the genes belong to the largest cluster (#2, **D**) in the TP group. Conversely, genes in the largest cluster (#2, D) of the TP group are co-expressed in the NC group (**C**) and again almost entirely belong to the largest cluster of the NC group. The second largest cluster of the NC group #1 (**E**) is the inversion of the #1 cluster (**a**) in the NC. Genes are almost entirely in the second largest cluster (#1, **F–H**) of the TP group. The opposite is seen in (**G** and H). In contrast with the NC, Clusters #1 and #2 in the TP are not the reverse reflections of each other.

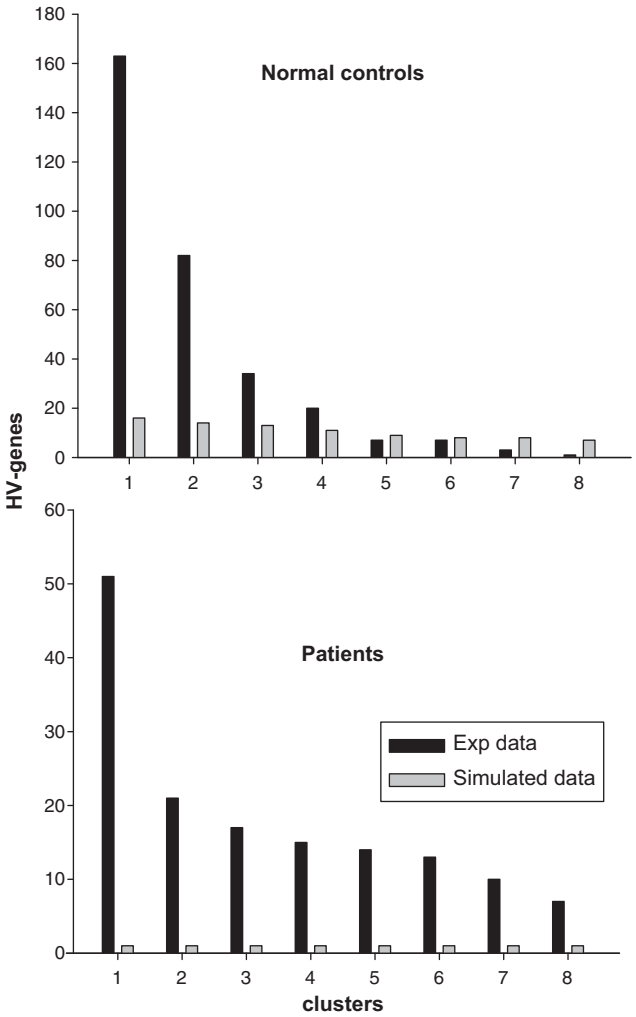

Figure 6. Contents of the eight biggest clusters in the NC and TP groups (Figure 5) (black bars) compared with the same for the simulated data (data obtained by substitution of the real gene expressions with random values having the same SD and means for each gene over all samples in groups).

interdependencies between genes can be visualized and differences between subgroups can be assessed. Correlation clustering is not just a procedure for gene partitioning into different compartments but is rather a combination of clustering and networking. This method provides a tool for quantitatively estimating interconnections between genes within clusters.

Gene networking based on partial correlation coefficients

Gene regulatory networks have become a major focus of interest in recent years. A number of reverse engineering approaches have been developed to help uncover these regulatory networks. Correlative mosaics demonstrate the existence of closely correlated modules, which are connected through positive or negative correlations. This type of presentation seems to be in good agreement with the widely discussed modularity of gene networks. In spite of this agreement some caution is necessary as the relatively high connectivities of gene clusters in correlation mosaic analysis mostly represent the indirect influences of a small number of regulatory elements. Information about direct interactions gives partial correlations that in turn enable to the distinguish of correlations between two variables that originate through direct influence versus correlation originated through the influence of intermediate variables. Partial correlation excludes many possibilities and usually significantly diminishes gene connectivity. We used this procedure for the networking of HVE-genes (18,20,21).

Résumé 4: Networking procedure based on partial correlations. The environmental circle for each gene is determined as a set of genes correlated with any given one having a correlation coefficient above threshold t_1.

The matrix of partial correlation coefficients within the environmental circle of genes is calculated. The elements of the matrix R_{ij} represent the partial correlation coefficients between the given gene and gene i with the removed influence of gene j. All genes are within the given gene's environmental circle.

The genes G_i are considered to be causally interconnected with the given gene if the row R_{ij} of the matrix does not have members below threshold t_1, and if the averaged value of the row is above threshold t_2. A Monte–Carlo simulation study is used to define the statistical thresholds (t_1 and t_2) below which partial correlation coefficients are likely due to chance.

One example of the networking of HVE-genes was obtained during comparative analysis of the response to stimulation of EBV-transformed B cells derived from SLE patients and normal unrelated controls. Pathway Analysis allowed us to establish model networks of functional gene expression important for B cell signaling and elucidate gene expression regulatory interconnections disrupted in B cells from individuals with lupus (Dozmorov I, Dominguez N, Sestak AL, Xu HM, Harley JB, James JA, Guthridge JM manuscript in preparation). Fragments of this network that include genes uniquely activated in only one of these groups (controls or patients) are shown in Figure 9. These unique network fragments reproduced in two independent experiments present functional fingerprints of activated B cells from lupus patients and normal controls. In this context, one should note that practically all genes uniquely activated in normal controls (Figure 9A) are known as being 'pro-apoptotic', while the genes uniquely activated in B cells from lupus patients (Figure 9B) are 'anti-apoptotic'. These results are in good agreement with the established defects of B cell apoptosis in lupus patients (27).

TNF pathway modulation

In another example this networking procedure was used to establish functional interconnections between HVE-genes in TRAPS pathology and normal control samples. HVE-genes demonstrating reproducible co-expression both in control and in TRAPS patients were selected (Supplementary Figure 3S). It is important to note that the majority of genes belonging to the largest cluster in

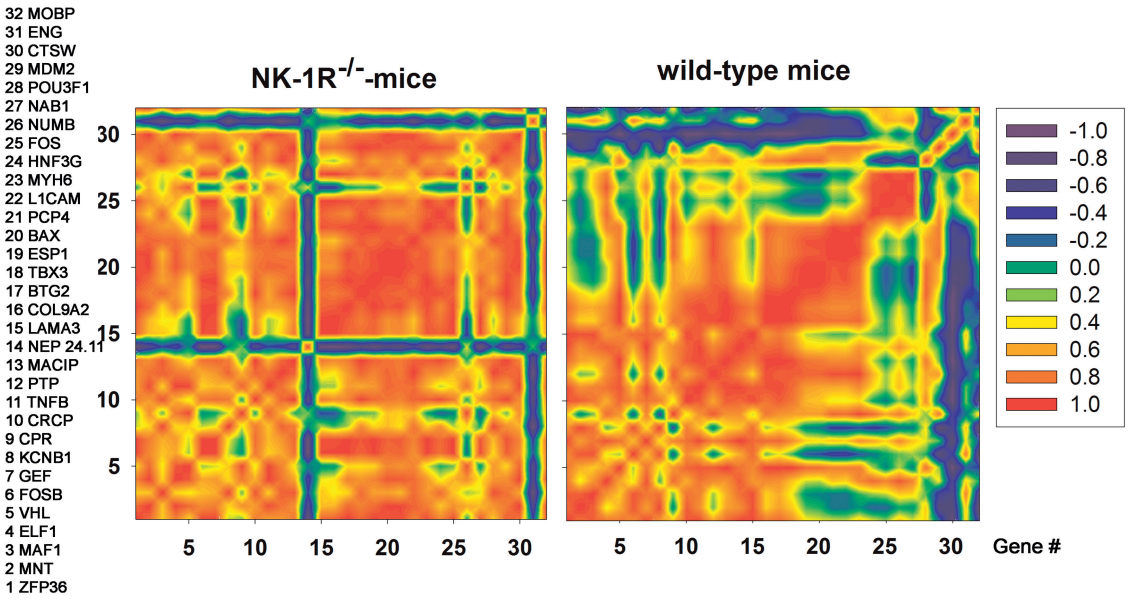

Figure 7. Mosaic of correlation coefficients of the HVE-genes in wild-type and NK1R–/– mice. The coordinates along axis are the numbers of genes listed in the left box. The white lines in A indicate the borders of three clusters of tightly interconnected genes. The colored lines and spots beyond the clusters represent positively linked genes (red) belonging to two or more clusters (Gene 5, for example), or negatively linked genes (blue). Genes that exhibited positive correlations over time were represented in graded shades of red, and genes negatively correlated are shown in graded shades of blue. Genes with an absence of correlation are indicated in green. Neprilysin is in the central position in the most prominent cluster found in wild-type mice, which includes a group of AP-1 responsive genes. In contrast, the association with these genes becomes negative in NK1R$^{-/-}$ mice, who fail to mount antigen induced bladder inflammation.

the control samples are also tightly clustered in the largest cluster of the patient samples. The close similarity of the contents of the largest clusters in two independently produced clustering procedures supports our hypothesis about common biological basis for such co-expression.

F-means clustering of some genes associated with the TNF pathway are shown in Figure 10. Partial correlation coefficients were calculated for each pair of 42 selected genes. Two thresholds were used to select significant interconnections. The threshold (t_1) 0.7 was used to select the unique connections, and 0.5 was used for connections reproduced in the networks of both groups. The results of these calculations are presented in Figure 10A and B. The connections obtained with this method appeared to be consistent with current knowledge about this TNF pathway (Supplementary Figure S4 shows the pathway obtained with the use of Ingenuity Pathway Analysis). Interleukin-6 (IL-6) interconnections were expected based on the altered function of this cytokine in TRAPS pathology (28). The appearance of the MEFV gene in the TRAPS network is also interesting because mutations in this gene characterize another periodic fever, Mediterranean fever.

DISCUSSION

Microarray technology has revolutionized the study of biology by allowing the simultaneous examination of the expression profile of the entire genome. Gene expression profiling enables rapid analysis of thousands of genes in parallel and has been used to establish many disease-specific fingerprints of pathology (29–31). Such profiling might facilitate the development of diagnostic strategies for complex diseases, although one has to bear in mind that among hundreds of differentially expressed genes, only a portion might play a critical role in pathology, while many others may have only bystander effects. The analysis of the disease processes requires methods that extend beyond comparing gene expression levels. The most exciting opportunity is to characterize pathology through changes in 'functional associations' among genes. Genes involved in such processes reveal extreme variability in their expression levels, thereby uncovering functional associations among them. As stated in the work from the Kauffman laboratory (1), random independent inputs (as chaotic environmental perturbations are) allow for better recognition of regulatory associations, and such identifications are more robustly resistant to noise. These properties make HVE-genes an important source of information about regulatory interconnections in biological systems.

The most renowned problem in HVE-gene research is the absence of adequate statistical methods for the selection and interpretation of HVE-genes (8). Among the most frequently employed statistical evaluations for HVE-genes are ANOVA methods, which are used to determine the fraction of genes significantly differentially expressed between individuals (32,33). These methods are simple and are based on commonly understood statistical principles. However, the problems of sensitivity and specificity prevent blindfolded application of these straightforward statistical methods to microarray analysis without previously determined corrections to the significance thresholds.

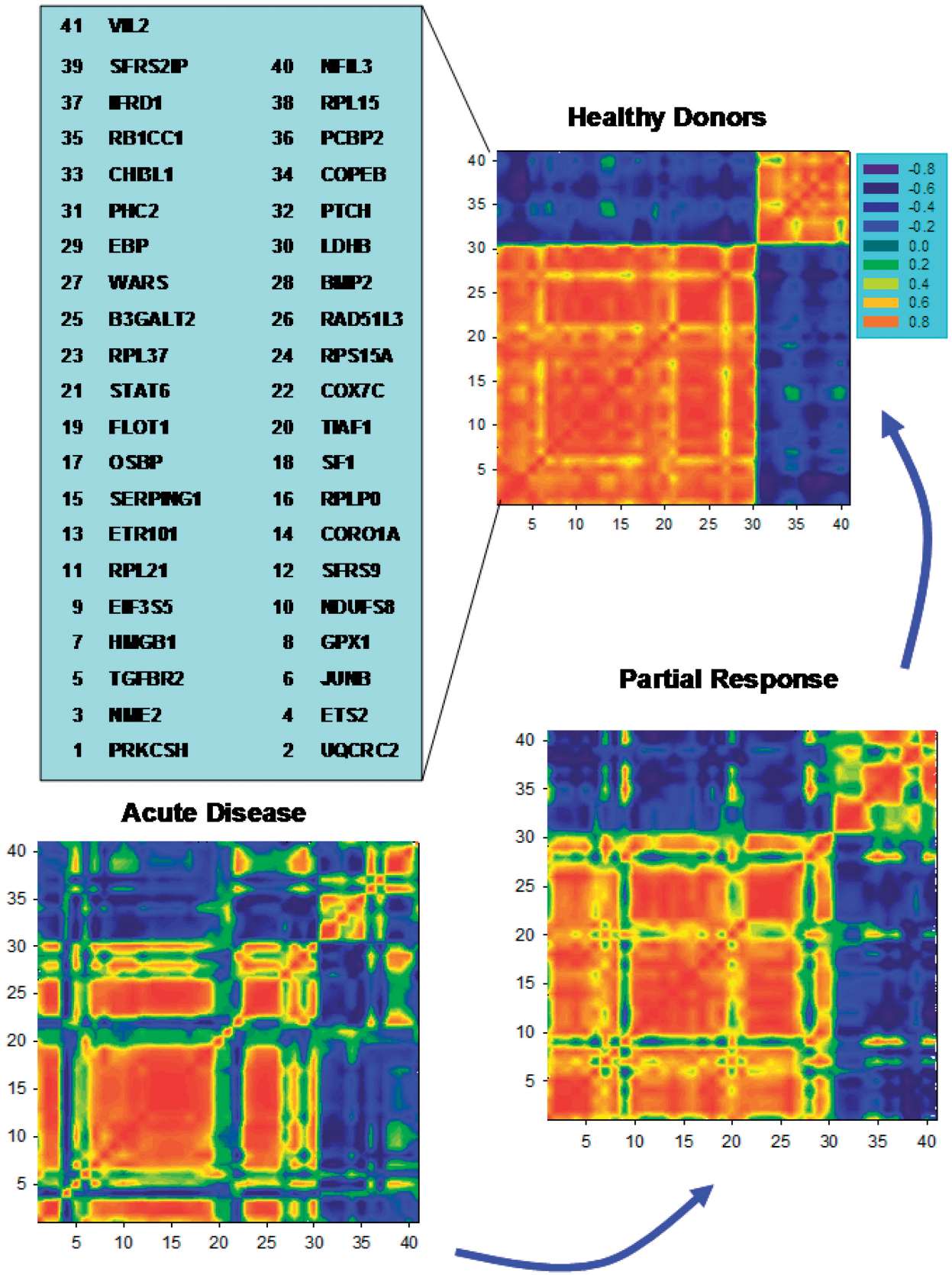

Figure 8. Correlation mosaics for genes from the two largest clusters in the control group (adopted from [Jarvis ea, 2003]). The designations are the same as in Figure 7. There is shown transformation of the mosaic created for patients group (Acute disease) to the Partial Response mosaic (patients who have been treated with corticosteroids or other anti-inflammatory drugs), and finally to the Healthy Donors mosaic.

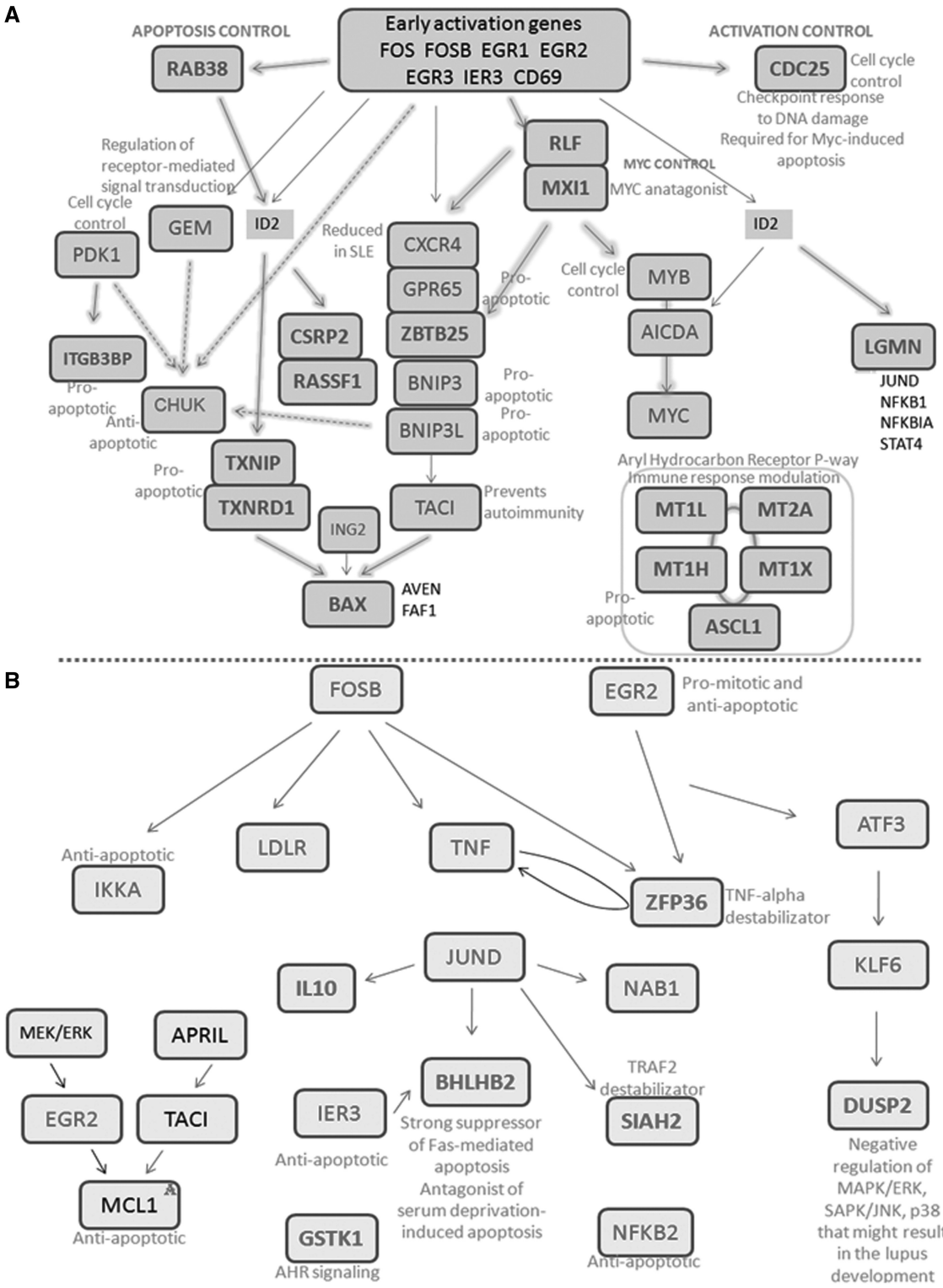

Figure 9. Networking of reproducibly variable genes after stimulation of EBV-transformed B cells from normal controls (**A**) and lupus patients (**B**). This network is a fragment of a gene network consisting of genes uniquely activated in normal (**A**) or lupus patient (**B**) groups. The gene network was built through the partial correlations method (as described in the 'Materials and Methods' section).

To address this issue, we have successfully implemented the Internal Standard strategy for differential gene expression analysis (9) and developed optimal power analysis, including the estimation of replication requirements. Although we have presented several experimental conclusions within each project presented in this communication, some of them appear to be of general validity, and in turn they become solid attributes of gene expression analysis.

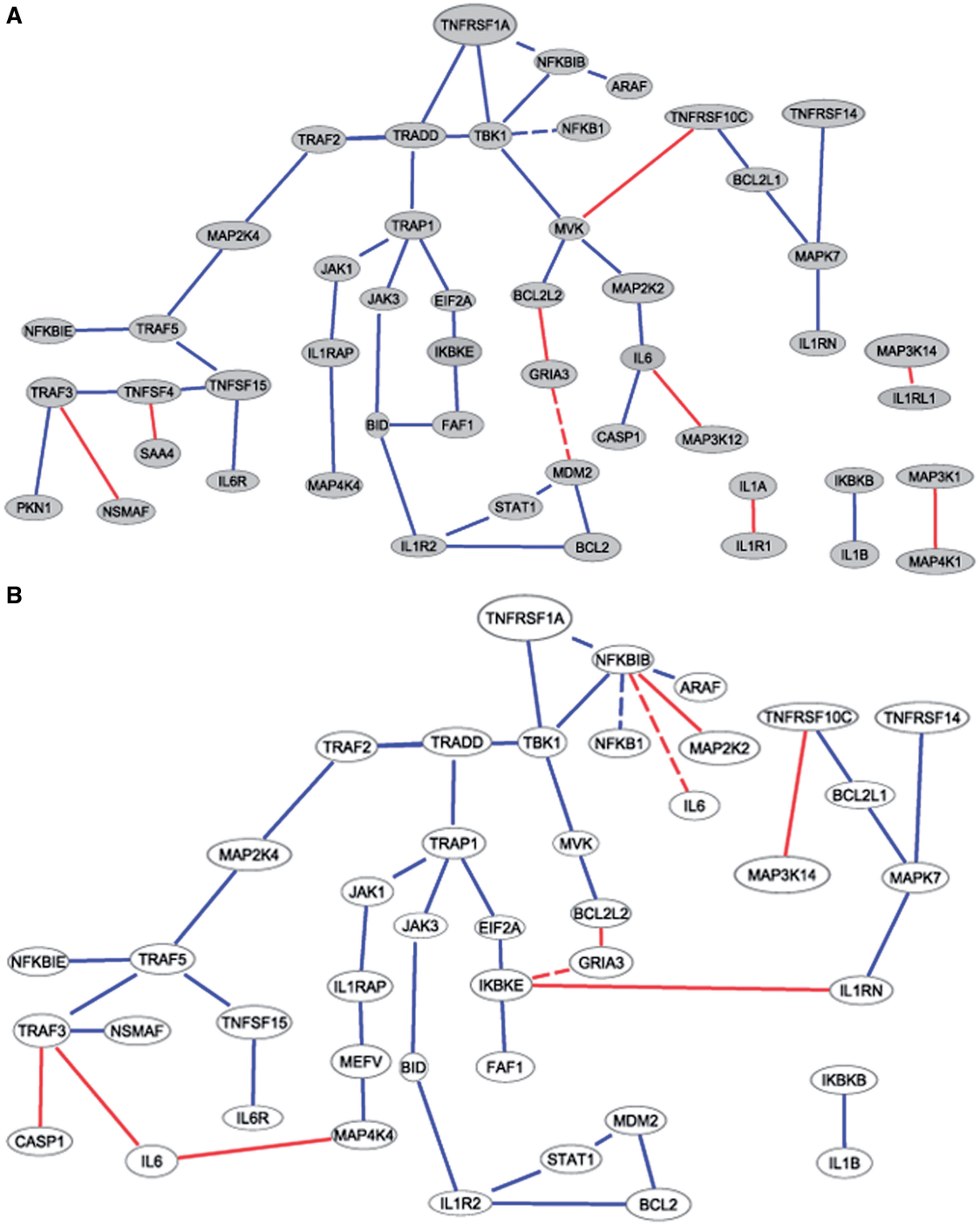

Figure 10. TNF pathway. Gene interconnection in both normal control (**A**) and TRAPS patients (**B**) obtained by calculating partial correlation coefficients. The solid lines represent positive interconnections with averaged partial correlation coefficients >0.7. The dashed lines represent interconnections with negative partial correlation coefficients with averaged values <–0.7. The red lines represent interconnections significantly unique in each of the populations.

We found that HVE-genes are true components of the process of gene expression regulation. Together, HVE-genes serve as an important source of information about the functional connectivity of the genome and about dynamical processes based on this connectivity.

The high incidence of expression variability, as well as the coherent appearance of this kind of expression, excludes the likelihood that this behavior occurs by chance (Figures 4–6). A striking feature of our findings is not only that a significant portion of genes are expressed hypervariably, but that the resulting patterns of variability are remarkably similar. These observations enable the application of standard clustering procedures to the analysis with the result that the contents of such clusters exceed

any chance coincidences. Additional evidence supporting the premises of our model includes the extraordinary high reproducibility of independently derived experimental sample groups (Figure 5 and Supplementary Figure S3).

Our finding that many genes with high expression variabilities are associated exclusively with pathologies while the same set of genes display stable expression in normal samples (Figure 2) suggests the possibility that the mentioned pathologies are associated with a loss of control in transcriptional processes. However, this problem is awaiting careful investigation. Another surprising aspect of our findings is a functional relatedness among many of the studied HVE-genes. As an example, we point out that most genes demonstrating unique variability in the periodic fever syndrome (TRAPS) are directly associated with inflammatory processes (Figure 2A and Supplementary Figure S1).

In addition, almost all of the genes that are uniquely variable in samples from lupus patients have anti-apoptotic activity, whereas genes uniquely variable in control samples all have distinct pro-apoptotic activity (Figure 9). This result is in strong agreement with the known fact that B cells of lupus patients have defects in apoptosis (27).

Application of the networking procedure to the HVE-genes selected from samples from TRAPS patients and normal controls produced remarkably reproducible associations among genes of the TNF pathway. The few differences between the 'pathological' and 'normal' networks are consistent with the established features of this pathology (6,34).

We are committed to the viewpoint that the biological reality of hyper-variations in gene expression forms a solid basis for the analysis of biological objects. For example:

- Statistically significant differences in the variabilities of HVE-genes as compared with the majority of relatively stable genes in an array (Figure 1A) exclude the possibility that such fluctuations are due to chance.
- Many HVE-genes have very similar expression profiles, thereby enabling the identification of large clusters of co-expressed genes (Figures 4 and 5). The sizes of such clusters significantly exceed the sizes of clusters in simulated random sets of data (Figure 6).
- Some groups of co-expressed genes are highly reproducible, appearing to be only slightly altered in different groups of samples (Figure 5 and Supplementary Figure S3).
- The clusters of co-expressed HVE-genes present groups of genes joined by their participation in regular biological processes (Figures 7–9).

As we have shown in various applications, these features of HVE-genes make them a very important source of information regarding functional interconnections in biological systems and processes.

Various pathologies associated with the stimulation of defense functions (e.g. inflammation and autoimmunity) increased the proportions of the HVE-genes in comparison with the relatively quiet control state (Figure 2). It is possible that an analogy with the temperature of physical bodies could be drawn with regard to the increased mobility of such pathologies.

Considering that HVE-genes are a presentation of internal dynamic processes, it is possible to employ the usual methods of analyses for these processes, including clustering and networking approaches usually applied to the study of temporal dynamics. Genes could be gathered into groups of co-expressed genes by conventional clustering procedures. Such clusters contain HVE-genes associated with common biological processes and signaling pathways. Loss or change of membership in these clusters by one or several genes could be a hallmark of pathology-associated alterations, as demonstrated in Figure 7.

We usually observe more than one large cluster of HVE-genes with possible functional associations, which substantiates the coexistence of different internal dynamics. For example, we often observe the presence of two large clusters with anti-correlated profiles (Figure 5, see also Figure 2C). Such anti-correlation indicates that these two dynamic processes exist not as independent phenomena but as compensatory reactions to mutual changes. Deviation from the stability of genes within one group is accompanied by a corresponding and opposite change by the genes in another cluster. Alterations in such compensatory reactions could also be important hallmarks of pathology.

The sum of two anti-correlated profiles is constant, and this invariability is maintained in the coordinated variations of the profiles, i.e. the changes in one profile are compensated by opposite changes in another. In this situation, it is possible that a more complicated form of compensatory reactions, incorporating the involvement of more than two clusters or HVE-genes with different dynamic profiles, is occurring. Examples of such associations were obtained through linear discriminatory analysis for the classification of sample groups. Dynamic discriminant function analysis was developed based on the concept that stable classification parameters (roots) can be derived from highly variable gene-expression data (35). We demonstrated earlier that the functional interconnections between HVE discriminatory genes can be presented in the form of functional networks that exhibit distinctive changes in pathology cases when compared to controls (35).

In conclusion, the analysis of the coordinated behavior of HVE-genes can resolve the very important clinical problem of non-homogeneity in sample groups that consist of patients with phenotypically similar syndromes. Such discrimination and exclusion of homogeneity is especially important in characterizing the phases of pathology development and the changes in the course of response to the treatment and in discriminating hidden pathologies when a disease with common clinical characteristics can include pathologies of different molecular mechanisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Robert Hurst, Yuhong Tang and Mikhail Dozmorov for fruitful discussions, Nicholas Knowlton and Shengguang Qian for help with programming, and Mark B. Frank for technical assistance with the microarray experiments.

FUNDING

The National Institutes of Health (P20 RR020143 to I.D., R01 AI045050 to I.D., P30 AR053483 to I.D. and J.G., P20RR016478 to I.D., R01 AI084200 to I.D. and J.J., CA106713 to D.B.); The Royal College of Pathology UK and The Journal of Experimental Pathology (to E.D. and P.J.T.); and The Jones Charitable Trust (to E.D. and P.J.T.). Funding for open access charge: P20RR016478.

Conflict of interest statement. None declared.

REFERENCES

1. Kauffman,K.J., Ogunnaike,B.A. and Edwards,J.S. (2006) Designing experiments that aid in the identification of regulatory networks. *Brief. Funct. Genomic Proteomic*, **4**, 331–342.
2. Pritchard,C., Coil,D., Hawley,S., Hsu,L. and Nelson,P.S. (2006) The contributions of normal variation and genetic background to mammalian gene expression. *Genome Biol.*, **7**, R26.
3. Lindberg,J., af Klint,E., Ulfgren,A.K., Stark,A., Andersson,T., Nilsson,P., Klareskog,L. and Lundeberg,J. (2006) Variability in synovial inflammation in rheumatoid arthritis investigated by microarray technology. *Arthritis Res. Ther.*, **8**, R47.
4. Akahoshi,M., Nakashima,H. and Shirakawa,T. (2006) Roles of genetic variations in signalling/immunoregulatory molecules in susceptibility to systemic lupus erythematosus. *Semin. Immunol.*, **18**, 224–229.
5. Jarvis,J.N., Dozmorov,I., Jiang,K., Frank,M.B., Szodoray,P., Alex,P. and Centola,M. (2004) Novel approaches to gene expression analysis of active polyarticular juvenile rheumatoid arthritis. *Arthritis Res. Ther.*, **6**, R15–R32.
6. Centola,M., Aksentijevich,I. and Kastner,D.L. (1998) The hereditary periodic fever syndromes: molecular analysis of a new family of inflammatory diseases. *Hum. Mol. Genet.*, **7**, 1581–1588.
7. Garge,N.R., Page,G.P., Sprague,A.P., Gorman,B.S. and Allison,D.B. (2005) Reproducible clusters from microarray research: whither? *BMC Bioinformatics*, **6(Suppl. 2)**, S10.
8. McShane,L.M., Radmacher,M.D., Freidlin,B., Yu,R., Li,M.C. and Simon,R. (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, **18**, 1462–1469.
9. Dozmorov,I. and Lefkovits,I. (2009) Internal standard-based analysis of microarray data. Part 1: analysis of differential gene expressions. *Nucleic Acids Res.*, **37**, 6323–6339.
10. Benbrook,D.M., Lightfoot,S., Ranger-Moore,J., Liu,T., Chengedza,S., Berry,W.L. and Dozmorov,I. (2008) Gene expression analysis of biological systems driving an organotypic model of endometrial carcinogenesis and chemoprevention. *Gene Regul. Syst. Bio.*, **2**, 21–42.
11. Chiorazzi,N., Hatzi,K. and Albesiano,E. (2005) B-cell chronic lymphocytic leukemia, a clonal disease of B lymphocytes with receptors that vary in specificity for (auto)antigens. *Ann. N Y Acad. Sci.*, **1062**, 1–12.
12. van der Heul-Nieuwenhuijsen,L., Padmos,R.C., Drexhage,R.C., de Wit,H., Berghout,A. and Drexhage,H.A. (2010) An inflammatory gene-expression fingerprint in monocytes of autoimmune thyroid disease patients. *J. Clin. Endocrinol. Metab.*, **95**, 1962–1971.
13. Morel,L., Croker,B.P., Blenman,K.R., Mohan,C., Huang,G., Gilkeson,G. and Wakeland,E.K. (2000) Genetic reconstitution of systemic lupus erythematosus immunopathology with polycongenic murine strains. *Proc. Natl Acad. Sci. USA*, **97**, 6670–6675.
14. Morel,L., Blenman,K.R., Croker,B.P. and Wakeland,E.K. (2001) The major murine systemic lupus erythematosus susceptibility locus, Sle1, is a cluster of functionally related genes. *Proc. Natl Acad. Sci. USA*, **98**, 1787–1792.
15. Subramanian,S., Yim,Y.S., Liu,K., Tus,K., Zhou,X.J. and Wakeland,E.K. (2005) Epistatic suppression of systemic lupus erythematosus: fine mapping of Sles1 to less than 1 mb. *J. Immunol.*, **175**, 1062–1072.
16. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
17. Dozmorov,I., Knowlton,N., Tang,Y., Shields,A., Pathipvanich,P., Jarvis,J.N. and Centola,M. (2004) Hypervariable genes–experimental error or hidden dynamics. *Nucleic Acids Res.*, **32**, e147.
18. Dozmorov,I., Saban,M.R., Knowlton,N., Centola,M. and Saban,R. (2003) Connective molecular pathways of experimental bladder inflammation. *Physiol. Genomics*, **15**, 209–222.
19. Nunlist,E.H., Dozmorov,I., Tang,Y., Cowan,R., Centola,M. and Lin,H.K. (2004) Partitioning of 5alpha-dihydrotestosterone and 5alpha-androstane-3alpha, 17beta-diol activated pathways for stimulating human prostate cancer LNCaP cell proliferation. *J. Steroid Biochem. Mol. Biol.*, **91**, 157–170.
20. Zimmerman,R.A., Dozmorov,I., Nunlist,E.H., Tang,Y., Li,X., Cowan,R., Centola,M., Frank,M.B., Culkin,D.J. and Lin,H.K. (2004) 5alpha-Androstane-3alpha,17beta-diol activates pathway that resembles the epidermal growth factor responsive pathways in stimulating human prostate cancer LNCaP cell proliferation. *Prostate Cancer Prostatic. Dis.*, **7**, 364–374.
21. Dozmorov,I., Saban,M.R., Gerard,N.P., Lu,B., Nguyen,N.B., Centola,M. and Saban,R. (2003) Neurokinin 1 receptors and neprilysin modulation of mouse bladder gene regulation. *Physiol. Genomics*, **12**, 239–250.
22. Szodoray,P., Alex,P., Jonsson,M.V., Knowlton,N., Dozmorov,I., Delaleu,N., Jonsson,R. and Centola,M. (2005) Distinct profiles of Sjorgen's syndrome patients with ectopic salivary gland germinal centers revealed by serum cytokines and BAFF. *Clin. Immunol.*, **117**, 168–176.
23. Knowlton,N., Dozmorov,I., Kyker,K.D., Saban,R., Cadwell,C., Centola,M.B. and Hurst,R.E. (2006) Template-driven gene selection procedure. *IEE Proc. Syst. Biol.*, **153**, 4–12.
24. Szodoray,P., Alex,P., Frank,M.B., Turner,M., Turner,S., Knowlton,N., Cadwell,C., Dozmorov,I., Tang,Y., Wilson,P.C. *et al.* (2006) A genome-scale assessment of peripheral blood B-cell molecular homeostasis in patients with rheumatoid arthritis. *Rheumatology*, **45**, 1466–1476.
25. Jarvis,J.N., Petty,H.R., Tang,Y., Frank,M.B., Tessier,P.A., Dozmorov,I., Jiang,K., Kindzelski,A., Chen,Y., Cadwell,C. *et al.* (2006) Evidence for chronic, peripheral activation of neutrophils in polyarticular juvenile rheumatoid arthritis. *Arthritis Res. Ther.*, **8**, R154.
26. Lawrence,S., Tang,Y., Frank,M.B., Dozmorov,I., Jiang,K., Chen,Y., Cadwell,C., Turner,S., Centola,M. and Jarvis,J.N. (2007) A dynamic model of gene expression in monocytes reveals differences in immediate/early response genes between adult and neonatal cells. *J. Inflamm.*, **4**, 4.
27. Veeranki,S. and Choubey,D. (2010) Systemic lupus erythematosus and increased risk to develop B cell malignancies: role of the p200-family proteins. *Immunol. Lett.*, **133**, 1–5.
28. McDermott,M.F. and Aksentijevich,I. (2002) The autoinflammatory syndromes. *Curr. Opin. Allergy Clin. Immunol.*, **2**, 511–516.
29. Kurella,M., Hsiao,L.L., Yoshida,T., Randall,J.D., Chow,G., Sarang,S.S., Jensen,R.V. and Gullans,S.R. (2001) DNA microarray analysis of complex biologic processes. *J. Am. Soc. Nephrol.*, **12**, 1072–1078.
30. Catarino,P.A. and Goldstraw,P. (2006) The future in diagnosis and staging of lung cancer: surgical techniques. *Respiration*, **73**, 717–732.

31. Bailey,W.J. and Ulrich,R. (2004) Molecular profiling approaches for identifying novel biomarkers. *Expert Opin. Drug Saf.*, **3**, 137–151.
32. Oleksiak,M.F., Churchill,G.A. and Crawford,D.L. (2002) Variation in gene expression within and among natural populations. *Nat. Genet.*, **32**, 261–266.
33. Turk,R., t Hoen,P.A., Sterrenburg,E., de Menezes,R.X., de Meijer,E.J., Boer,J.M., van Ommen,G.J. and den Dunnen,J.T. (2004) Gene expression variation between mouse inbred strains. *BMC Genomics*, **5**, 57.
34. McDermott,M.F., Aksentijevich,I., Galon,J., McDermott,E.M., Ogunkolade,B.W., Centola,M., Mansfield,E., Gadina,M., Karenko,L., Pettersson,T. *et al.* (1999) Germline mutations in the extracellular domains of the 55 kDa TNF receptor, TNFR1, define a family of dominantly inherited autoinflammatory syndromes. *Cell*, **97**, 133–144.
35. Dozmorov,I.M., Centola,M., Knowlton,N. and Tang,Y. (2005) Mobile classification in microarray experiments. *Scand. J. Immunol.*, **62(Suppl. 1)**, 84–91.