AD_____

WARD NUMBER: W81XWH-11-1-0709

TITLE: Genome-Wide Association Study of a Validated Case Definition of Gulf War Illness in a Population-Representative Sample

PRINCIPAL INVESTIGATOR: Robert W. Haley, M.D.

CONTRACTING ORGANIZATION: University of Texas Southwestern Medical Center
Dallas, Texas 75390-8874

REPORT DATE: September 2013

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
September 2013	Final	1 September 2011 – 31 August 2013

4. TITLE AND SUBTITLE
Genome-Wide Association Study of a Validated Case Definition of Gulf War Illness in a Population-Representative Sample

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0709

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Robert W. Haley, M.D.

Email: Robert.Haley@UTSouthwestern.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)

University of Texas Southwestern Medical Center
5323 Harry Hines Blvd.
Dallas, Texas 75390-7208

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The specific aim of this study was to perform gene expression profiling and group comparison analyses of differential gene expression in RNA-stabilized blood samples of the approximately 150 veterans meeting the Factor case definition and Gulf War-era veteran controls. From the banked blood samples collected previously and stored frozen at -80°C in PAXgene tubes from 2 population subsamples of Gulf War veteran cases and controls, we were able to obtain sufficient high quality peripheral blood mononuclear cell (PBMC) RNA from 142. Our Genomics Core Laboratory performed Next Generation Transcriptome Sequencing (RNA-Seq), using Illumina's multiplexing mRNA-Seq to generate full sequence libraries from the poly-A tailed RNA to a read depth of 30 million reads. Bioinformatics analyses for replicable differences in the expression levels among the 4 clinical groups were completed with the Dozmorov method to test for multiple comparisons-corrected group differences. In the tests for differentially expressed single target genes, no significant group differences were found, but the second stage using Ingenuity Pathway Analysis to test for common regulator genes identified 6 transcription regulators whose downstream target genes differed between syndrome group 2 and controls in the Developmental subsample and the Replication subsample. Most noteworthy was the finding of apparent instability in expression of the STAT3 regulatory oncogene, which, if confirmed, could explain increased rates of brain cancer in Gulf War veterans. Future research for a Gulf War illness diagnostic test should employ methods, such as pharmacologic stimulation and analysis of pure cell types, to reduce high background variability of gene expression in peripheral blood cells.

15. SUBJECT TERMS
Gulf War illness; RNA; gene expression; next-generation sequencing; RNA-seq; regulator genes

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	81	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

INTRODUCTION

Many veterans with chronic illness following deployment to the 1991 Gulf War appear to have a chronic encephalopathy associated epidemiologically with exposure to cholinesterase-inhibiting chemicals during deployment.[1] In past research we used principal components factor analysis to identify a large nucleus of these veterans whose symptoms suggest a unique encephalopathic illness with at least 3 phenotypic variants: syndrome 1 (impaired cognition), syndrome 2 (confusion-ataxia) and syndrome 3 (central neuropathic pain).[2] Syndrome 2, which has repeatedly been shown to be the most disabling,[3,4] has been linked epidemiologically with exposure to low-level nerve agent during the 1991 Gulf War.[5,6] This Factor case definition, which is a subset of the CDC case definition, has been validated by structural equation modeling in 3 validation samples.[4] We recently completed a large nested case-control study in a population-representative sample of Gulf War veterans which identified objective autonomic,[7] electroencephalographic,[8,9] and neuroimaging[10,11] measures of brain dysfunction in those meeting the Factor case definition and a strong gene-environment interaction of Gulf War illness (GWI) with the *PON1* gene and having heard nerve gas alarms in the war.[6] This strong finding with a candidate gene indicates a high likelihood that an unbiased genomewide gene-expression study in this sample would identify group differences useful in diagnosis and treatment. Thus, we proposed to study gene expression in the same veteran sample, comprised of **two separate samples** suitable for hypothesis development and validation. The **Developmental Sample** comprises 59 veterans selected as a nested case-control sample from a larger study of a Naval Reserve construction battalion that we have studied extensively since 1995,[2,5,12] and the **Replication Sample** comprises 93 Gulf War-era veterans selected randomly as a nested case-control sample from a nationwide telephone interview survey of a random sample of Gulf War-era veterans, the U.S. Military Health Survey (USMHS).[4] The objective of this study was to identify differences in gene expression profiles of the human transcriptome expressed in peripheral blood mononuclear cells (PBMCs) associated with the validated Factor case definition of GWI in a population-representative sample of Gulf War-era veterans to identify new targets for rational development of new diagnostic and treatment approaches.

BODY

I. The sample of Gulf War veterans

Two previously published papers that describe the validation of the Factor case definition,[4] and the selection and composition of the two subsamples[7] are provided in **Appendix A** of this report. From our archival tissue bank, we located frozen whole blood in PAXgene tubes from 145 members of the Developmental and Replication samples and transferred them to the UT Southwestern Genomics and Microarray Core Laboratory for processing of mRNA.

II. Laboratory Procedures Carried Out

Isolation and quality assessment of RNA. All procedures were performed in the IIMT UT Southwestern Genomic and Microarray Core using standard protocols (http://microarray.swmed.edu/). Briefly, total RNA was extracted from whole blood obtained in PAXgene tubes following the manufactures' protocol for manual extraction (PAXgene Blood RNA Kit Handbook 2). Adequate amounts of RNA were extracted from the PAXgene tubes of

144 of the 145 samples. Isolated total RNA was quantified by absorbance at 260 nm and quality assessed using an Agilent Bioanalyzer. All 244 mRNA samples had Nano drop absorbance ratios >1.9. All but 2 samples had RNA integrity number (RIN) values ≥7.0; the 2 samples falling below this standard, having RIN values of 6.7 and 6.8, were excluded, leaving 142 RNA samples for sequencing and bioinformatics analysis (**Table 1**). Isolated RNA was aliquoted in storage buffer and stored at -80℃ until use.

Table 1. Sample sizes of Gulf War-era veterans from whom sufficient high-quality mRNA was available for study of gene expression

	Extensively phenotyped independent samples		
Case definition	Develop-mental sample*	Validation sample*	Total
Syndrome 1 (impaired cognition)	9	19	28
Syndrome 2 (confusion-ataxia)	17	22	39
Syndrome 3 (central neuropathic pain)	11	20	31
Controls (No GWI)	14	30	44
Deployed controls	7	15	22
Non-deployed controls	7	15	22
Total subjects	51	91	142

*The developmental sample was selected from an epidemiologic study of a Naval Reserve construction battalion,[12] and the validation sample from a national survey of a random sample of Gulf War-era veterans.[4]

Preparation of transcriptome sequence dataset. All procedures were performed by personnel in the IIMT UT Southwestern Genomics and Microarray Core using standard protocols. More detailed information about these procedures is available on our website (http://genomics.swmed.edu/). Briefly, 1 µg aliquots of total RNA were used for the preparation of RNA-SEQ libraries using standard Illumina protocols and TruSeq indexed adaptors. Sequencing libraries were quantified by picogreen, and quality and size distributions were determined by Bioanalyzer analysis. The indexed samples were processed for sequencing in groups of 4 (7 pM loaded), and individual groups were sequenced on a HISEQ 2000 flow cell lane using a single-end 50 bp protocol. Approximately 30 million reads were obtained. Real time run quality assessment indicated that all samples yielded high quality sequence (>90% Q30). After the completion of the sequencing run, samples were demultiplexed using standard algorithms in the Genomics and Microarray Core and processed into individual sample Illumina single read sequence files.

Primary alignment and analysis of transcriptome sequence dataset. Sample sequence datasets were processed initially using CLC-Biosystems Genomic Workbench following our established bioanalytical pipeline for RNA-SEQ data (http://genomics.swmed.edu/). Briefly, sequence data from each sample was initially processed and trimmed for quality and subsequently processed through the RNA-SEQ alignment module of CLC-Biosystems (details available through http://www.clcbio.com/) using the most current human reference genome (HG19) from NCBI. The output of this analysis is: 1) sequence data imported into CLC-Biosystems; 2) trimmed sequence data ready for analysis 3) aligned sequence data set with interactive table containing quantitation of message levels (i.e., RPKM, unique reads, "putative exons", etc); and 4) quantitative data for isoform expression for annotated isoforms relative

expression levels. For comparative studies of gene expression among the individual samples, the table containing extensive quantitative data for all genes and isoforms was exported as an Excel file for bioinformatics analysis.

III. Bioinformatics statistical analysis of mRNA gene expression

The statistical analysis, still in progress, has so far involved two phases: the analysis of group differences in RNA levels of hypervariably expressed (*HVE*) *target genes*, and an IPA (Ingenuity Pathway Analysis) analysis for group differences in upstream *regulator genes*. The statistical techniques used represent among the most statistically powerful approaches currently available to identify group differences in gene expression.

A. Analysis for Group Differences in HVE Target Genes

The first approach has involved analysis of the Developmental Sample first to identify the group of HVE target genes and then to analyze the HVE target genes to identify individual target genes with significantly different expression between the syndrome groups and the control group that replicates in the Replication Sample. In **Appendix B** we have included 2 published papers giving a detailed explanation of the methods of HVE target gene analysis[13,14] but will summarize them briefly here. HVE target gene analysis, developed and popularized by our collaborator Dr. Igor Dozmorov,[13-16] exploits the idea that genes responsible for a disease process and thus expressed differentially in ill and well groups will show greater variability in the total population than genes not involved in the disease process (which will show low variability).[15] This idea is exploited to greatly reduce the loss of statistical power from the multiple-comparisons corrections necessary to avoid type I errors.

After all mRNA gene expression values are normalized to a mean of 0 and SD of 1 and residualized by subtraction from the group mean, a group of genes with low variability is constructed by iteratively excluding genes with expression variability greater than 2 standard deviations (SD) of the mean of all genes until no further exclusions result. The group of low-variability genes remaining is called the **Reference Group**, and the group of genes that were excluded is called the **HVE Target Gene Group**. Of the approximately 33,309 genes found to be expressed in PBMCs in this study, we found 6,034 to be HVE genes, a percentage well within the expectation.[14,15] Since there are many causes for high variability (e.g., very low expression, methodologic perturbation, etc.[15]), most of the HVE genes are not involved in the disease process. The following steps are to separate these from likely disease-related genes.

1. To identify group differences in gene expression we next performed a standard 2-group t-test of the normalized residual gene expression of each gene in the HVE Target Gene Group between a syndrome group and the control group, using the usual significance threshold of $p < 0.05$. This identifies group differences uncorrected for multiple comparisons, thus containing virtually all true group differences (high sensitivity) but many false positives (low specificity) as well.

2. The distribution of each HVE target gene found in step 1 to differ significantly from the control group is then compared with the distribution of the expression means of all genes in the Reference Group with an associative t-test having a modified Bonferroni multiple-comparisons-corrected threshold of $p < 1/N$ where N is the number of HVE target genes to be evaluated (associative analyses). This multiple-comparisons correction eliminates the false positives, but the fact that the comparison is with the distribution of the mean residual expression of the very large number of genes in the Reference Group counteracts the loss of power from the multiple-comparisons correction.[13]

3. Leave-one-out cross-validation analysis is then applied to each surviving gene's group difference analysis to exclude genes whose expression passed the tests in steps 1 and 2 due to bias from high or low outliers.

4. Finally, the ratio of gene expression in the syndrome and control groups is calculated to exclude statistically significant but biologically trivial differences.

This statistical approach has been shown to identify useful group differences in gene expression that are not detected by the type of traditional group comparisons used in genomewide association studies.[14]

Results of this first step in the analysis has thus far been negative, that is, although a number of genes were found to differ significantly between any of the 3 syndrome groups and the control group, none proved replicable from the Developmental Sample to the Replication Sample. Specifically, in the comparison between syndrome 1 and the control group in the *Developmental sample* we found 32 target genes significantly up-regulated in syndrome 1 and 4 significantly down-regulated; in syndrome 2 versus controls, we found 2 genes significantly up-regulated and none down-regulated; and in syndrome 3, we found none to be significantly up- or down-regulated. However, none of these significant group differences was replicated in the *Replication sample*, so they were all rejected, yielding negative results for the first step of the analysis. (See next step for addressing this problem in the Conclusion section below.)

B. Analysis for Group Differences in Upstream Regulator Genes

The second approach of the analysis involved a search for evidence among the expression of many HVE target genes for patterns that indicate abnormal up- or down-regulation of *regulator genes* upstream from the target genes. This type of analysis, performed with the *Ingenuity Pathway Analysis (IPA) software*, is necessary because most often the changes in expression of regulator genes is so small that, although they may alter the expression of many downstream target genes, their own expression is changed too little to be detected. The IPA software system contains constantly updated information from the scientific literature showing the patterns of downstream target gene expression changes produced by up- or down-regulation of all known regulator genes. It also contains powerful pattern-recognition algorithms that recognize patterns of expression of target genes that identify specific alterations of normal expression of regulator genes. IPA analysis is explained more fully at http://www.ingenuity.com.

Step 1. Identifying gene subgroups to control for regulator gene instability. One possible reason for not finding group differences in gene expression in the initial analyses is that the disease may have been caused by environmental exposures that caused regulator genes to become unstable so that their effects fluctuate back and forth. Our first step in the IPA analysis then was to analyze the gene expression data for evidence of *instability in gene regulation*, that is, for subgroups of subjects within the 4 clinical groups where one subgroup of subjects shows strong up-regulation of a group of target genes and another subgroup shows strong down-regulation of the same target genes. This phenomenon can be seen when some pathology causes instability of a regulator gene so that its function fluctuates between up-regulation and down-regulation. In such a circumstance, a study of blood drawn at a single point in time will show approximately half the subjects with up-regulation of the downstream target genes and the other half with down-regulation of the same genes, with some in between. With the effects on gene expression going in opposite directions tends to average away the differences so that no differential effects are seen.

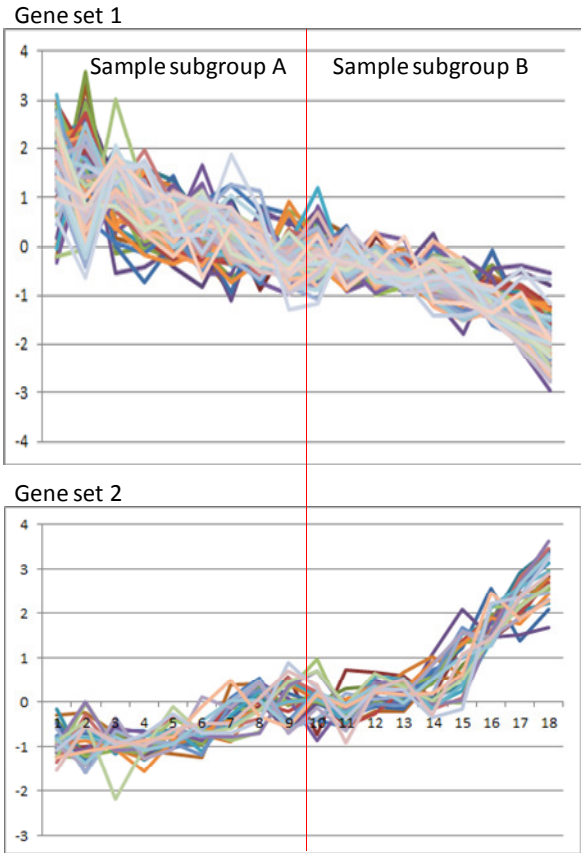


Fig. 1. Normalized residual expression (vertical axis) of genes of the largest cluster in 18 representative syndrome 2 subjects (horizontal axis). The lines represent different genes separated into gene sets 1 and 2. The vertical red line separates sample subgroup A (up-regulated in top set of genes in the cluster and down-regulated in the bottom set of genes in the cluster) and subgroup B (vice versa).

To detect instability of regulator genes, we analyzed the HVE target genes by two cluster analysis techniques: *correlative clustering* and *F-means clustering*.[14] The largest cluster attained a size (number of genes having a similar expression profile across all participating samples) that significantly exceeded that which would have been obtained by chance (according to simulation experiments). This cluster was composed 2 sets of target genes whose expression levels varied inversely; that is, when one of these sets of target genes was up-regulated, the other set tended to be down-regulated, and vice versa (**Figs. 1 and 2**). The samples were then divided into 2 subgroups: samples showing up-regulation of gene set 1 and down-regulation of gene set 2 were classified into Patient Subgroup A; and samples showing down-regulation of gene set 1 and up-regulation of gene set 2 were classified into Patient Subgroup B (**Fig. 1**). The resulting subgroup designation A and B was then introduced into the IPA analyses as the equivalent of an interaction term, thus allowing the identification of genes differentially expressed in either direction due to instability of regulator genes. An example of the results is shown in **Fig. 2**.

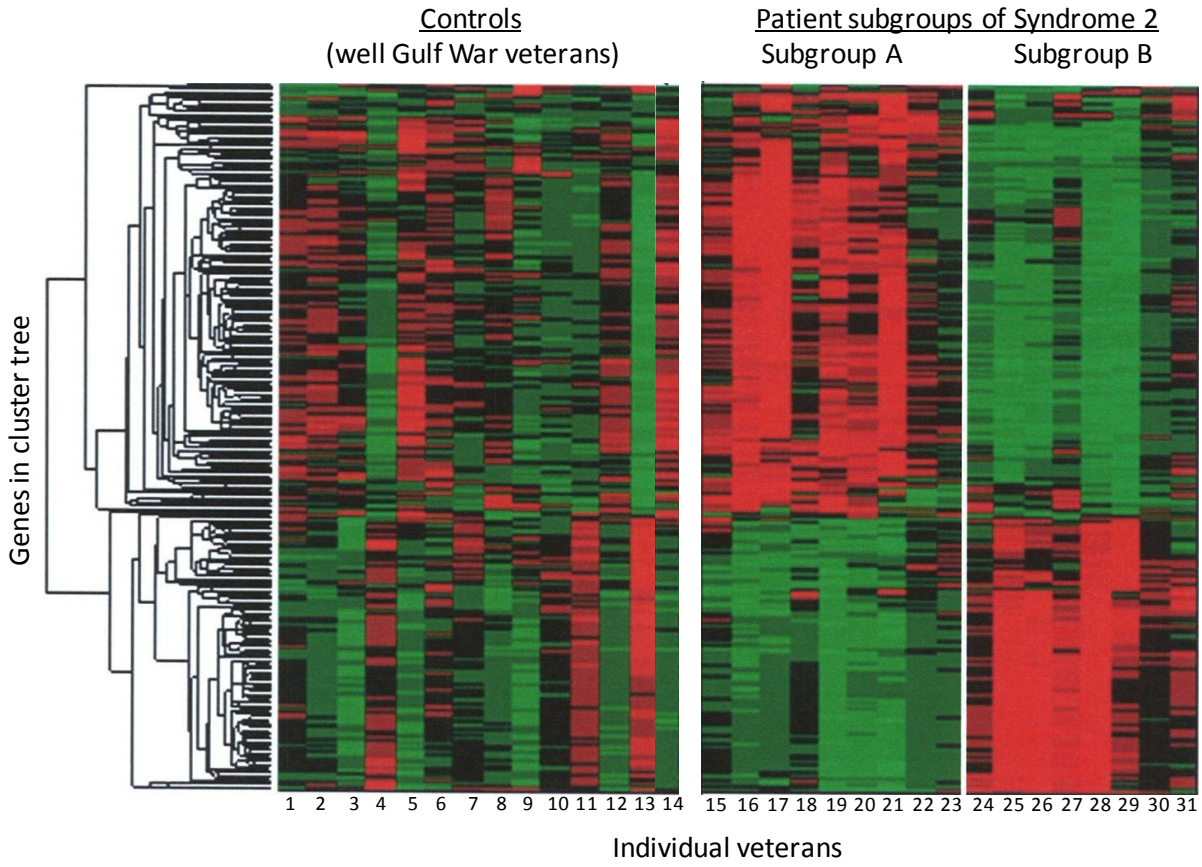


Fig. 2. Heat map of gene expression illustrating the comparative groups resulting from the cluster analysis. The comparative groups in this example include a group of well controls and the syndrome 2 subjects of the Developmental Sample, each stratified into the A and B subgroups on the basis of the predominant direction of residual gene expression (see Fig. 1). The vertical axis lists the genes of the predominant cluster arranged by a dendrogram (tree diagram) generated by the cluster analysis. The heat scale shows gene expression maximally up-regulated in red and maximally down-regulated in green, with black being neither up- nor down-regulated. Subjects are numbered sequentially along the horizontal axis within the 3 comparison groups.

Step 2. Identifying individual differentially expressed target genes within stability-controlled subgroups. Using the same approach as in section IIIA above, we reran the analyses to identify individual HVE target genes differentially expressed between the syndrome 2 subjects in Patient Subgroup A compared with controls and then in Patient Subgroup B compared with controls in the Developmental Sample and accepted those genes as truly differentially expressed that were verified by the same analyses in the Replication Sample (see the significantly differentially expressed genes in the last column in **Table 2** below).

Step 3. Identifying upstream regulator genes potentially responsible for the patterns of differential expression of the individual target genes. The expression levels of target genes identified in Step 2 were entered into the IPA Upstream Regulator Analysis software and analyzed to identify their potential upstream regulator genes. This analysis was run separately for the target genes identified in Patient Subgroups A and B in the Developmental and the Replication samples, thus yielding 4 sets of results. **Table 2** summarizes the regulator genes that

showed significantly different patterns of gene expression in the A and B Patient Subgroups of the syndrome 2 group compared with the control group, which were identified in the Developmental Sample and verified in the Replication Sample. These differences were identified by the following two statistical tests: *Proper Regulation* and *Enrichment* analyses, shown as columns 5 and 6 in Table 2.

Analysis of the **"proper regulation"** of the targets uses a bias-corrected z score to test for *activation* of the regulator gene [indicated by up-regulation (red in Fig. 2) of targets positively regulated by a given regulator gene and down-regulation (green) of target genes negatively regulated by it] or *inhibition* of the regulator gene [indicated by down-regulation (green) of targets positively regulated by a given regulator gene and up-regulation (red) of target genes negatively regulated by it]. The bias correction refers to constraining the z test for the direction of control (activation or inhibition) of particular target gene by a particular regulator gene established in the scientific literature. For example, in the first line of **Table 2**, the bias-corrected z score for identification of the STAT3 regulator gene was z = 1.88 for syndrome 2 versus the control group, indicating that a significant proportion of the target genes for STAT3 were included in our selection of HVE target genes and that most of them showed changes in expression in the direction consistent with reports of STAT3 regulation in the scientific literature.

In the **enrichment method**, we compared the proportion of the target genes within the list of our HVE genes with the proportion in the total list of all genes in the array. For example, in the first line of **Table 2**, the analysis identified a difference in gene expression between the syndrome 2 and the control group of the target genes for the regulator gene, STAT3, with an enrichment p value of p = 1.15E-03.

Of great interest is that the identified regulator genes that replicated in the Developmental and Replication Samples were the same in the A and B subgroups, although they showed opposite directions of control over their target genes (**Table 2**). This finding suggests that environmental exposures rendered these regulator genes unstable, so that they are exerting exaggerated effects that vacillate in direction, the particular direction observed here being merely what was captured in a single observation. If so, this predicts that a second measurement from these same individuals would show the same list of significant regulator genes but randomly differing in direction of the effects—a prediction that can be tested by measuring gene expression again in another blood sample from some of the subjects.

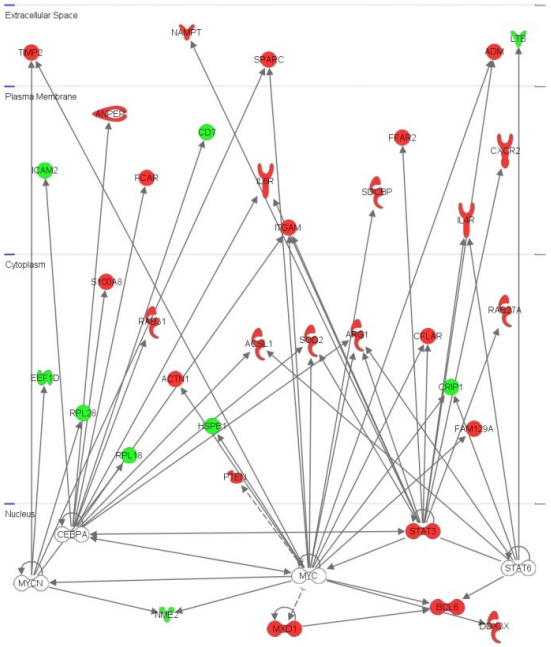
The pathway diagrams shown in **Fig. 3**, panels A-D, portray the relationships between each regulator gene, identified in this analysis, and its downstream target genes.

Table 2. Identification of upstream regulator genes from the patterns of their target genes differentially expressed in patient subgroups of syndrome 2 vs controls

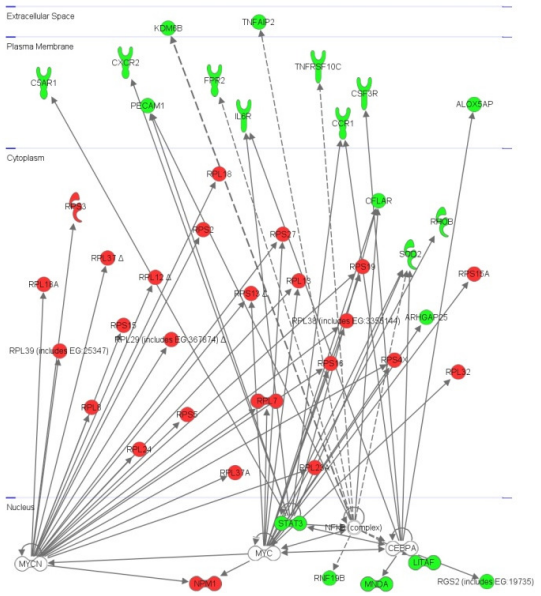
Patient subgroup	Upstream Regulator gene	Molecule Type	Predicted Activation State	Bias-corrected z-score	Enrichment p-value of overlap	Target molecules in dataset
Developmental Sample						
A	STAT3	transcription regulator	Activated	1.883	1.15E-03	ADM, ARG1, CFLAR, CXCR2, FFAR2, IL4R, IL6R, ITGAM (includes EG:16409), NAMPT, RAB27A, SOD2, STAT3
	CEBPA	transcription regulator	Activated	2.270	4.43E-03	ACSL1, ANPEP, ARG1, CD7, FCAR, ICAM2, IL6R, ITGAM (includes EG:16409), RAB31, S100A8, SOD2
	STAT6	transcription regulator	Activated	2.060	3.21E-02	ACSL1, ARG1, BCL6, CRIP1, IL4R, LTB
	MYC	transcription regulator	Inhibited	-2.266	1.45E-02	ACTN1, ADM, ARG1, BCL6, CFLAR, CRIP1, DDX3X, FAM129A, HSPB1, ITGAM (includes EG:16409), NME2, PTEN, SDCBP, SOD2, SPARC, TIMP2 (includes EG:21858)
	MYCN	transcription regulator	Inhibited	-2.698	6.86E-02	EEF1D, NME2, RPL18, RPL28, SPARC, TIMP2 (includes EG:21858)
B	MYCN	transcription regulator	Activated	4.975	2.12E-18	NPM1, RPL12, RPL13, RPL18, RPL18A, RPL23A, RPL24, RPL29 (includes EG:367874), RPL37, RPL37A, RPL38 (includes EG:3355144), RPL39 (includes EG:25347), RPL7, RPL8, RPS13, RPS15, RPS16, RPS19, RPS2, RPS27, RPS3, RPS4X, RPS5
	MYC	transcription regulator	Activated	3.301	1.42E-03	ARHGAP25, CFLAR, NPM1, RHOB, RPL13, RPL32, RPL38 (includes EG:3355144), RPL7, RPS13, RPS15A, RPS16, RPS19, RPS27, RPS4X, SOD2
	STAT3	transcription regulator	Inhibited	-2.133	1.16E-02	C5AR1, CCR1, CFLAR, CXCR2, IL6R, PECAM1, SOD2, STAT3
	CEBPA	transcription regulator	Inhibited	-2.553	1.35E-02	ALOX5AP, CCR1, CSF3R, IL6R, LITAF, MNDA, RGS2 (includes EG:19735), SOD2
	NFkB (complex)	complex	Inhibited	-2.098	4.40E-02	CFLAR, FPR2, KDM6B, LITAF, PECAM1, RNF19B, SOD2, TNFAIP2, TNFRSF10C
Replication Sample						
A	STAT3	transcription regulator	Activated	2.587	2.81E-06	ADM, CFLAR, CXCR2, FFAR2, IL4R, NAMPT, PROK2, SOD2, STAT3
	CEBPA	transcription regulator	Activated		2.91E-02	ACSL1, ANPEP, ICAM2, SOD2
	STAT6	transcription regulator	Activated	1.938	4.11E-03	ACSL1, BCL6, CRIP1, IL4R
	MYC	transcription regulator	Inhibited	-2.280	2.52E-05	ACTN1, ADM, BCL6, CFLAR, CRIP1, FAM129A, NME2, PTEN, SDCBP, SOD2, SPARC
B	MYCN	transcription regulator	Activated	3.582	3.92E-11	RPL18, RPL18A, RPL23A, RPL29 (includes EG:100039782), RPL37, RPL37A, RPL39 (includes EG:100361661), RPL7, RPL8, RPS16, RPS2, RPS27, RPS4X
	MYC	transcription regulator	Activated	1.667	6.88E-03	ARHGAP25, CFLAR, IFI35, RHOB, RPL32, RPL7, RPS16, RPS27, SOD2
	STAT3	transcription regulator	Inhibited	-2.039	4.30E-05	C5AR1, CCR1, CFLAR, CXCR2, IFI35, IL6R, PECAM1, SOD2, STAT3
	CEBPA	transcription regulator	Inhibited	-2.260	4.95E-06	ALOX5AP, CCR1, CSF3R, IL6R, LITAF, MNDA, MT2A, PTAFR, RGS2 (includes EG:19735), SOD2
	NFkB (complex)	complex	Inhibited	-2.361	3.30E-03	CFLAR, FPR2 (includes EG:100426968), KDM6B, LITAF, PECAM1, SOD2, TNFAIP2, TNFRSF10C

Note: Other regulator genes (not shown) were statistically significant in the analysis but did not replicate across the developmental and replication samples.

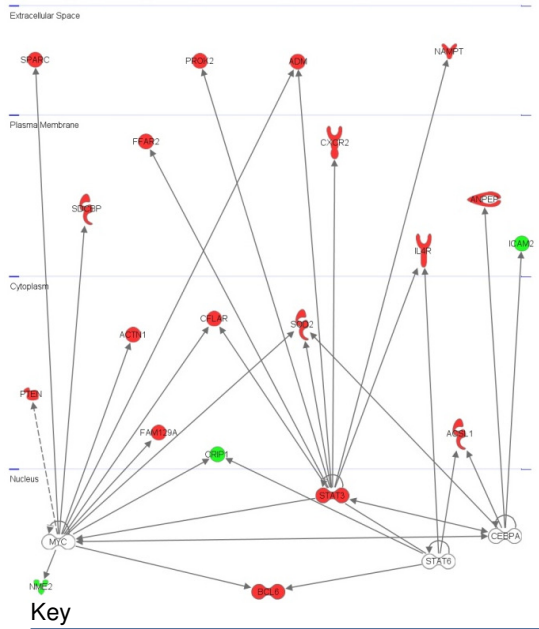
A. Patient subgroup A in Developmental Sample (PaTFd.jpg)



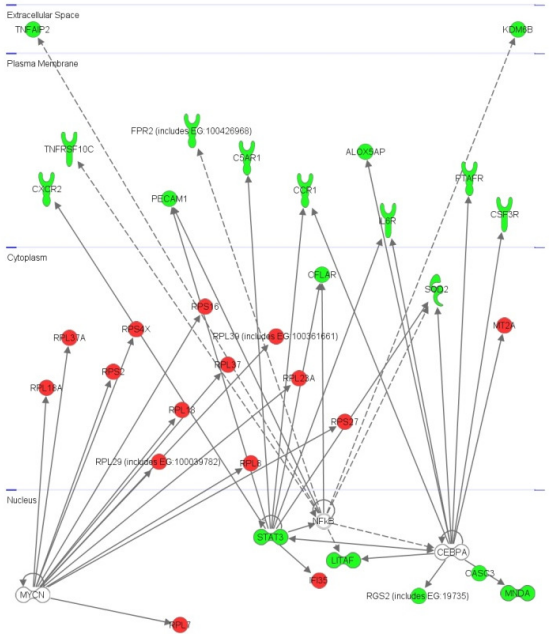
B. Patient subgroup B in Developmental Sample (PbTFd.jpg)



C. Patient subgroup A in Replication Sample (PaaTFd.jpg)



D. Patient subgroup B in Replication Sample (PbbTFdnew.jpg)



Key

Transcription regulator
Enzyme
Complex
G-protein coupled receptor
Transmembrane receptor

Fig. 3. Pathway diagrams showing patterns of downstream genes controlled by regulator genes and found to be differentially expressed within patient subgroups A and B of syndrome 2 compared with the control group in the Developmental Sample and the Replication Sample. Green symbols are up-regulated and red ones are down-regulated. Full size versions of the 4 pathway diagrams are reproduced in **Appendix C**.

KEY RESEARCH ACCOMPLISHMENTS

- Obtained human subjects protection approval from the HSRRB.
- Located the PAXgene blood samples for processing.
- Extracted high quality mRNA from the PAXgene blood samples.
- Produced mRNA sequenced libraries for each of the samples.
- Sequenced the PRMC transcriptome for each sample.
- In the analysis we first identified high variably expressed (HVE) genes and produced a Reference Group of low expressed genes to increase power of multiple comparison-corrected analyses.
- Analyzed the HVE genes for differential expression of single genes that replicate; the initial analysis was negative due to high background variation from variable WBC differential counts.
- Completed cluster analysis that identified 2 groups of subjects with mirror-image gene expression patterns, suggesting instability of upstream regulator genes in the Syndrome 2 group.
- Completed IPA Upstream Regulator Analysis of the Syndrome 2 group vs controls that identified in the Development Sample apparent differential function of 6 regulator genes from the gene expression levels of downstream target genes known to be controlled by these regulators and reproduced the findings in the Replication Sample.
- Began the literature reviews to interpret these findings.
- Have undertaken further analyses to identify differentially expressed single genes by controlling for differences in WBC differential counts.

REPORTABLE OUTCOMES

- Created a large RNA-seq database containing the sequencing information on all PBMC genes in the 144 research subjects studied
- Created a gene expression database containing the gene expression levels of all PBMC genes in the 144 research subjects studied.
- Submitted a new grant proposal to CDMRP to increase the sensitivity of gene expression analysis by collecting LPS- and acetylcholine-stimulated blood samples from the same groups of subjects in the present study and separating the different types of WBCs for separate RNA-seq analysis.

CONCLUSION

Under DoD funding for this project, we successfully completed the laboratory work, yielding excellent quality mRNA and sequencing data and performed the planned analyses The IPA Upstream Regulator Analysis identified patterns of differential up- and down-regulation between the syndrome 2 group and controls that implicate instability in several transcription regulator genes in the syndrome 2 group of subjects, most prominently the STAT3 regulator gene. These group differences survived the multiple comparisons correction, were replicable between our Development and Replication samples, and closely conformed to the regulator-target gene networks described in the literature. The knowledgebase of the IPA software, which reflects a thorough synthesis of scientific literature, indicates that the implicated regulator-target pathways have been shown in past studies to be involved in general inflammation, neurodegeneration, brain ischemia, neuronal injury, encephalomyelitis, post-infective chronic fatigue syndrome, and neuroinflammation—all neuropathologic conditions with similarities to pathologic processes thought to underlie GWI.

Next Steps. The lack of significant, replicable group differences in expression of single genes may be due to any or a combination of the following: 1) the mRNA extracted from the PAXgene tubes in our sample of subjects contains a very high degree of background variability of gene expression due to wide inter-subject variation in the WBC differential count; 2) instability of regulator genes produces fluctuating up- and down-regulation; or 3) the lack of genomic correlates of Gulf War illness.

We have noticed a high degree of background variation in gene expression in these PBMCs from peripheral blood samples—greater than we would see, say, in gene expression data from tissue samples—that is likely to be obscuring subtle group differences. This excessive background variation is probably due mostly to the fact that PBMCs represent a mixture of RNA from the 12 different cell types of WBCs (e.g., B lymphocytes, T lymphocytes, monocytes, neutrophils, etc.) found in routinely collected whole blood and the wide inter-individual differences in the distribution of these cell types (WBC differential count). After finding these negative results in the initial analysis, we designed 2 ways of overcoming the problem.

First, due to concerns that the large differences in WBC differential counts among the subjects increased variation and reduced the power of the analysis, we performed initial re-analyses of the gene expression data, this time standardizing the expression values by the subjects' WBC differential distributions. Our initial attempts at this identified 10-15 genes that survive the multiple-comparisons correction and are replicable between the two samples, but the funding for this award was exhausted before we could complete this analysis. We plan, however, to pursue this with non-DoD funds after the expiration of this grant.

Second, in future studies this can be definitively overcome, and far more powerfully, by stimulating the whole blood with lipopolysaccharide (LPS) or acetylcholine (ACh) and then separating the WBCs of the buffy coat into pure suspensions of individual cell types, such as lymphocytes, neutrophils or monocytes, before stopping RNA synthesis and extracting the RNA for sequencing. Pharmacologic stimulation has been shown to identify group differences in gene expression not demonstrable without stimulation,[17] and isolation of pure cell types has also been shown to increase the ability to show group differences.[18,19] Since there were insufficient resources in the present grant budget to undertake this step, we proposed this approach in a new grant submission in the 2012 round of grant solicitations from CDMRP. Given the findings we have made in this phase of the study, we are optimistic that combining pharmacologic stimulation and isolation of pure cell types will identify group differences in gene expression

among the 3 syndrome groups and the control group that can be used to make objective clinical diagnoses of Gulf War illness variants. Since our proposal for funding this additional approach was not funded, we plan to pursue it with non-DoD funding.

"So What Section." Our provisional finding of an apparent instability in the STAT3 regulatory gene in the Syndrome variant 2 group vs controls, which proved replicable across the 2 study samples, suggests an explanation for the higher rate of brain cancer in Gulf War veterans compared with non-deployed veterans.[20,21] Specifically environmental exposure to chemical toxicants, particularly low-level nerve agents,[22,23] may have damaged the STAT3 gene or other genes that regulate or stabilize its function. STAT3 is a well studied oncogene, linked to brain cancer among many,[24] and thus, instability in its function could increase the risk of developing brain cancer. Were this to be confirmed, ongoing research into countering the oncogenic effects of STAT3[24] could lead to prevention or treatment of brain cancers in Gulf War veterans and chemically exposed military personnel in general.

The lack of success of our initial analyses to identify a gene, or group of genes, whose expression levels distinguish cases from controls and differentiate among the 3 syndrome variant groups indicates the need for further statistical analyses of this dataset, applying additional approaches to identify discriminant functions that might serve as diagnostic tests for Gulf War illness. Failing that, new studies should be undertaken with pharmacologic stimulation of gene expression to increase the chances of successful discrimination.

REFERENCES

1. VA Research Advisory Committee on Gulf War Veterans' Illnesses. *Gulf War illness and the health of Gulf War veterans*. Washington, DC: Department of Veterans Affairs; 2008.

2. Haley RW, Kurt TL, Hom J. Is there a Gulf War Syndrome? Searching for syndromes by factor analysis of symptoms. *JAMA.* 1997;277:215-222.

3. Haley RW, Maddrey AM, Gershenfeld HK. Severely reduced functional status in veterans fitting a case definition of Gulf War syndrome. *Am J Publ Health.* 2002;92:46-47.

4. Iannacchione VG, Dever JA, Bann CM, et al. Validation of a research case definition of Gulf War illness in the 1991 U.S. military population. *Neuroepidemiol.* 2011;37:129-140.

5. Haley RW, Kurt TL. Self-reported exposure to neurotoxic chemical combinations in the Gulf War. A cross-sectional epidemiologic study. *JAMA.* 1997;277:231-237.

6. Haley, RW, Kramer, GL, Xiao, J, Teiber, JF. Nerve agent exposure associated with Gulf War encephalopathy through gene-environment interaction with the Q192R polymorphism of *Paraoxonase 1 (PON1)*. 49th Annual Meeting of the Society of Toxicology, Salt Lake City, UT, March 7-11, 2010; Neurotoxicity and Neurodegenerative Disease Session, abstract no. 2242.

7. Haley, RW, Charuvastra, E, Shell, WE, et al. *Cholinergic autonomic dysfunction in veterans with Gulf War illness: confirmation in a population-based sample*. 2012; published online first, Nov. 26. Archives of Neurology. Accessed September 20, 2013.

8. Tillman GD, Green TA, Ferree TC, et al. Impaired response inhibition in ill Gulf War veterans. *J Neurol Sci.* 2010;297:1-5.

9. Tillman GD, Calley CS, Green TA, et al. Event-related potential patterns associated with hyperarousal in Gulf War illness syndrome groups. *Neurotoxicol*. In press.

10. Li X, Spence JS, Buhner DM, et al. Hippocampal dysfunction in Gulf War veterans: investigation with ASL perfusion MR imaging and physostigmine challenge. *Radiol.* 2011;261:218-225.

11. Gopinath K, Gandhi P, Goyal A, et al. FMRI reveals abnormal central processing of sensory and pain stimuli in ill Gulf War veterans. *Neurotoxicol.* 2012;33:261-271.

12. Haley RW, Hom J, Roland PS, et al. Evaluation of neurologic function in Gulf War veterans. A blinded case-control study. *JAMA.* 1997;277:223-230.

13. Dozmorov I, Lefkovits I. Internal standard-based analysis of microarray data. Part 1: analysis of differential gene expressions. *Nucleic Acids Research.* 2009;37:6323-6339.

14. Dozmorov IM, Jarvis J, Saban R, et al. Internal standard-based analysis of microarray data2--analysis of functional associations between HVE-genes. *Nucleic Acids Research.* 2011;39:7881-7899.

15. Dozmorov I, Knowlton N, Tang Y, Shields A, Pathipvanich P, Jarvis JN, Centola M. Hypervariable genes--experimental error or hidden dynamics. *Nucleic Acids Research.* 2004;32:e147.

16. Dozmorov MG, Guthridge JM, Hurst RE, Dozmorov IM. A comprehensive and universal method for assessing the performance of differential gene expression analyses. *PLoS ONE.* 2010;5:2010.

17. Spijker S, van de Leemput JC, Hoekstra C, Boomsma DI, Smit AB. Profiling gene expression in whole blood samples following an in-vitro challenge. *Twin Res.* 2004;7:564-570.

18. Jarvis JN, Petty HR, Tang Y, et al. Evidence for chronic, peripheral activation of neutrophils in polyarticular juvenile rheumatoid arthritis. *Arthritis Research & Therapy.* 2006;8:R154.

19. Wang A, Guilpain P, Chong BF, et al. Dysregulated expression of CXCR4/CXCL12 in subsets of patients with systemic lupus erythematosus. *Arthrits Rheum.* 2010;62:3436-3446.

20. Bullman TA, Mahan CM, Kang HK, Page WF. Mortality in US Army Gulf War veterans exposed to 1991 Khamisiyah chemical munitions destruction. *Am J Publ Health.* 2005;95:1382-1388.

21. Barth SK, Kang HK, Bullman TA, Wallin MT. Neurological mortality among U.S. veterans of the Persian Gulf War: 13-year follow-up. *Am J Ind Med.* 2009;52:663-670.

22. Tuite JJ, Haley RW. Meteorological and intelligence evidence of long-distance transit of chemical weapons fallout from bombing early in the 1991 Persian Gulf war. *Neuroepidemiol.* 2012;40:160-177.

23. Haley RW, Tuite JJ. Epidemiologic evidence of health effects from long-distance transit of chemical weapons fallout from bombing early in the 1991 Persian Gulf war. *Neuroepidemiol.* 2013;40:178-189.

24. de la Iglesia N, Puram SV, Bonni A. STAT3 regulation of glioblastoma pathogenesis. *Current Molecular Medicine.* 2009;9:580-590.

Appendix A.

Two published scientific articles explaining the validation of the case definition of Gulf War illness

Methods in Neuroepidemiology

Neuro
epidemiology

Neuroepidemiology 2011;37:129–140
DOI: 10.1159/000331478

Received: June 7, 2011
Accepted: August 3, 2011
Published online: October 7, 2011

Validation of a Research Case Definition of Gulf War Illness in the 1991 US Military Population

Vincent G. Iannacchione[a] Jill A. Dever[a] Carla M. Bann[a] Kathleen A. Considine[a]
Darryl Creel[a] Christopher P. Carson[a] Heather Best[a] Robert W. Haley[b]

[a]RTI International, Research Triangle Park, N.C., and [b]Division of Epidemiology, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Tex., USA

Key Words
Epidemiological methods · Case definition · Diagnostic criteria · Confirmatory factor analysis · Health surveys · Persian Gulf syndrome · Validation studies

Abstract
Background: A case definition of Gulf War illness with 3 primary variants, previously developed by factor analysis of symptoms in a US Navy construction battalion and validated in clinic veterans, identified ill veterans with objective abnormalities of brain function. This study tests prestated hypotheses of its external validity. ***Methods:*** A stratified probability sample (n = 8,020), selected from a sampling frame of the 3.5 million Gulf War era US military veterans, completed a computer-assisted telephone interview survey. Application of the prior factor weights to the subjects' responses generated the case definition. ***Results:*** The structural equation model of the case definition fit both random halves of the population sample well (root mean-square error of approximation = 0.015). The overall case definition was 3.87 times (95% confidence interval, 2.61–5.74) more prevalent in the deployed than the deployable nondeployed veterans: 3.33 (1.10–10.10) for syndrome variant 1; 5.11 (2.43–10.75) for variant 2, and 4.25 (2.33–7.74) for variant 3. Functional status on SF-12 was greatly reduced (effect sizes, 1.0–2.0) in veterans meeting the overall and variant case definitions. ***Conclusions:*** The factor case definition applies to the full Gulf War veteran population and has good characteristics for research.

Introduction

A substantial proportion, perhaps 25% [1], of veterans of the 1991 Persian Gulf War continue to experience a pattern of symptoms that has become known as 'Gulf War illness'. Initial investigations of ill soldiers in units reporting high rates of illness by military medical teams soon after the war documented a list of symptoms, most prominently chronic fatigue, memory/attention problems, personality change and body pain, which began during or soon after the war. Finding little evidence of diagnosable physical or psychiatric illness, including posttraumatic stress disorder, initial medical investigations were unable to define the illness and thus drew no clear associations with environmental conditions in the war [2, 3]. Subsequent medical examinations of tens of thousands of veterans in military and US Department of

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0251–5350/11/0372–0129$38.00/0

Accessible online at:
www.karger.com/ned

Prof. Robert W. Haley, MD
Division of Epidemiology, Department of Internal Medicine
University of Texas Southwestern Medical Center
5323 Harry Hines Blvd, Dallas, TX 75390-8874 (USA)
Tel. +1 214 648 3075, E-Mail Robert.Haley@UTSouthwestern.edu

Veterans Affairs (VA) Persian Gulf War registries likewise yielded no case definition of the illness and thus no evidence of etiology [4].

Subsequently, a study in a US Navy reserve battalion that served in the Gulf War used principal components factor analysis to identify unique symptom patterns suggesting at least 3 primary syndrome variants comprising a definable Gulf War illness [5]. The syndrome variants were strongly associated with different sets of self-reported environmental exposures [6]; objective testing identified different patterns of altered brain biochemistry and function associated with the syndrome variants [7], and confirmatory factor analysis of the survey questionnaire reproduced the factor structure in a replication sample of primarily US Army veterans [8]. Since the case definition was developed and validated in relatively small samples, however, the acceptance of the case definition has been limited by questions of external validity.

This article describes the design, implementation and primary findings of the US Military Health Survey (USMHS), a computer-assisted telephone interview (CATI) survey designed to test prestated validation hypotheses in a large statistically representative sample of the US military population at the time of the 1991 Gulf War. Evidence that would support its validity would include finding a good fit of the latent factor structure to the symptom data of the Gulf War veteran population, low prevalence of veterans meeting the case definition in the nondeployed military population and substantially higher prevalence in the deployed population, and a strong inverse association with measures of health-related quality of life.

Materials and Methods

The Factor Case Definition

To enable research, one of the authors (R.W.H.) developed a survey questionnaire of typical symptoms of ill Gulf War veterans soon after the war expressly to derive a case definition [5]. Since the illness resembled many psychiatric diseases in being composed of patterns of symptoms without objective signs or laboratory findings, the survey questionnaire was designed to be analyzed by a two-stage principal components factor analysis to resolve ambiguities in common symptom complaints and detect symptom patterns that might represent illness variants linked to specific environmental exposures. Similar approaches have been used to define psychiatric diseases listed in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition [9]. The investigators administered the questionnaire in controlled group settings to 249 members of a naval reserve battalion deployed to the 1991 Gulf War [3]. The analysis yielded evidence of 6 unique symptom patterns suggestive of syndrome variants, and having any one of these patterns constituted an *overall factor case definition*. With variants 4–6 overlapping variant 2, variants 1–3 were considered *primary syndrome variants* for further study [5].

Syndrome variant 1 ('impaired cognition') was comprised of mild cognitive deficits, including distractibility, forgetfulness, depression and chronic fatigue (daytime sleepiness) – not limiting employment appreciably. Variant 2 ('confusion/ataxia') included reduced intellectual functioning, confusion, vertigo and disorientation, resulting in substantial limitations of employment. Variant 3 ('central neuropathic pain') involved chronic, widespread joint and muscle pains and other sensory abnormalities such as paresthesias and numbness but, as with variant 1, carried little limitation of employment [5].

Epidemiological analysis identified strong associations of the 3 primary syndrome variants with self-reported environmental exposures to different chemical toxins [6]. Repeat administration of the questionnaire to 335 primarily US Army veterans of the 1991 Gulf War replicated the principal component structure, tested by confirmatory factor analysis [8]. Subsequent studies of representative ill veterans and well controls from the naval reserve battalion differentiated the 3 syndrome variants and controls on neuropsychological [10, 11], neurophysiological [10, 12], autonomic [13], brain imaging [7, 14–16] and functional status [17] measures, with abnormalities severest and most widespread in factor syndrome variant 2 (table 1). The 4 subgroups were particularly well differentiated by a discriminant function of changes in regional cerebral blood flow from a cholinergic pharmacological challenge, measured by single-photon emission computed tomography [7] (fig. 1). Of particular note is that the 3 factor syndrome variant groups tended to deviate from the control group in different directions, e.g. syndrome variant 1 being abnormally lower than the controls and syndrome variant 2 higher, so that the composite of the 3 syndrome variant groups would not differ significantly from the controls [7] (fig. 1) – emphasizing the importance of a subclassification of the case definition.

Main Objectives

The USMHS was designed primarily as a confirmatory test of the null hypothesis of no difference in the prevalence rates of the overall factor case definition between the US military personnel deployed to the Kuwaiti Theater of Operations (KTO) during the conflict and those who were medically able but were not deployed to the KTO (the deployable nondeployed). This required estimation of the prevalence of the overall factor case definition, and the individual factor syndrome variant definitions, within a set of predetermined subgroups of interest (reporting domains). The KTO included Saudi Arabia, Iraq, Kuwait, Bahrain, Qatar, United Arab Emirates, and ships in the Persian Gulf. Secondary objectives were to test the fit of a structural equation model of the factor case definition to the survey data, and to assess the association of the case definition with a measure of health-related quality of life.

Sampling Design

The *sampling frame* from which the survey sample was randomly selected was constructed by merging the following two databases:

(1) The Desert Shield/Storm file and Defense Manpower Data Center (DMDC) Operation Mission/Contingency file (Seaside, Calif., USA) contained one record for each person on ac-

Table 1. Results (means with SEM in parentheses) of medical tests from previously published studies showing differences in overall physical functioning and brain function and metabolism among well veteran controls and the 3 primary syndrome variants defined by the factor case definition

Test	Well veteran controls (n = 20)	Primary factor syndrome variants			p value
		1 (n = 5)	2 (n = 12)	3 (n = 5)	
Health-related quality of life (SF-36) [17]					
Physical component summary	56.1 (1.5)	41.7 (2.6)	31.0 (1.7)	34.4 (2.6)	<0.001
Mental component summary	55.1 (2.4)	46.7 (4.3)	31.7 (2.8)	47.0 (4.3)	<0.001
Circadian variation in parasympathetic nervous system activity (night-day difference in high-frequency heart rate variability) [13]: mean, ms^2	90.9 (21.3)	–6.8 (15.9)	22.9 (15.3)	4.3 (19.0)	<0.001
Chemical analysis of deep brain centers (N-acetylaspartate/creatine ratio in right basal ganglia) by ^{1}H-magnetic resonance spectroscopy) [14]: mean ratio	4.08 (0.13)	3.95 (0.24)	3.35 (0.11)	3.90 (0.18)	0.003
Integrity of acetylcholine receptors in brain (response of regional cerebral blood flow to physostigmine challenge measured by SPECT brain scan, least significant interval) [7]: mean difference in rCBF, ml/mg/min	–1.43 (2.93)	–4.30 (5.25)	4.26 (3.5)	–4.56 (5.17)	0.005

Results are data from a nested case-control study of 22 cases and 20 age-sex-education-matched controls selected from a 1995 epidemiological survey of 249 members of a Naval Reserve construction battalion deployed in the combat zone of the 1991 Persian Gulf War [5]. The studies were performed in 1998 with the subjects residing in the General Clinical Research Center at the University of Texas Southwestern Medical Center, Dallas, Tex., USA. p values from 4-group ANOVA. The physical and mental component summary scores were calculated from the 8 SF-36 scales as previously published [53]. Values are T scores with the mean of the 1998 US population approximately 50 and SD 10.

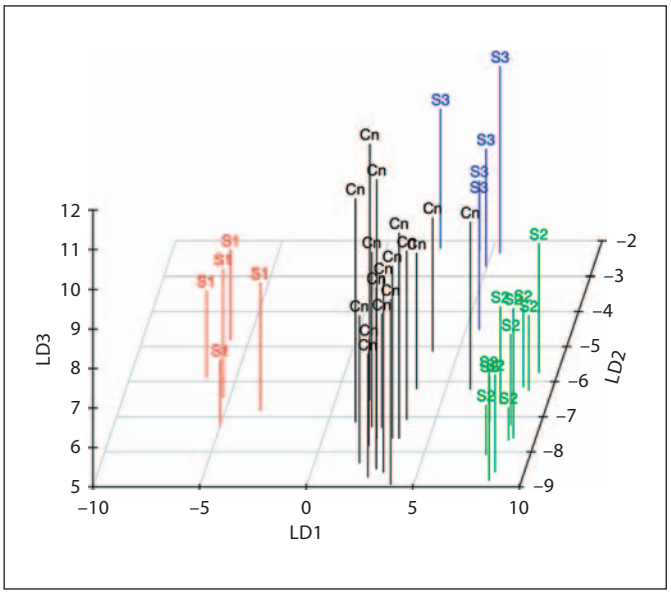


Fig. 1. Results of a previously published [7] linear discriminant analysis to identify a subset of the deep brain regions whose mean normalized regional cerebral blood flow measured by single-photon emission computed tomography scanning under the baseline or physostigmine-stimulated condition would jointly classify subjects into the 4 clinical groups defined by the factor case definition. The discriminant model of normalized regional cerebral blood flow from 17 brain regions from either the baseline session or the physostigmine-stimulated session yielded 3 linear discriminant functions (LD1–LD3) that best classified the subjects. The 3 linear discriminant functions provided clear separation of the 4 groups, with factor syndrome variant 1 in red, factor syndrome variant 2 in green, factor syndrome variant 3 in blue, and the control (Cn) group in black. The 3 primary syndrome variants had brain imaging abnormalities deviating from the control group in different directions so that the composite of the 3 syndrome variant groups would not differ significantly from the controls, emphasizing the importance of a subclassification of the case definition. Reproduced with permission from *Psychiatry Research and Neuroimaging* [7].

tive duty, in the reserves, and in the National Guard on August 2, 1990 [18, 19]. This file included historical data and updated information on characteristics such as decedent status, current military status, and last known residence or duty station.

(2) The US Army Services Center for Unit Records Research database contained records of the geographic location (longitude and latitude) of most military units that served in the Gulf War for each day during the conflict period and afterward.

Under a research protocol approved by the institutional review boards of the US Army and the authors' research institutions, DMDC provided the personally identifying contact information for only the members of the final survey sample. A certificate of confidentiality protecting the privacy of the survey participants was obtained from the National Institute of Environmental Health Sciences prior to the start of data collection.

The *inferential population* for the USMHS comprised all Gulf War era veterans who were living in the 50 United States and Washington, D.C., at the time of data collection and who were physically and mentally able to complete a telephone interview. The inferential population was divided into two subpopulations:

- The *deployed* subpopulation consisted of all active-duty and ready-reserve military personnel (including Coast Guard) who served in the KTO any time from August 2, 1990, through July 31, 1991. This was defined by the binary deployment flag in DMDC's Desert Shield/Storm file updated by a series of deployment questions in the CATI.
- The *deployable nondeployed* subpopulation consisted of the complement of active-duty and ready-reserve personnel serving in August 1990 in any location other than the KTO, excluding persons who were not deployable because of illness *(medically nondeployable)*. Medically nondeployable personnel – identified during the interview from questions on illnesses, other than pregnancy, in the 2 years before the war that precluded deployment – were excluded from the referent population to avoid the 'healthy-warrior effect' [20–23].

Allocation and Selection of the Sample

The sample was allocated to ensure adequate numbers of observations in the domains hypothesized to be associated with symptoms of Gulf War illness. The frame was stratified into 229 sampling strata by crossing the variables in each of the following three major strata.

(1) Not deployed to KTO:

- Age group as of January 1, 2007 (<49, ≥49 years) [5, 18, 24]
- Gender [18, 24]
- Race/ethnicity (non-Hispanic White, other race/ethnicity) [18, 25]
- Military component (active duty, reserve/National Guard) [18, 24, 26–28]
- Military occupation (air flight crew, aircraft maintenance, army special operations, other) [29]

(2) Deployed to KTO:

- Location on January 20, 1991 [18, 24]
- Age group as of January 1, 2007 (<49, ≥49 years) [5, 18, 24]
- Gender [18, 24]
- Race/ethnicity (non-Hispanic White, Black/other) [18, 25]
- Military component (active duty, reserve/National Guard) [18, 24, 26–28]
- Military occupation (air flight crew, aircraft maintenance, army special operations, other) [29]
- Stationed at Camp Doha, Kuwait, between July and November 1991 [30]

(3) Groups for special studies:

- Member of a twin pair (one or both siblings deployed to KTO)
- Member of 24th Reserve Naval Mobile Construction Battalion (Seabees) [5, 24]
- Parent of a child with Goldenhar complex birth defect [31]

To test the confirmatory hypothesis, sample size requirements (before and after attrition) were estimated to detect a difference in syndrome prevalence of 5 percentage points for the domains within the deployed population and 10–15 percentage points for comparisons between deployed and deployable nondeployed domains at a one-tailed significance level of 5% with 80% power. Application of a one-tailed test was justified by the prestated confirmatory purpose of the survey and the findings of all prior surveys of Gulf War era veterans, including the pilot phase of this survey, of higher symptom rates in deployed than nondeployed samples [18, 26–28]. Estimated prevalence rates for the domains used in the sample allocation were obtained from prior studies [32] and based on illnesses with definitions closely associated with components of the factor syndromes, e.g. symptoms of fibromyalgia with prevalence estimates of 18–24% in the deployed populations versus 9–13% in nondeployed veterans.

To allow for the increased efficiency of hypothesis testing with the planned multivariable analysis, a logistic regression analysis of pilot survey data was performed to predict the overall factor case definition adjusting for age, gender, race/ethnicity and active/reserve status resulting in an R^2 of 0.12. Using this result, a compression factor of 0.88 ($1 - R^2$) was applied to the expected variances of the prevalence rates to reflect the expected gain in precision produced by the model [33].

The final allocation of the sample among the strata was optimized with the Sample Planning Tool software developed by RTI for DMDC [34]. This software uses a nonlinear algorithm satisfying the Karush-Kuhn-Tucker necessary conditions [35] for optimally minimizing the variable costs of data collection subject to constraints set on the precision of the key survey estimates. The data collection cost model was expressed as a linear, convex function while the equality constraints were defined with respect to the sample design through a set of concave functions. These parameters provide the sufficient conditions needed to ensure an optimal allocation for the USMHS [36, 37]. In addition to the survey design, the precision was conditioned on the actual stratification affected by unequal stratum weighting (which increases variances of final parameter estimates) and sample inefficiencies associated with nonlocation and nonresponse. After inflating the allocation solution for expected response rates, a stratified random sample of 14,817 Gulf War era veterans was selected (fig. 2). Sample members were selected with equal probabilities within each design stratum. All members of the Seabees battalion and parents of Goldenhar children were selected for the study (certainty strata).

Questionnaire Content

The CATI questionnaire comprised three modules administered to all participants in the following order:

(1) The *Symptoms* module included the questions for constructing the syndrome variant factors and overall factor case definition [5, 8], as well as supplementary symptom information for comparison with other research case definitions (i.e. CDC multisymptom illness [38] and modified Kansas [18] definitions) and similar conditions (e.g. chronic fatigue syndrome, fibromyalgia). Medically nondeployable personnel were identified during the interview from questions on prewar illnesses that had precluded their deployment and were excluded from the referent population to avoid a 'healthy-warrior effect' [20–23].
(2) The *Exposure* module measured environmental and other risk factors related to Gulf War illness. The locations of persons deployed to the KTO were determined to ensure that the effects of exposure to areas with suspected chemical warfare releases could be evaluated. Questionnaire skip patterns were used to avoid asking nondeployed participants questions about exposures encountered only during deployment.
(3) The *Family Issues* module covered health issues of the respondents' families that could possibly have resulted from war-related chemical exposures. For example, questions were included to ascertain the numbers of pregnancies, miscarriages, stillbirths, live births, birth defects and learning disabilities in offspring conceived by or born to the subjects and their partners, as well as problems with infertility.

Data Collection

Interviewing for a pilot survey of 200 veterans randomly selected from the target population occurred in 2005–2006 to test the CATI content and interviewing procedures and to generate parameters used to estimate the main study sample size. Telephone interviewing for the full USMHS ran from May 22, 2007, through April 26, 2009, with a dormant period of no outbound calling between June 1 and October 27, 2008.

Current telephone numbers of sampled veterans were initially sought by batch tracing of name, birth date and social security number through the National Change of Address file and other online resources. Unlocated sample members were periodically traced through an interactive search of multiple open and commercial sources. The final unlocatable veterans were sought with addresses submitted with 2007 US income tax returns provided by the Internal Revenue Service.

Sample members with a locatable address were mailed a packet that included the purpose of the study, the importance of their responses, the voluntary nature of their participation, materials to facilitate the interview, an endorsement letter from the American Legion, the internet address of a project website containing additional background information, a 10-dollar bill, and a promised USD 40 upon completion of the interview. Because federal funding of the survey precluded offering financial incentives to the 6% of sample members still on active military duty, they were offered a study-engraved pen and keychain.

Interviewers, certified after a 4-day training course, contacted and interviewed sample members by telephone, primarily during evening and weekend hours. Average interview times varied from approximately 60 min for the nondeployed nonill to 2.5 h for the deployed ill. All participants were informed about the usual length of the interview and were offered multiple sessions if they became fatigued. Since the symptom questions appeared first in the interview, they should not have been affected by interview length or continuation sessions. Only after initial refusal conversion techniques had failed, such as stressing the importance of the study, were non-active-duty sample members offered an additional USD 25 for a total of USD 65 to complete the survey.

Table 2. Effective sample sizes required for testing the association of the overall factor case definition with deployment

Reporting domain	Deployable nondeployed			Deployed		
	target	actual	difference	target	actual	difference
Males[1]	111	591	480	111	2,903	2,792
Females[1]	200	174	–26	200	539	339
Age <49[1]	111	469	358	111	1,123	1,012
Age ≥49[1]	200	383	183	200	534	334
Non-Hispanic white[1]	111	537	426	111	819	708
Black/other[1]	298	304	6	298	344	46
Active duty[1]	111	422	311	111	2,597	2,486
Reservists[1]	200	374	174	200	1,027	827
Flight crews[2]	46	56	10	46	153	107
Aircraft maintenance[2]	46	58	12	46	198	152
Army special forces[2]	46	49	3	46	86	40

Effective sample sizes: required for comparisons of the overall factor case definition between deployed and deployable nondeployed reporting domains at the 0.05 one-tailed significance level with 80% power; the target effective sample sizes were based on the USMHS Pilot Survey.

[1] Detectable difference of 10%. [2] Detectable difference of 15%.

Statistical Power

A total of 8,020 persons, who completed at least the *Symptoms* module, were included in the analysis file, constituting an overall response rate of 60.1%, using the AAPOR response rate RR4 definition [39] (fig. 2). The effective number of respondents required to test the main hypothesis, as estimated during the design phase of the study, was exceeded in all domains except for deployed females (table 2). (The effective sample size is the estimated sample size adjusted for unequal weighting. It can be thought of as the sample size equivalent to that drawn by simple random sampling.) In spite of the shortfall, the desired power to detect a difference in the overall factor case definition between the deployed and nondeployed females was achieved because of the larger than expected number of nondeployed female respondents.

Construction of Analysis Weights for Bias Reduction

We developed an analysis weight variable to correct bias from nonproportional sampling of strata and from inability to locate (nonlocation) or obtain participation (noncooperation) from veterans selected into the sample. The ability to locate and to obtain consent for study participation from subjects exploits different processes in most household interview surveys [40]. The USMHS process for locating veterans was largely dependent on the success of the tracing activities. By contrast, the likelihood of participa-

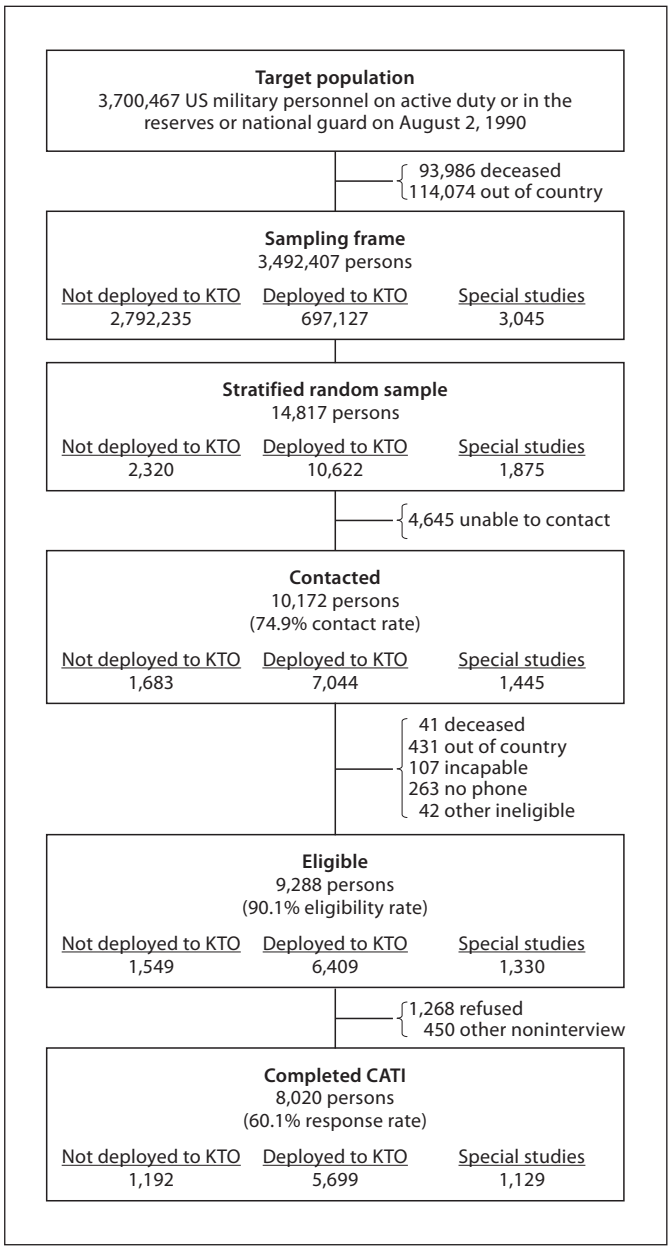


Fig. 2. Sample selection process for the USMHS. 'Not deployed to KTO' includes medically nondeployable personnel. 'Special studies' included twin pairs, members of the 24th Reserve Naval Construction Battalion (Seabees) and parents of children with Goldenhar complex. Counts for subgroups are suppressed to maintain confidentiality according to terms of the Certificate of Confidentiality. The 'contact rate' includes in the base the number of known eligible cases and the estimated number of eligible cases among the undetermined cases. The 'eligibility rate' is among sample members with known survey eligibility. The 'response rate' is the American Association for Public Opinion Research Response Rate 4 (RR4) and includes in the base the estimated number of eligible cases among those initially selected for the CATI phase of the study [39].

tion was likely affected by numerous factors including the sampled veterans' military experience. As a result, we constructed survey analysis weights by combining adjustment factors for the sampling design, nonlocation and nonparticipation by the following five-step process [41].

Step 1. A survey design weight was calculated for each sample member as the inverse of the selection probability within the respective design stratum. The design weights had the following form:

$$d_{hi} = \frac{N_h}{n_h} \tag{1}$$

where n_h is the number of sample members selected within stratum h (h = 1, …, 229) and N_h is the total number of veterans on the sampling frame within stratum h. The design weight d_{hi} is the same for every sample member in the same sampling stratum.

Step 2. The design weights were first adjusted to minimize bias associated with nonlocation using adjustment classes defined by a classification and regression trees algorithm [42] and applied to data available on both located and nonlocated sample veterans. The classes were formed using variables such as service, service component and age group in addition to certain paradata variables (the complete list of variables has been withheld in accordance with conditions in the Certificate of Confidentiality). The adjustments ($\hat{\lambda}_{hi}$) can be written in terms of a logistic model containing the classification and regression tree adjustment classes:

$$\hat{\lambda}_{hi} = P\left[L_{hi} = 1 | d_{hi}, \mathbf{X}_{hi}, \hat{\boldsymbol{\beta}}_1\right] = \left[1 + \exp\left(d_{hi}\, \mathbf{X}'_{hi}\, \hat{\boldsymbol{\beta}}_1\right)\right]^{-1}$$

where d_{hi} is the design weight given in equation 1, $\mathbf{X}_{hi}$ ($l \times 1$) is a vector of indicator variables that identify membership in one of the l adjustment classes, $\hat{\boldsymbol{\beta}}_1$ ($l \times 1$) is a vector of estimated model parameters, and L_{hi} = 1 if sample member hi was located (zero otherwise). The resulting location-adjusted analysis weight is written as:

$$w_{1hi} = \begin{cases} d_{hi}\, \hat{\lambda}_{hi} & \text{for located sample members} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Step 3. A subsequent adjustment was applied to the weights to address any potential nonparticipation bias among sample members who were contacted. Procedures similar to those discussed in step 2 resulted in the following weight adjustment:

$$\hat{\rho}_{hi} = P\left[P_{hi} = 1 | w_{1hi}, \mathbf{Z}_{hi}, \hat{\boldsymbol{\beta}}_2\right] = \left[1 + \exp\left(w_{1hi}\, \mathbf{Z}'_{hi}\, \hat{\boldsymbol{\beta}}_2\right)\right]^{-1}$$

where w_{1hi} is the adjusted weight given in equation 2, $\mathbf{Z}_{hi}$ ($p \times 1$) is a vector of indicator variables that identify membership in one of the p adjustment classes, $\hat{\boldsymbol{\beta}}_2$ ($p \times 1$) is a vector of estimated model parameters, and P_{hi} = 1 if sample member hi was located *and* participated in the USMHS (zero otherwise). The weight adjusted for nonlocation and nonparticipation was then computed as:

$$w_{2hi} = \begin{cases} w_{1hi}\, \hat{\rho}_{hi} & \text{for located sample members who participate in the USMHS} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Step 4. Extreme values of w_{2hi} (falling outside the median ± 2 × interquartile range) were trimmed to ensure that excessive variation in the weights would not unnecessarily degrade the precision of the survey estimates.

Step 5. The trimmed weights were ratio adjusted to sum to the number of persons on the sampling frame within the key reporting domains (table 2). The final USMHS analysis weight was constructed as:

$$w_{3hi} = w_{2hi}\, a_{1hi}\, a_{2hi} \quad (4)$$

where a_{1hi} is the weight trimming adjustment discussed in step 4 ($a_{1hi} = 1$ for most responding sample members), and $a_{2hi} = f(w_{2hi}\, a_{1hi})$, the poststratification adjustment calculated as a function of the trimmed weights.

Classification of Syndromic versus Nonsyndromic

The 6 syndrome factor scales were generated by summing the responses to symptom questions multiplied by the scoring weights from the original exploratory factor analysis [5, 8]. The resulting factor scales were then dichotomized, as before, at 1.5 standard deviations where values 1.5 and above were classified as *syndromic* [5, 8]. The cutpoint of 1.5 standard deviations, originally selected from inspection of the scale distributions only [5, 8], was later found to identify syndromic groups with severer functional disability on the Medical Outcomes Study 36-Item Short Form (MOS SF-36) physical component score and mental component score [17]. Persons met the overall factor case definition if they met any of the 6 dichotomized component definitions.

Confirmatory Factor Analysis

A cross-validation approach was used to split the sample (excluding the special samples of twins, Goldenhar parents and Seabees) into two halves. The halves were selected using stratified random sampling to ensure adequate representation of the sampling strata in each half. Confirmatory factor analysis, performed with the M-Plus software [43] incorporated the sampling strata and analysis weights.

Statistical Analyses

Individual subjects' scores on the MOS SF-12 physical and mental summary scales (version 1) were calculated from the 12 questionnaire items by the standard Medical Outcome Trust's scoring algorithm, using the May 2006 SAS program (Janel Hanmer), checked against scores published in the 2001 nationally representative sample of the noninstitutionalized general US population in the Medical Expenditure Panel Survey (http://meps.ahrq.gov) [44]. Statistical analyses of the survey data were performed with SUDAAN© programs [45] allowing for the complex stratified random sampling design and applying the analysis weights to adjust for unequal selection probabilities and minimize nonresponse bias. Odds ratios and standard errors of the association of the case definition with deployment were obtained with SUDAAN proc rlogist, and mean SF-12 summary scores in clinical groups defined by the case definition were obtained with SUDAAN proc regress – both analyses controlling for age, gender and race/ethnicity. Seventy-one nondeployed respondents, found by their questionnaire responses to have been medically nondeployable, were excluded from all analyses to avoid bias from the 'healthy-warrior effect' [20–23].

Results

External Validity of the Factor Case Definition

The confirmatory factor analysis found that a parsimonious structural equation model, previously developed to express the 3 primary syndrome factors and a second-order overall Gulf War illness [8] (fig. 3), fit the two random split halves of the survey database (n = 3,408 each) well and showed invariance of fit (forced equal loadings) across the two halves (table 3). The goodness-of-fit statistics indicated that the model fit the two random halves of the survey database as well as the same model fit a previous validation sample of US Army veterans recruited from the clinic of a VA medical center [8] (table 3).

Association with Deployment

The prevalence of illness by the overall factor case definition was 3.98% in the nondeployed and 13.59% in the deployed, for a deployment odds ratio adjusted for age, gender and race/ethnicity of 3.87 (95% confidence interval, 2.61–5.74). The deployment odds ratio was highest for factor syndrome variant 2 among the 6 individual factor syndrome variant case definitions (table 4). The rate of the overall case definition was significantly greater in the deployed than the deployable nondeployed in all groups studied except those serving as Air Force aircraft maintenance (table 4).

Association with Functional Status

Even though the factor analysis detected nonrandom symptom patterns without regard to severity of illness, deployed veterans meeting the overall case definition and its component syndrome variant case definitions had significantly lower mean functional status, adjusted for age, sex and race/ethnicity and the analysis weights, than the nonsyndromic veterans on the SF-12 physical component and mental component summary scales [46] (table 5). The effect sizes (difference from the nonsyndromic group divided by the standard deviation of that group) of all groups meeting the case definition ranged between 1.0 and 2.0, indicating very large losses of health-related quality of life in both physical and mental functioning [47] (table 5).

Discussion

The findings from this national survey provide evidence supporting the usefulness of the original factor analysis-derived case definition with 3 primary variants

Table 3. Goodness-of-fit statistics for structural equation model of Gulf War illness with 3 first-order factors (syndrome variants) and a second-order factor (overall Gulf War illness)[1], by study and sample within study

Study and sample within study	Sample size	Goodness-of-fit statistics			
		SRMR	RMSEA	CFI	TLI
Criteria for a good fit [49]		≤0.080	≤0.060	≥0.950	≥0.950
Deployed US Navy Seabees battalion (developmental sample) [8]	249	0.043	0.023	0.992	0.988
Deployed US Army veterans (first validation sample) [8]	335	0.043	0.044	0.975	0.964
USMHS[2]					
Random half 1	3,408	0.054	0.018	0.968	0.954
Random half 2	3,408	0.048	0.017	0.972	0.967
Both halves combined	6,816	0.048	0.017	0.970	0.958
Forced equal loadings across both halves	6,816	0.054	0.015	0.972	0.967

SRMR = Standardized root mean-square residual, an absolute fit index, analogous to R^2 for a linear model, and the most sensitive to misspecification of factor covariances or latent structures; the remaining 3 fit indexes are most sensitive to misspecification of factor loadings [49]. Hu and Bentler [49] reported that the combination of SRMR >0.09 and root mean-square error of approximation >0.06 for rejection results in the least sum of type I and type II model rejection errors; RMSEA = root mean-square error of approximation, an absolute fit index that adjusts fit by the number of model parameters estimated to prevent large complex model structures from inflating the fit [49]; CFI = comparative fit index, a type 3 incremental fit index that estimates the improvement in fit over a baseline null model where all measured variables are uncorrelated [49]; TLI= Tucker-Lewis index (also Bentler-Bonett nonnormed fit index), a type 2 incremental fit index [49].

[1] Corresponds to model 3 developed in the earlier confirmatory factor analysis study [8], pictured in figure 3. [2] All results are population estimates adjusted to correct for unequal selection probabilities and minimize bias from nonlocation and nonparticipation by application of the survey analysis weights.

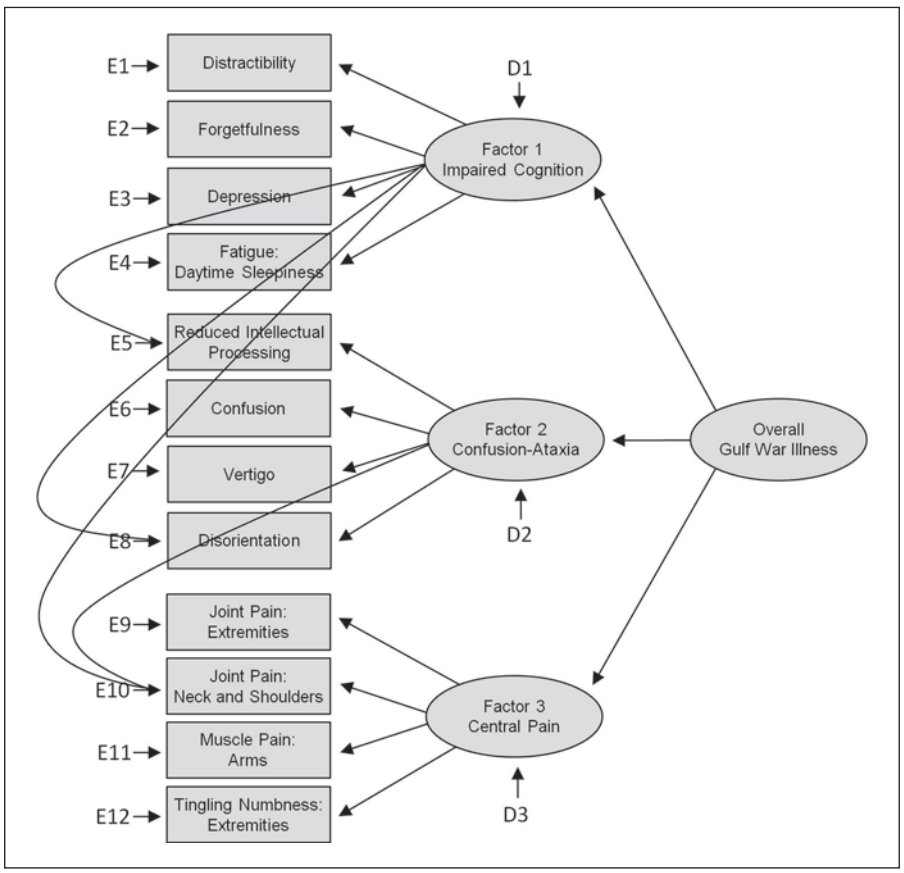


Fig. 3. A structural equation model of Gulf War illness with 3 first-order factors (syndrome variants), each with 4 symptom scales loading on it, and a second-order factor (overall Gulf War illness). The model was developed by a 1995 exploratory factor analysis in 249 members of a US Navy construction battalion [5], validated by a later confirmatory factor analysis in 335 primarily US Army veterans from a VA medical center [8], and in the current study validated with population estimates from the 2007–2009 USMHS, excluding military personnel selected for special studies (see fig. 2). The model shown in the figure corresponds to model 3 validated in the earlier factor analysis study [8].

Table 4. Adjusted odds ratio (aOR) for meeting the case definition in deployed versus deployable nondeployed populations, by case definition and by demographic and military characteristics

Case definition and domain	Percent meeting case definition		aOR	95% CI	Tests of the prestated hypothesis
	deployable nondeployed	deployed			
Overall factor case definition	3.98	13.59	3.87	2.61–5.74	0.001
Factor syndrome variant case definitions					
Syndrome variant 1: impaired cognition	0.59	1.76	3.33	1.10–10.10	0.033
Syndrome variant 2: confusion-ataxia	1.12	6.10	5.11	2.43–10.75	0.001
Syndrome variant 3: central neuropathic pain	1.22	4.58	4.25	2.33–7.74	0.001
Syndrome variant 4: phobia-apraxia	1.41	5.31	3.44	1.75–6.76	0.001
Syndrome variant 5: fever-adenopathy	0.98	1.78	2.06	1.02–4.12	0.042
Syndrome variant 6: weakness-incontinence	0.70	1.20	1.48	0.55–3.94	0.437
Overall factor case definition by domain					
Age					
<49 years	2.72	13.68	5.50	3.09–9.81	0.001
≥49 years	5.88	13.37	2.55	1.49–4.35	0.001
Gender					
Male	3.23	13.40	4.27	2.65–6.90	0.001
Female	7.61	15.37	2.30	1.24–4.27	0.008
Race/ethnicity					
Non-Hispanic White	3.33	9.77	3.56	2.25–5.63	0.001
Black/other	6.78	20.95	4.32	2.10–8.89	0.001
Component					
Active duty	3.23	13.51	4.39	2.43–7.92	0.001
Reserve/guard	4.82	13.91	3.53	2.06–6.06	0.001
Occupation					
Air flight crew	0.35	1.68	10.04	1.87–53.81	0.007
Aircraft maintenance	3.63	6.37	1.56	0.34–7.21	0.568
Army special forces	2.08	17.92	10.15	1.19–86.60	0.034

CI = Confidence interval. All results are population estimates adjusted to correct for unequal selection probabilities and minimize bias from nonlocation and nonparticipation by application of the survey analysis weights. Odds ratios adjusted for age, gender and race/ethnicity using the logistic regression procedure (proc rlogist) in SUDAAN®. The overall factor case definition is defined as satisfying any of the 6 factor syndrome variant case definitions. Factor syndrome variants 1–3 are considered the primary syndrome variants because factor syndrome variants 4–6 overlap strongly with factor syndrome variant 2, which is associated with the greatest reduction in functional status and the severest neuropsychological and neuroimaging abnormalities [7, 10–17].

as a case definition for research on Gulf War illness. That the case definition was originally developed by factor analysis of symptoms in a single battalion was shown to fit well a validation sample drawn from ill Gulf War veterans attending a VA clinic, and was associated with objective tests of illness in similarly small pilot studies suggested its usefulness but left open the questions of its external validity. Our present findings address this question in 3 ways.

First, they show that the structural equation model of the case definition fits well the symptom data collected in our survey of a stratified random sample of the Gulf War era US veteran population. The survey incorporated state-of-the-art survey techniques to ensure a representative sample of Gulf War era veterans from whom to obtain reports of symptoms, exposures and family effects. The distribution of the factor case definition throughout the target population was demonstrated by showing a good fit of the complex syndromic structure to both random halves of the sample with confirmatory factor analysis by structural equation modeling. The quantitative criteria used to indicate a good fit are evidence-based

Table 5. Health-related quality of life measured by the MOS SF-12 physical and mental component summary scores in deployed Gulf War veterans, by the overall factor case definition and its component syndrome variant case definitions

Clinical groups defined by the case definition	SF-12 physical component summary score		SF-12 mental component summary score	
	mean	effect size	mean	effect size
Does not meet the overall case definition	47.5 (0.3)	reference	54.2 (0.9)	reference
Meets the overall factor case definition	34.7 (1.0)	1.3	39.5 (0.9)	1.5
Meets factor syndrome variant case definitions				
Syndrome variant 1: impaired cognition	35.1 (1.6)	1.2	36.9 (1.2)	1.7
Syndrome variant 2: confusion-ataxia	34.9 (1.9)	1.3	33.8 (1.1)	2.0
Syndrome variant 3: central neuropathic pain	32.7 (1.7)	1.5	42.4 (2.3)	1.2
Syndrome variant 4: phobia-apraxia	32.8 (1.3)	1.5	35.7 (1.3)	1.9
Syndrome variant 5: fever-adenopathy	37.6 (1.5)	1.0	43.0 (1.7)	1.1
Syndrome variant 6: weakness-incontinence	31.4 (1.7)	1.6	35.9 (1.9)	1.8

Results are means, with standard error of the mean in parentheses. The two component scores are T scores with a US reference population mean of approximately 50 and a standard deviation of 10. All results are population estimates adjusted to correct for unequal selection probabilities and minimize bias from nonlocation and nonparticipation by application of the survey analysis weights. Least squares mean and standard error of the mean were calculated with the linear modeling procedure of SUDAAN (proc regress) adjusting for age, gender and race/ethnicity, allowing for the complex stratified sampling design, and weighting by the analysis weights. Effect size is the difference in means of the nonill (not meeting the case definition) and ill groups divided by the standard deviation in the nonill group; an effect size of 0.2 is considered clinically a small effect, one of 0.5 a moderate effect and of 0.8 a large effect [47].

thresholds that maximize the rejection of poorly fitting models and the acceptance of good fitting ones [48–50], particularly the combination of the standardized root mean-square residual and root mean-square error of approximation fit indices [49]. Analysis weights incorporating corrections for unequal selection probabilities among strata and for nonlocation and nonparticipation were applied to the structural equation model analyses to facilitate the unbiased estimation of inferential population parameters from the survey sample.

Second, we found the prevalence of veterans meeting the overall case definition to be low in the deployable nondeployed Gulf War era veteran population and approximately 4-fold higher in the deployed force, as would be expected from an illness caused by exposures in the war theater. In making this comparison, we addressed the well-known selection bias from the fact that only the healthiest soldiers are deployed to a war zone ('healthy-warrior effect') [20–23] by omitting from analysis veterans who were nondeployed because of a definable health problem, leaving the deployable nondeployed as the comparable referent group.

Third, the case definition identified groups of deployed Gulf War veterans with greatly reduced functional capacity typical of other serious chronic diseases, as measured by the MOS SF-12 functional status scales [46, 51]. The effect size of the difference between the groups meeting the case definition and those not meeting it was large, 1.0–2.0. According to Cohen's rule of thumb, an effect size of 0.2 is considered a small effect, one of 0.5 a moderate effect and of 0.8 a large effect [47]. The SF-36 is the most widely used multi-item instrument for assessing health-related quality of life, and the SF-12 is a short version containing a 12-item subset that strongly predicts the results of the full SF-36 and that is better suited for large surveys. Application of the full SF-36 in a small sample nested in an epidemiological survey of a single battalion previously found that the ill Gulf War veterans meeting the factor syndrome case definitions had reductions in scores on the physical and mental component summary scores that were both statistically and medically significant [17]. The present study confirms that finding in a representative population sample of deployed Gulf War veterans.

Basing a case definition on empirically derived combinations of symptoms shown to occur uncommonly in the deployable nondeployed population and far more commonly in the deployed population and derived from a relatively high cut point (1.5 standard deviations) on the syndrome variant factor scales, maximizes its specificity,

which, by misclassifying few noncases as cases, is optimal for research on etiology, pathogenesis and treatment. However, high specificity is often accompanied by lower sensitivity, in this instance by excluding cases just below the syndrome variant factor scale cut point and ones with rare or unique symptom patterns. Consequently, this research case definition may not prove optimal for eventual clinical screening, treatment and decision-making on service connection of disabilities, where objective biological measures may be preferable.

To identify such objective measures, we have selected two sequential nested case-control subsamples from the participants in this national survey for more efficient application of expensive clinical research techniques, pursuing hypotheses developed in prior research on smaller convenience samples [7, 10–17]. The first subsample is comprised of all subjects meeting the overall factor case definition or the modified Kansas [18] or CDC [38] case definitions (cases) and a random subsample of those not meeting any case definition (controls). These subjects were contacted and asked to provide a blood sample for banking of serum, plasma, DNA and RNA, primarily for testing gene-environment interactions relevant to inferring the original causes and pathogenetic mechanisms of the illness [25, 52, 53]. The second is a smaller random subsample of the cases and controls who provided a blood sample; they have undergone extensive clinical testing including multimodal brain imaging, high-resolution electroencephalography and other biomarker measurements [15, 16]. The results of these clinical studies, to be described in future articles, should eventually form the basis for an objective, optimally sensitive and specific disease definition for medical use developed in statistical samples of the target population of Gulf War veterans.

Acknowledgements

This work was supported by US Army Medical Research and Materiel Command grants DAMD17-97-2-0725 and DAMD17-01-1-0741 and contract VA549-P-0027 (Robert W. Haley, PI) administered by the US Department of Veterans Affairs Medical Center, Dallas, Tex., USA. The content does not necessarily reflect the position or the policy of the Federal government or the sponsoring agencies, and no official endorsement should be inferred. A large research team of survey specialists from RTI International contributed importantly to the design and performed the field work for the national CATI survey of Gulf War era veterans. Research leaders included Kathleen A. Considine, Vincent G. Iannacchione, Jill A. Dever, Christopher P. Carson, Heather Best, Carla Bann, Darryl Creel, Barbara Alexander, Amanda Lewis-Evans, Lily Trofimovich, Kirk Pate, Anne Kenyon, Jeremy Morton, Craig Hill and Robert E. Mason. E. William Byrd Jr., Michael E. Murray, Helen Koo and a team of RTI staff contributed to the design of reproductive and child development issues of the CATI questionnaire. Rick Thompson, Eric Cordell and Jennifer Escobar provided program management at UT Southwestern.

References

1 Research Advisory Committee on Gulf War Veterans' Illnesses: Gulf War illness and the health of Gulf War veterans. Washington, Department of Veterans Affairs, 2008. http://www1.va.gov/rac-gwvi/.

2 DeFraites RF, Wanat RR, Norwood AE, William S, Cowan D, Callahan T: Investigation of a suspected outbreak of an unknown disease among veterans of Operation Desert Shield/Storm, 123rd Army Reserve Command, Fort Benjamin Harrison, Indiana, April 1992. Washington, Walter Reed Army Institute of Research, 1992. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA301076&Location=U2&doc=GetTRDoc.pdf.

3 Berg SW: Post-Persian Gulf medical findings in military reservists. Presentation to the NIH Technology Assessment Conference on the Persian Gulf experience and health, 27–29 April 1994, Bethesda, MD, 1994. http://www.gulflink.osd.mil/seabee/med_270494.pdf.

4 Joseph SC: A comprehensive clinical evaluation of 20,000 Persian Gulf War veterans: Comprehensive Clinical Evaluation Program evaluation team. Milit Med 1997;162:149–155.

5 Haley RW, Kurt TL, Hom J: Is there a Gulf War syndrome? Searching for syndromes by factor analysis of symptoms. JAMA 1997;277:215–222.

6 Haley RW, Kurt TL: Self-reported exposure to neurotoxic chemical combinations in the Gulf War: a cross-sectional epidemiologic study. JAMA 1997;277:231–237.

7 Haley RW, Spence JS, Carmack PS, Gunst RF, Schucany WR, Petty F, Devous MD Sr, Bonte FJ, Trivedi MH: Abnormal brain response to cholinergic challenge in chronic encephalopathy from the 1991 Gulf War. Psych Res Neuroimag 2009;171:207–220.

8 Haley RW, Luk GD, Petty F: Use of structural equation modeling to test the construct validity of a case definition of Gulf War syndrome: invariance over developmental and validation samples, service branches and publicity. Psychiatr Res 2001;102:175–200.

9 American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, ed 4 (DSM-IV). Washington, American Psychiatric Association, 1994.

10 Haley RW, Hom J, Roland PS, Bryan WW, Van Ness PC, Bonte FJ, Devous MD Sr, Mathews D, Fleckenstein JL, Wians FH Jr, Wolfe GI, Kurt TL: Evaluation of neurologic function in Gulf War veterans: a blinded case-control study. JAMA 1997;277:223–230.

11 Hom J, Haley RW, Kurt TL: Neuropsychological correlates of Gulf War syndrome. Arch Clin Neuropsychol 1997;12:531–544.

12 Roland PS, Haley RW, Yellin W, Owens K: Vestibular dysfunction in Gulf War syndrome. Otolaryngol Head Neck Surg 2000;122:319–329.

13 Haley RW, Vongpatanasin W, Wolfe GI, Bryan WW, Armitage R, Hoffmann RF, Petty F, Callahan TS, Charuvastra E, Shell WE, Marshall WW, Victor RG: Blunted circadian variation in autonomic regulation of sinus node function in veterans with Gulf War syndrome. Am J Med 2004;117:469–478.

14 Haley RW, Marshall WW, McDonald GG, Daugherty M, Petty F, Fleckenstein JL: Brain abnormalities in Gulf War syndrome: evaluation by ^{1}H magnetic resonance spectroscopy. Radiology 2000;215:807–817.
15 Briggs RW, Cheshkov S, Lu H, Li X, McColl RW, Buhner D, Ferree T, Haley RW: Objective brain abnormalities by 3T MRI and EEG in Gulf War illness. 49th Annual Meeting of the Society of Toxicology, Salt Lake City, March 2010, Biomarkers Session, abstract No 2293.
16 Ringe W, Briggs RW, Gopinath K, Kraut M, Rypma B, Odegard T, Bartlett J, Crosson B, Hart J, Haley RW: Functional neuroimaging shows abnormalities in brain function underlying symptoms of Gulf War illness. 49th Annual Meeting of the Society of Toxicology, Salt Lake City, March 2010, Biomarkers Session, abstract No 2294.
17 Haley RW, Maddrey AM, Gershenfeld HK: Severely reduced functional status in veterans fitting a case definition of Gulf War syndrome. Am J Publ Health 2002;92:46–47.
18 Steele L: Prevalence and patterns of Gulf War illness in Kansas veterans: association of symptoms with characteristics of person, place, and time of military service. Am J Epidemiol 2000;152:992–1002.
19 Poblete PP, Araneta MRG, Sato PA, Hiliopoulos KM, Kamens DR, Morn CB, Zau AC, Gray GC: National study on reproductive outcomes: a reliability study of self-administered survey vs telephone interview. Conference of Federally Sponsored Gulf War Veterans' Illnesses Research, Washington, June 1998, p 46.
20 Haley RW: Point: bias from the 'healthy-warrior effect' and unequal follow-up in three government studies of health effects of the Gulf War. Am J Epidemiol 1998;148:315–323.
21 Larson GE, Highfill-McRoy RM, Booth-Kewley S: Psychiatric diagnoses in historic and contemporary military cohorts: combat deployment and the healthy warrior effect. Am J Epidemiol 2008;167:1269–1276.
22 McLaughlin R, Nielsen L, Waller M: An evaluation of the effect of military service on mortality: quantifying the healthy soldier effect. Ann Epidemiol 2008;18:928–936.
23 Wilson J, Jones M, Fear NT, Hull L, Hotopf M, Wessely S, Rona RJ: Is previous psychological health associated with the likelihood of Iraq War deployment? An investigation of the 'healthy warrior effect'. Am J Epidemiol 2009;169:1362–1369.
24 Defense Science Board: Report of the Defense Science Board Task Force on Persian Gulf War health effects. Washington, Office of the Under Secretary of Defense for Acquisition and Technology, 1994. http://www.gulflink.osd.mil/dsbrpt/.
25 Haley RW, Billecke S, La Du BN: Association of low PON1 type Q (type A) arylesterase activity with neurologic symptom complexes in Gulf War veterans. Toxicol Appl Pharmacol 1999;157:227–233.
26 Kang HK, Mahan CM, Lee KY, Murphy FM, Simmens SJ, Young HA, Levine PH: Evidence for a deployment-related Gulf War syndrome by factor analysis. Arch Environ Health 2002;57:61–68.
27 Cherry N, Creed F, Silman A, Dunn G, Baxter D, Smedley J, Taylor S, Macfarlane GJ: Health and exposures of United Kingdom Gulf war veterans. I. The pattern and extent of ill health. Occup Environ Med 2001;58:291–298.
28 Blanchard MS, Eisen SA, Alpern R, Karlinsky J, Toomey R, Reda DJ, Murphy FM, Jackson LW, Kang HK: Chronic multisymptom illness complex in Gulf War I veterans 10 years later. Am J Epidemiol 2006;163:66–75.
29 Horner RD, Kamins KG, Feussner JR, Grambow SD, Hoff-Lindquist J, Mitsumoto H, Pascuzzi R, Spencer P, Tim R, Howard D, Smith TC, Ryan MA, Coffman CJ, Kasarskis EJ: Occurrence of amyotrophic lateral sclerosis among Gulf War veterans. Neurology 2003;61:742–749.
30 Institute of Medicine: Gulf War and Health. Washington, National Academies Press, 2004, vol 1: Depleted Uranium, Pyridostigmine Bromide, Sarin, Vaccines; vol 2: Insecticides and Solvents – Updated Literature Review on Sarin; vol 3: Fuels, Combustion Products, and Propellants. http://www.iom.edu/report.asp?id=24236.
31 Araneta MR, Moore CA, Olney RS, Edmonds LD, Karcher JA, McDonough C, Hiliopoulos KM, Schlangen KM, Gray GC: Goldenhar syndrome among infants born in military hospitals to Gulf War veterans. Teratology 1997;56:244–251.
32 Iowa Persian Gulf Study Group: Self-reported illness and health status among Gulf War veterans: a population-based study. The Iowa Persian Gulf Study Group. JAMA 1997;277:238–245.
33 Nagelkerke NJD: A note on a general definition of the coefficient of determination. Biometrika 1991;78:691–692.
34 Dever JA, Mason RE: DMDC Sampling Planning Tool (Version 2.1). Arlington, Defense Manpower Data Center, 2003.
35 Kuhn WW, Tucker AW: Nonlinear Programming. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, 1951, pp 481–492.
36 Mason RE, Wheeless SC, George BJ, Dever JA, Riemer RA, Elig TW: Sample Allocation for the Status of the Armed Forces Surveys. Proceedings of the Section on Survey Research. Alexandria, American Statistical Association, 1995, pp 769–774.
37 Hillier FS, Lieberman GJ: Operations Research, ed 2. San Francisco, Holden-Day, 1974.
38 Fukuda K, Nisenbaum R, Stewart G, Thompson WW, Robin L, Washko RM, Noah DL, Barrett DH, Randall B, Herwaldt BL, Mawle AC, Reeves WC: Chronic multisymptom illness affecting Air Force veterans of the Gulf War. JAMA 1998;280:981–988.
39 American Association for Public Opinion Research: Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, ed 5. Lenexa, AAPOR, 2008.
40 Groves R, Couper M: Nonresponse in Household Interview Surveys. New York, Wiley & Sons, 1998.
41 Iannacchione V: Sequential weight adjustments for location and cooperation propensity for the 1995 National Survey of Family Growth. J Offic Statist 2010;19:31–43.
42 Breiman L, Friedman J, Stone C, Olshen R: Classification and Regression Trees. New York, CRC Press, 1984.
43 Muthen LK, Muthen BO: Mplus User's Guide, ed 5. Los Angeles, Muthen & Muthen, 2007.
44 Ware J Jr, Kosinski M, Keller SD: SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. Boston, New England Medical Center, 1995.
45 SUDAAN Language Manual, Release 10.0. Research Triangle Park, Research Triangle Institute, 2008.
46 Ware J Jr, Kosinski M, Keller SD: A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220–233.
47 Hays RD, Farivar SS, Liu H: Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD J Chron Obstruct Pulm Dis 2005;2:63–67.
48 Hu L-T, Bentler PM: Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol Methods 1998;3:424–453.
49 Hu L-T, Bentler PM: Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equat Modeling 1999;6:1–55.
50 Byrne BM: Structural Equation Modeling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming. Thousand Oaks, Sage Publications, 1994.
51 Ware JE, Snow KK, Kosinski M, Gandek B: SF-36 Health Survey Manual and Interpretation Guide. Boston, Health Institute, 1997.
52 Haley RW, Kramer GL, Xiao J, Teiber JF: Nerve agent exposure associated with Gulf War encephalopathy through gene-environment interaction with the Q192R polymorphism of paraoxonase 1 (PON1). 49th Annual Meeting of the Society of Toxicology, Salt Lake City, March 2010, Neurotoxicity and Neurodegenerative Disease Session, abstract No 2242.
53 Hays RD, Sherbourne CD, Spritzer KL, Dixon WJ: A microcomputer program (sf36.exe) that generates SAS code for scoring the SF-36 health survey. Proceedings of the 22nd Annual SAS Users Group International Conference, 1996, pp 1128–1132. http://gim.med.ucla.edu/FacultyPages/Hays/util.htm.

ONLINE FIRST

Cholinergic Autonomic Dysfunction in Veterans With Gulf War Illness

Confirmation in a Population-Based Sample

Robert W. Haley, MD; Elizabeth Charuvastra, RN†; William E. Shell, MD; David M. Buhner, MD; W. Wesley Marshall, MD; Melanie M. Biggs, PhD; Steve C. Hopkins, BS; Gil I. Wolfe, MD; Steven Vernino, MD, PhD

Background: The authors of prior small studies raised the hypothesis that symptoms in veterans of the 1991 Gulf War, such as chronic diarrhea, dizziness, fatigue, and sexual dysfunction, are due to cholinergic autonomic dysfunction.

Objective: To perform a confirmatory test of this prestated hypothesis in a larger, representative sample of Gulf War veterans.

Design: Nested case-control study.

Setting: Clinical and Translational Research Center, University of Texas Southwestern Medical Center, Dallas.

Participants: Representative samples of Gulf War veterans meeting a validated case definition of Gulf War illness with 3 variants (called syndromes 1-3) and a control group, all selected randomly from the US Military Health Survey.

Main Outcome Measures: Validated domain scales from the Autonomic Symptom Profile questionnaire, the Composite Autonomic Severity Score, and high-frequency heart rate variability from a 24-hour electrocardiogram.

Results: The Autonomic Symptom Profile scales were significantly elevated in all 3 syndrome groups ($P<.001$), primarily due to elevation of the orthostatic intolerance, secretomotor, upper gastrointestinal dysmotility, sleep dysfunction, urinary, and autonomic diarrhea symptom domains. The Composite Autonomic Severity Score was also higher in the 3 syndrome groups ($P=.045$), especially in syndrome 2, primarily due to a significant reduction in sudomotor function as measured by the Quantitative Sudomotor Axon Reflex Test, most significantly in the foot; the score was intermediate in the ankle and upper leg and was nonsignificant in the arm, indicating a peripheral nerve length–related deficit. The normal increase in high-frequency heart rate variability at night was absent or blunted in all 3 syndrome groups ($P<.001$).

Conclusion: Autonomic symptoms are associated with objective, predominantly cholinergic autonomic deficits in the population of Gulf War veterans.

Arch Neurol. Published online November 26, 2012. doi:10.1001/jamaneurol.2013.596

FEW MEDICAL CONDITIONS ARE as vexing as Gulf War illness to the veterans who experience it, the physicians who are charged with caring for them, and the policy makers who determine the institutional attitudes and level of resources to be directed at the problem. In 1991, the US military deployed 700 000 of the highest-performing members of the all-volunteer army to the Middle East for a 5-week air bombing campaign and a 5-day ground operation involving tank battles and little traditional combat. Yet, an estimated 25% of the force returned with a chronic, often disabling illness involving symptoms of multiple organ systems without obvious physical signs or laboratory abnormalities,[1] variously ascribed to fibromyalgia, somatization, deployment stress, chronic fatigue syndrome, adult-onset attention-deficit disorder, or simply multisymptom illness. Evidence from epidemiological and clinical studies suggests a chronic neurotoxic encephalopathy from exposure to cholinesterase-inhibiting chemicals.[1,2] A similar chronic illness has been described in pesticide-exposed agricultural workers[3] and in survivors of the 1995 subway sarin attack in Tokyo, Japan.[4]

Among the most troubling reports of the ill veterans are symptoms suggesting autonomic nervous system dysfunction. These include chronic fatigue, pathogen-free diarrhea, delayed gastric emptying and reflux, dizziness, light sensitivity, night sweats, unrefreshing sleep, sexual dysfunction, and an unusually high rate of cholecystitis and cholecystectomy in atypically young male veterans.[1] A 2004 study by Haley et al[5] measured autonomic function in 21 veterans who fit

Author Affiliations are listed at the end of this article.
†Deceased.

a factor case definition of 3 syndrome variants and in 17 veteran control subjects (all male) who were matched by age, sex, and education, drawn from an epidemiological survey of a naval reserve unit.[6] Spectral analysis of 24-hour Holter electrocardiography demonstrated significant blunting of the normal nocturnal increase in high-frequency heart rate variability (HF HRV), suggesting impaired central control of parasympathetic tone,[5] but test results of baroreceptor function, sleep architecture by polysomnography, and sensory and motor nerve conduction were normal. Stein et al[7] reported reduced circadian variation in HF HRV among 12 veterans of the Gulf War meeting a modified case definition of multisymptom illness[8] recruited from a rheumatology clinic compared with 36 healthy civilian volunteers, but HF HRV reduction was present only in 5 female veterans and not in 6 male veterans with usable HRV measurements. In an evaluation of neuromuscular function in 49 ill British Gulf War veterans and in 26 healthy controls, Sharief et al[9] found no differences in quantitative test results of sensory detection thresholds, Valsalva and standing heart rate ratios, and thermoregulatory control of sweating; however, 24-hour HF HRV was not measured. None of these studies provided a thorough description of autonomic symptoms, and none was performed among a population-representative sample of veterans with the full spectrum of Gulf War illness symptoms. The 3 studies[5,7,9] are compatible with the possibility of a selective abnormality of central cholinergic parasympathetic control with preserved sympathetic adrenergic and cardiovagal baroreceptor function.

Therefore, we designed a study to test this prestated hypothesis. We evaluated a population-representative sample of Gulf War veterans meeting a validated case definition of Gulf War illness, with a control group and 3 syndrome variants representing the full spectrum of the condition.[10]

METHODS

STUDY DESIGN

We studied 97 Gulf War–era veterans, including 66 case veterans with Gulf War illness and 31 control veterans. The participants, randomly selected as a nested case-control study by a 3-stage sample from the US Military Health Survey (**Figure 1**), were representative of the entire Gulf War–era veteran population (eAppendix; http://www.archneurol.com). The 66 case veterans met the standardized factor case definition of Gulf War illness, which was previously validated in a clinical sample[10] and in a large nationally representative sample.[12] Specifically, we studied 21 veterans meeting the factor case definition of Gulf War syndrome 1 (impaired cognition), 24 veterans with syndrome 2 (confusion-ataxia), and 21 veterans with syndrome 3 (central neuropathic pain). The 31 control veterans included 16 who did not meet the factor case definition of Gulf War illness but were deployed to the Kuwaiti theater of operations (deployed controls) and 15 who were in the military during the 1991 Gulf War but were not deployed (nondeployed controls). The demographic characteristics and comorbidities of the final sample are given in **Table 1**.

CLINICAL RESEARCH PROTOCOL

All participants were admitted to the University of Texas Southwestern Medical Center's Clinical and Translational Research Center located in Parkland Memorial Hospital, Dallas, where coffee drinking and smoking were allowed to continue. All participants gave written informed consent according to a protocol approved by the institutional review boards of the university. Because all the participants of this nationally representative sample traveled to Dallas for the study, medications could not be discontinued safely until they arrived in the Clinical and Translational Research Center under medical supervision; therefore, medications could be discontinued for only 24 to 48 hours (not necessarily for a full 5 half-lives) before autonomic testing. Whereas full washout is critical for clinical testing of individual participants, potential biasing effects of medication use on group comparisons were tested by multivariable analyses. An experienced clinical psychologist (M.M.B.) interviewed all participants following administration of the Structured Clinical Interview for *DSM-IV-TR* (SCID)[13] and the Clinician-Administered PTSD [posttraumatic stress disorder] Scale (CAPS).[14] All investigators who performed or interpreted (E.C., M.M.B., S.C.H., G.I.W., and S.V.) test results were blinded to the participants' case-control status.

Participants initially completed the self-administered Autonomic Symptom Profile (ASP) questionnaire measuring autonomic symptoms, which has been validated in healthy individuals and in patients with autonomic failure.[15] Standard weights were applied to construct the Composite Autonomic Symptom Scale (COMPASS) and the subscales of autonomic symptom domains.[15] After a 12-hour fast and abstention from alcohol and caffeine, at 7 AM participants underwent the following objective tests of autonomic function in an autonomic laboratory: pupillometry, lacrimation test, the Quantitative Sudomotor Axon Reflex Test,[16] heart rate response to deep breathing and Valsalva maneuver, quantitative sensory testing of cooling and heat pain thresholds,[17] and blood pressure and heart rate response to head-up tilt with a tilt table (Finapres Monitor; Ohmeda). Details of these tests are provided in the eAppendix.

The Composite Autonomic Severity Score (CASS), a standardized semiquantitative score measuring the severity of autonomic dysfunction from 0 (no deficit) to 10 (maximal deficit), was calculated by combining the results of the 3 subsets of the objective autonomic tests and adjusting to standard age and sex. These included sudomotor (range, 0-3), cardiovagal (range, 0-3), and adrenergic (range, 0-4) subsets.[16]

Twenty-four–hour Holter electrocardiography recordings, performed at home, were digitized at high resolution, and all QRS complexes were reviewed (Pathfinder 710; Reynolds Medical) by a skilled technician who censored aberrant complexes and artifacts. The normal-to-normal R-R intervals in a 5-minute epoch every 15 minutes were analyzed in the frequency domain using a fast Fourier transform algorithm based on the Lomb-Scargle method of spectral analysis[18] to produce the standard measures of HF (0.15 to <0.40 Hz), low frequency (0.04 to <0.14 Hz), and very low frequency (0.003 to <0.04 Hz) spectral power, expressed in milliseconds.[2] High-frequency HRV is an index mainly of vagal parasympathetic influence on cardiac rhythm and is reproducible over time.[19,20]

STATISTICAL ANALYSIS

P values are 2-tailed. The reported results were adjusted for age, sex, and race/ethnicity (black vs other). Analyses were rerun to test for confounding by the following covariates: smoking, creatinine clearance, diagnosis of heart disease, glycated hemoglobin level, officer rank during the war, CAPS diagnosis of PTSD, indicators of deconditioning (body mass index and resting pulse rate), and SCID diagnoses of alcohol or other drug abuse or dependence and major depressive disorder, as well as medications the participants were taking, including anticholinergic medications and tricyclic antidepressants.

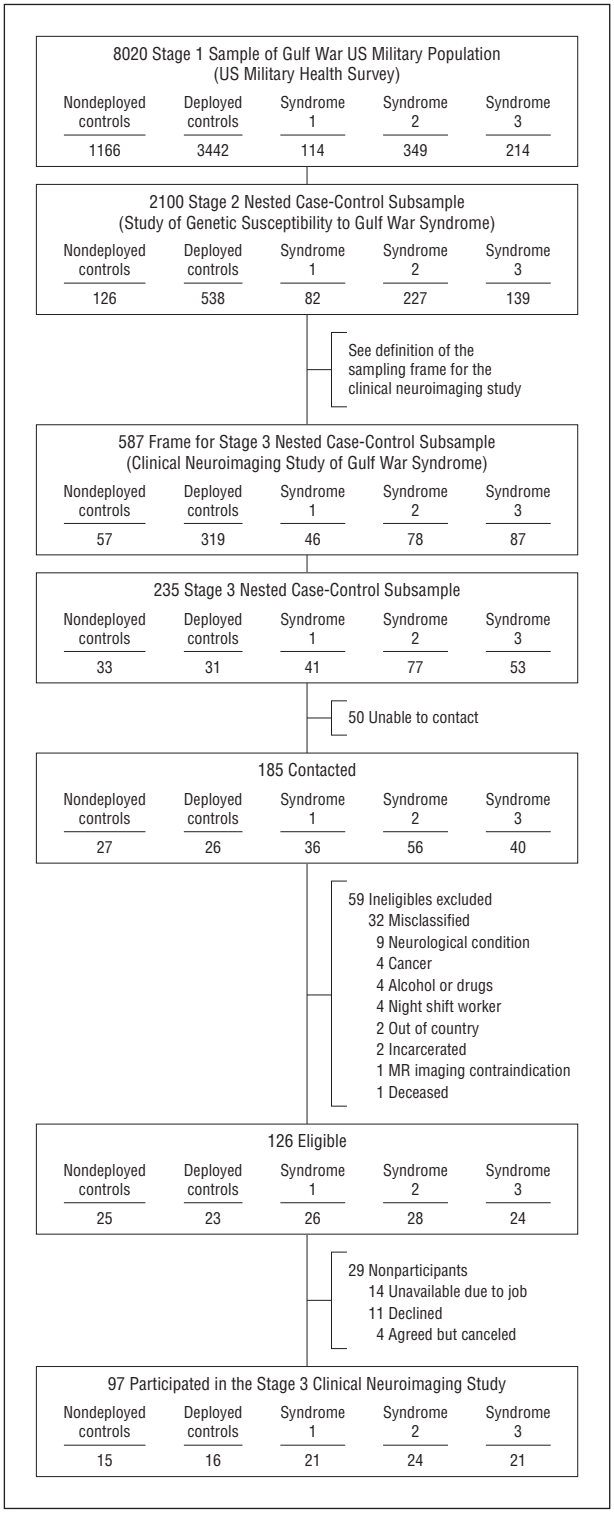


Figure 1. Process for selecting the nested case-control sample of Gulf War veterans suitable for the clinical neuroimaging study of Gulf War illness from the population sample of the US Military Health Survey. Nondeployed control subjects included those who were medically deployable personnel in the US military during the Gulf War but who were not deployed to the Kuwaiti theater of operations and did not meet any of the case definitions for Gulf War illness. In the stage 1 and stage 2 boxes, the differences between the total and the sum across its 5 comparative groups are due to subsyndromic subjects or members of special strata.[10] Numbers in the age by sex, race/ethnicity, and officer rank during the war strata in each clinical group are suppressed according to terms of the certificate of confidentiality. The 32 veterans excluded from group misclassification included 31 classified in one of the syndrome groups whose symptoms reported on the survey were not verified by the medical history taken by telephone and 1 classified as a control subject who had omitted symptoms of Gulf War illness on the survey. The 9 excluded for neurological conditions included 5 with a history of traumatic brain injury and 1 each with cerebrovascular disease, Parkinson disease, Guillain-Barré syndrome, and an unspecified chronic disease. The response rate is the American Association for Public Opinion Research response rate 4 and includes in the base the estimated number of eligible cases among those initially selected from the sampling frame to be contacted.[11] MR indicates magnetic resonance.

RESULTS

AUTONOMIC SYMPTOMS

All 3 Gulf War illness variant groups reported significantly more autonomic symptoms, assessed by the ASP, than the control group (**Figure 2** and **Table 2**). The COMPASS scores were significantly elevated for all 3 syndrome groups compared with the controls and were most elevated for syndrome 2 (Figure 2).

In the various symptom domains of the ASP (Table 2), the syndrome 2 group had the highest autonomic symptom scores, but the pattern of symptom score elevations was similar among the 3 syndrome groups. The differences between cases and controls explained more variance ($R^2 \geq 0.20$) in the orthostatic intolerance, secreto-

Table 1. Demographic and Comorbidity Measures in Controls and Gulf War Illness Variant Groups

Characteristic	Nondeployed Controls (n = 15)	Deployed Controls (n = 16)	Gulf War Illness Variant Group[a]			*P* Value[b]
			Syndrome 1 (n = 21)	Syndrome 2 (n = 24)	Syndrome 3 (n = 21)	
Age, mean (SD), y	51.9 (7.8)	47.8 (7.9)	48.2 (8.6)	49.8 (8.0)	51.0 (7.9)	.42
Female sex, No. (%)	3 (20)	3 (19)	7 (33)	7 (29)	4 (19)	.78
Black race/ethnicity, No. (%)	2 (13)	4 (25)	3 (14)	4 (17)	3 (14)	.91
Officer rank during the war, No. (%)	2 (13)	2 (13)	2 (10)	1 (4)	3 (14)	.80
Education scale, mean (SD)	5.5 (1.7)	5.1 (1.8)	5.8 (1.8)	4.6 (1.6)	5.0 (2.0)	.30
BMI, mean (SD)	30.1 (3.2)	29.6 (4.7)	29.0 (5.0)	28.4 (4.7)	30.7 (5.8)	.66
Resting pulse rate, mean (SD), beats/min	75.9 (9.8)	75.9 (14.1)	75.4 (15.1)	73.3 (12.3)	74.4 (14.7)	.80
Glomerular filtration rate from two 24-h urine samples, mean (SD)	129 (42)	132 (35)	113 (35)	125 (23)	121 (31)	.65
Taking anticholinergic medications or tricyclic antidepressants, No. (%)	1 (7)	0	2 (10)	3 (13)	1 (5)	.62
Diabetes by history or glycated hemoglobin level ≥7% on admission, No. (%)	1 (7)	0	1 (5)	1 (4)	1 (5)	.95
CDC definition of multisymptom illness, No. (%)	0	0	21 (100)	24 (100)	21 (100)	<.001
MOS SF-12 *t* score, mean (SD)						
Physical component	51.5 (9.4)	51.6 (7.7)	37.8 (12.3)	26.2 (7.6)	32.4 (9.2)	<.001
Mental component	57.8 (3.5)	58.4 (7.4)	34.5 (12.0)	39.6 (9.3)	45.6 (12.4)	<.001
CDC definition of chronic fatigue syndrome, No. (%)	0	0	1 (5)	2 (8)	4 (19)	.19
ACR survey definition of fibromyalgia, No. (%)	0	0	5 (24)	14 (58)	18 (86)	<.001
SCID diagnosis, No. (%)						
Active major depressive disorder	0	0	5 (24)	1 (4)	3 (14)	.04
Active alcohol abuse or dependence	3 (20)	1 (6)	6 (29)	10 (42)	5 (24)	.15
Active drug abuse or dependence or admission urine test	2 (13)	2 (13)	4 (19)	5 (21)	1 (5)	.58
CAPS diagnosis of active posttraumatic stress disorder, No. (%)[c]	0	0	8 (38)	9 (38)	5 (24)	.002

Abbreviations: ACR, American College of Rheumatology; BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); CAPS, Clinician-Administered PTSD [posttraumatic stress disorder] Scale; CDC, Centers for Disease Control and Prevention; MOS SF-12, Medical Outcomes Study 12-Item Short Form Health Survey; SCID, Structured Clinical Interview for *DSM-IV-TR*.

SI conversion factor: To convert glycated hemoglobin level to proportion of total hemoglobin, multiply by 0.01.

[a] Syndrome 1 is impaired cognition, syndrome 2 is confusion-ataxia, and syndrome 3 is central neuropathic pain.

[b] By 5-group Fisher exact test or Wilcoxon rank sum test.

[c] Among 22 participants with PTSD by CAPS, the inciting event was a horrifying or life-threatening experience in 7 of them.

motor, upper gastrointestinal dysmotility, sleep dysfunction, and urinary symptom domains and explained less variance ($R^2 < 0.20$) in the pupillomotor, autonomic constipation, vasomotor, male sexual dysfunction, and reflex syncope symptom domains, suggesting deficits related more to cholinergic than adrenergic autonomic systems. Moreover, the group difference on the male sexual dysfunction subscale was mainly due to erectile dysfunction, possibly related to parasympathetic cholinergic control, and not ejaculatory failure, a sympathetic adrenergic function.

OBJECTIVE AUTONOMIC TESTS

On objective autonomic tests, participants with Gulf War illness had significantly more evidence of autonomic deficits than the controls (**Table 3**). The CASS varied significantly across the clinical groups ($P = .045$) and was higher in the syndrome 2 group than in the controls ($P = .02$).

Compared with the controls, all 3 syndrome groups showed significantly reduced distal postganglionic sudomotor function, most significant in the foot, intermediate in the ankle and upper leg, and nonsignificant in the arm, indicating nerve length–related damage to the peripheral autonomic nervous system affecting the distal small cholinergic sudomotor fibers (Table 3). In a multivariable linear model of sudomotor function in the foot controlling for age and race/ethnicity, the case-control difference was significant ($P = .02$) and did not vary by sex ($P = .78$ for group × sex interaction). Controlling for the covariates did not alter these findings.

In contrast, no group differences were statistically significant in tests of tear production (Schirmer test), in sympathetic adrenergic function (including the blood pressure responses to Valsalva maneuvers and tilt), or in any of the pupillary measures. These results are summarized in Table 3.

QUANTITATIVE SENSORY TESTS

The syndrome 2 and syndrome 3 groups had increased cooling detection threshold, which was statistically sig-

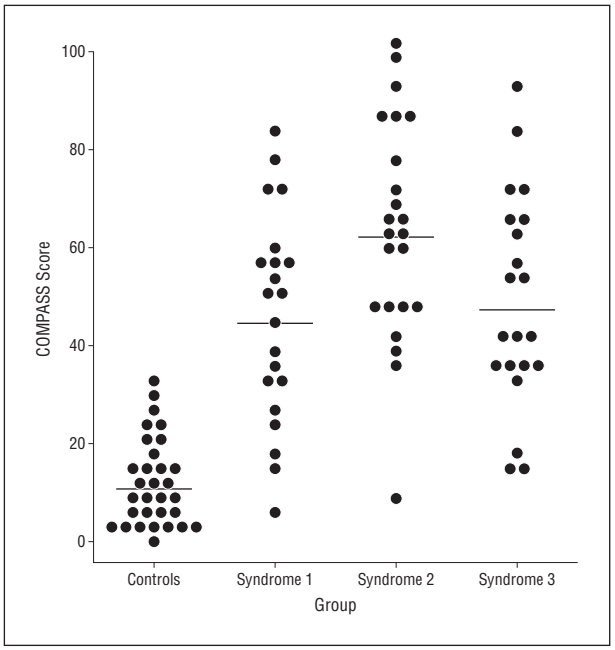


Figure 2. Distribution of values on the Composite Autonomic Symptom Scale (COMPASS) in the control subjects and in the 3 Gulf War illness variant groups. The horizontal bars represent the mean scores. The differences in COMPASS scores across the 4 groups are statistically significant ($R^2 = 0.59$, $P < .001$).

nificant only for the syndrome 2 group (Table 3). None of the 3 syndrome groups differed significantly from controls on the heat pain threshold.

CIRCADIAN VARIATION IN PARASYMPATHETIC TONE

From spectral analysis of 24-hour electrocardiogram monitoring, HF HRV increased normally at night in the control group but not in the 3 syndrome groups (**Figure 3**A and **Table 4**). In a repeated-measures mixed-effects linear model of log HF HRV, the case-control × day minus night interaction was statistically significant ($P < .001$), but the 3-way interaction with sex was not ($P = .88$), indicating that the loss of circadian variation in the 3 syndrome groups compared with the controls was found in both men and women veterans. Controlling for the covariates did not alter these findings.

When analyzed by group, all 3 syndrome groups showed significant blunting or loss of the normal nocturnal increase (Figure 3B and Table 4). During the day, HF HRV of the syndrome 1 group did not differ from that of the controls, but the syndrome 2 group had significantly lower HF HRV than the controls, and the syndrome 3 group had significantly higher HF HRV than the controls, particularly during the morning hours (Figure 3B and Table 4).

High-frequency HRV at night was moderately inversely correlated with the CASS index of objective autonomic test results ($r = –0.41$, $P < .001$) (**Figure 4**A). High-frequency HRV during the day was weakly correlated with the CASS ($r = –0.22$, $P = .04$) (Figure 4B).

ASSOCIATION OF AUTONOMIC SYMPTOMS AND OBJECTIVE TEST RESULTS

The COMPASS of all autonomic symptoms was inversely correlated with HF HRV and was directly correlated with the CASS subscales. The correlation was highest with HF HRV during the day and with the CASS sudomotor subscale, and the correlation was lowest with the CASS cardiovagal and adrenergic subscales (**Table 5**).

The individual symptom domains tended to be correlated with HF HRV or with the CASS sudomotor subscale but not both (Table 5). Specifically, the vasomotor, secretomotor, upper gastrointestinal dysmotility, and pupillomotor symptom domains were most strongly correlated with the CASS sudomotor subscale. The orthostatic intolerance symptom domain was also correlated with the CASS sudomotor subscale, and it was the only symptom domain to be significantly correlated with the CASS adrenergic subscale. In contrast, the upper gastrointestinal dysmotility and sleep dysfunction symptom domains were most strongly associated with HF HRV at night, and the autonomic diarrhea, male sexual dysfunction, and urinary symptom domains were most strongly correlated with HF HRV during the day. Of the 2 components of male sexual dysfunction, erectile dysfunction, a parasympathetic function, was highly correlated with HF HRV during the day, while ejaculatory failure, an adrenergic function, was not. Like ejaculatory failure, reflex syncope was not associated with any of the objective autonomic measures, and these were the only autonomic symptom domains not associated with the 3 syndrome groups (Table 2).

COMMENT

In a nested case-control sample drawn from a national survey in a large representative sample of the Gulf War–era US military population, this study found that a well-validated research case definition of Gulf War illness was strongly associated with standard scales of autonomic symptoms and with objective tests of autonomic dysfunction. Autonomic symptom scores and objective test results were most abnormal compared with the controls in the syndrome 2 group. This reflects the findings of several prior studies in which syndrome 2 consistently was the most disabling[10,21] and had the most prominent abnormalities on various objective tests of brain function.[22-29]

The ASP autonomic symptom domains most strongly associated with the case definition tended to be those related predominantly to cholinergic autonomic control, and these symptom domains tended to be most strongly associated with HF HRV measures or with the CASS sudomotor subscale but not with the CASS cardiovagal or adrenergic subscales. On the objective autonomic tests, the 3 syndrome groups differed most from controls on sudomotor testing (Quantitative Sudomotor Axon Reflex Test). The degree of difference on the Quantitative Sudomotor Axon Reflex Test was related to peripheral nerve length, typical of a length-dependent neuropathy of small-caliber, unmyelinated, peripheral nerve fibers. The increased cooling detection thresholds observed in

Table 2. Scores on the Autonomic Symptom Profile Domains Among Control and Gulf War Illness Variant Groups

Autonomic Symptom Profile Domain	Maximum Possible Score	Reference Means for Controls/Patients With NAF[a]	Controls, Mean (SEM) Score (n = 31)	Gulf War Illness Variant Group, Mean (SEM) Score			R^2[c]	*P* Value[d]
				Syndrome 1 (n = 21)	Syndrome 2 (n = 23)[b]	Syndrome 3 (n = 21)		
Orthostatic intolerance	40	3.6/21.6	2.4 (1.3)	12.9 (1.6)	22.2 (1.6)	13.7 (1.6)	0.44	<.001
Secretomotor	20	0.9/6.5	0.9 (0.6)	4.7 (0.7)	6.2 (0.7)	5.2 (0.7)	0.39	<.001
Upper gastrointestinal dysmotility	10	0.5/2.4	0.2 (0.3)	2.1 (0.4)	3.0 (0.4)	1.7 (0.4)	0.30	<.001
Urinary	20	0.8/2.9	1.0 (0.5)	3.7 (0.7)	4.0 (0.6)	4.8 (0.7)	0.25	<.001
Sleep dysfunction	15	0.8/2.4	1.5 (0.5)	4.8 (0.6)	5.0 (0.6)	4.6 (0.6)	0.24	<.001
Autonomic diarrhea	20	1.5/4.2	1.7 (1.0)	6.1 (1.2)	8.2 (1.1)	6.7 (1.2)	0.16	<.001
Pupillomotor	5	0.4/1.6	0.9 (0.3)	2.1 (0.3)	2.4 (0.3)	1.7 (0.3)	0.15	.002
Autonomic constipation	10	0.6/2.5	0.3 (0.3)	2.1 (0.4)	2.1 (0.4)	1.6 (0.4)	0.15	.002
Vasomotor	10	0.4/2.2	0.1 (0.4)	1.6 (0.5)	2.2 (0.5)	2.1 (0.5)	0.13	.006
Male sexual dysfunction	30	0.6/9.5	1.6 (1.0)	5.4 (1.2)	6.0 (1.1)	5.4 (1.2)	0.13	.009
Erectile dysfunction[e]	20	. . .	1.4 (0.8)	4.7 (1.0)	4.7 (0.9)	4.9 (1.0)	0.13	.01
Ejaculatory failure[e]	10	. . .	0.2 (0.4)	0.7 (0.5)	1.3 (0.4)	0.5 (0.5)	0.04	.24
Reflex syncope	20	0.0/0.9	0.2 (0.2)	0.2 (0.3)	1.0 (0.3)	0.2 (0.3)	0.08	.07

Abbreviation: NAF, neurogenic autonomic failure.
[a] Previously published reference means for controls and patients with neurogenic autonomic failure.[15]
[b] This value is 23 in Tables 2 and 3 because 1 participant with syndrome 2 did not complete the autonomic testing.
[c] Percentage of variance explained by the 4-group variable in an analysis of variance performed on the rank-transformed scores.
[d] By the Kruskal-Wallis nonparametric 4-group test.
[e] These are subdomains of the male sexual dysfunction domain; no separate reference values were given for them.[15]

Table 3. Objective Autonomic and Quantitative Sensory Tests in Controls and Gulf War Illness Variant Groups[a]

Variable	Controls, Mean (SEM) (n = 31)	Gulf War Illness Variant Groups, Mean (SEM)			*P* Value[b]
		Syndrome 1 (n = 21)	Syndrome 2 (n = 23)	Syndrome 3 (n = 21)	
CASS	0.71 (0.27)	1.15 (0.32)	1.90 (0.31)[c]	0.57 (0.32)	.045
QSART sudomotor quantitative sweat production, μL					
Foot	0.79 (0.07)	0.53 (0.08)[c]	0.40 (0.07)[d]	0.55 (0.08)[c]	.005
Ankle	1.33 (0.12)	1.16 (0.16)	0.78 (0.15)[e]	0.92 (0.16)[c]	.04
Upper leg	0.90 (0.09)	0.60 (0.11)	0.49 (0.10)[e]	0.68 (0.10)	.02
Arm	1.09 (0.14)	1.01 (0.18)	0.96 (0.16)	1.34 (0.17)	.24
Schirmer test tear production at 5 min, mm	6.0 (1.2)	6.2 (1.5)	4.4 (1.4)	3.5 (1.5)	.50
Ratio of expiration to inspiration for R-R intervals	1.25 (0.02)	1.23 (0.29)	1.25 (0.03)	1.24 (0.03)	.78
Valsalva ratio	1.81 (0.05)	1.83 (0.64)	1.67 (0.06)	1.77 (0.06)	.28
Change in systolic blood pressure from baseline at 3-min tilt, mm Hg	0.12 (1.33)	2.70 (1.64)	0.67 (1.52)	–0.13 (1.60)	.64
Maximum pupillary constriction velocity, mm/s[f]					
Left eye	4.85 (0.17)	4.71 (0.21)	4.52 (0.20)	4.95 (0.21)	.69
Right eye	4.96 (0.16)	4.58 (0.20)	4.51 (0.19)	4.95 (0.21)	.29
Quantitative sensory in the dominant hand					
Cooling detection threshold					
Just noticeable difference units	8.5 (0.7)	9.4 (0.8)	11.1 (0.8)[c]	10.7 (0.8)	.12
Percentile	70.1 (4.4)	84.6 (5.2)	93.0 (5.0)[c]	86.8 (5.2)	.13
Heat pain threshold					
Just noticeable difference units	23.0 (0.3)	22.5 (0.4)	22.7 (0.3)	22.3 (0.3)	.38
Percentile	38.7 (5.0)	27.5 (6.2)	31.1 (5.8)	27.1 (6.0)	.38
Basal corticotropin level, pg/dL	28.9 (2.9)	27.7 (3.4)	23.7 (2.7)	25.9 (3.2)	.49
Basal cortisol level, μg/dL	0.71 (0.08)	0.56 (0.08)	0.47 (0.06)	0.77 (0.11)	.10

Abbreviations: CASS, Composite Autonomic Severity Score; QSART, Quantitative Sudomotor Axon Reflex Test.
SI conversion factors: To convert corticotropin level to picomoles per liter, multiply by 0.22; to convert cortisol level to nanomoles per liter, multiply by 27.588.
[a] Means (SEMs) are standardized for age, sex, and race/ethnicity (black vs other).
[b] By the Kruskal-Wallis nonparametric 4-group test.
[c] $P \leq .05$.
[d] $P \leq .001$ for difference from controls by Wilcoxon rank sum test. The sudomotor group differences remained significant after controlling for whether participants were taking anticholinergic medications or tricyclic antidepressants.
[e] $P \leq .01$.
[f] No significant group differences were observed in resting pupillary diameter, dilation velocity, or constriction amplitude responses to 30-millisecond or 1-second light flash.

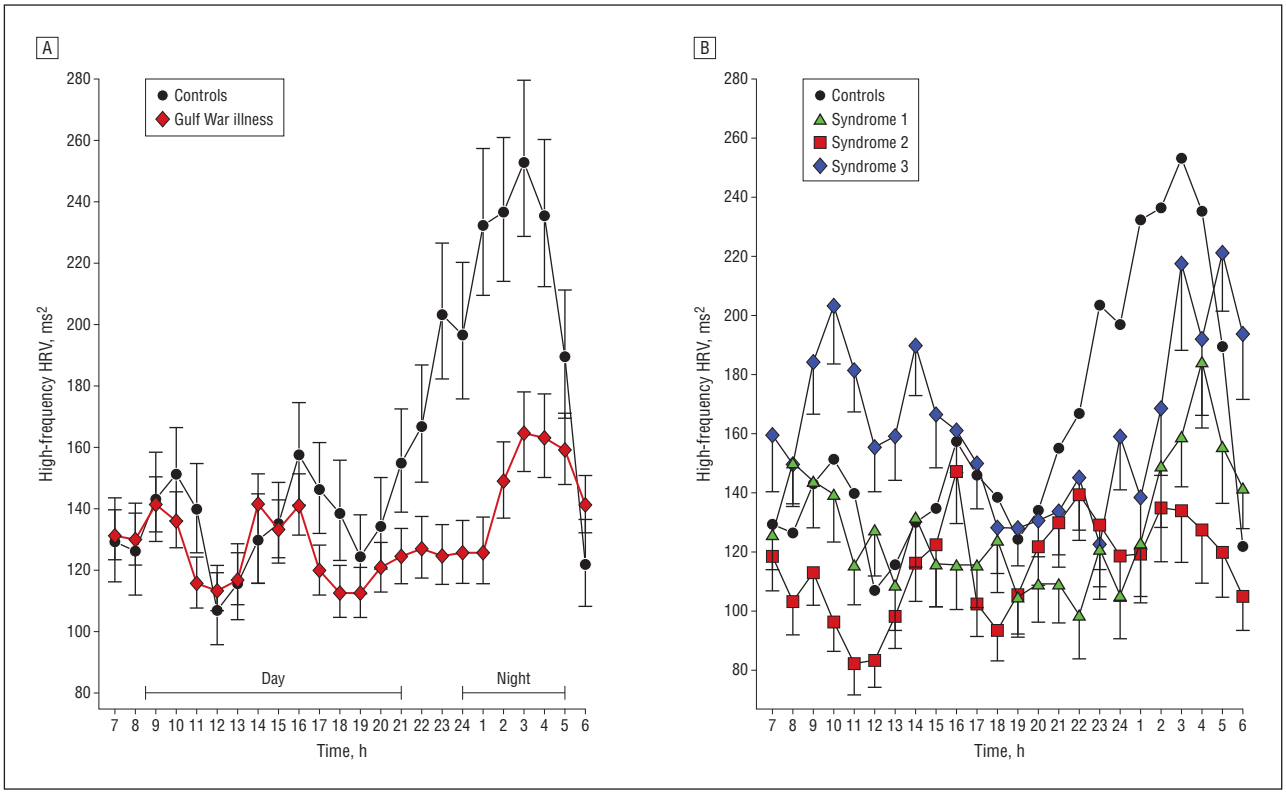


Figure 3. Difference in circadian variation of parasympathetic cardiovagal tone between control subjects and 3 Gulf War illness variant groups, measured by spectral analysis of high-frequency heart rate variability (HRV) in 5-minute epochs every hour from 24-hour Holter monitoring electrocardiography. The control group (black circles) is compared with all cases of Gulf War illness (red diamonds) (A) and with syndrome 1 (green triangles), syndrome 2 (red squares), and syndrome 3 (blue stars) (B). The control group showed the expected low cardiovagal tone during the day and a large increase at night. The syndrome 2 group showed depressed tone throughout the 24-hour period, with no evidence of a nocturnal increase. Syndrome 1 and syndrome 3 showed a blunted, delayed increase at night, and syndrome 3 had elevated tone during the day. The statistical test results of the effects in these graphs are given in Table 4.

Table 4. Difference in Circadian Variation of Parasympathetic Cardiovagal Tone Measured by 24-Hour Holter Monitoring Among Gulf War Illness Variant and Control Groups[a]

Group	Spectral Power of High-frequency HRV, Mean (SEM), ms²			
	Day, 8 AM to 9 PM	Night, 12 AM to 5 AM	Circadian Difference, Night Minus Day	*P* Value[b]
Model 1[c]				
Controls	135 (9)	226 (19)	91 (21)	<.001
Cases	131 (6)	139 (8)	8 (10)	.36
Controls minus cases	5 (10)	87 (22)	83 (25)	<.001
Model 2[c]				
Controls	135 (9)	226 (19)	91 (21)	<.001
Syndrome 1	133 (11)	125 (13)	8 (17)	.60
Syndrome 2	106 (8)	129 (13)	23 (15)	.07
Syndrome 3	160 (12)	165 (17)	4 (21)	.82
Controls minus syndrome 1	2 (13)	101 (24)	99 (28)	<.001
Controls minus syndrome 2	29 (11)	97 (24)	68 (27)	<.001
Controls minus syndrome 3	–25 (15)	61 (27)	87 (31)	.004

Abbreviation: HRV, heart rate variability.
[a] Apparent discrepancies in reported differences are due to rounding.
[b] From repeated-measures mixed-effects linear model predicting log-transformed high-frequency HRV measured in 5-minute epochs every hour, from the fixed effects of group, day minus night, and their interaction, with participants as random effects and the Dunnett correction for multiple comparisons. Significance was not altered by controlling for age, sex, race/ethnicity, body mass index, officer rank during the war, glycated hemoglobin level, glomerular filtration rate, major depressive disorder, active posttraumatic stress disorder, alcohol abuse or dependence, smoking, and anticholinergic medication or tricyclic antidepressant use.
[c] Model 1 (all cases combined) tests the effects seen in Figure 3A. Model 2 (cases analyzed by Gulf War illness variant group) tests the effects seen in Figure 3B.

the syndrome 2 group and the syndrome 3 group and described in a previous study[30] may also reflect underlying small-fiber impairment.

The autonomic impairment was most clearly demonstrated in the blunting of the normal rise in HF HRV at night. Because peripheral vagal baroreflex function was not

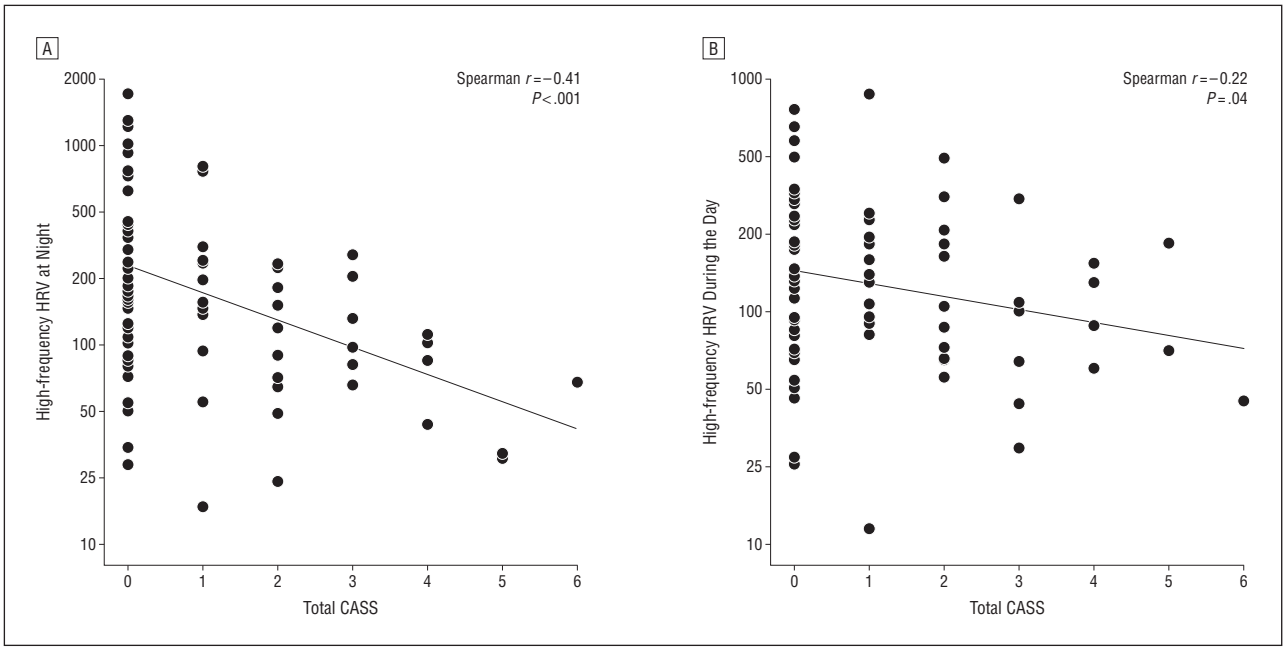


Figure 4. Correlation of high-frequency heart rate variability (HRV) at night (A) and during the day (B) with the total Composite Autonomic Severity Score (CASS). Normal R-R intervals from a 24-hour Holter monitor electrocardiogram were analyzed by spectral analysis in a 5-minute epoch each hour. For each participant, the hourly measures of the high-frequency spectral component (0.15 to <0.40 Hz) were averaged across the nighttime hours (12 AM to 5 AM) and across the daytime hours (8 AM to 9 PM), and both measures were log transformed. Measurements on the battery of objective autonomic tests were combined to calculate the CASS, on which higher scores indicate greater autonomic impairment.

Table 5. Partial Spearman Rank Order Correlations of the Total COMPASS Score and Autonomic Symptom Profile Domains With Objective, Laboratory-Based Measures of Autonomic Function[a]

	Spectral Power of High-frequency HRV		CASS			
Variable	Night, 12 AM to 5 AM	Day, 8 AM to 9 PM	Total	Sudomotor	Cardiovagal	Adrenergic
Total COMPASS score	−0.20[b]	−0.26[c]	0.20[b]	0.21[d]	0.10	0.11
Autonomic Symptom Profile Domain						
Orthostatic intolerance	−0.17	−0.13	0.22[d]	0.19[b]	0.12	0.22[d]
Vasomotor	−0.10	−0.14	0.14	0.22[d]	−0.01	−0.02
Secretomotor	−0.12	−0.16	0.12	0.22[b]	−0.04	−0.01
Upper gastrointestinal dysmotility	−0.22[d]	−0.21[d]	0.28[c]	0.26[c]	0.16	0.10
Autonomic diarrhea	−0.12	−0.26[c]	0.04	0.14	0.03	−0.13
Autonomic constipation	0.03	−0.05	0.04	0.01	0.15	0.01
Male sexual dysfunction	−0.13	−0.27[c]	0.17	0.10	0.04	0.13
Erectile dysfunction	−0.16	−0.33[e]	0.13	0.05	0.05	0.15
Ejaculatory failure	0.06	0.02	0.14	0.16	−0.05	−0.02
Urinary	−0.12	−0.31[e]	0.12	0.11	0.03	0.05
Pupillomotor	−0.08	−0.18[b]	0.15	0.23[d]	0.06	−0.04
Sleep dysfunction	−0.23[d]	−0.19[b]	0.11	0.08	0.08	0.08
Reflex syncope	−0.04	0.06	0.06	0.08	0.06	0.09

Abbreviations: CASS, Composite Autonomic Severity Score; COMPASS, Composite Autonomic Symptom Scale; HRV, heart rate variability.
[a] Partial Spearman rank order correlations are adjusted for age, sex, and race/ethnicity (black vs other).
[b] $P \leq .10$.
[c] $P \leq .01$.
[d] $P \leq .05$.
[e] $P \leq .005$.

significantly impaired, this abnormality of circadian variation in HF HRV suggests dysfunction in the central nervous system control of parasympathetic outflow. The sample size of this study was also sufficient to demonstrate significant, although more subtle, differences in HF HRV among the 3 syndrome groups during the day. Multivariable statistical analyses demonstrated that the objective findings of peripheral sudomotor neuropathy and impaired HF HRV were not explained by smoking, creatinine clearance, psychiatric comorbidity, diagnosis of heart disease, glycated hemoglobin level, officer rank during the war, indicators of deconditioning (body mass index and

resting pulse rate), or medications the participants were taking during the period of the study, including anticholinergic medications and tricyclic antidepressants.

The pattern of autonomic symptoms and objective test findings points predominantly to dysfunction of both central and peripheral cholinergic functions, possibly from neurotoxic damage to cholinergic neurons or cholinergic receptors. This proposed explanation is compatible with prior studies[27,28] showing that, compared with control subjects, regional cerebral blood flow in veterans with Gulf War illness responds abnormally to cholinergic challenge with physostigmine, suggesting chronic alteration of cholinergic receptors in the brain. Experiments in rodents, undertaken to model the possible chronic effects of sarin in low doses to which Gulf War veterans were exposed in the war, have identified persisting alterations of cholinergic receptors[31,32] and of autonomic responses.[33]

These findings and this explanation are compatible with a prior study[9] of neurologic function in ill Gulf War veterans, which found no associations with tests of adrenergic autonomic function and nerve conduction investigations of large-caliber peripheral nerves but generally did not test for circadian variation in HF HRV. Our findings did not confirm the interaction of blunted circadian variation in HF HRV with sex (blunted in women but not in men) reported by Stein et al,[7] which may have resulted from their studying a small sample drawn from health care–seeking clients.

This study has several strengths built into the design to avoid weaknesses in past research on Gulf War illness. In contrast to the exploratory nature of a prior study[9] of autonomic function in Gulf War veterans, this study was designed as a confirmatory test of a prestated hypothesis raised by previous investigations. The robust sample size and external validity afforded by the nested case-control design drawn from a survey in a large population-representative sample add greater confidence to the findings from prior small studies[7,22-28] performed in samples from single military units or from clinic volunteers. Particularly important for studying a disease defined by symptoms alone, the case definition of Gulf War illness used in this study is the only one that has been empirically validated by demonstrating a statistically good fit in other Gulf War veteran populations.[10,12] Its 3 syndrome variants provide homogeneous clinical groups to maximize statistical power and represent the full spectrum of the illness to determine whether autonomic dysfunction spans the entire spectrum or is limited to part of it. The extensive work by Suarez et al,[15] Low,[16] and Low et al[34] in developing the ASP and the CASS testing systems, used in this study, provided validated measures of autonomic symptom domains and objective autonomic function testing. As in a previous study of autonomic symptoms measured by the ASP,[15] the validity of the veterans' symptom reports was supported by correlations of the COMPASS and its domains with the appropriate CASS subscales of objective autonomic test results.

The greatest challenge in our study was the logistical difficulty of selecting and obtaining participation in a lengthy clinical evaluation of non–treatment-seeking veterans with the full spectrum of the Gulf War illness and representative of the population of Gulf War veterans. To accomplish this, the cases and controls sampled from the nationally representative US Military Health Survey were screened by a physician (R.W.H.) who called them by telephone to ensure correct classification on the case definition before participants were enrolled. The medical screening found that 31 of 132 cases (23.5%) and 1 of 53 controls (1.9%) who were selected and contacted were misclassified on the case definition. While some degree of misclassification is present in any epidemiological case definition, minimizing it through advance medical interviews greatly reduced its adverse effect on the statistical power of this study.

The autonomic measures that differed between cases and controls in this study may prove useful in a strategy for clinical diagnosis of Gulf War illness. Of the objective tests used, the one showing the clearest discrimination among all 3 syndrome groups and the control group was the measurement of circadian variation in HF HRV. When tested with the repeated-measures mixed-effects linear model, which appropriately manages variance of the fixed and random effects, the group discrimination is extremely good. However, when HF HRV measurements in multiple epochs are combined to form a single measure of nighttime HF HRV for each participant, the resulting participant-level means display enough residual variance to reduce the usefulness in clinical diagnosis. Additional research should attempt to reformulate the measure of circadian variation in HF HRV to reduce the variance. Measures of the central nervous system mechanisms upstream from the autonomic dysfunction, such as neuroimaging or electroencephalography of brain function,[26-29] may also be combined with autonomic testing to improve clinical diagnosis.

Perhaps the most important implications of the findings are those bearing on the long-standing debate about the nature of the Gulf War illness. These results confirm dysfunction among Gulf War veterans of both central control of parasympathetic function and peripheral cholinergic autonomic nerves, further implicating underlying damage to the cholinergic components of the central and peripheral nervous systems.

Accepted for Publication: July 3, 2012.
Published Online: November 26, 2012. doi:10.1001/jamaneurol.2013.596
Author Affiliations: Epidemiology Division, Department of Internal Medicine (Drs Haley, Buhner, Marshall, and Biggs), and Neuromuscular Section, Department of Neurology and Neurotherapeutics (Mr Hopkins and Drs Wolfe and Vernino), University of Texas Southwestern Medical Center, Dallas; and Targeted Medical Pharma, Inc, Los Angeles, California (Ms Charuvastra and Dr Shell). Dr Buhner is now with the Department of Mathematical Sciences, University of Texas at Dallas. Dr Marshall is now with Wesmar Solutions, Inc, Frisco, Texas. Dr Wolfe is now with the Department of Neurology, University of Buffalo, The State University of New York.
Correspondence: Robert W. Haley, MD, Epidemiology Division, Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-8874 (robert.haley@UTSouthwestern.edu).

Author Contributions: *Study concept and design:* Haley, Wolfe, and Vernino. *Acquisition of data:* Haley, Charuvastra, Shell, Buhner, Marshall, Biggs, Hopkins, Wolfe, and Vernino. *Analysis and interpretation of data:* Haley, Shell, Biggs, Wolfe, and Vernino. *Drafting of the manuscript:* Haley, Shell, Hopkins, and Vernino. *Critical revision of the manuscript for important intellectual content:* Haley, Charuvastra, Shell, Buhner, Marshall, Biggs, Wolfe, and Vernino. *Statistical analysis:* Haley and Shell. *Obtained funding:* Haley. *Administrative, technical, and material support:* Haley, Charuvastra, Buhner, Marshall, Biggs, Hopkins, and Wolfe. *Study supervision:* Haley.

Conflict of Interest Disclosures: Dr Haley received an honorarium from Targeted Medical Pharma, Inc, for critical review of a Food and Drug Administration new drug application for a nonpharmaceutical medication to treat fatiguing illness of possible benefit to Gulf War veterans.

Funding/Support: This study was supported by Indefinite Delivery Indefinite Quantity contract VA549-P-0027, awarded and administered by the Department of Veterans Affairs Medical Center, Dallas, Texas; by grant DAMD17-01-1-0741 from the US Army Medical Research and Materiel Command; and by grant UL1RR024982-05, titled North and Central Texas Clinical and Translational Science Initiative, from the National Center for Research Resources, a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research.

Role of the Sponsor: The funding agencies had no involvement in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Disclaimer: The content does not necessarily reflect the position or the policy of the federal government or the sponsoring agencies, and no official endorsement should be inferred.

Online-Only Material: The eAppendix is available at http://www.archneurol.com.

REFERENCES

1. US Department of Veterans Affairs. Research Advisory Committee on Gulf War Veterans' Illnesses. http://www1.va.gov/rac-gwvi/. Accessed October 1, 2012.
2. Golomb BA. Acetylcholinesterase inhibitors and Gulf War illnesses. *Proc Natl Acad Sci U S A*. 2008;105(11):4295-4300.
3. Ecobichon DJ. Organophosphorus ester insecticides. In: Ecobichon DJ, Joy RM, eds. *Pesticides and Neurological Diseases*. 2nd ed. Boston, MA: CRC Press, Inc; 1994:171-250.
4. Yokoyama K, Araki S, Murata K, et al. Chronic neurobehavioral and central and autonomic nervous system effects of Tokyo subway sarin poisoning. *J Physiol Paris*. 1998;92(3-4):317-323.
5. Haley RW, Vongpatanasin W, Wolfe GI, et al. Blunted circadian variation in autonomic regulation of sinus node function in veterans with Gulf War syndrome. *Am J Med*. 2004;117(7):469-478.
6. Haley RW, Kurt TL, Hom J. Is there a Gulf War syndrome? searching for syndromes by factor analysis of symptoms [published correction appears in *JAMA*. 1997;278(5):388]. *JAMA*. 1997;277(3):215-222.
7. Stein PK, Domitrovich PP, Ambrose K, et al. Sex effects on heart rate variability in fibromyalgia and Gulf War illness. *Arthritis Rheum*. 2004;51(5):700-708.
8. Fukuda K, Nisenbaum R, Stewart G, et al. Chronic multisymptom illness affecting Air Force veterans of the Gulf War. *JAMA*. 1998;280(11):981-988.
9. Sharief MK, Priddin J, Delamont RS, et al. Neurophysiologic analysis of neuromuscular symptoms in UK Gulf War veterans: a controlled study. *Neurology*. 2002;59(10):1518-1525.
10. Iannacchione VG, Dever JA, Bann CM, et al. Validation of a research case definition of Gulf War illness in the 1991 US military population. *Neuroepidemiology*. 2011;37(2):129-140.
11. American Association for Public Opinion Research. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 5th ed. Lenexa, KS: American Association for Public Opinion Research; 2008.
12. Haley RW, Luk GD, Petty F. Use of structural equation modeling to test the construct validity of a case definition of Gulf War syndrome: invariance over developmental and validation samples, service branches and publicity. *Psychiatry Res*. 2001;102(2):175-200.
13. First MB, Spitzer RL, Gibbon M, Williams J. *Structured Clinical Interview for Axis I* DSM-IV *Disorders*. New York, NY: Biometrics Research Dept; 1996.
14. Blake DD, Weathers FW, Nagy LM, et al. The development of a Clinician-Administered PTSD Scale. *J Trauma Stress*. 1995;8(1):75-90.
15. Suarez GA, Opfer-Gehrking TL, Offord KP, Atkinson EJ, O'Brien PC, Low PA. The Autonomic Symptom Profile: a new instrument to assess autonomic symptoms. *Neurology*. 1999;52(3):523-528.
16. Low PA. Composite autonomic scoring scale for laboratory quantification of generalized autonomic failure. *Mayo Clin Proc*. 1993;68(8):748-752.
17. Dyck PJ, O'Brien PC, Kosanke JL, Gillen DA, Karnes JL. A 4, 2, and 1 stepping algorithm for quick and accurate estimation of cutaneous sensation threshold. *Neurology*. 1993;43(8):1508-1512.
18. Press WH, Rybicki GB. Fast algorithm for spectral analysis of unevenly sampled data. *Astrophys J*. 1989;338:277-280.
19. Kamalesh M, Burger AJ, Kumar S, Nesto R. Reproducibility of time and frequency domain analysis of heart rate variability in patients with chronic stable angina. *Pacing Clin Electrophysiol*. 1995;18(11):1991-1994.
20. Burger AJ, Charlamb M, Weinrauch LA, D'Elia JA. Short- and long-term reproducibility of heart rate variability in patients with long-standing type I diabetes mellitus. *Am J Cardiol*. 1997;80(9):1198-1202.
21. Haley RW, Maddrey AM, Gershenfeld HK. Severely reduced functional status in veterans fitting a case definition of Gulf War syndrome. *Am J Public Health*. 2002;92(1):46-47.
22. Haley RW, Hom J, Roland PS, et al. Evaluation of neurologic function in Gulf War veterans: a blinded case-control study. *JAMA*. 1997;277(3):223-230.
23. Hom J, Haley RW, Kurt TL. Neuropsychological correlates of Gulf War syndrome. *Arch Clin Neuropsychol*. 1997;12(6):531-544.
24. Roland PS, Haley RW, Yellin W, Owens K, Shoup AG. Vestibular dysfunction in Gulf War syndrome. *Otolaryngol Head Neck Surg*. 2000;122(3):319-329.
25. Haley RW, Marshall WW, McDonald GG, Daugherty MA, Petty F, Fleckenstein JL. Brain abnormalities in Gulf War syndrome: evaluation with ^{1}H MR spectroscopy. *Radiology*. 2000;215(3):807-817.
26. Tillman GD, Green TA, Ferree TC, et al. Impaired response inhibition in ill Gulf War veterans. *J Neurol Sci*. 2010;297(1-2):1-5.
27. Haley RW, Spence JS, Carmack PS, et al. Abnormal brain response to cholinergic challenge in chronic encephalopathy from the 1991 Gulf War. *Psychiatry Res*. 2009;171(3):207-220.
28. Li X, Spence JS, Buhner DM, et al. Hippocampal dysfunction in Gulf War veterans: investigation with ASL perfusion MR imaging and physostigmine challenge. *Radiology*. 2011;261(1):218-225.
29. Tillman GD, Calley CS, Green TA, et al. Event-related potential patterns associated with hyperarousal in Gulf War illness syndrome groups. *Neurotoxicology*. 2012;33(5):1096-1105. doi:10.1016/j.neuro.2012.06.001.
30. Jamal GA, Hansen S, Apartopoulos F, Peden A. The "Gulf War syndrome": is there evidence of dysfunction in the nervous system? *J Neurol Neurosurg Psychiatry*. 1996;60(4):449-451.
31. Jones KH, Dechkovskaia AM, Herrick EA, Abdel-Rahman AA, Khan WA, Abou-Donia MB. Subchronic effects following a single sarin exposure on blood-brain and blood-testes barrier permeability, acetylcholinesterase, and acetylcholine receptors in the central nervous system of rat: a dose-response study. *J Toxicol Environ Health A*. 2000;61(8):695-707.
32. Henderson RF, Barr EB, Blackwell WB, et al. Response of rats to low levels of sarin. *Toxicol Appl Pharmacol*. 2002;184(2):67-76.
33. Morris M, Key MP, Farah V. Sarin produces delayed cardiac and central autonomic changes. *Exp Neurol*. 2007;203(1):110-115.
34. Low PA, Denq JC, Opfer-Gehrking TL, Dyck PJ, O'Brien PC, Slezak JM. Effect of age and gender on sudomotor and cardiovagal function and blood pressure response to tilt in normal subjects. *Muscle Nerve*. 1997;20(12):1561-1568.

Appendix B

Two published scientific articles explaining the Dozmorov statistical method of analyzing hypervariably expressed genes

Published online 31 August 2009

Nucleic Acids Research, 2009, Vol. 37, No. 19 ***6323–6339***
doi:10.1093/nar/gkp706

Internal standard-based analysis of microarray data. Part 1: analysis of differential gene expressions

Igor Dozmorov[1,*] and Ivan Lefkovits[2]

[1]Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA and [2]Department of Biomedicine, University Clinics Basel, Vesalianum, Vesalgasse 1, CH-4051 Basel, Switzerland

Received February 25, 2009; Revised August 7, 2009; Accepted August 10, 2009

ABSTRACT

Genome-scale microarray experiments for comparative analysis of gene expressions produce massive amounts of information. Traditional statistical approaches fail to achieve the required accuracy in sensitivity and specificity of the analysis. Since the problem can be resolved neither by increasing the number of replicates nor by manipulating thresholds, one needs a novel approach to the analysis. This article describes methods to improve the power of microarray analyses by defining internal standards to characterize features of the biological system being studied and the technological processes underlying the microarray experiments. Applying these methods, internal standards are identified and then the obtained parameters are used to define (i) genes that are distinct in their expression from background; (ii) genes that are differentially expressed; and finally (iii) genes that have similar dynamical behavior.

INTRODUCTION

Microarray technology provides a genome-wide screening and monitoring of expression levels for thousands of genes simultaneously, and has been extensively applied to a broad range of biological and medical problems in order to identify changes in expression between different biological states. The immense amount of information that can be obtained from microarray studies enables us to address a variety of different research aims but still presents a challenge for data analysis, especially in terms of mutually exclusive parameters such as sensitivity and specificity. Many excellent reviews have been written on this subject (1–4). Our intention is, rather than providing an overview of available approaches, to offer a presentation of our methodological approaches with the main emphasis of using internal standards as means of robust evaluation strategy. Some of the methods have been published at least in part, others are completely new.

Methods based on conventional *t*-tests estimate the probability (*P*) that a difference in gene expression occurred by chance. If the threshold for probability chosen as significant in the context of a small sized experiment is applied in another microarray experiment, it can have a high false positive rate. For example, if the *P* threshold is 0.01, then even a set of random data satisfying the null hypothesis will result in one false positive per every 100 genes tested. A microarray containing tens of thousands of genes will generate hundreds of false positive results.

Two of the most popular approaches to address this problem are to make adjustment of thresholds or to use various combinatory calculations in order to improve the power (sensitivity) and specificity of the statistical conclusions. Due to its simplicity, the Bonferroni adjustment was used frequently despite its well-known conservativeness. The correction of *P* threshold by dividing the desired significance by the total number of statistical tests performed, ensures the achievement of a desired false positive rate over the entire set of genes, but conversely sets a criterion that can be too strict for each individual gene. Specificity is gained at the expense of sensitivity. Thus, the method does not reject hypotheses as often as it should and therefore it lacks power. This is of course a paradoxical situation, since the statistical significance for each individual measurement apparently depends on the total number of unrelated measurements.

None of the various attempts to improve Bonferroni adjustments has helped to resolve the problem. The most popular of such adjustments, the so called false discovery rate (FDR) control (5,6) that has been introduced into microarray analysis by Benjamini and Hochberg (7) enables to estimate the measure of the proportion of rejected null hypotheses. All genes are ranked according to their *P*-values and tested against individualized thresholds: the smallest observed *P*-value is tested against the strictest threshold, and the remaining *P*-values against successively more relaxed thresholds. In other tests, e.g. in the popular significance analysis

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

of microarrays (SAM) method (8,9), the use of individualized thresholds improves the conservativeness of the Bonferroni test, though the improvement is only partial and often minor.

The relative difference in gene expressions computed from replicated hybridizations provides a control for random fluctuations, the power of which depends essentially on the number of replicates. To improve statistical significance of biological variation without increasing the number of replicates, additional controls are needed. In the aforementioned methods, like SAM, 'instead of performing more experiments', which are expensive and labor intensive, Tusher *et al.* (8) generated a large number of controls using re-sampling methods such as bootstrap or permutation to estimate the underlying distribution from the observed data. However, generation of larger number of controls by using combinatory approaches instead of performing more experiments is somewhat illusory in that it does not truly increase the amount of information being analyzed.

Fortunately, there exists an adequate resource to increase the power of statistical tests by using the massive quantity of information inherently obtainable in each microarray experiment. We introduce here an approach in which the paired comparison of gene expression in two different situations is accompanied by the associative test—checking the hypothesis that each given gene in the experimental group has common features and can be associated with an internal standard. Internal standard in this context is considered as a large family of genes sharing some useful features for analysis, which in turn are neither dependent on the particular gene sequence nor on the level of expression, and are also not dependent on the coordinate position in the chip.

The methodology of the evaluation described in this communication will serve us as a stepping stone to our further effort of using internal standards for analysis in a statistically robust manner, functional associations through clustering and networking genes having similar dynamical behavior. These methods are equally applicable to time course dynamics initiated by various treatments and to natural variations of genes involved in essential dynamical processes in biological systems as well. This we intend to describe in the follow-up article (in preparation).

Early variants of some procedures described here were first included in the Matlab toolbox for microarray data analysis MDAT described in Knowlton *et al.* (10), while the improved and modified version exists now and is available on request.

MATERIALS AND METHODS

Gene expression datasets

This work uses a wide spectrum of experimental data that were only partially published.

The expression datasets were obtained with the use different sources of mRNA and different microarray technologies. They include Mouse Atlas 1.2 membranes and Mouse plastic 5K arrays Human Cancer Atlas 1.2 membranes (Clontech, Palo Alto, CA). Most data were obtained with the use of high-density microarrays.

Custom microarrays were prepared at the Oklahoma Medical Research Foundation Microarray Core Facility using commercially available libraries of oligonucleotides: Human Genome Oligo Ser Version 2.0 and mouse genome set, version 2.0 (Qiagen, Valencia, CA).

All data of recent years were obtained with the use of Affimetrix U133 Plus 2.0 and U95 GeneChips (Human) and Mouse genome 430 2.0 arrays, and the BedArray technology—Illumina Sentrix® Expression BeadChip microarrays.

Microarray data analysis

Our methods of data normalization and analysis are based on the use of internal standards that characterize some aspects of system behavior such as technical variability. In general, an internal standard is constructed by identifying a large family of similarly behaving genes. These internal standards are used to estimate in a robust manner those parameters that describe some state of the experimental system such as the identification of genes expressed distinctly from background, differentially expressed genes and genes having similar dynamical behavior. This will be elaborated in detail in the Results section.

Résumé 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 six sequential step-by-step résumés.

Step-by-step Résumé 1: individual normalization of the microarray data to background.
Step-by-step Résumé 2: determination of parameters and adjustment of the normalized profiles.
Step-by-step Résumé 3: two-sample data adjustment.
Step-by-step Résumé 4: multi- sample data adjustment.
Step-by-step Résumé 5: reference group of equally expressed genes.
Step-by-step Résumé 6: gene expression analysis.

RESULTS

Statistical monitoring of weak spots

Among the most controversial aspects of the treatment of data that are related to low-intensity signals, is the procedure that enables to distinguish between true (specific) hybridization signals and technological noise. In this context, we consider the genes either as 'expressed' or 'non-expressed' though this discrimination is not based on biological but rather on technological difference. Depending on the sensitivity of the used technology and on technical quality of experiments, the same low-expression level genes could be treated in high-quality experiments as being expressed (distinctively from nonspecific noise), while in 'soiled' experiments (with

high level of non-specific hybridizations and/or background noise) they would fall in the category of non-expressed genes. The importance of discrimination of these genes is related to their different information content for subsequent analytical procedures.

Ratio of expressed to non-expressed genes is not a meaningful term. Ratio analysis is commonly employed to determine expression differences between two samples. However, any procedure that uses raw intensities to infer relative expression is imperfect due to the fact that accuracy is signal-level-dependent, with variations increasing dramatically for low intensity signals (9,11,12). Besides, only those ratios that are based on expressed genes are meaningful. The best demonstration of this statement could be obtained with array consisting of duplicated spots for each gene (13). Figure 1 presents results of such an analysis with the use of data from Clontech membrane array (analogous results were obtained also with Perkin-Elmer Micromax cDNA arrays of 2400 human genes spotted in duplicates—not shown). The histogram for the distribution of all spots on the array is presented in Figure 1A. Ratios of duplicated spots that should be equal to 1 with some systematic variations are depicted in Figure 1B. However, this appeared to be the case only for genes expressed above certain threshold level (in this particular set, the threshold being 3). Below this threshold, the ratios are highly variable, demonstrating the absence of any agreement with the duplicate expressions. It follows that the removal of the background level spots should precede any microarray data analysis based on the use of expression ratios.

Technologically non-expressed genes represent non-correlated noise. The distribution of the ratios similar to that presented in Figure 1B could also be obtained with expression profiles of samples from a homogenous group, where one expects equal expression of the vast majority of genes. Drastically distorted ratios below a certain level of expression suggest that low levels of gene expression lack any correlation (Figure 1C). A sharp border that discriminates correlated expressions from non-correlated noise is obtained when 'sliding window' approach for comparison of the ratio variations (Figure 2) is used. In the presented comparison one set is sorted, while keeping gene association with the second set. Thereafter, an *F*-test is performed for the standard deviation (SD) of ratios of genes in the 'window' (the 10th lowest one is sample one) compared with the SD of ratios of all remaining genes with highest expression. When the window moves like a stencil along the data stream, one obtains comparative characteristics of ratio variability depending upon expression level. There is a sharp border for the *P*-value (probability for identity of SD in *F*-test) in this dependence as shown in Figure 2B. Above this threshold, there are all possible levels of *P*-values from 0 to 1 (10 sequential genes could have very similar levels of expression when the majority of genes in homogenous group of samples are equally expressed), however there are no exclusions for low-expression levels, i.e. all *P*-values here are close to zero indicating absence of any correlation in the noise level expressions. The border obtained for background noise appears to be in good agreement with the method for obtaining the zone of normally distributed background noise through iterative procedure described below.

Normally distributed additive noise is a convenient internal standard for 'non-expressed genes'. Several methods have been developed to select 'non-expressed genes' and hence to diminish the influence of background noise. One such solution is to ignore genes that yield low total abundance transcripts, another one is to exclude weak spots arbitrarily [in the work of Kooperberg *et al*. (11)] an intensity cutoff was chosen such that the relative error in ratios was <25%) and still other one is to compare spot expressions with local background level (see Dozmorov *et al*., 2004 (13) for review). Those procedures for flagging and excluding weak spots that are not based on robust statistical criterion remain problematic since potentially valuable data might be discarded. This issue is compounded by the fact that in biological systems a number of key regulators might be expressed at low levels presumably to ensure a tight control of the expression of regulatory entities (14,15).

The work of Churchill *et al*. (14) is the first example of solving the problem efficiently with the use of an internal standard. The two main sources of heterogeneity in gene expression variations are indicated in Rocke and Durbin (16) by including the 'additive component', prominent at low-expression levels, and the 'multiplicative component', prominent at high-expression levels. The intensity measurement $y_{i,j}$ for gene $I \in I = \{i_1, \ldots, i_n\}$ in sample $j \in J = \{j_1, \ldots, j_m\}$ is modeled by the equation $y_{i,j} = a_{i,j} + (m_{i,j} e^h + e_{i,j})$, where $a_{i,j}$ is the normal background, m_{ij} is the expression level in arbitrary units, $e_{i,j}$, is the additive error term within a spot, and h is the second error term, which represents the multiplicative component. Gene expression data obtained with the standard procedure of local background subtraction will not exclude spot intensities $e_{i,j}$, which present additive noise above background levels. The distribution of the spots with $e_{i,j}$, as predominant member of intensity depends on the array technology used and on the quality of data. Atlas arrays (Clontech) are a good example of high-quality membrane-based arrays exemplifying high specificity and low levels of background. Background spots comprise up to 50% of all spots on the array. The nearly normal distribution of this noise can be seen in a histogram of all intensity values (Figure 3A and B). Parameters of this distribution were estimated with the use of the multi-step iterative procedure.

First—the expressed genes are excluded one by one as their values exceed the mean $\pm$ 2SD of the core of non-discarded genes. This procedure is repeated in an iterative manner until no additional spot is excluded and the resulting non-discarded values represent the set of non-expressed genes (Figure 3C).

Second—the parameters of the additive noise are estimated by non-linear fitting of a normal distribution function to the core of non-expressed values. The

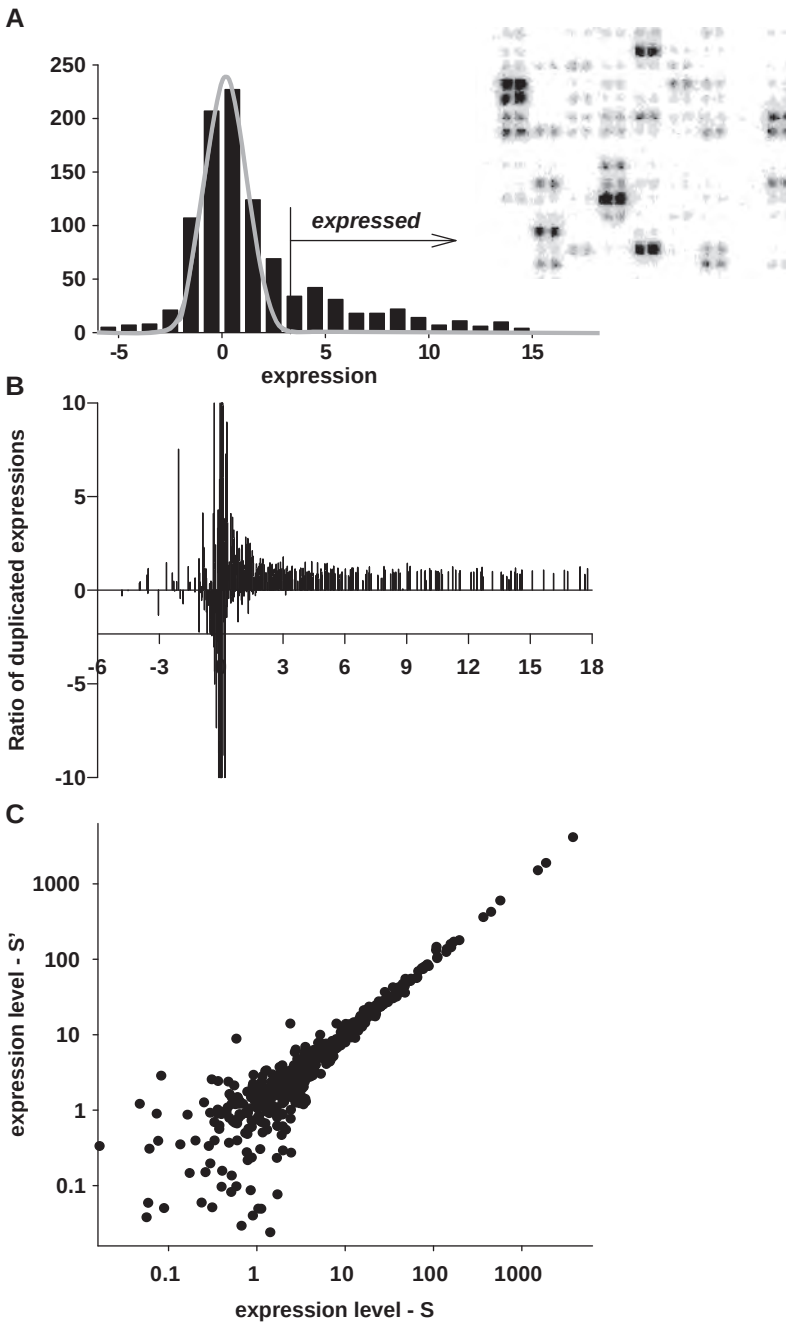


Figure 1. Ratio of the duplicated spots in the area of background noise is meaningless. (A) Localization of the normally distributed background noise in the histogram of all microarray gene expressions using iterative exclusion procedure (see Figure 2 and explanations in text). (**B**) Ratio of the expression levels of the duplicated spots demonstrates increased variability in the area of low-intensity expressions. Fragment of array with duplicated spotting is shown in the right-upper corner. (**C**) Lack of correlation between the intensities of duplicated spots of low intensities. The axes present intensities of the duplicated spots.

parameters of this distribution [average (Av), SD and the number of members] completely characterize this internal standard of 'absence of expression'. After that data normalization proceeds by assigning to each experimental value, a normalized score S using the formula $S' = (S - \text{Av})/\text{SD}$. As a result, the internal standard of the 'absence of expression' has a mean of zero and SD = 1 and all gene expressions on array are presented in the SD units of this internal standard.

The iterative procedure described above for discarding the gene expression that alters the normality of the background noise is efficient only with array technology that yields a major gap between the value range of this noise distribution and the set of values of the expressed genes. This was the case with the data obtained with high-quality Clontech membrane array using very sensitive radioactive probes and ensuring that for the probe synthesis only gene-specific primers are used. With these

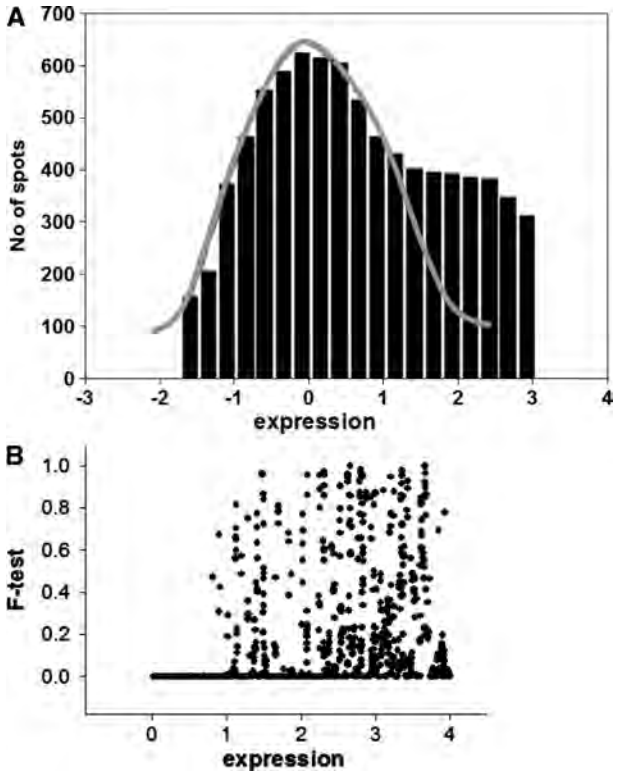


Figure 2. Selection of the normally distributed background noise in the presence of low expressed genes. (**A**) Histogram of the low-spot distribution after iterative cutting off the expressed genes (see details in text). The presence of low-expressed genes causes in some instances skewing the right side of the background distribution even in high-quality microarrays. For this case, only the left-portion residual after trimming is not distorted by the presence of expressed genes. For estimation of the parameters of the noise distribution, a new histogram is created by substituting the right portion of the background distribution with the mirror image of the left portion. The parameters of the noise distribution are estimated by non-linear fitting of a normal distribution function to this histogram. (**B**) The sliding window method for estimation of the changes in correlation between gene expressions depending on the level of expression. The *F*-test is performed for SD of ratios of genes in the 'window' (for 10 genes with lowest expression in sample one) compared with SD of ratios of all remaining genes with highest expression. The appearance of the sharp decrease of the *P*-values (probability for identity of SD in *F*-test) evidences about the existance of the area of low expression whose variations exceeded significantly the variations of the majority of the rest gene expressions. The position of the sharp decrease of the *P*-values shows the border for the non-correlated background noise.

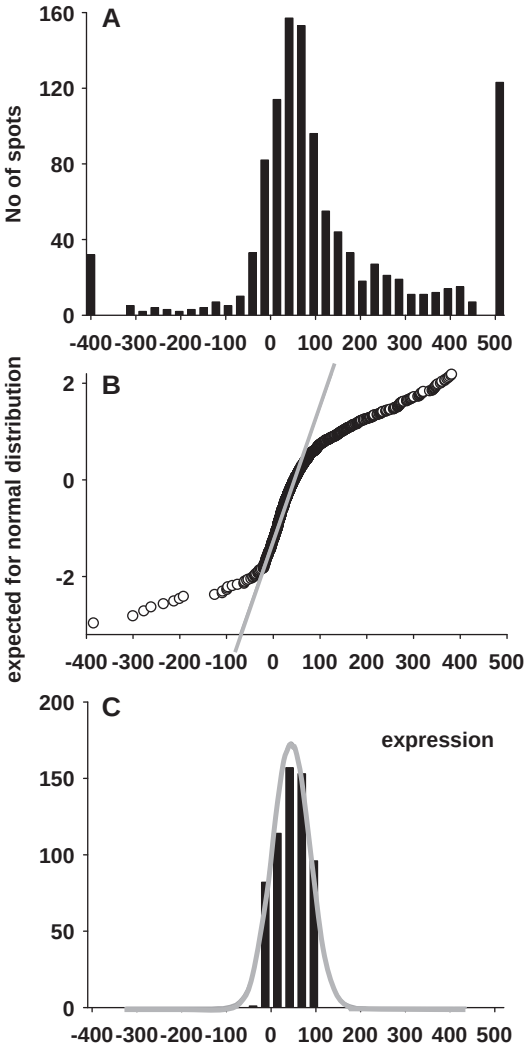


Figure 3. Procedure of normalization of the gene expression profile. (**A**) The histogram of overall gene expressions fits poorly to a normal distribution, with noticeably extended left and right tails. Values at the left tail results from the background correction procedure, while values at the right tail correspond to genes expressed above background. (**B**) Normal probability plot demonstrates deviations from normality in the tails of the A-distribution. (**C**) The results of iterative removal of residual background spots demonstrate a good fit to normal distribution. This histogram is used for the estimation of the parameter of normal distribution through the non-linear least-squares curve fitting procedure. Once the parameters of the normally distributed background noise are determined, all expression data are transformed, yielding mean = 0, SD = 1 for background distribution. All gene expressions are presented now in the SD units of the background distribution.

measures, the distance between normally distributed additive noise and majority of low-expressed genes in the histogram is promoted (Figure 3A).

This is not always the case when oligo or random primers are employed. Even in high-quality fluorescently labeled oligonucleotide microarrays (Affimetrix), the distribution of low-intensity noise spots might turn out to be unsatisfactory. The right side of the distribution is often skewed by the abundance of low-expressed genes. This skewness of the distribution can be present even in the histogram obtained upon application of the iterative procedure as shown in Figure 2A. For this case, only the left side of the histogram is used for the estimation of the parameters of the noise distribution. A new histogram is created substituting the right portion of the background distribution with the mirror image of the left portion. Curve fitting is then applied to the new histogram in order to obtain parameters of the noise distribution for subsequent normalization of the array data. This approach to the characterization of the noise distribution seems to be more adequate than attempts to approximate the distorted distribution with artificial combination of overlapping distributions (17,18).

Microarray profiles with relatively low content of non-expressed genes generate another type of problem for localization of the background distribution. The background level spots represent only a relatively small portion of all spots (<30%) in these arrays, thus their distribution is not as prominent as in the previous examples when viewed in a histogram of all spots. The automated iterative procedure for selection of background described above will not locate the background distribution. Therefore, it is necessary to perform a special preliminary step intended to increase the area of the background distribution and focus the iteration procedure onto this area—initial selection of the lowest 30th percentile of data. Then, the new sub-set is trimmed and subsequently curve fitted (see above).

Statistical significance of gene expression—signal/noise discrimination. As we demonstrated earlier (13) the additive noise distribution is quite homogenous over the whole chip after the background correction procedure that makes it possible to use weak spots from the entire chip for estimation parameters of its distribution and use them as a united internal standard for non-expressed genes. Discrimination of 'expressed' from 'non-expressed genes' is based upon the use of recognized statistical criteria instead of subjective cutoff rules. The power of this statistical criterion is determined by the content of the internal standard—normally several thousand members—and this enables to use relatively high-statistical thresholds without loss of the sensitivity of the selection.

In a replicated experiment, genes that are expressed distinctively above the background noise are readily identified by paired analysis. As it is demonstrated below, data from a replicated experiment upon proper normalization can be used for statistical discrimination of even very weakly expressed genes from the normally distributed noise. Genes with low-level signals—even within background area—could also be identified distinctively from the background due to their higher stability (low SD in replicate measurements).

Step-by-step Résumé 1: individual normalization of the microarray data to background.

The mean and SD are calculated. Using these as a starting point, data beyond +2SD above the mean are cut and discarded. The mean and SD are recalculated and data beyond −2SD below the mean are cut and discarded. This trimming of outlier values above and below is continued, further refining the SD estimate, until no additional cuts can be made.

The rest of data are used for creation of the 10 bar histogram of expression distribution.

Interactive curve fitting for the trimmed data is performed. Using final trimmed data mean and SD are estimated. Theoretical normal distribution is established with estimated mean and SD. Using the theoretical estimate, a non-linear least square curve fitting procedure is performed in order to improve the SD estimate. The quality of fitting is determined visually. If there is some visual distortion of the right tail (proposed presence of weak gene expression) the procedure is repeated using a new user-defined mean (Histogram bars 1–5) and estimating the new distribution on the bars to the left of the chosen one.

In case of low-quality arrays with the abundance of weak expressions distributed too close to background noise the initial choice of the lowest 30th percentile of data is selected to eliminate highly expressed values. Then, the new sub-set is trimmed and subsequently curve is fitted as described above.

Once an appropriate fit is achieved and parameters of the normally distributed background noise is determined as m and s then all the data is Z-transformed $Z = ((x - \mu)/\sigma)$ yielding Mean = 0, SD = 1 for background distribution. All gene expressions are presented now in the SD units of the background distribution.

Finally, the data are log-transformed in such a manner that the negative values are substituted with the log of the minimum positive value.

The follow-up is given in the Step-by-step Résumé 2.

Data adjustment

Individual normalization of data from each chip to their background is not sufficient for making their profiles comparable, because first—backgrounds are often different in different experiments, and second—there might be several additional reasons for systemic differences in the expression profiles that can be compensated only by two-parametric regression procedure. This procedure is described in details in the next section and the important feature of it is that this procedure is based on the comparison of potentially equivalent gene expression correlated in compared profiles. The background non-correlated noise could be a serious obstacle for such procedure as it is shown in Figure 1. Knowledge of the background distribution parameters enables to remove the non-correlated noise from correlation adjustments. The threshold 3SD above the mean of background excluded the noise with excess before the final adjustment is made.

The observed variations of the intensity of spots result from biological changes in gene expressions and also due to stochastic and systemic variations that occur in every microarray experiment. In order to accurately and precisely measure gene expression changes, it is important to minimize systemic variations and to estimate the contribution of stochastic variations. Systemic variations appear due to differences in experimental conditions and come from many sources such as procedures of sample handling, methods of cell cultures, methods for mRNA isolation, extraction and amplification, hybridization conditions and labeling efficiencies, as well as due to contamination by genomic DNA [major sources of fluctuations in microarray experiments were listed and discussed in several publications (19)]. The purpose of normalization is to minimize systematic variations in the measured gene expression levels of replicative experiment. Once this is achieved, estimation of the parameters of the stochastic variations the biological differences can be more readily accomplished.

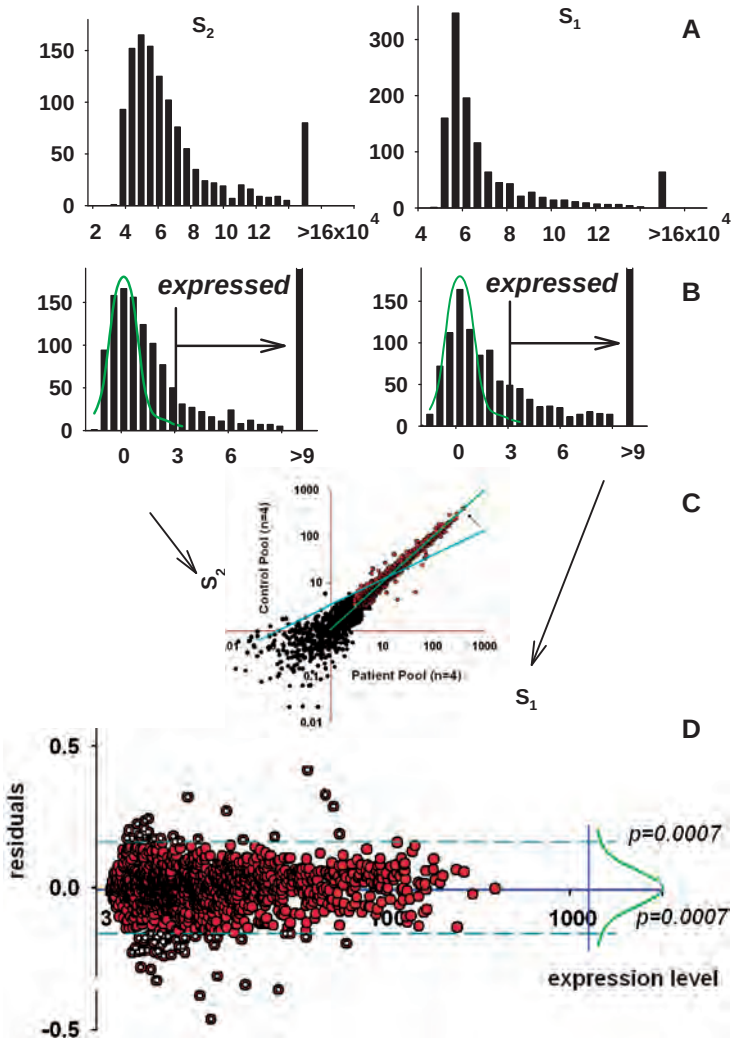


Figure 4. Two-sample data adjustment. (**A**) Histograms for the spots in two arrays. (**B**) Histograms for normalized to background and log-transformed data. Normal distribution curves fitted to the truncated histograms (as in Figure 3) are shown in green. (**C**) Profiles of control pool (*y*-axis) and patient pool (*x*-axis) adjusted to each other through linear regression with excluded background noise (black spots) and potential outliers. Blue line, position of the regression line before adjustment; green line, position after adjustment. (**D**) Data of the plot presented in the transformed coordinates. Right side shows the nearly normal distribution of the deviations from equity of expression. The use of the majority of equally expressed genes as an internal standard presents opportunity to select differentially expressed genes as outliers from this standard beyond of some statistical thresholds. These genes are shown as open circles.

In several excellent reviews, there were proposed different methods of normalization that relieve us from the necessity to discuss them in details (1–3,20). We will note only that the two independent sources of systemic variability in microarray data (additive and multiplicative) need normalization procedures.

Two sample data adjustment

The regression analysis of duplicates from the same array (Figure 1C) presents an excellent example of data having only stochastic variations. Neither multiplicative variations due to differences in hybridization or due to labeling conditions nor additive variations due to non-compensated background noise occur. Both these sources of systemic variations are equal for duplicated spots at the same chip. After exclusion of the area of common non-correlated noise and log-transformation of the data, gene pairs are presented in the scatter plot as dots close to the straight line intercepting the origin with slope 1. The log-transformation is the simplest one making individual gene spots deviating from regression line independently on the level of gene expression and which is normally distributed. The normality can be proven graphically (normal probability plot) or analytically—applying Kolmogorov–Smirnov criteria.

A scatter plot of data from two independent arrays will demonstrate additional systemic variations: 'additive'—due to differences in background (leading to the deviation of the regression line from the coordinate origin; position of the initial regression line is shown as blue straight line in Figure 4C) and 'multiplicative'—due to the overall difference in the spot densities (leading to the change of the slope of regression line)—Figure 4C. Transformation

of one of these datasets will minimize this differences and make the scatter plot similar to the one obtained for duplicated spots where additional multiplicative and additive differences are absent (compare Figures 1C and 4C).

An array of gene expression profiles may be conceptualized as a vector of outcomes in the scatter plot of data. Let $Y_k = (Y_{1k}, Y_{1k}, \ldots, Y_{Jk})$ denote the array, where Y_{jk} denotes the expression of the j-th gene in the k-th sample ($j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$).

$$Y_{jk} = \partial_k + \lambda_k(a_j + b_j x_k) + \varepsilon_{jk},$$

in which (a_j, b_j) are gene-specific additive and multiplicative factors, (∂_k, λ_k) are the sample-specific regression coefficients, and ε_{jk}, is used to depict variations due to all unknown sources. Estimated regression factors are used for overall adjustment of the expression levels in one sample to another as $(Y_{jk} - \hat{\partial}_k)/\hat{\lambda}$. After these adjustment relations of the expressions in two samples presented as $Y_{jk} = a_j + b_j x_k$ are obtained where a_j presents the difference in local background and b_j—multiplicative factor. For data acquisition with local background subtraction the a_j are minimized or even disappear and a log-transformation produces expressions differing by the additive close to normal distribution noise $\log(b_j)$ that is an unified measure variation in gene expression essentially unrelated to the influence of level of expression.

The described adjustment leads to maximal similarity of expression of all genes in two arrays. This procedure, however, will be incorrect in the presence of differentially expressed genes, because it will aspire to make them equally expressed also. It means that the presence of differentially expressed genes can seriously impede the adjustment procedure. Generally, their influence could be minimal if they are distributed more or less symmetrically around regression line. However, the presence of not compensated outliers might influence the bias adjustment drastically, especially when such unbalanced outliers are present in the area of very high expressions—usually area less populated with spots. These outliers violate the assumption of normally distributed residuals in least squares regression. They tend to pull the least squares fit too much in their direction by getting considerably more 'weight' than they deserve.

Various methods were proposed to diminish the distorting influence of differentially expressed genes. They were based mainly on arbitrary estimations of permissible distances from equity line. The procedure of revealing and down weighting could be produced on the strong statistical basis using another internal standard—family of equally expressed genes. Fortunately, in any normal experiment, the majority of genes are equally expressed, and their variations around regression line have prominent distribution that can be elicited by the iteration procedure described earlier for background data analysis. Such stochastic distribution of the deviations of gene positions looks very similar to the distribution obtained for duplicated spots in Figure 1C. A histogram of these deviations (Figure 4D) includes the normal distribution with tails distorted by the presence of differentially expressed genes that could be selected and excluded once the parameters of the normal distribution are determined.

The stochastic distribution of the random variations is typically unknown. In our practice of making hundreds of analyses using different technological platforms, we were never confronted with a violation of the normality assumption, nevertheless, if hypothetically the assumptions of normality are violated, some non-parametric criteria will be more reliable for making statistical inferences—as. For example, Thomas *et al.* (21) proposed to use Z-scores that is closely connected with Wilcoxon rank sum statistics (22). Z-scores do not require any distributional assumptions or homogeneity of deviations. In practice, Z-scores are expected to be similar to t-test statistics, when the distribution of expression levels can be approximated by the normal distribution. When these assumptions are violated, Z-scores will differ from t-statistics and will be more reliable for making statistical inferences.

Step-by-step Résumé 2: determination of parameters and adjustment of the normalized profiles.

The first step is the determination of the parameters of the background of the array—Av and SD of normally distributed low-level expressions in array with subsequent normalization of all expressions in 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 SD of the set of background spots. The distribution of S has mean of zero and SD = 1 over the set of background genes in the normalized array. We accept $S = 3SD$ above the mean background level as the preliminary criterion for distinguishing expressed from non-expressed genes. Only genes expressed above background are used for the second step 'adjustment' as described below.

The second step is the adjustment of the normalized profiles to each other by robust regression analysis of genes expressed above the 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 regression line. The parameters of this distribution are obtained by iterative procedures similar to the one used before for the selection of the kernel part of normally distributed background noise. Outliers are thereafter determined as having deviations not associated with this internal standard of equity in expression including thousands of members (Figure 4D).

The follow-up is given in the Step-by-step Résumé 3.

Nonlinear regression. Linear regression analysis will be valid only if (i) the hybridization signal is linearly related to target concentration and (ii) the majority of the genes expressed in both samples are expressed equally. Bias adjustment transforms the dependence between two samples into a simple multiplicative model (see above). Sometimes, however such a model is inadequate. Such cases can be identified on the scatter plot when a straight line fits the data poorly and instead a curved shape results. The use of straight line for

normalization can lead to a high rate of false positive results. A variety of approaches to normalize such gene expression data have been proposed, including a cubic spline transformation (23,24), and locally weighted linear regression [Lowess; see for review Do *et al*. (2006) (2), Bolstad *et al*. (2003) (20) and Wu (2001) (25)].

Remarkably, the assumption that non-linear transformation is always beneficial for tests for differentially expressed genes has never been properly tested. Making the choice in favor of the non-linear normalization procedures, it is necessary to keep in mind that serious problems might occur in cases where the non-linearity is the result of non-homogenous distribution of differentially expressed genes of opposite directions. From this perspective, the non-linear transformation can be beneficial for the adjustment of profiles of samples from a homogenous group. However in a comparative analysis, this method bears a definite danger of losing sensitivity of discrimination of the differences in gene expression.

The examination of examples of non-linear distribution of gene expression in the regression plot indicates that in most cases essential non-linearity is present in the area of low-gene expressions. The exclusion of the background area and of the closely associated low-expressed genes is able to diminish considerably the influence of such non-linearity.

The residual essential non-linearity is an evidence of the low quality of the technological procedure and the best way to correct it is to avoid it in the first place. Examination of the quality of the data from high-throughput platforms 'prior to interpretative analysis' is a critical step that will help researchers to avoid contaminating their otherwise well-conducted study with samples harmful to overall analysis and interpretation.

Step-by-step Résumé 3: two-sample data adjustment.

- Regression analysis of two-sample data gives residuals (deviations from regression line) for each gene expressed >3 ($=0.477$ after log-transformation) in both samples.
- The mean and SD of all residuals are calculated. Using these values as a starting point for data trimming as described above, the parameters of the normal distribution of the majority of residuals are obtained.
- The probability of belonging to the normal distribution of the majority of residuals (for equally expressed genes) is estimated for each gene (each residual).
- Genes having probability less than $1/N$ (N—number of all genes expressed >3 in both samples) are excluded and the regression analysis for the rest of them is used for estimation and exclusion of additive and multiplicative factors.
- The result of adjustment can be presented in transformed coordinates with indicated borders $\pm(1/N)$ for differentially expressed genes (Figure 4D).

The follow-up is given in the Step-by-step Résumé 4.

Multiple-sample data adjustment

Many of the issues that we discussed in the two-sample case, such as bias correction, remain important for replicate experiments, although we will not discuss them further. Often the two-sample methods can be generalized to handle replicate experiments. For example, we can extend the methods for bias correction by normalizing across a series of N samples, rather than one sample against another. In this case, the solution involves fitting a normalization curve in an N-dimensional space. However, in practice, we successfully use different iterative procedures of normalization to common averaged profile as detailed in Figure 5. In this multi-step procedure, we use averaged profile for bias adjustment of each individual profile with subsequent recalculation of the averaged profile and repetitive adjustment.

Step-by-step Résumé 4: Multi-sample data adjustment.

- Averaged profile is calculated and each sample is adjusted to the averaged profile using robust regression procedure described earlier for two-sample adjustment.
- 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 enough.
- The result of the adjustment can be presented in transformed coordinates in form of Mean + SD of multiplicated residuals for each gene (Figure 6A).

The follow-up is given in the Step-by-step Résumé 5.

Reference group—an internal standard for replicate experiment

One of the problems in performing a reliable *t*-test from microarray data is to obtain accurate estimates of the SDs of individual gene measurements based on only a few measurements. It has been, however, observed that an overall reciprocal relationship exists between variance and gene expression levels, and that genes expressed at similar levels exhibit similar variance (26). Beside that, there were obtained transformations depriving variance dependence on the gene expression levels (27). Log-transformation is one of the simplest examples of such transformation. Therefore, it is possible to use this prior knowledge to obtain more robust estimates of variance for any gene by examining the expression levels of other genes within a single experiment.

After normalization, the residuals from the calibration data are used to provide prior information on variance components in the analysis of comparative experiments. After adjustment of the each array profile to the averaged profile for the control group, we obtain two new standards joined by the common name 'reference group'.

First, all genes are represented here by their residuals (relatively averaged profile) that after normalization and log-transformation loose their sample dependent and expression level dependent individualities (Figure 6A and C). As soon as absolute majority of genes in homogenous group are equally expressed, their residuals demonstrate very similar to normal distribution (Figure 6E).

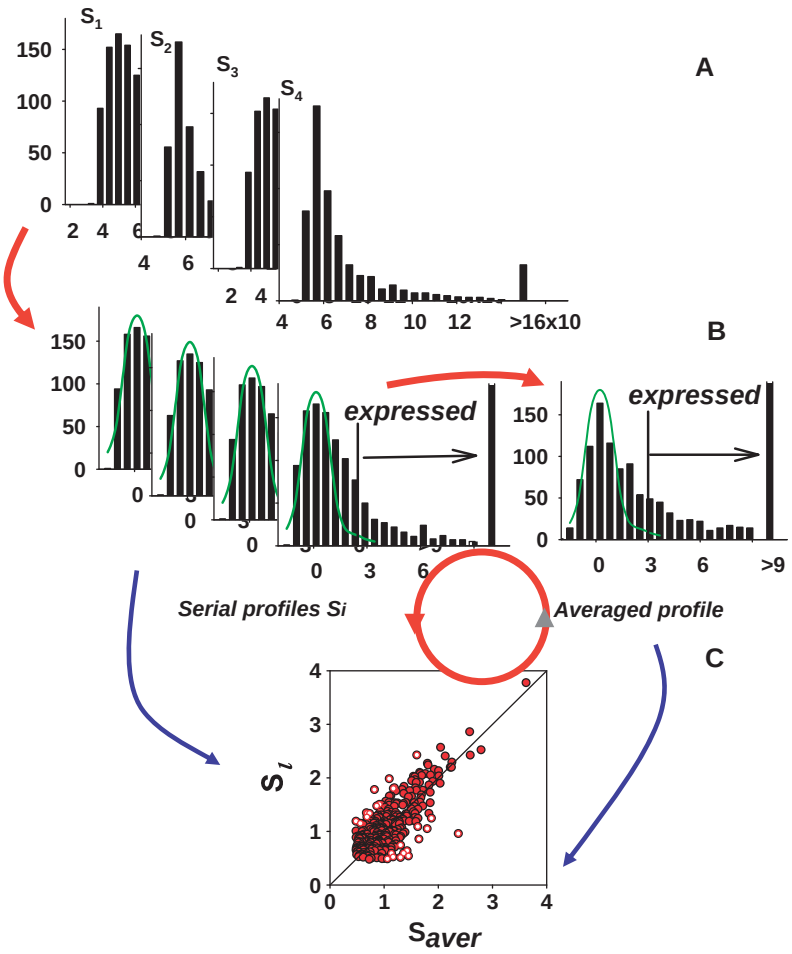


Figure 5. Multi-sample data adjustment to the averaged profile using robust regression procedure. Data first normalized to background with procedure described above (**A**). The averaged profile (**B**) is created and data at each sample adjusted to this average profile with robust regression procedure (**C**). After that, a new averaged profile from transformed data is created. Several subsequent cycles may be necessary for the best result, however, for the data initially normalized to background, two-steps adjustment is usually enough.

Second, the residuals of these genes in the replicated experiment could be presented as mean ± SD. For the majority of genes, their replicate variations are relatively small and homogenous following to the standard *F*-distribution. The small portion of genes having enormously high (statistically distinctive from the rest) variation present so called hypervariable genes (HV-genes), whose nature was discussed elsewhere (28,29). To get the internal standard for gene variability, HV-genes should be excluded by iterative procedure similar to described above (for normally distributed background events and for normally distributed residuals of equally expressed genes). The only difference is that in this procedure, the *F*-test is used as a criterion for the exclusion of outliers. To perform the *F*-test, we compare two estimates of variance, one from the variability of expression levels of the entire group, and the other from the variability of the expression level of every given gene. If the gene variability estimate is much higher than the total-group estimate, we have evidence that the given gene does not share the same stability as a majority of genes and should be excluded from the reference group. The procedure continues until no more genes could be excluded in this manner. The result of all these exclusions is a new internal standard—the reference group, composed of genes expressed above background in control samples with normal low variability of expression (as determined by an *F*-test) and whose residuals approximate a normal distribution.

Very similar standards for equity of expression and stable variability were introduced earlier by Rocke and Durbin (16). However, none of them were cleaned from HV-gene contamination, with the consequence that the standards were biased, thus decreasing significantly the sensitivity of the criteria.

Step-by-step Résumé 5: reference group of equally expressed genes.

In course of normalization with bias adjustment

(i) residuals as differences between final normalized expression and the average before last adjustment are calculated;
(ii) SD of all residuals taken together are calculated;
(iii) SD for all genes individually are calculated;

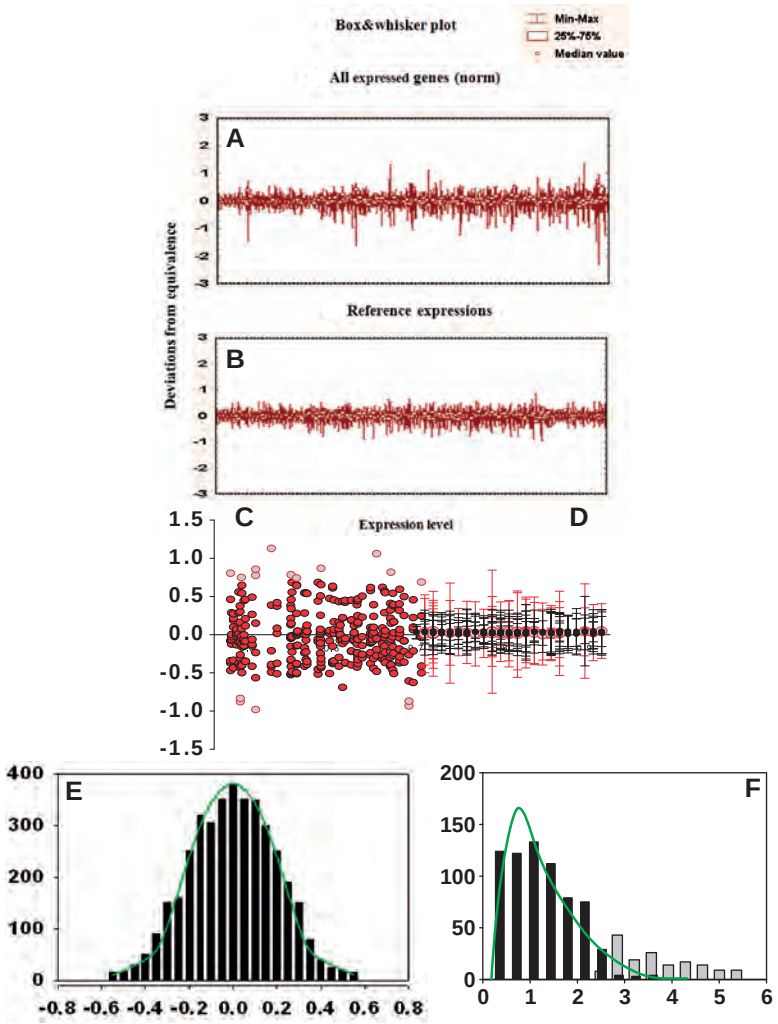


Figure 6. Reference group—the main internal standard for Associative Analysis of differentially expressed genes. The reference group (**B**) is created from initial distribution of the residuals (**A**) after trimming of hyper-variably expressed genes (HVE genes) with use *F*-test. (**C** and **D**) Reference group as an internal standard for equity in expression [normally distributed deviations in the left part (**E**)] and for stability of expression [*F*-distributed SDs in the right side (**F**)].

(iv) *F*-test is performed on every gene to determine if the variability is higher than that of all genes;
(v) all genes whose SD is higher than in step (ii) and/or fail *F*-test are excluded;
(vi) SD for all remaining genes are recalculated;
(vii) steps (iv)–(vi) are repeated until no further genes can be excluded

The follow-up is given in the Step-by-step Résumé 6.

Associative analysis—identification of differentially expressed genes

The use of the reference group created in the previous section, as an internal standard enables to carry out differential gene expression analysis, and what is of utmost importance, it solves the problem of mutually exclusive characteristics of sensitivity and specificity. For this purpose, we use an associative *t*-test (30) developed as a modification of the 'General Error Model' (16) in which the replicated residuals for each gene of the experimental group are compared with the entire set of residuals from the reference group. The null hypothesis is checked to determine if gene expression in the experimental group is associated with the reference group. The significance threshold is corrected to make the appearance of false positive determinations improbable.

Selecting differentially expressed genes relies on five statistical steps.

- Assume Group 1 has n samples and k genes and Group 2 has m samples and k genes. A Student's t-test is performed, with $(n + m - 2)$ degrees of freedom, in order to determine if the genes are equally expressed.
- Then an associative t-test is performed, with $(m + k - 2)$ degrees of freedom to see if the gene belongs to the group of equally expressed genes with stabile variability. Selections passing through both tests have high sensitivity (Student's t-test

with normal low threshold $P < 0.05$) and high specificity (subsequent associative *t*-test with corrected threshold $P < 1/k$ excludes all false positive determinations).
- Another two Student's *t*-tests are used to establish the distinction from technical noise—discrimination of 'expressed' from 'non-expressed' genes.
- Finally, the ratio of gene expressions in Groups 2 and 1 is used to help exclude statistically significant but not biologically significant changes.

Clearly, simple discriminations based on 'fold changes' or ratios are insufficient for drawing proper conclusions. But, we use foldness restrictions as an addition to the statistical analysis of differentially expressed genes to concentrate attention on the most prominent differences first of all.

The *t*-test assumes that the replicate data have an underlying normal distribution. This assumption is reasonable, especially if the replicate samples are relatively homogeneous. Note that the assumption of normality is different in these two subsequent steps of the analysis. In the first step—paired comparison—in most cases, we have relatively few replicate samples and it is difficult to test for normality having only a few data points. Therefore, we often adopt the assumption of normality because it is hard to prove otherwise. In the second step—associative analysis—we use the reference group as an internal standard and proved that after log-transformation and exclusion of outliers with iterative procedure the rest of residuals has a distribution whose normality is confirmed by statistical and graphical criterions.

The two step procedure allows the use of traditional low-level significance cutoffs ($P < 0.05$) at the first step without the risk of including false positive selections. These false positives are excluded in subsequent second step—associative analysis having extreme statistical power enabling to use the significance cutoff corrected to the number of comparisons without risk to loose sensitivity. The use of the reference group enables to receive all benefits of the thousands replicates of technical variations—deviations from equity—to increase statistical power of the comparative analysis. This analysis is based on an idea, which is opposite to the commonly held view that large-scale array experiments suffer from compensatory tradeoffs in sensitivity and specificity. In fact, the procedures presented herein demonstrate that large scale datasets are extraordinary information-rich and provide means for discrimination of common technical variation from individual biological variability. More evidence of this is presented in a power analysis (Figure 7).

Step-by-step Résumé 6: in this step, gene expression analysis is described.

- Selection with a Student's *t*-test for replicates using the commonly accepted significance threshold of $P < 0.05$. It keeps the commonly accepted sensitivity level, however a significant proportion of genes identified at this threshold level as differentially expressed will be false positive determinations.

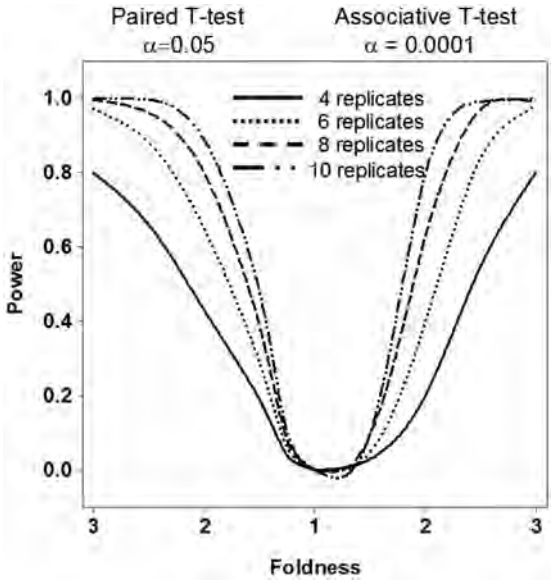


Figure 7. Power analysis. Estimation of the number of microarray experiments required to obtain reliable results from a comparison of data from patients and controls. The sample size was estimated using PASS 2005 (Keysville, Utah). Our experience with different array technologies (including 'Illumina', which is used here) indicates that a coefficient of variation between 0.25 and 0.5 is typical among expressed genes. The left portion of graph demonstrates the dependence of the power of analysis on the number of replicates for a paired *T*-test with a statistical threshold of $\alpha = 0.05$. On the right portion of the graph, power analysis results from an associative analysis are estimated. An associative analysis with threshold of $\alpha = 0.0001$ has power comparable with a paired *T*-test using a threshold of $\alpha = 0.05$. Results of this analysis will be used for estimating the number of replicate experiments required for selection of differentially expressed genes. For example 2- to 3-fold difference can be observed with power $1 - \beta = 0.8$ with a 6-replicate experiment.

- An associative *t*-test in which the replicated residuals for each gene of the experimental group are compared with the entire set of residuals from the reference group defined above. Ho hypothesis is checked if gene expression in experimental group presented as replicated residuals (deviations from averaged control group profile) is associated with highly representative (several hundreds members) normally distributed set of residuals of gene expressions in the reference group. The significance threshold is corrected to make the appearance of false positive determinations improbable. Only genes that passed through both tests were presented in the result tables.
- Genes expressed distinctively from background were determined by analysis of the association of each replicated gene expression with normally distributed background having Av = 0 and SD = 1. Genes expressed distinctively from background in one group and not distinctive from background in another group are given as further example of differentially expressed genes.

Data filtration and error exclusion procedures

Selection of 'bad' samples. The local errors in the data acquisitions will be able to produce significant increase

of the SD for given gene in replicated experiment. It is possible to use the *F*-test for selection of such errors, however the problem of the sensitivity/specificity alternative will prevent from accurate estimation of outliers. At the same time, the summary estimation of such outliers in every given sample will enable to characterize overall quality of the array data in every chip. We propose a simple program for the chip quality estimation. In a homogenous group of samples for each gene in the array, we estimate the changes in its variability by comparing the SD of the total set of expressions with the SD obtained after exclusion of one replicate after another. If the *F*-test results in probability for no difference being <0.05 then this gene expression in the given sample is considered as an outlier. Finally, the resulting outliers estimated for every sample are considered as of bad quality and are excluded from analysis. The use of non-corrected low threshold $P < 0.05$ produces massive presence of false positive selections. The sum of these false selections should be comparable for all good quality samples presenting internal standard of good quality sample that can be used for statistical selection of bad samples with significantly elevated number of outliers. The program EFILTER produces the histogram of the numbers of outliers in samples for the visual inspection of the group quality.

Ranking of selections. Another method of data filtration is based on the comparison of the results of many differential expression analyses produced with sequential exclusion of samples one by one to determine the dependence of the conclusion about any selection on the exact group's content. This method determines the robustness of the differential gene expression selections and deliberates them from the influence of singular experimental errors.

The analysis uses standard Associative Analysis algorithms (30). The 'leave-one-out' approach excludes one sample from the group—one by one until all possible singular exclusions are produced—with estimation of the frequency of positive selection for every gene. This approach produces accurate ranking estimation of the robustness for most selections. However, it is not safe from the effect of singular errors of measurements, because the presence of one such outlier within any of the replicate is able to mask its difference of expression and diminishes the rank of otherwise ideal selection. The next modification of the procedure makes it defended from this effect of singular errors. Note that the exclusion of the 'bad' replicate and re-estimation of the robustness of the given selection will produce results devoid of the outlier influence. The new algorithm can be named as 'leave-two-out', because includes preliminary step—exclusion of one sample—with subsequent application 'leave-one-out' procedure for the rest of the samples in consideration. For an experiment having total number of samples equal n (sum of samples in both compared groups), this algorithm will produce a set of n ranks for each excluded sample and highest of them will be the one most independent on the worse replicate. Compared with previous EFILTER procedure, the TWOEX algorithm provides the opportunity to benefit even from a relatively bad sample, incorporating only expressions and excluding erroneous measurements. Based on the use of standard program for associative analysis, this algorithm enables to produce ranking estimation with selected restriction on the minimal expression and foldness being an adequate addition to the standard associative analysis.

Estimation of the quality of differential expression analyses

For the estimation of quality, we use 'artificial' data with controlled differences in gene expressions. The presumably homogenous group of samples was divided into two sub-groups. One of them was used as a control, whereas in another sub-group (experimental) artificial changes in gene expressions were introduced. Towards this aim, all data were sorted according to the averaged gene expression in experimental group. The entire data set was split into 1000 gene blocks, and thereafter controlled balanced (±) changes were introduced into 20% of data of experimental group. Within each block (1000 genes) 100 genes received positive changes—multiplied by 'foldness', and 100 genes received negative changes—divided by 'foldness'). One such block is presented in Figure 8. After applying the analysis procedure, the resulting number of selections is compared with true selections for determination of the 'Sensitivity' and 'Specificity' of the given analysis as it is shown in Figure 8.

The presented system enables to compare different methods of data normalization, and it enables also to estimate the role of restrictions made in course of differential gene expression analysis. The following designations were used in this analysis.

- Fd—'foldness' of controlled changes in the data;
- Fa—minimal 'foldness' of Associative Analysis;
- Em—minimal expression for genes selected as being expressed distinctly from background in Associative Analysis.

Results using data obtained with mRNA collected from peripheral blood mononuclear cells from healthy donors with the use of 'Illumina microarray' technology are presented in Figure 9. Quality of analysis is estimated here by the two parameters: sensitivity is determined as a proportion of true positive selections within all introduced changes, and specificity determined as 1—portion of false positive selections among all not changed expressions (31).

Figure 9A demonstrates the dependence of sensitivity and specificity in terms of the relationship between Fa and Fd. When $Fa < Fd$, the Associative Analysis of normalized data selects more that 80% of changes. Sensitivity drops down sharp when the Fa becomes comparable or even higher than the 'foldness' of introduced changes Fd.

The number of replicates is the most essential parameter form the output quality. Figure 9B shows a sharp decrease of the sensitivity of analysis for the number of replicates <4. Five to six replicates could be recommended as

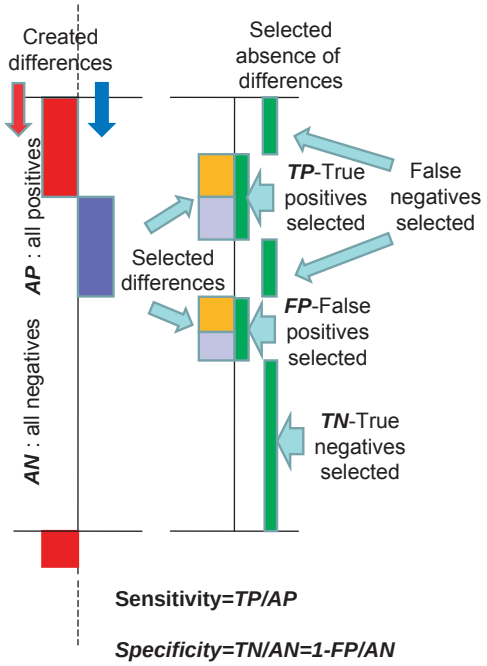


Figure 8. Test system for determination of the sensitivity and specificity of the differential gene expression analyses. The presumingly homogeneous group of samples was divided to two equal sub-groups one of which not changed used as a control and another one used as a experimental group with introduced changes. Here is shown a fragment of these experimental dataset with introduced positive (red) and negative (blue) changes in the 20% portion of gene expression—left part. Right part presents differences selected by the differential gene expression analysis (left of the vertical axis) with indication on the right side from the axis which of this selection is true (co-incidenting with the artificially made selections) and which are false. The sensitivity of selections is determined here as a proportion of true positive selections within all produced changes, whereas a specificity determined as a proportion of true negative selections. The fragments with artificial changes presented here are evenly distributed along all experimental group.

minimal size of the groups, whereas the usually used four replicated experiments might loose up to 20% of true differences. These numbers could vary in different microarray technologies and with the use samples from different sources. This result could be used as an alternative of the standard methods for the estimation of the power of analysis and the number of replicates necessary to achieve desired quality of analysis. The advantage of this approach is in the use of real data with practically not distorted infrastructure (variations and their distributions over expression levels) for estimation of the quality of the future analysis.

The method presented here enables the comparison of the quality of different types of analysis and influence of different normalization methods. In Figure 9C, we compared results of associative analysis with use different methods of normalization. It appeared that the use of our two-step normalization procedure and two popular methods Quantile (Q) and Lowess (L) (32) produced very comparable results except the area of highly expressed genes (first, thousand genes with highest expression) where quality of analysis based on the use Q- and L-normalizations significantly worse compared with two-step normalization presented here as it is shown in Figure 9C. Quite obvious that the same difference in quality was presented in comparison of our Associative Analysis based on the use two-step normalization and SAM analysis that used Quantile and Loess normalizations (32; Figure 9D).

To estimate the stability of the obtained estimations, we modified the quality analysis by using several variants of arbitrary splitting of the total dataset to two equal sub-groups (column permutation with subsequent splitting). The averaged result of five permutations presented in Figures 9A and B demonstrates relative stability of these estimations.

DISCUSSION

Current statistical methods do not adequately address mutually exclusive characteristics of sensitivity and specificity in microarray experiments monitoring the expression levels of thousands genes simultaneously. The common practice to use low-significance thresholds ($P < 0.05$) will result in a large number of false positive selections. Attempts to increase stringency by raising the threshold of significance above this value will cause a compensatory decrease in sensitivity and a resultant increase in false negative selections.

In measurements of gene expressions, the biological component is accompanied with variations of non-biological origin coming from a number of different sources. Normalization reduces systemic variations, while not affecting random variations. Common practice is to obtain information about random variation from replicated measurements. The number of replicates is critical for the accuracy of estimation of random variation and biological component as well. The use of large numbers of replicates is able to improve the situation in microarray experiments as well (33,34), although it can be rather expensive and labor intensive. Fortunately, there is a real resource to increase the power of statistical tests by using the enormous mass of information coming from each microarray experiments. We introduce here an approach based on the use of internal standards—large families of genes sharing some important features, while not being dependent on any particular gene sequence, level of expression, or coordinate position on the chip. Here were discussed standards for equity in gene expression, stability, standard for expressions below the sensitivity of the system (standard for 'non-expressed' genes). Deprived with dependence on the level of expression elements such standards bear information about experimental variation replicated thousands times by the count of the elements in the standard. This is an alternative to replications for increasing the power of statistical criterions. The increase of the power from such huge 'replication' should be tremendous.

The two main problems should be resolved before using this approach. Is the distribution of the elements of the internal standard normal and how to determine parameters of this distribution? Usually, each internal standard is contaminated with outliers. For example,

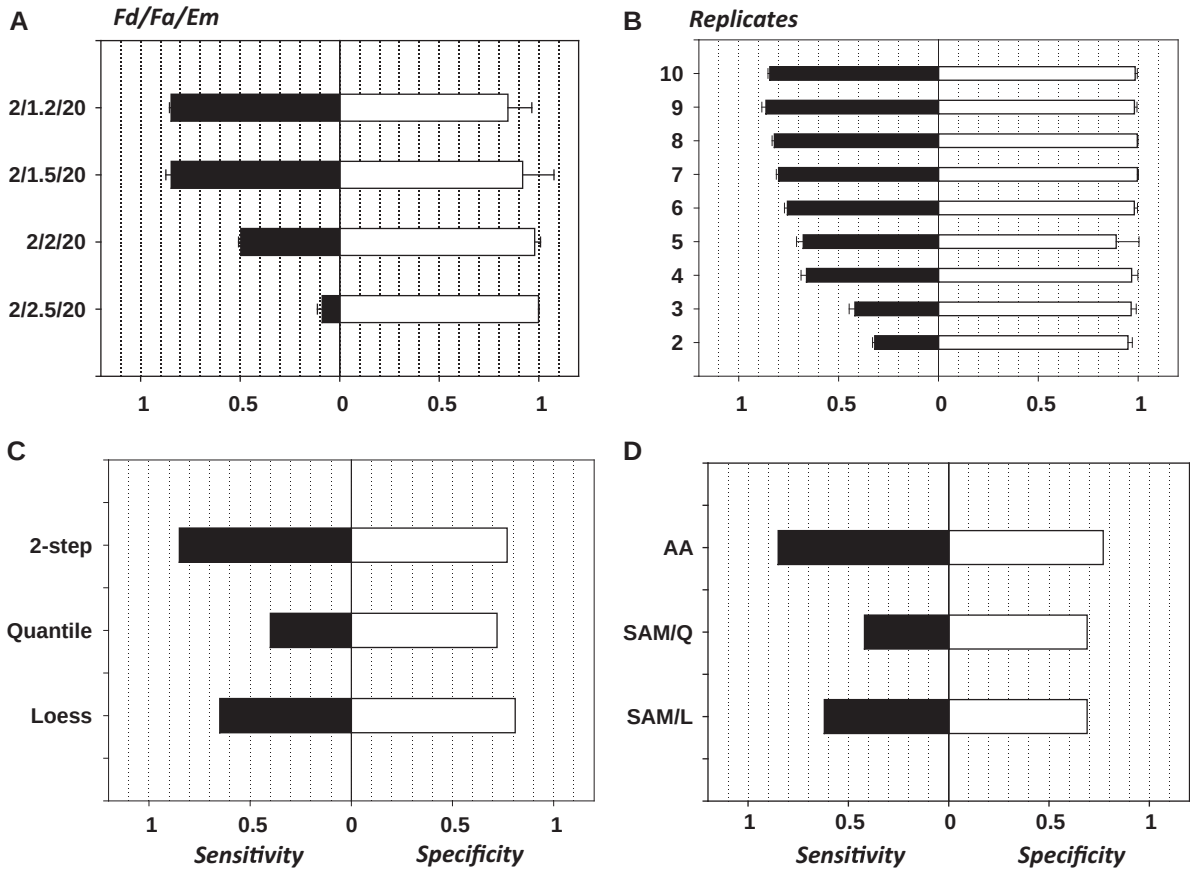


Figure 9. Sensitivity/specificity (Sn/Sp) characteristics of the normalization and analyses procedures. (**A**) Dependence of the analysis quality of the foldness of changes in gene expression: along ordinates Fd-foldness of controlled changes of data/Fa-minimal foldness accepted for results of differential gene expression analysis/Em = 20-minimal expression above background. (**B**) Dependence of the analysis quality of the number of replicates. Fd/Fa/Em = 2/1.5/20. (**C**) Comparison of normalization methods: two-step analysis (presented above) versus Quantile normalization versus Lowess normalization. (**D**) Comparison of analysis methods: associative analysis—SAM with Quantile normalization—SAM with Loess normalization. Abscise—sensitivity and specificity of the analysis as described in text.

majority of genes are equally expressed in any homogenous group and have a relatively small variability, however there are always some genes that does not share these features. Reduction of the influence of outliers is a critical step in the analyses based on the use of internal standards. Fortunately, this contamination with outliers is always relatively small and can be selected and removed with simple procedures.

The problem of normality is solved for this standard in several different ways. The selection of the normally distributed additive noise (background) is solved by using only the left portion of the non-distorted part of distribution for fitting to normal distribution. Standard of equity of expression and standard of the stability (reference group) appeared to have normal distribution after exclusion of outliers in the simple iterative procedure. It means that the rest of the distribution obtained after sequential truncation steps was always satisfactory fitted to the normal distribution. Even if there is some contamination with not normally distributed members, it is not essential and does not interfere with the normality of the rest.

Procedures similar to associative analysis have been previously proposed by Newton *et al.* (35); Rocke and Durbin (16); Tseng *et al.* (36). However, there are critical differences between these methods and ours. For example, in Rocke and Durbin (16), all genes were used as a reference group without exclusion of HV-genes. The presence of HV-genes increases the SD of the residuals in the reference group, thus reducing the power of the associative analysis.

There were versatile assumptions about the distribution of the background level signals and additive error term in the literature. Rocke and Durbin (16) were the first to suggest the use of iterative procedure for estimation of background parameters similar to the procedure presented here. Our approach goes one step further and demonstrates that the apparent deviation of the additive noise distribution from normality is produced by the presence of the weak signals overlapping with the noise. These results enable the skewed distribution presented in Figure 2 to be treated as a normally distributed additive noise distorted on its right side by the presence of low gene expressions.

The estimation of the performance of microarray data analysis demonstrated an advantage of the proposed here normalization and analysis methods over the popular normalization (Quantile, Loess) and analysis

(SAM) procedures. The application of the methods presented here to various biological and clinical problems demonstrated their ability to reveal essential features of the systems under investigations [see for example (28–30,37–43)], confirmed by the subsequent analysis of signaling pathways involved, transcription factor analysis and comparison with other publications. In some applications, the parallel use of different approaches to the analysis of the same data demonstrated advantage of the internal standard based methods over others in the selection of the gene sets reasonably associated with the studied phenomenon [see for example Dozmorov and Centola (2003) (30)].

Internal standard-based analysis enables to improve the power of microarray analysis at several levels. In the next part, we will demonstrate that the knowledge of the parameters governed by internal standards can be used for analysis in a statistically robust manner also for functional associations through clustering and networking genes having similar dynamical behavior.

ACKNOWLEDGEMENTS

Authors thank Michael Centola, Richard Miller, Edward Wakeland and Nicholas Chiorazzi for fruitful discussions, Nicholas Knowlton and Shengguang Qian for help with programming and Jonathan Wren and Teodor Ene for the help in the preparation of this article.

FUNDING

National Institutes of Health (grants P20 RR020143, R01 AI045050 and P30 AR053483); National Institutes of Health/National Center for Research Resources – Centers of Biomedical Research Excellence (grant IRG-05-066-04); American Cancer Society (grant IRG-05-066-04). Funding for open access charge: American Cancer Society (grant IRG-05-066-04).

Conflict of interest statement. None declared.

REFERENCES

1. Lee,N.H. and Saeed,A.I. (2007) Microarrays: an overview. *Methods Mol. Biol.*, **353**, 265–300.
2. Do,J.H. and Choi,D.K. (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol. Cells*, **31**, 254–261.
3. Hua,J., Balagurunathan,Y., Chen,Y., Lowey,J., Bittner,M.L., Xiong,Z., Suh,E. and Dougherty,E.R. (2006) Normalization benefits microarray-based classification. *J. Bioinform. Syst. Biol.*, 43056.
4. Saviozzi,S. and Calogero,R.A. (2003) Microarray probe expression measures, data normalization and statistical validation.Comp. *Funct. Genomics.*, **4**, 442–446.
5. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statistics.*, **6**, 65–67.
6. Cheng,C. and Pounds,S. (2007) False discovery rate paradigms for statistical analyses of microarray gene expression data. *Bioinformation*, **1**, 436–446.
7. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, **57**, 289–300.
8. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
9. Lin,D., Shkedy,Z., Burzykowski,T., Ion,R., Göhlmann,H.W., Bondt,A.D., Perer,T., Geerts,T., Van den Wyngaert,I. and Bijnens,L. (2008) An investigation on performance of Significance Analysis of Microarray (SAM) for the comparisons of several treatments with one control in the presence of small-variance genes. *Biom. J.*, **50**, 801–823.
10. Knowlton,N., Dozmorov,I. and Centola,M. (2004) Microarray data analysis toolbox (MDAT) for normalization, adjustment and analysis of gene expression data. *Bioinformatics*, **20**, 3687–3690.
11. Kooperberg,C., Fazzio,T.G., Delrow,J.J. and Tsukiyama,T. (2002) Improved background correction for spotted DNA microarrays. *J. Comput. Biol.*, **9**, 55–66.
12. Attoor,S., Dougherty,E.R., Chen,Y., Bittner,M.L. and Trent,J.M. (2004) Which is better for cDNA-microarray-based classification: ratios or direct intensities. *Bioinformatics*, **20**, 2513–2520.
13. Dozmorov,I., Knowlton,N., Tang,Y. and Centola,M. (2004) Statistical monitoring of weak spots for improvement of normalization and ratio estimates in microarrays. *BMC Bioinformatics*, **5**, 53.
14. Churchill,G.A. and Oliver,B. (2001) Sex, flies and microarrays. *Nat. Genet.*, **29**, 355–356.
15. Wei,J., Riley,R.M., Wolfinger,R.D., White,K.P., Passador-Gurgel,G. and Gibson,C. (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat. Gen.*, **2**, 389–395.
16. Rocke,D.M. and Durbin,B.A. (2001) A model for measurement error for genhe expression analysis. *J. Comput. Biol.*, **8**, 557–569.
17. Takeya,M., Matsuda,T., Iwamoto,M., Tsumura,M., Nakaguchi,T. and Miyake,Y. (2007) Noise analysis of duplicated data on microarrays using mixture distribution modeling. *Opt. Rev.*, **14**, 97–104.
18. Sidorov,I.A., Hosack,D.A., Gee,D., Yang,J., Cam,M.C., Lempicki,R.A. and Dimitrov,D.S. (2002) Oligonucleotide microarray data distribution and normalization. *Inform. Sci.*, **146**, 67–73.
19. Schuchhardt,J., Beule,D., Malik,A., Wolski,E., Eickhoff,H., Lehrach,H. and Herzel,H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.
20. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
21. Thomas,J.G., Olson,J.M., Tapscott,S.J. and Zhao,L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.
22. Cheadle,C., Vawter,M.P., Freed,W.J. and Becker,K.G. (2003) Analysis of microarray data using Z score transformation. *J. Mol. Diagn.*, **5**, 73–81.
23. Workman,C., Jensen,L.J., Armer,H., Berka,R., Gautier,L., Nielser,H.B., Saxild,H.H., Nielsen,C., Brunak,S. and Knudsen,S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.*, **3**
24. Fujita,A., Sato,J.R., Rodrigues,L.O., Ferreira,C.E. and Sogayar,M.C. (2006) Regulatory dendritic cells protect against cutaneous chronic graft-versus-host disease mediated through CD4 + CD25 + Foxp3 + regulatory T cells. *BMC Bioinformatics*, **7**, 469.
25. Wu,T.D. (2001) Analysing gene expression data from DNA microarrays to identify candidate genes. *J. Pathol.*, **195**, 53–65.
26. Hatfield,G.W., Hung,S.P. and Baldi,P. (2003) Differential analysis of DNA microarray gene expression data. *Mol. Microbiol.*, **47**, 871–877.
27. Durbin,B.P. and Rocke,D.M. (2004) Variance-stabilizing transformations for two-color microarrays. *Bioinformatics*, **20**, 660–667.
28. Dozmorov,I., Knowlton,N., Tang,Y., Shields,A., Pathipvanich,P., Jarvis,J. and Centola,M. (2004) Hypervariable genes – experimental error or hidden dynamics. *Nucleic Acids Res.*, **32**, e147.
29. Dozmorov,I.M., Centola,M., Knowlton,N. and Tang,Y.-H. (2005) Mobile-classification in microarray experiments. *Scand J. Immunol.*, **62(Suppl. 1)**, 84–91.

30. Dozmorov,I.M. and Centola,M. (2003) An associative analysis of gene expression array data. *Bioinformatics*, **19**, 204–211.
31. Khodarev,N.N., Park,J., Kataoka,Y. *et al*. (2003) Receiver operating characteristic analysis: a general tool for DNA array data filtration and performance estimation. *Genomics*, **81**, 202–209.
32. Chiogna,M., Massa,M.S., Risso,D. and Romualdi,C. (2009) A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bioinformatics*., **13**, 61.
33. Pavlidis,P., Li,Q. and Noble,W.S. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.
34. Glynne,R.J., Ghandour,G. and Goodnow,C.C. (2000) Genomic-scale gene expression analysis of lymphocyte growth, tolerance and malignancy. *Curr. Opin. Immunol*., **12**, 210–214.
35. Newton,M.A., Kendziorski,C.M., Richmond,C.S., Blattner,F.R. and Tsui,K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol*., **8**, 37–52.
36. Tseng,G.C., Oh,M.K., Rohlin,L., Liao,J.C. and Wong,W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*., **29**, 2549–2557.
37. Torgerson,T.R., Genin,A., Chen,C., Zhang,M., Zhou,B., Anover,S., Frank,M.B., Dozmorov,I., Ocheltree,E., Kulmala,P. *et al*. (2009) FOXP3 Inhibits Activation-Induced NFAT2 Expression in Human T Cells. *J. Immunol*., **183**, 907–915.
38. Saban,M.R., O'Donnell,M.A., Hurst,R.E., Wu,X.-R., Simpson,C., Dozmorov,I., Davis,C., Anant,S., Vadigepalli,R. and Saban,R. (2008) Molecular networks discriminating mouse bladder responses to intravesical bacillus calmette-guerin (BCG), LPS, and TNF-α. *BMC Immunol*., **9**, 4.
39. Jorgensen,E.D., Dozmorov,I., Frank,M.B., Centola,M. and Albino,A.P. (2004) Global gene expression analysis of human bronchial epithelial cells treated with tobacco condensates. *Cell Cycle*, **3**, 1154–1168.
40. Dozmorov,I.M., 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.
41. Dozmorov,I.M., Saban,M.R., Knowlton,N., Centola,M. and Saban,R. (2003) Connective molecular pathways of experimental bladder inflammation. *Physiol. Genomics*, **15**, 209–222.
42. Jarvis,J., Dozmorov,I., Jiang,K., Frank,M.B., Szodoray,P., Alex,P. and Centola,M. (2003) Novel approaches to gene expression analysis of active polyarticular juvenile rheumatoid arthritis. *Arth. Res. Therapy*, **6**, R15–R31.
43. Kurella,S., Yaciuk,J.C., Dozmorov,I., Frank,M.B., Centola,M. and Farris,A.D. (2005) Transcriptional modulation of TCR, Notch and Wnt signaling pathways in SEB anergized CD4+ T cells. *Genes Immunity*., **6**, 596–608.

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

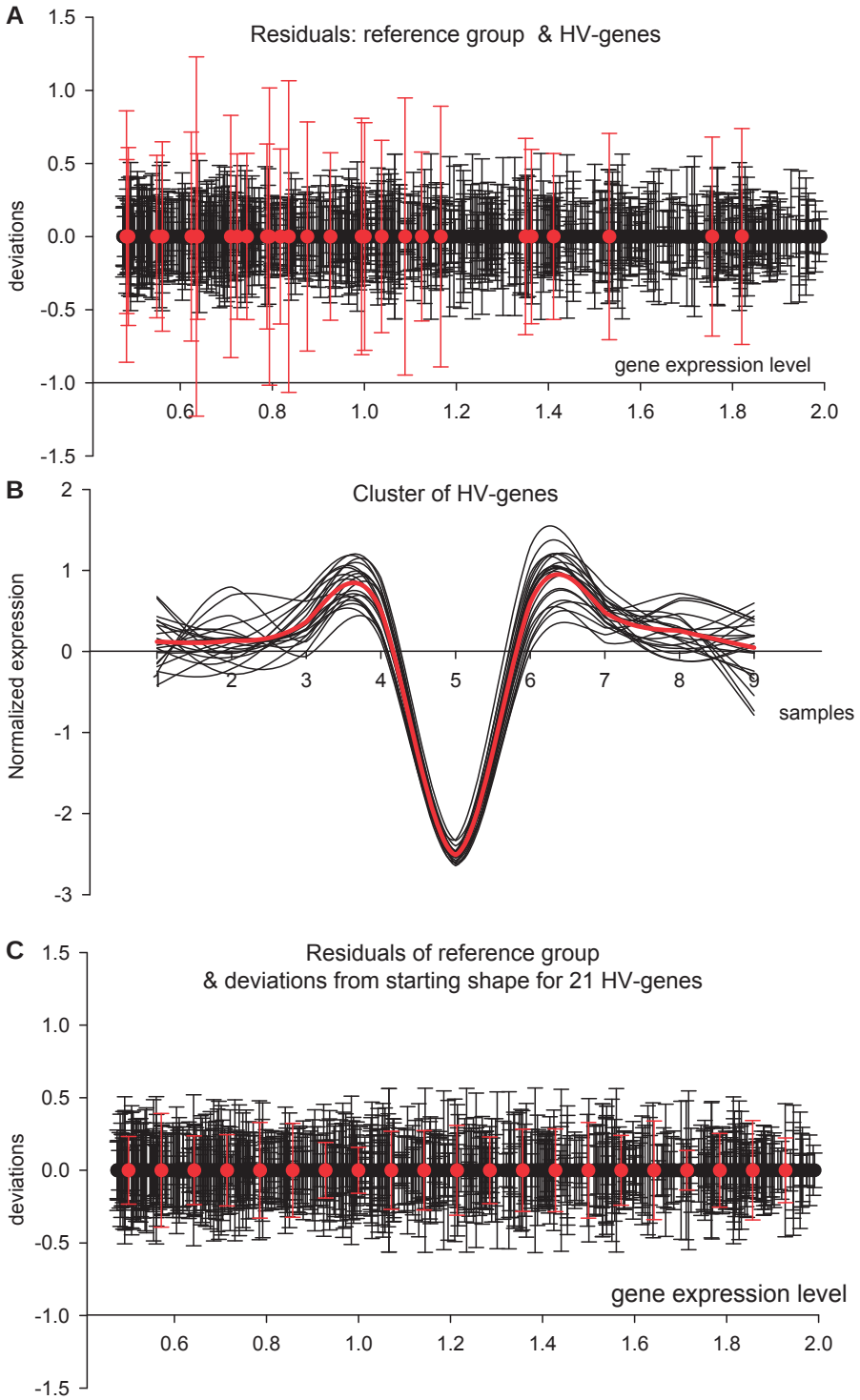


Figure 1. *F*-means clustering procedure. (**A**) The standard deviations of genes from the Reference Group, with HVE-genes (red bars) included. (**B**) Gene content of the cluster with seeding profile shown as a red line. (**C**) Deviations of genes' profiles from the seeding profile (shown as red SD 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).

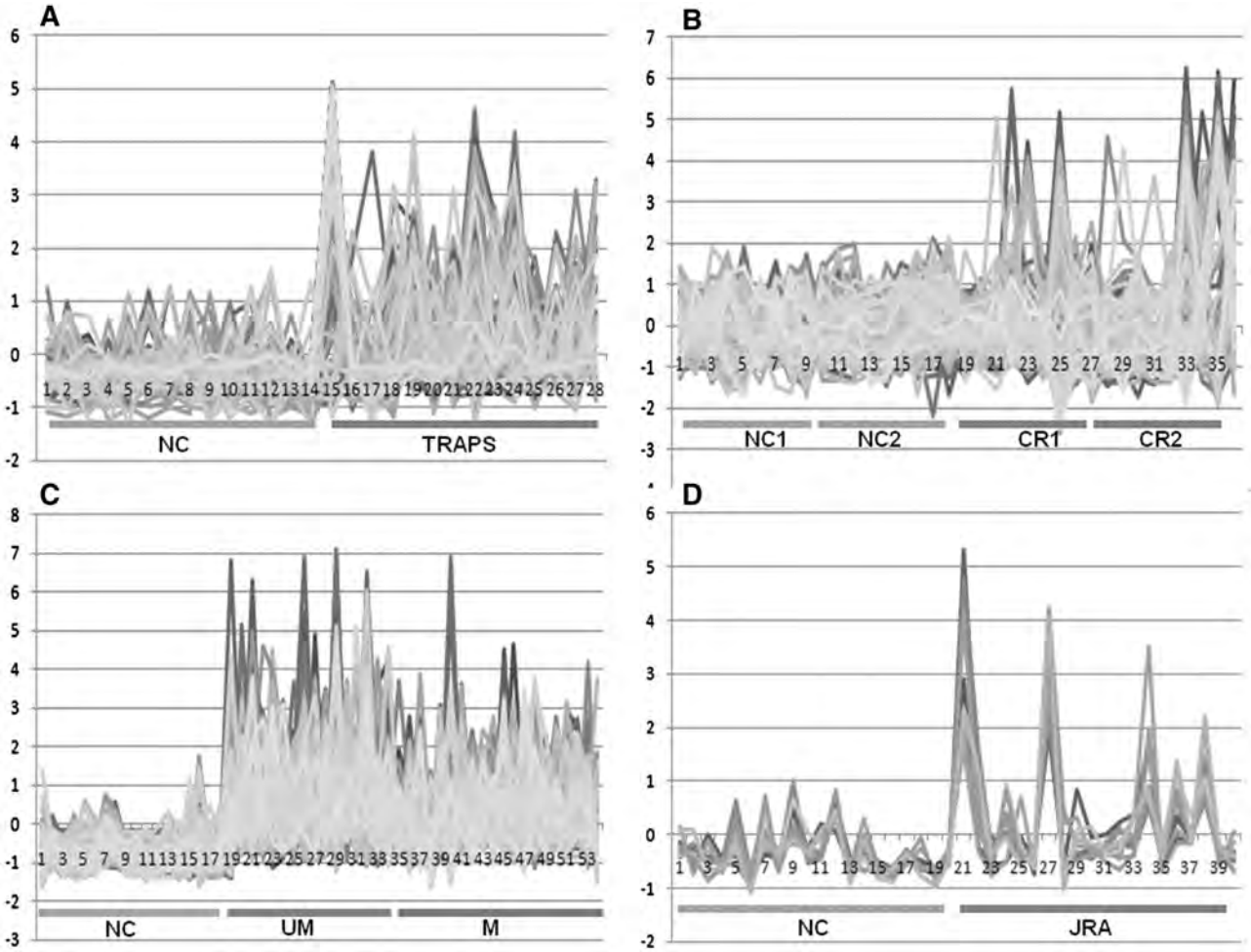


Figure 2. Increase in gene variability associated with different pathologies. Expression data normalized to make the overall Average = 0, SD = 1. Abscissa: the sample numbers. Ordinate: the normalized expression level. mRNA for the transcription study was obtained from various samples: (**A**) Samples 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

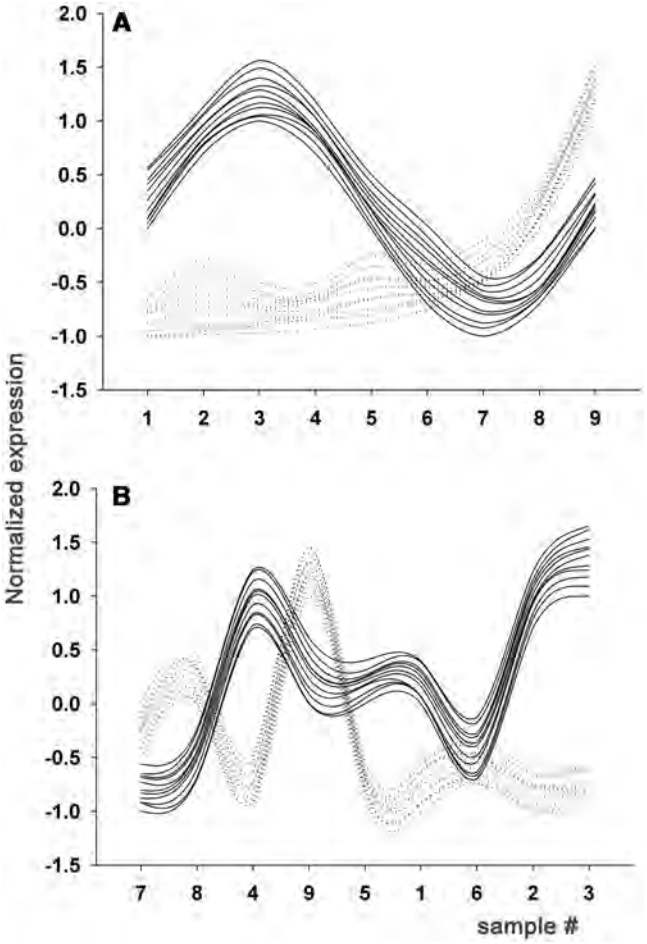


Figure 3. Shapes of the HVE gene expression profiles does not have sense. Diagrams illustrating the formation of the cluster profiles of HVE-genes in a homogeneous group. (**A**) Possible assortment of nine samples representing two dynamical processes with participation of several genes, each of whose 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

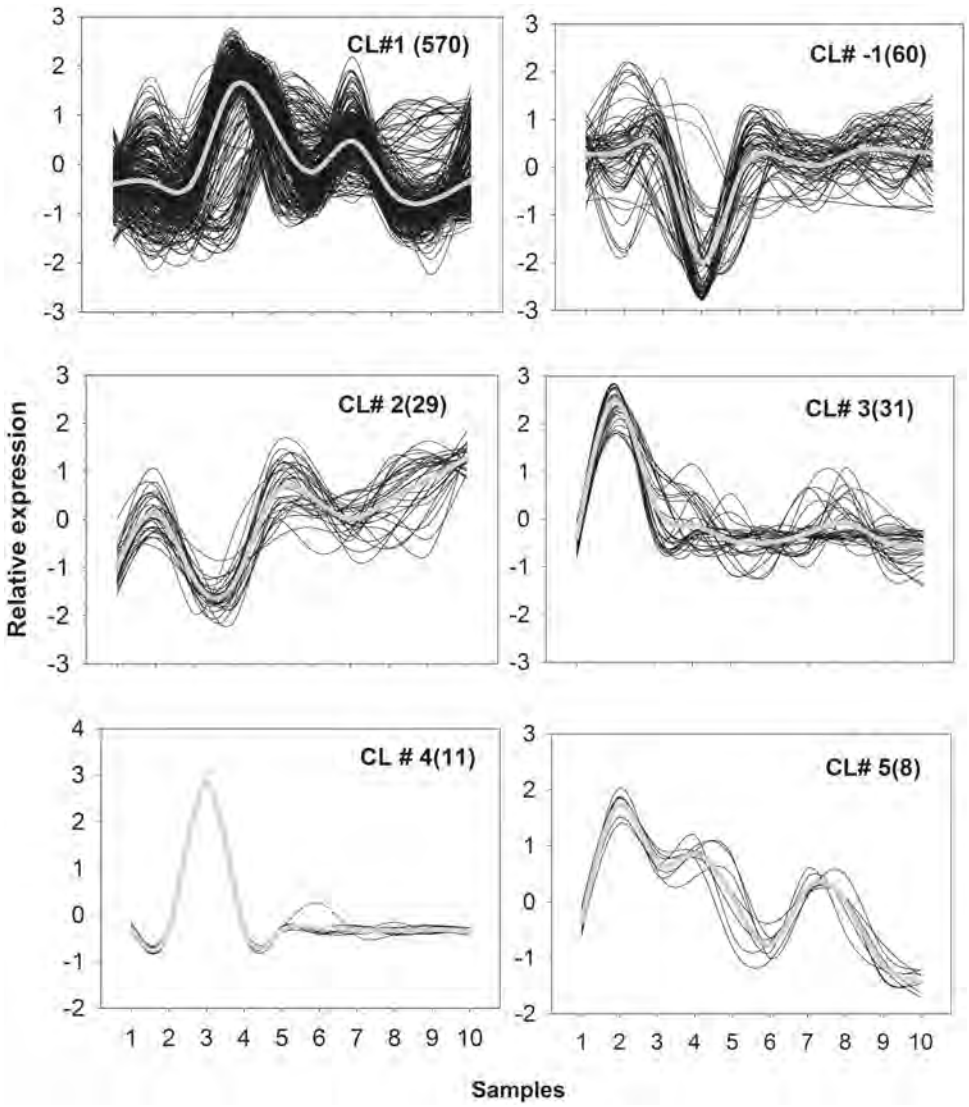


Figure 4. *F*-means clustering of gene expressions in T cells from B6 mice. The six largest clusters are shown. Abscissa: cluster numbers derived from 10 samples from 10 different mice. Ordinate: the normalized expression levels. Figures in brackets: the numbers of genes in each cluster.

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

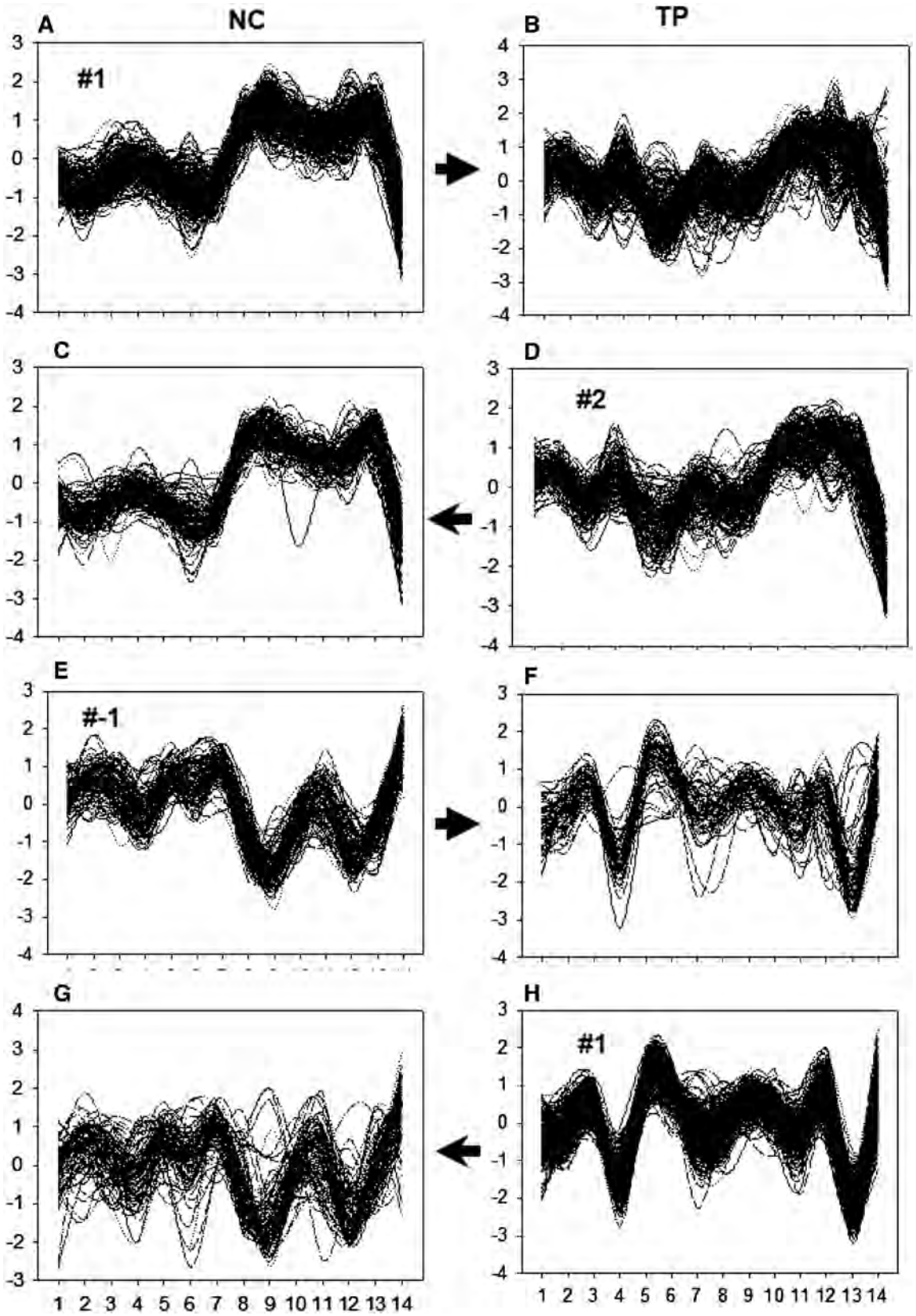


Figure 5. Reproducibility of the HVE gene co-expression in two unrelated sample groups: NC (normal controls) and TP (TRAPS patients). Normalized expression levels (ordinate) are presented against the numbers of samples in each group. Genes in the largest cluster (#1, **A**) in the NC group are also co-expressed 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.

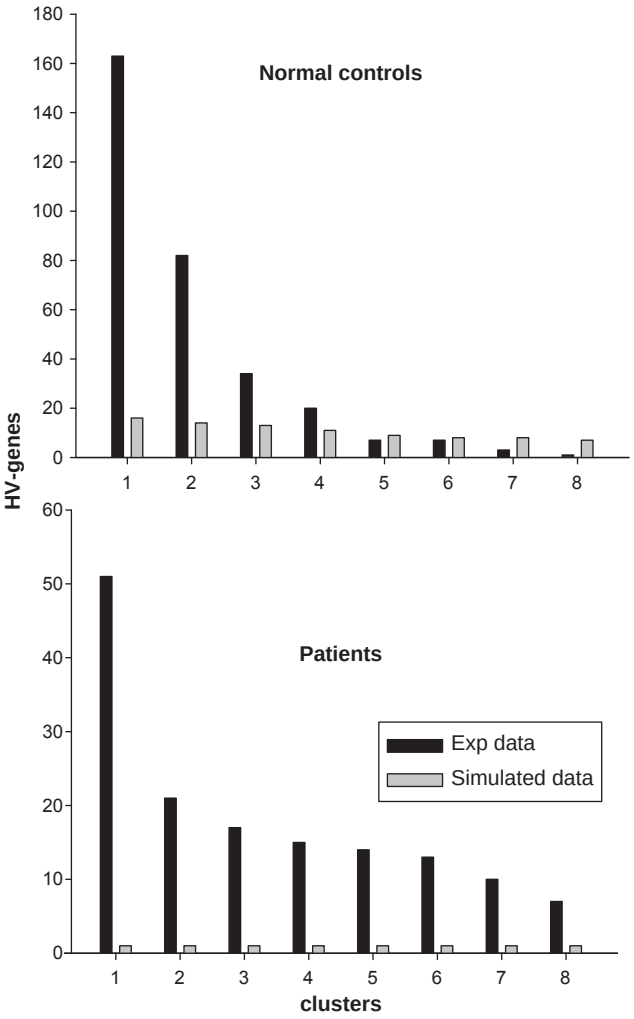


Figure 6. Contents of the eight biggest clusters in the NC and TP groups (Figure 5) (black bars) compared with the same for the simulated data (data obtained by substitution of the real gene expressions with random values having the same SD and means for each gene over all samples in groups).

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

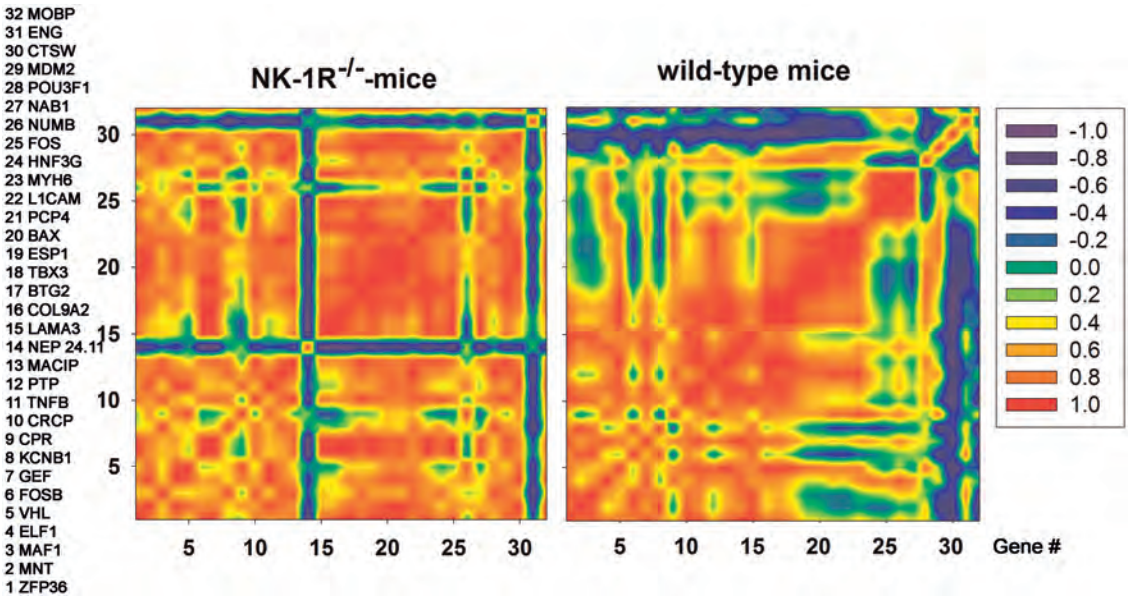


Figure 7. Mosaic of correlation coefficients of the HVE-genes in wild-type and NK1R–/– mice. The coordinates along axis are the numbers of genes listed in the left box. The white lines in A indicate the borders of three clusters of tightly interconnected genes. The colored lines and spots beyond the 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.

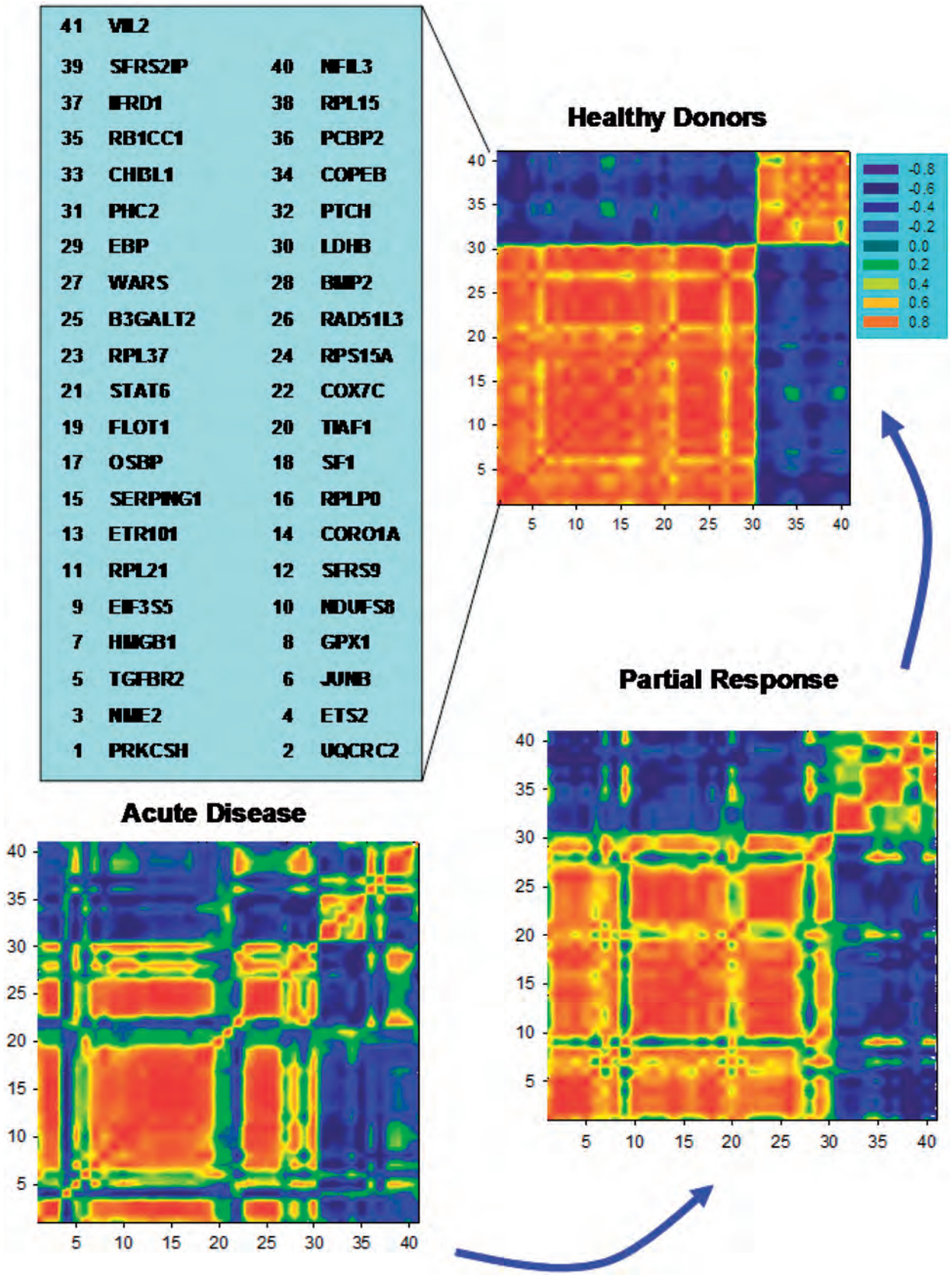


Figure 8. Correlation mosaics for genes from the two largest clusters in the control group (adopted from [Jarvis ea, 2003]). The designations are the same as in Figure 7. There is shown transformation of the mosaic created for patients group (Acute disease) to the Partial Response mosaic (patients who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

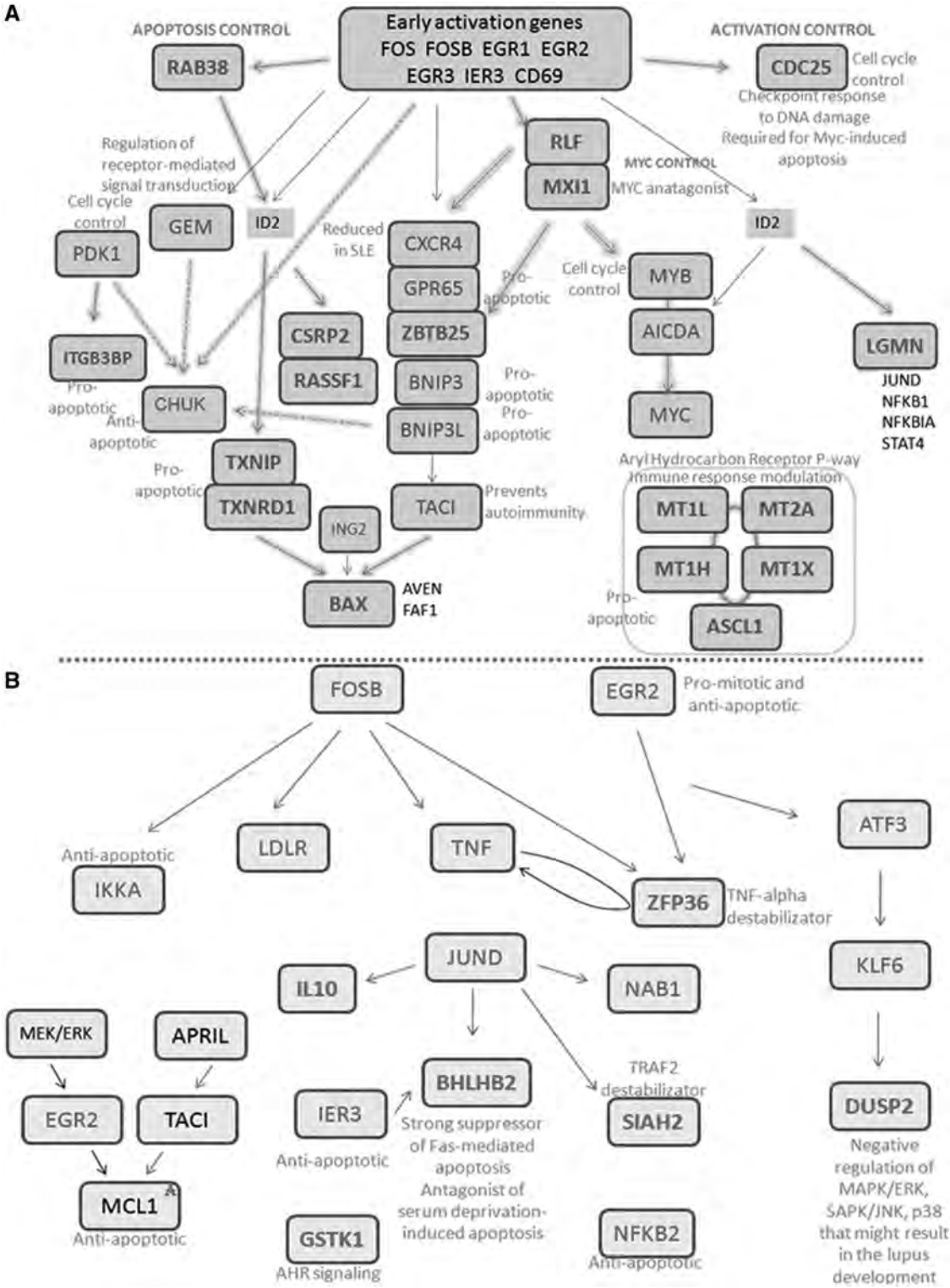


Figure 9. Networking of reproducibly variable genes after stimulation of EBV-transformed B cells from normal controls (**A**) and lupus patients (**B**). This network is a fragment of a gene network consisting of genes uniquely activated in normal (**A**) or lupus patient (**B**) groups. The gene network 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.

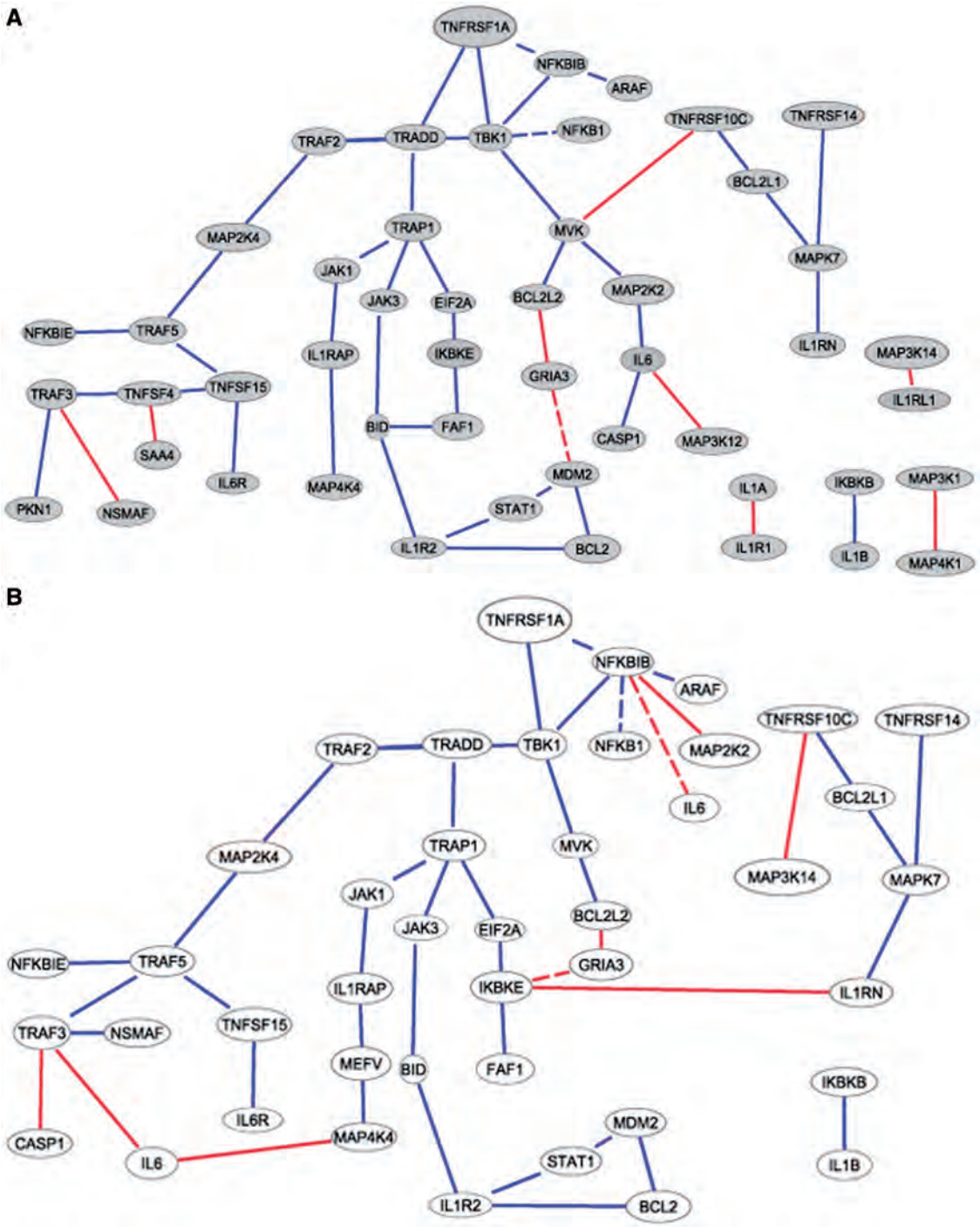


Figure 10. TNF pathway. Gene interconnection in both normal control (**A**) and TRAPS patients (**B**) obtained by calculating partial correlation coefficients. The solid lines represent positive interconnections with averaged partial correlation coefficients >0.7. The dashed lines represent interconnections 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

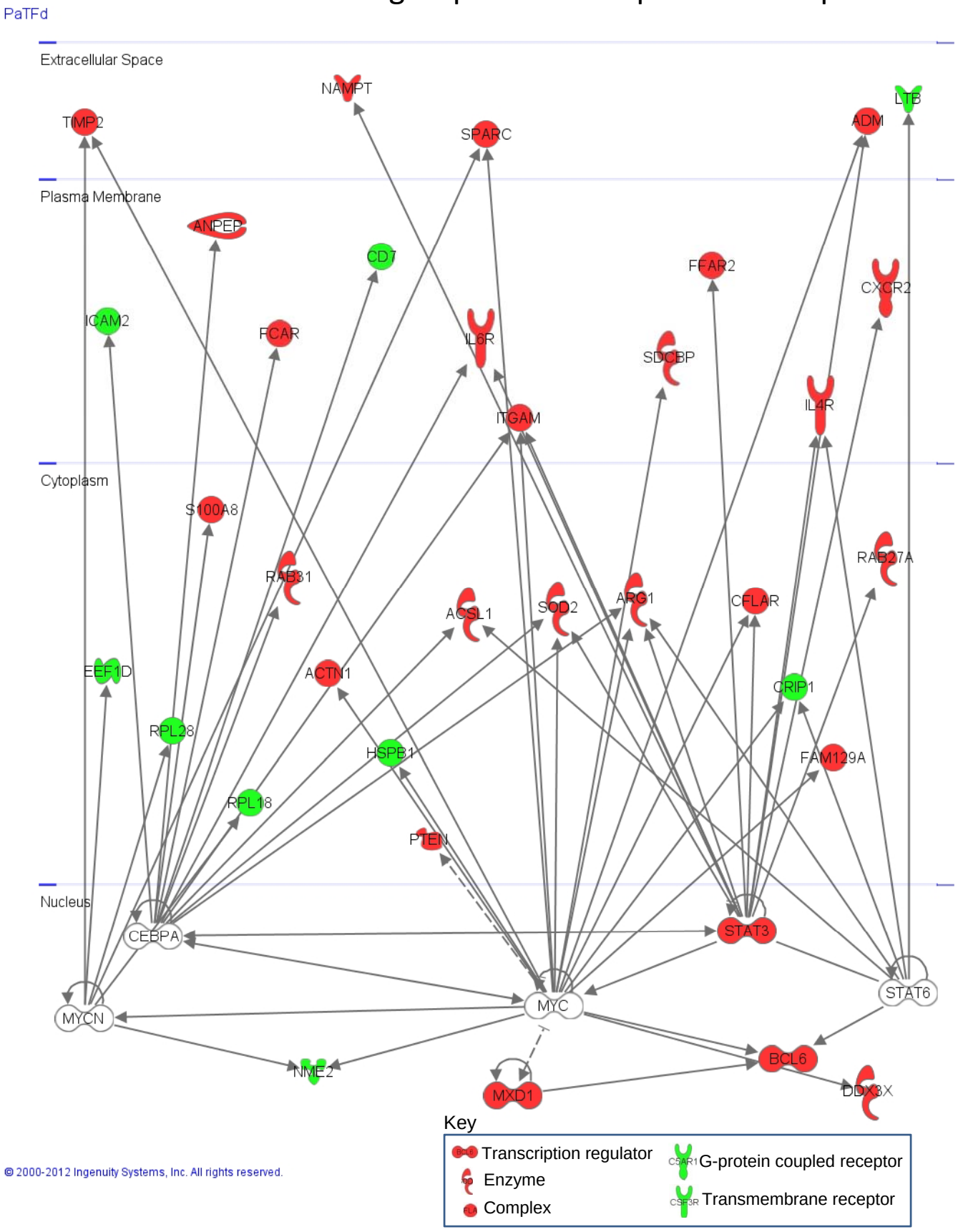
1. Kauffman,K.J., Ogunnaike,B.A. and Edwards,J.S. (2006) Designing experiments that aid in the identification of regulatory networks. *Brief. Funct. Genomic Proteomic*, **4**, 331–342.
2. Pritchard,C., Coil,D., Hawley,S., Hsu,L. and Nelson,P.S. (2006) The contributions of normal variation and genetic 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.

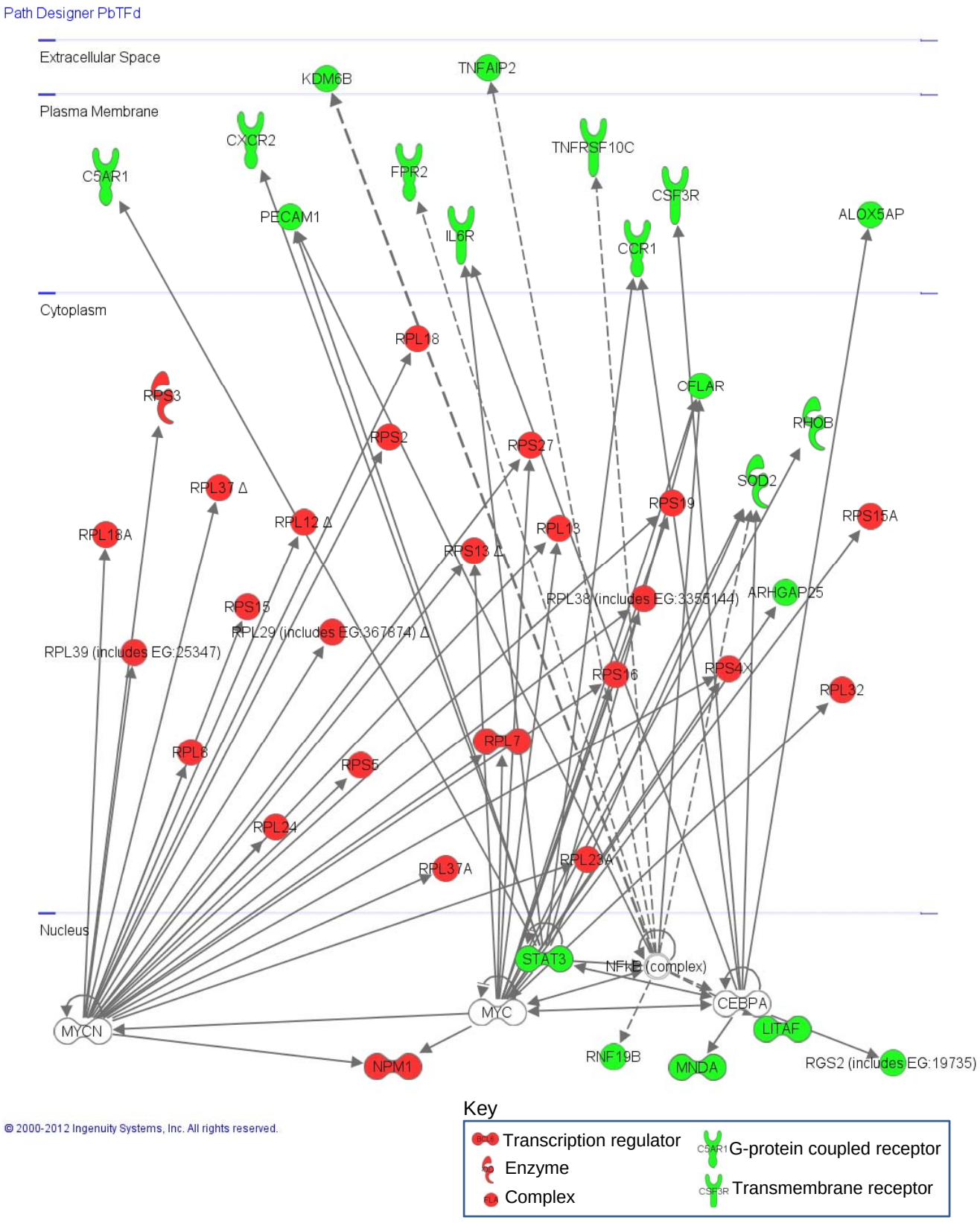
Appendix C

Diagrams illustrating the interaction of regulator-target gene pathways found to be differentially expressed in Syndrome 2 vs controls

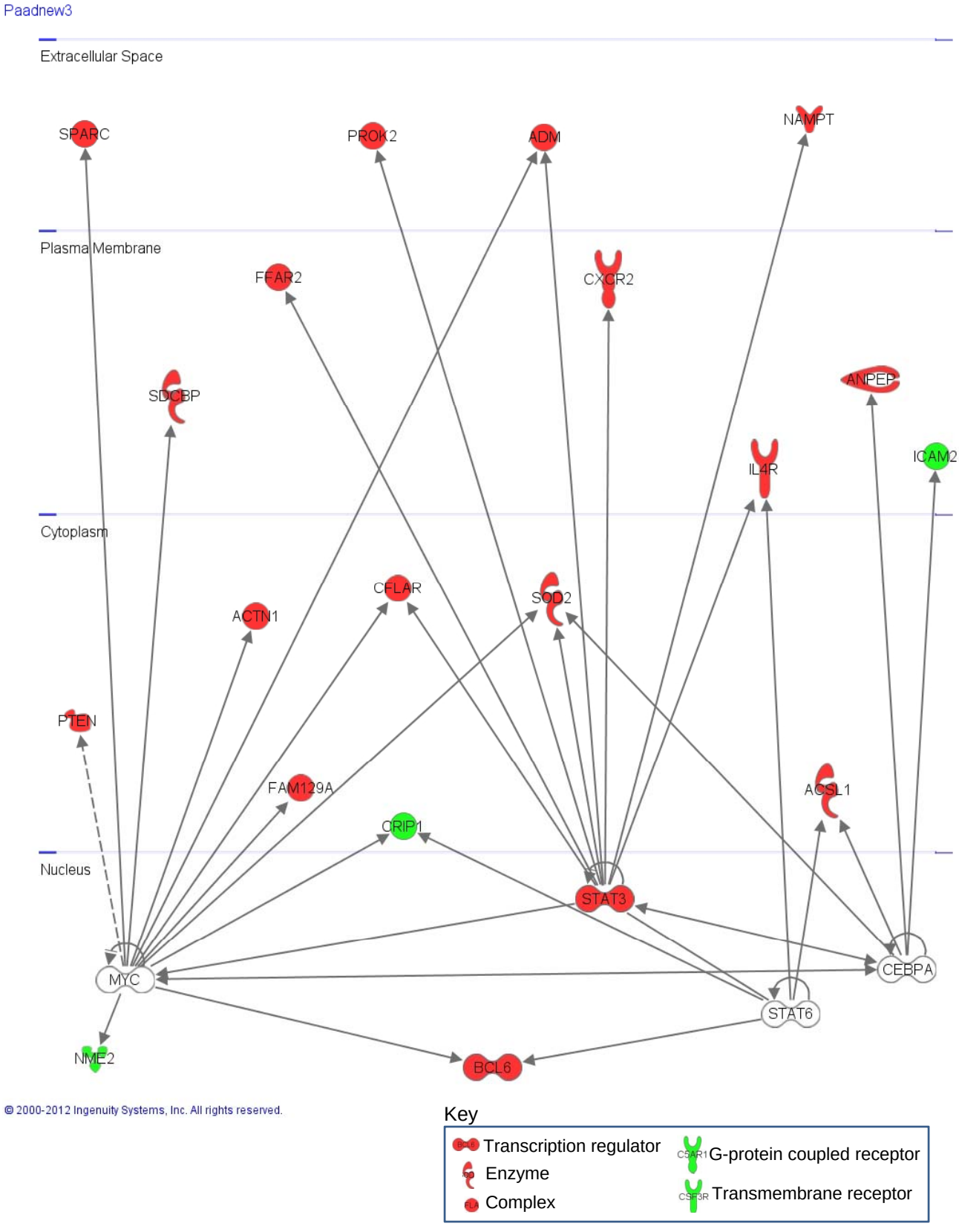
Panel A. Patient subgroup A in Developmental Sample



Panel B. Patient subgroup B in Developmental Sample



Panel C. Patient subgroup A in Replication Sample



Panel D. Patient subgroup D in Replication Sample