



Identifying illness parameters in fatiguing syndromes using classical projection methods

Gordon Broderick^{1†},
R Cameron Craddock²,
Toni Whistler²,
Renee Taylor³,
Nancy Klimas⁴ &
Elizabeth R Unger²

[†]Author for correspondence
¹University of Alberta,
Institute for Biomolecular
Design, Edmonton, Alberta,
T6G 2H7, Canada
Tel.: +1 780 492 6902;
Fax: +1 780 492 9394;
E-mail: gordon.broderick@
ualberta.ca

²Centers for Disease Control
and Prevention,
Viral Exanthems and
Herpesvirus Branch,
Atlanta, GA, 30333, USA

³University of Illinois at
Chicago,
Department of Occupational
Therapy, Chicago,
IL, 60612, USA

⁴University of Miami,
Miami Veterans Affairs
Medical Center, Miami,
FL, 33125, USA

Objectives: To examine the potential of multivariate projection methods in identifying common patterns of change in clinical and gene expression data that capture the illness state of subjects with unexplained fatigue and nonfatigued control participants.

Methods: Data for 111 female subjects was examined. A total of 59 indicators, including multidimensional fatigue inventory (MFI), medical outcome Short Form 36 (SF-36), Centers for Disease Control and Prevention (CDC) symptom inventory and cognitive response described illness. Partial least squares (PLS) was used to construct two feature spaces: one describing the symptom space from gene expression in peripheral blood mononuclear cells (PBMC) and one based on 117 clinical variables. Multiplicative scatter correction followed by quantile normalization was applied for trend removal and range adjustment of microarray data. Microarray quality was assessed using mean Pearson correlation between samples. Benjamini-Hochberg multiple testing criteria served to identify significantly expressed probes. **Results:** A single common trend in 59 symptom constructs isolates of nonfatigued subjects from the overall group. This segregation is supported by two co-regulation patterns representing 10% of the overall microarray variation. Of the 39 principal contributors, the 17 probes annotated related to basic cellular processes involved in cell signaling, ion transport and immune system function. The single most influential gene was *sestrin 1 (SESN1)*, supporting recent evidence of oxidative stress involvement in chronic fatigue syndrome (CFS). Dominant variables in the clinical feature space described heart rate variability (HRV) during sleep. Potassium and free thyroxine (T4) also figure prominently. **Conclusion:** Combining multiple symptom, gene or clinical variables into composite features provides better discrimination of the illness state than even the most influential variable used alone. Although the exact mechanism is unclear, results suggest a common link between oxidative stress, immune system dysfunction and potassium imbalance in CFS patients leading to impaired sympatho-vagal balance strongly reflected in abnormal HRV.

While great strides continue to be made, many aspects of chronic fatigue syndrome (CFS) etiology and pathophysiology remain poorly understood [1]. Subtle differences in immune system function [2], hypothalamic-pituitary-adrenal axis function [3] and psychological profiles [4] have been observed between CFS patients and nonfatigued control subjects. While these studies are revealing in their own right, they are generally hypothesis-driven and, as such, focus on a narrow set of disease indicators. Moreover, subjects are generally recruited from a research registry [5] or clinic [6] and constitute a very specific subset of the general population. Therefore, it may not be surprising to find that no single consistent distinguishing feature or overarching mechanism has yet been confirmed or widely agreed upon for CFS.

In an effort to broaden the scope of CFS study, an extensive population-based study was recently conducted by the Centers for Disease

Control and Prevention (CDC). As described in this issue [7], the resulting dataset contains a highly comprehensive spectrum of detailed clinical and laboratory measures describing patients from the general population. Encouraged by previous results obtained at the proof-of-concept scale (<4000 genes) [8,9], high-throughput measures of gene expression for 20,000 genes in peripheral blood mononuclear cells (PBMC) were also recorded. The challenge with such a diverse and rich dataset is to exploit its enormous potential for improving our understanding of CFS and its many aspects, potentially clarifying some of the relationships between intracellular processes and system-wide manifestations. As a result, our goal in this work was not to test hypotheses made *a priori*, but rather to identify and examine patterns in the data that might lead to the formulation of new candidate hypotheses.

Keywords: CD74, CFS, fatigue, free T4, gene expression, HRV, immune response, IRF5, MAPK, MFI, mTOR, partial least squares, PLS, potassium, projection methods, SESN1, Wnt

To accommodate a data structure, in which relatively few subjects have an extremely large number of descriptive attributes, it was of primary importance to reduce data complexity while retaining as much biological information of potential importance to CFS as possible. We avoided using individual gene responses and clinical measurements as independent contributors to disease classification. Instead, we chose to use classical multivariate statistical methods, namely principal components analysis (PCA) and partial least squares (PLS), for data reduction. Using these methods, the partial redundancy in individual clinical measurements and the co-regulation of genes was exploited to reduce the dimensionality of the problem and enhance solution robustness. To highlight the patterns in variability captured by these composite features that best correlate with CFS, we constructed a target space comprised of phenotypic measures of the illness (multidimensional fatigue inventory [MFI], Short Form 36 [SF-36], Cambridge Neuropsychological Test Automated Battery [CANTAB], and Zung depression scale and symptom inventory scores) and rotated candidate features toward this target. Extensions of conventional statistics were used to evaluate the performance of the resulting models and we applied multivariate contribution measures to rank genes and laboratory parameters most significantly correlated to this target space. The role of these influential genes and laboratory parameters will of course require additional study. However, it is hoped that these indicators will at the very least provide a basis for the formulation of new hypotheses regarding CFS pathogenesis or to refinements in the definition of the symptom space to best reflect all dimensions of CFS.

Methods

Data are derived from the population-based study of CFS described in the introductory paper of this issue [7]. The subset of data used and methods of data processing are described below.

Subjects

To avoid the potential impact that sex has on gene expression, we excluded the 38 male subjects. We further excluded the 23 subjects with medical or psychiatric conditions (other than major depressive disease with melancholic features) considered exclusionary for the research case definition of CFS. Thus, our analysis focused on data from the 112 female subjects for whom microarray data had been collected. Using

disease classification based on the CFS research case definition as measured by scores on the symptom inventory, MFI and SF-36 instruments [10,11], our study population included 40 CFS, 37 nonfatigued (NF), and 35 individuals suffering unexplained fatigue with symptoms or severity short of the case definition (ISF). One additional subject was excluded due to inadequate microarray data (see below), resulting in a final analysis group of 111 subjects.

Microarray data processing

Replicate microarray data were available and used for quality control. The data were normalized with multiplicative scatter correction [12,13] followed by quantile normalization [14] which allowed for sample-to-sample trend removal and range adjustment. This normalization was performed from the raw data between each stage of the following quality assessment procedure. Microarrays were assessed by computing the mean Pearson correlation between each array and every other array in the dataset [15]. Arrays with a mean correlation $r \leq 0.6$ were excluded, which resulted in the removal of a single microarray. Replicates were evaluated by Pearson correlation for each pair and removed if $r \leq 0.9$. None of the replicates were removed and the raw replicated data were averaged. The final 111 microarrays were normalized and \log_2 transformed for analysis. The proportion of false discoveries incurred when comparing over 20,000 gene probes was controlled using a sequential Bonferroni-type procedure proposed by Benjamini and Hochberg [16]. Reproducibility was estimated from the pooled variance of paired replicate samples collected for approximately one out of ten treatment conditions. Using p-values associated with a standard F-test and a false discovery rate (FDR) of 0.001, the number of genes carried forward for analysis was reduced from 19,760 to 15,136.

Laboratory & clinical variables

We excluded time course samples (nightly sleep data, oxygen saturation and salivary cortisol), date, time and text data not amenable to numeric conversion (gynecological history and gynecological surgery), and data with a large number of missing values (Wechsler Abbreviated Scale of Intelligence Wide Range Achievement Test [WASRI_WRAT]). Several new variables were generated: waist:hip ratio, total number of medications, number of medications acting on the CNS, number of immunological modulating

medications and CDC symptom inventory scores as defined in Wagner and colleagues [11]. The final clinical and laboratory dataset included 196 variables. Missing values were coded as missing. Variables with more than 50% missing entries were excluded from analysis.

Principal component analysis for dimensionality reduction & feature identification

We used the basic model for principal component analysis (PCA) shown in matrix notation in the equation below, where X is the original data set of n subjects described in k variables, T is the same data described in terms of A composite features, and E is the residual error:

$${}_n[X]^k = {}_n[T]^A \quad {}_A[P']^k + {}_n[E]^k$$

The composite features of the gene expression profiles were computed using noniterative partial least squares (NIPALS) [17]. Each coordinate value of a subject in the feature or score space T consists of a weighted sum of the coordinates in the original variable space. The contribution of each original variable to a given feature is captured by its weight or loading stored in the array P . Estimates of the standard error associated with each loading were calculated using a standard jackknifing technique [18]. The selection of A significant features was performed using a 'leave-one-out' cross-validation technique [19] based on the prediction error sum of squares (PRESS).

Partial least squares for rotation of co-regulated expression patterns (features) toward illness parameters (symptom space)

We used projection to latent structures technique (PLS) [20] to align the features that capture the variability in a data set with those of a specific outcome set Y . The PLS model extends the PCA model as described in equations 1, 2 and 3 below. This alignment is achieved by computing features in the outcome space (U) and gene expression space (T) simultaneously and exchanging information between both feature spaces iteratively through the inner relation in equation 3.

$$\begin{aligned} (1) \quad & {}_n[X]^k = {}_n[T]^A \quad {}_A[P']^k + {}_n[E]^k \\ (2) \quad & {}_n[Y]^m = {}_n[U]^A \quad {}_A[C']^m + {}_n[F]^m \\ (3) \quad & {}_n[U]^A = {}_n[T]^A + {}_n[H]^A \end{aligned}$$

The relative contribution of each gene probe and clinical variable k to the overall model is quantified by the sum of the contributions w_{ak} in each feature a weighted by the fraction of total variation SSR_a captured by that feature. This

measure is termed the variable importance in the projection (VIP, equation below). Significance of a VIP result was estimated with a measure akin to a standard t-test. A pseudo-t ratio was defined as the ratio of the VIP for a model term divided by the corresponding standard error of estimation based on jackknifing [18]. All PLS and PCA calculations were conducted with the SimcaP software (Umetrics, NJ, USA).

$$VIP_{Ak} = \sum_{a=1}^A SSR_a w_{ak}^2$$

We used phenotypic measures of all dimensions included in the case definition of CFS to form the target space. Specifically, a total of 59 variables were used, including MFI, SF-36, symptom inventory scores, Zung depression scale and CANTAB measures. The overall performance of a composite feature model for gene expression or clinical variables was based on its ability to capture patterns of change in this target space. Three basic statistics were used:

- The fraction of total sum of squares captured by the model or R^2
- The fraction of the total variance modeled or adjusted R^2
- The proportion of total sum of squares captured in cross validation or Q^2

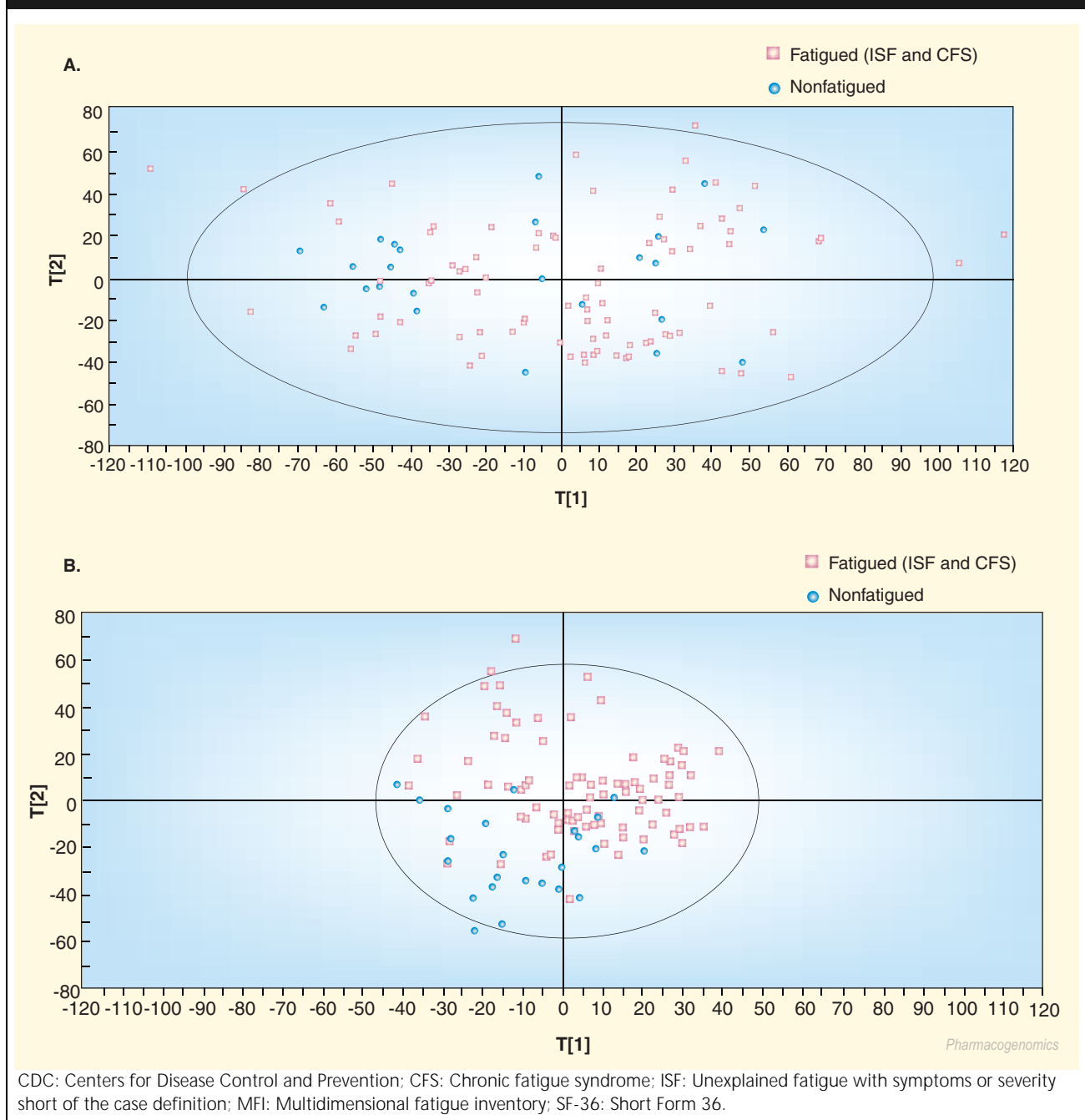
The latter is similar to the R^2 measure, but uses the predicted residual sum of squares (PRESS) instead of the standard residual sum of squares. The Q^2 is therefore an indication of the predictive ability of a model, rather than just its level of fit to training data. These measures are all applied in their multivariate form across all target variables as a block. For a more visual evaluation of model performance, we examined the scatter in a new feature space of subjects pre-assigned to each of three diagnostic groups: non-fatigued, ISF and CFS. To describe this scatter in coarse terms, we simply fit the coordinates of the subjects in each group to a two-dimensional (2D) normal distribution. We performed an analysis of these empirical distributions as a cursory evaluation of how cohesive and distinct each pre-assigned diagnostic group remained in different feature spaces.

Results

PCA of gene expression & symptom space data

In a PCA model of the gene expression data, 55 features were used to capture over 80% of the overall variation in microarray response. The first two features captured 16% of the variability

Figure 1. Subjects projected in the space defined by the first two composite features of gene expression when these features are (A) unrotated, and (B) rotated toward the full symptom space including MFI, SF-36, CDC symptom inventory, cognitive and sleep variables.



in 15,136 variables. Figure 1A shows the subjects located in a 2D space defined by the two most significant features (i.e., co-regulation patterns). Clearly the nonfatigued subjects are not well discriminated on the basis of gene expression alone.

Applying the same technique to the symptom space comprised of MFI scores, SF-36, CDC symptom inventory, and cognitive response, we found that roughly a quarter of

the overall variability in 59 symptom constructs is captured in a single feature. While subjects form a continuous distribution, this composite indicator allows a clear segregation of the nonfatigued subjects (data not shown). As the symptom space includes measures of the CFS phenotype, this segregation is expected and indicates that the symptom space captures meaningful measures of CFS.

Table 1. Projection model summary statistics.

Target space	Feature space	Features	Overall fit and predictive power					
			Target space			Feature space		
			R ²	Adj. R ²	Q ²	R ²	Adj. R ²	Q ²
MFI, SF-36, CDC SI, Zung, CANTAB	15,136 gene probes	2	0.170	0.155	0.0023	0.102	0.085	N/A
MFI, SF-36, CDC SI, Zung, CANTAB	Blood evaluation, catecholamines, endocrine, sleep general, sleep heart rate variability, tilt blood pressure, urine profile, cytokines, medications	2	0.152	0.136	0.0237	0.118	0.083	N/A

CANTAB: Cambridge Neuropsychological Test Automated Battery; CDC SI: CDC symptom inventory scores; MFI: Multidimensional fatigue inventory; SF-36: Short Form 36; Zung: Zung self rating depression scale index.

Rotated projections of gene expression space to symptom space

In an attempt to isolate the components of gene expression variability associated with the disease state, PLS was used to align gene features (co-regulation patterns) with shared trends in the symptom space. In Figure 1B, the subjects are projected into the feature space defined by the first two gene features that best align with the symptom space. The nonfatigued subjects group naturally in the lower left quadrant. As expected, the expression patterns responsible for this distinction is subtle, involving less than 10% of the overall variability (Table 1) in gene response.

The significance of a gene probe's overall contribution to the feature space in Figure 1B was evaluated by means of a pseudo-t; namely the ratio of the VIP to its corresponding standard error as determined by jackknifing. Out of 15,136 candidates, 6081 genes were estimated to contribute to the model at the 95% significance level (pseudo-t > 1.98) based on a standard parametric t-test. In a closer examination of the data distribution, a rather narrow distribution of signal to noise values indicated a large number of marginally significant contributors and a deviation from statistical normality. As a result, the empirical distribution was used directly to narrow the candidates to those genes with the highest significance. We selected the 95th percentile resulting in 766 highly significant genes (pseudo-t > 3.93). We further restricted the selection to the 95th percentile of the distribution of VIP results and identified 39 genes that were not only significant, but also highly influential. These are listed in Table 2. Only 17 of the 39 genes were functionally annotated in Pathway Miner software [101].

Six genes are not mapped to the current Genbank numbers, and two are hypothetical proteins. Four (indicated by [*] in Table 2) were also

identified as correlating with fatigue using quantitative trait analysis [21]. The mitogen-activated protein kinase (MAPK) signaling pathway is well represented, as is the Wnt signaling pathway. Also identified is the mammalian target of rapamycin (mTOR) pathway associated with ribosome function. The highest gain is that belonging to sestrin 1 (*SESN1*), a gene associated with maintaining the cell's antioxidant 'firewall'. Interferon regulatory factor 5 (*IRF5*) encodes a member of the IRF family, which plays a multitude of diverse roles, including virus-mediated activation of interferon, and modulation of cell growth, differentiation, apoptosis and immune system activity. Also associated with immune function is the gene presenting the CD74 antigen, an invariant polypeptide of the class II major histocompatibility complex.

The individual contribution of the 39 most influential gene probes to each composite feature is presented in Figure 2. These results show subtle patterns of synergistic and antagonistic interaction. In feature 1, while most of the 39 probe responses share a similar positive weight, a subset of six genes stand out as oppositely weighted, including dual specificity phosphatase 14 (*DUSP14*). The second feature appears to add a layer of detail to the first by altering the close alignment of the positively-weighted subset in feature 1 and essentially removing probes with rank 13, 31 and 37 in Table 2 from contributing in this second dimension of the feature space.

Rotated projections of clinical variables to symptom space

In an attempt to isolate the components of variability in the diverse clinical measures that are associated with the disease state, PLS was used to align clinical features with shared trends in the symptom space. As with gene expression, using the first two

Table 2. Top gene probe contributors to symptom space (95th percentile) .

Rank	Accession	VIP	SE	Pseudo t	Gene symbol	Description
1	AF033122	3.40	0.82	4.14	<i>SESN1</i>	Sestrin; GADD family; p53-regulated protein PA26; induced by genotoxic stress
2	AK056531*	3.11	0.62	4.99	<i>ODZ4</i>	Odd Oz/ten-m homolog 4 (<i>Drosophila</i>)
3	BC005853*	3.04	0.49	6.25	<i>ANKRD40</i>	Ankyrin repeat domain 40
4	AF286904*	2.80	0.66	4.24	<i>EPC2</i>	Enhancer of polycomb homolog 2
5	XM_031348	2.77	0.63	4.38		Unmapped
6	AF118637	2.77	0.48	5.74	<i>FLVCR</i>	Feline leukemia virus subgroup C receptor
7	NM_002200	2.65	0.36	7.40	<i>IRF5</i>	The information-processing pathway at the IFN-β enhancer
8	AK026341	2.58	0.50	5.17	FLJ22688	Hypothetical protein FLJ22688
9	AL023653	2.57	0.57	4.49		Unmapped
10	AK056277	2.55	0.63	4.07	FLJ31715 <i>GOLGA4</i>	Hypothetical protein FLJ31715 golgi autoantigen, golgin subfamily α, 4
11	L21998	2.46	0.55	4.49	<i>MUC2</i>	Trefoil factors initiate mucosal healing
12	NG_0000080	2.45	0.47	5.18		Unmapped
13	NM_004636	2.44	0.60	4.07	<i>SEMA3B</i>	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B
14	AF284765	2.43	0.37	6.60	<i>KLHL4</i>	Kelch-like 4 (<i>Drosophila</i>)
15	NM_032255	2.43	0.52	4.64	<i>ZNF541</i>	Zinc finger protein 541
16	NM_003436	2.43	0.50	4.82	<i>ZNF78L1</i> <i>ZNF135</i>	Zinc finger protein 78-like 1 (pT3) Zinc finger protein 135 (clone pHZ-17)
17	AF188745	2.42	0.51	4.78	<i>KLK2</i>	Erk1/Erk2 MAPK signaling pathway; nerve growth factor pathway (NGF); phosphorylation of MEK1 by cdk5/p35 down regulates the MAP kinase pathway; Trka receptor signaling pathway
18	XM_172332	2.41	0.33	7.23		Unmapped
19	AF403384	2.41	0.39	6.23	<i>LGR8</i>	Leucine-rich repeat-containing G protein-coupled receptor 8
20	AL022726	2.39	0.60	3.95		Unmapped
21	BC009891	2.37	0.55	4.32	<i>TCL1A</i>	T-cell leukemia/lymphoma 1A
22	BC001259	2.36	0.43	5.47	<i>AP4S1</i>	Adaptor-related protein complex 4, σ1 subunit
23	AF090189	2.35	0.55	4.30	<i>CER1</i>	Wnt signaling pathway
24	BC014930	2.33	0.52	4.43	<i>EIF4G2</i>	mTOR signaling pathway; internal ribosome entry pathway; regulation of eIF4e and p70 S6 kinase
25	AF026692*	2.32	0.28	8.42	<i>SFRP4</i>	Wnt signaling pathway
26	D13631	2.31	0.33	7.09	<i>ARHGEF6</i>	Integrin-mediated cell adhesion; agrin in postsynaptic differentiation
27	NM_032126	2.30	0.38	6.10	C1orf49	Chromosome 1 open reading frame 49
28	AF038852	2.30	0.52	4.45	<i>CACNB4</i>	Hs_calcium_channels

*Also identified as significant common contributors to all five MFI metrics by Whistler and colleagues [21].

CACNB4: Calcium channel, voltage-dependent, β4 subunit; *CER1*: Cerberus 1, cysteine knot superfamily, homolog; *DUSP14*: Dual specificity phosphatase 14; *EIF4G2*: Eukaryotic translation initiation factor 4 γ, 2; *GADD*: Growth arrest and DNA-damage; *MAPK*: Mitogen-activated protein kinase; *MBTPS1*: Membrane-bound transcription factor peptidase, site 1; *MFI*: Multidimensional fatigue inventory; *mTOR*: Mammalian target of rapamycin; *SE*: Standard error; *SFRP4*: Secreted frizzled-related protein 4; *SREBP*: Sterol regulatory element-binding protein; *VIP*: Variable importance in projection.

Table 2. Top gene probe contributors to symptom space (95th percentile) (cont.).

Rank	Accession	VIP	SE	Pseudo t	Gene symbol	Description
29	NM_015967	2.29	0.53	4.34	<i>PTPN22</i>	Protein tyrosine phosphatase, nonreceptor type 22 (lymphoid)
30	BC026330	2.29	0.46	5.02	<i>MBTPS1</i>	SREBP control of lipid synthesis
31	NM_025263	2.28	0.45	5.08	<i>PRR3</i>	Proline rich 3
32	BC024272	2.28	0.49	4.64	<i>CD74</i>	Antigen processing and presentation
33	NM_004366	2.24	0.39	5.71	<i>CLCN2</i>	Chloride channel 2
34	AF175206	2.24	0.50	4.51	<i>KLRF1</i>	Killer cell lectin-like receptor subfamily F, member 1
35	AF120032	2.22	0.51	4.39	<i>DUSP14</i>	MAPK signaling pathway
36	NM_021648	2.21	0.36	6.19	<i>TSPYL4</i>	TSPY-like 4
37	AC004991	2.19	0.46	4.81		Unmapped
38	AF065314	2.19	0.48	4.57	<i>CNGA3</i>	Cyclic nucleotide gated channel α 3
39	L10334	2.18	0.45	4.80	<i>RTN1</i>	Reticulon 1

*Also identified as significant common contributors to all five MFI metrics by Whistler and colleagues [21].

CACNB4: Calcium channel, voltage-dependent, β 4 subunit; *CER1*: Cerberus 1, cysteine knot superfamily, homolog; *DUSP14*: Dual specificity phosphatase 14; *EIF4G2*: Eukaryotic translation initiation factor 4 γ 2; *GADD*: Growth arrest and DNA-damage; *MAPK*: Mitogen-activated protein kinase; *MBTPS1*: Membrane-bound transcription factor peptidase, site 1; *MFI*: Multidimensional fatigue inventory; *mTOR*: Mammalian target of rapamycin; *SE*: Standard error; *SFRP4*: Secreted frizzled-related protein 4; *SREBP*: Sterol regulatory element-binding protein; *VIP*: Variable importance in projection.

composite features of the clinical variable block defined a space with a continuous distribution of subjects where nonfatigued subjects formed one end of the continuum (data not shown).

The VIP values for the contribution of 45 of the original 117 clinical variables exceeds twice the associated estimates for standard error (equivalent to approximately 60% significance). Computing the 95% confidence interval about the mean contribution in this subset, we obtain 17 clinical variables that display contributions significantly greater than average ($VIP > 1.41$). Of these only five are highly significant with a pseudo-t greater than 4.4 (approximately 95% significant). These variables are shown in Table 3. The variable with the single highest contribution to defining the two-component clinical feature space is the number of medications being taken by the subject that target CNS function. Nine variables describe heart rate variability during sleep, making this the most highly represented subgroup. Two indicators relate to the tilt table test namely, the heart rate measured 5 min after returning to a standing position and the recumbent heart rate at 30 min. Free or unbound T4 also ranks prominently with lower than average levels pointing toward fatigue-related illness.

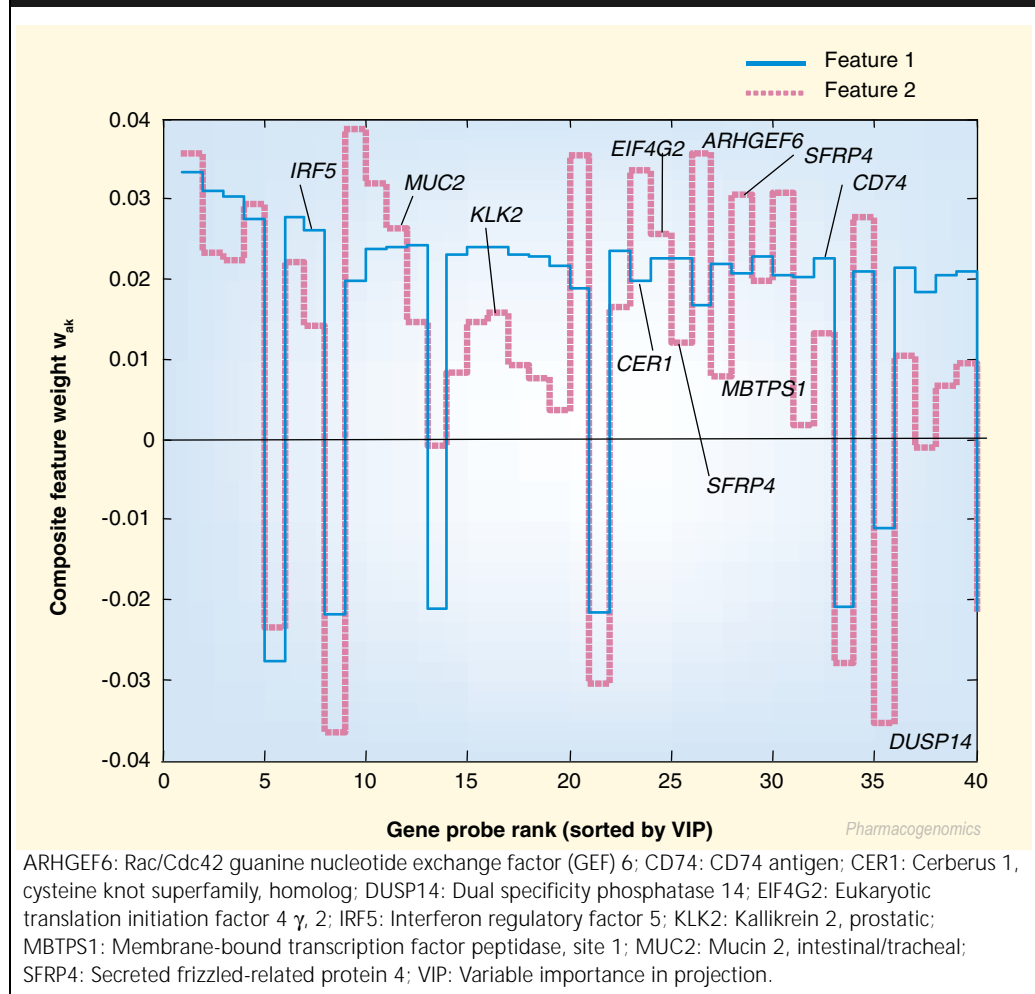
Discussion

In this study, we used classical multivariate projection methods to examine the CFS computational challenge data [7]. We used this unsupervised

approach to evaluate natural structure in the data. As gene regulatory networks display a high level of organization there tend to be significantly fewer distinct co-regulation patterns than there are individual genes. A denoising or averaging effect results from co-regulated genes becoming in essence repeated measures of the same underlying phenomenon. Linear methods for extracting these co-regulation patterns or composite features offer several advantages, not the least of which is that of a unique solution. Furthermore, by forcing these features to be mutually independent, overall variation in the original data can be formally decomposed, making it possible to gauge the relative importance of each model component. A review and comparison of multivariate projection methods can be found in work by Jackson [22]. As a *post hoc* evaluation of the model, we used the distribution of subjects within the 2D space of the first two composite features to determine if their spatial distribution correlated with disease classification.

We applied PCA to 59 variables known to contribute to the multiple dimensions of illness that make up CFS. We could identify a single common trend in these symptom constructs, representing roughly a quarter of the overall variation. The subjects were distributed as a continuum with the nonfatigued control subjects at one end (data not shown). This suggests that these variables manifest in a coordinated fashion and that the symptom space they define is a good representation of the illness.

Figure 2. Weights w_{ak} assigned to each of the influential gene probes k in Table 2 encoded in feature $a = 1$ and $a = 2$.



The PCA analysis of the gene expression data alone (unrotated, Figure 1A), does not capture illness effectively, as fatigued and nonfatigued subjects are intermingled. This suggests that the vast majority of the variation in the gene expression data are attributable to factors other than illness. However, using PLS to rotate the gene expression features to the symptom space allows us to extract the relatively subtle proportion of gene expression variation that does correlate with illness (as measured by the symptom space). PLS analysis identified two co-regulation patterns which support a distribution of subjects that reflect fatigue status (rotated, Figure 1B, Figure 3). Representing roughly 10% of the overall variation in gene expression, these co-regulation patterns are quite subtle and are easily obscured by the myriad of mainstream regulatory mechanisms.

Of the 39 genes with the highest influence (strength and statistical significance), only 17 are functionally annotated (Table 2). Four of these

genes were identified as correlating with fatigue applying quantitative trait analysis to the same data [21]. This is not surprising when one considers the high level of commonality linking MFI, SF-36 and the other components of the symptom space. However, the model structure used here captures coordinated changes in symptom space that include many dimensions of CFS. This is quite different from addressing the elements of a single measurement tool such as MFI individually or relying on an *a priori* classification of subjects required in analyses of differential gene expression. This may explain why only two (protein kinase C-like 1 [*PRKCL1*] and KH-type splicing regulatory protein [*KHSRP*]) of the 35 genes identified by Kaushik and colleagues [23], reached statistical significance in our model, and neither contributed much to the definition of the shared gene-symptom feature space; VIP values of 0.81 and 0.82, respectively. Similarly, only three of the candidate genes identified by Vernon

Table 3. Ranking of clinical variables in contribution to symptom space.

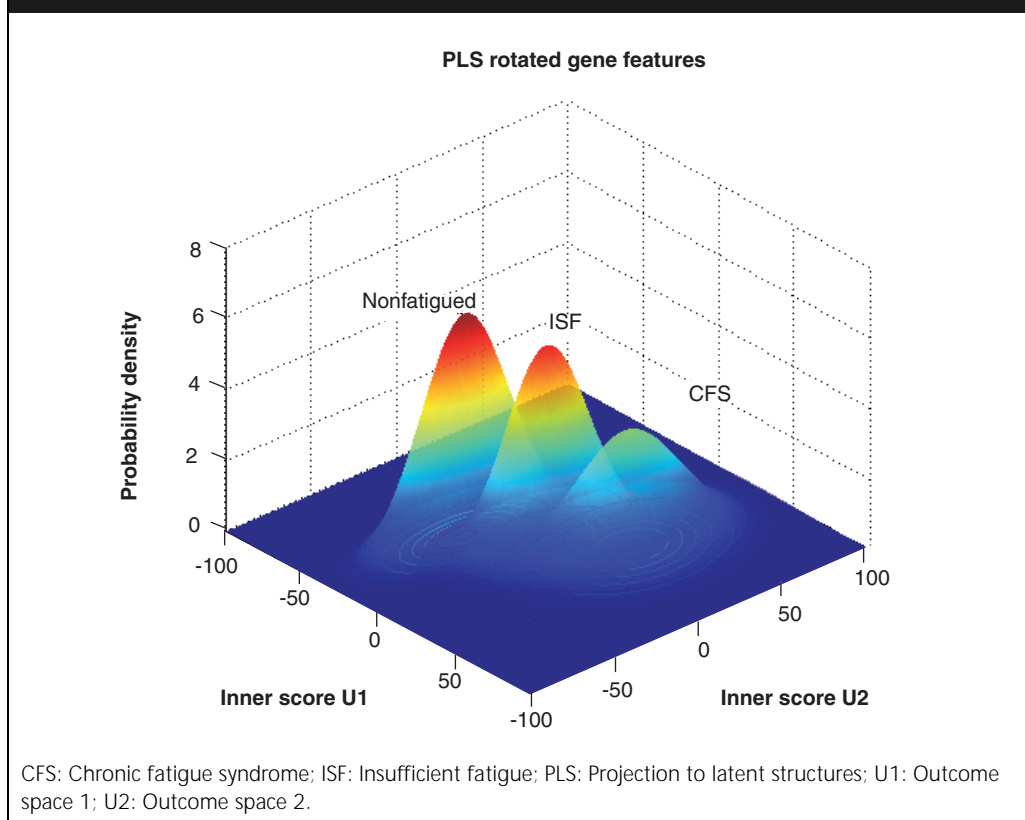
Rank	Variable tag	Variable block ID	VIP	SE	pseudo t
Better than average contributors					
1	num_cns	Medications	2.45	0.69	3.57
2	Epworth	Sleep general	2.04	0.63	3.26
3	tp	Sleep heart rate variability	2.03	0.35	5.84
4	vlf	Sleep heart rate variability	1.99	0.60	3.30
5	rr_interval	Sleep heart rate variability	1.97	0.24	8.37
6	free_thyroxine_free_t4	Endocrine	1.94	0.78	2.50
7	hrv_triangle	Sleep heart rate variability	1.85	0.41	4.48
8	nn50	Sleep heart rate variability	1.84	0.36	5.06
9	num_total_meds	Medications	1.82	0.54	3.40
10	Potassium	Blood_eval	1.81	0.59	3.09
11	pnn50	Sleep heart rate variability	1.80	0.43	4.15
12	hr_5min_stand	Tilt blood pressure	1.73	0.60	2.91
13	Lf	Sleep heart rate variability	1.72	0.20	8.50
14	hr_recum_30min	Tilt blood pressure	1.62	0.47	3.46
15	hf	Sleep heart rate variability	1.60	0.45	3.57
16	sdnn	Sleep heart rate variability	1.55	0.45	3.43
17	specific_gravity	Urine profile	1.49	0.67	2.22
Average contributors					
18	hr_imm_stand	Tilt blood pressure	1.38	0.34	4.09
19	num_sorems	Sleep general	1.34	0.61	2.18
20	sdnn_index	Sleep heart rate variability	1.34	0.38	3.52
21	alpha_noted	Sleep general	1.28	0.63	2.03
22	progesterone	Endocrine	1.24	0.58	2.14
23	albumin	Blood_eval	1.21	0.35	3.45
24	CO ₂	Blood_eval	1.20	0.42	2.86
25	aldosterone	Endocrine	1.20	0.45	2.65

albumin: Blood albumin content; aldosterone: Blood aldosterone concentration; alpha_noted: a electroencephalogram (EEG) arousal disturbance during nonrapid eye movement sleep; CO₂: Blood CO₂ concentration; epworth: Epworth Sleepiness Total Scale Score; free_thyroxine_free_t4: Free thyroxine T4 in the blood; hf: Power of electrocardiogram signature in the high frequency range (0.15-0.4 Hz); hr_5min_stand: Heart rate after 5 minutes standing; hr_imm_stand: Heart rate immediately upon standing after standard tilt test; hr_recum_30min: Recumbent heart rate after 30 minutes standing; hrv_triangle: Heart rate variability (HRV) triangular index is the total number of NN intervals divided by the height of the histogram of all NN intervals measured on a discrete scale with bins of 7.8125 ms (1/128 seconds); ID: Identification; lf: Power of electrocardiogram signature in the low frequency (LF) range (0.04-0.15 Hz); nn50: NN50 count is the number of pairs of adjacent normal-to-normal (NN) intervals with difference exceeding 50ms; num_cns: Number of CNS medications; num_sorems: Number of segments of rapid eye movement (REM) sleep; num_total_meds: Total number of medications being taken by the patient; pnn50: NN50 divided by the total number of NN intervals; potassium: Potassium level in blood; progesterone: Blood progesterone content; rr_interval: RR Interval is the mean duration of the interval between two normal QRS complexes in the electrocardiogram; sdnn: Standard deviation of all NN intervals; sdnn_index: SDNN index is the mean of the standard deviation of all NN intervals for all 5-minute segments; specific_gravity: Urine specific gravity; SE: Standard error; tp: Total power or the variance of all normal-to-normal heart beat (NN) intervals (~=<0.4 Hz); VIP: Variable importance in projection; vlf: Power in very low frequency (VLF) range (~=<0.04 Hz) of the electrocardiogram signature.

and colleagues [8], nuclear factor of activated T-cells (*NF-AT4c*), α 2A adrenoceptor and lymphocyte cytosolic protein 1 (*LCP-1*), were significant in our model and none were influential contributors; VIPs of 1.22, 1.05 and 1.22, respectively. However, it is interesting to note that from a functional standpoint many of the candidate genes identified in this work align well with

the results of a recent study where analysis was based on gene groups as defined by ontology [9]. Indeed, results in Table 2 also point to changes in ion transport and ion channel activity, as well as immune response. In addition, recent evidence suggests that excessive free radical generation may also be involved in CFS pathogenesis [24]. Once again, the highest gain reported in Table 2 is that

Figure 3. Empirical probability density functions describing the scatter of nonfatigued, ISF and suspected CFS subjects in the rotated gene feature space of Figure 1B.



belonging to *SESNI*, a gene associated with cellular response to oxidative stress. It may be interesting to note that immune response to persistent infection has been linked to excessive generation of free radicals by activated white cells [25]. Increased cell membrane permeability as a consequence of oxidative attack has also been reported [26] with a consequent increase in intracellular calcium. This may help explain the appearance in Table 2 of calcium channel, voltage-dependent $\beta 4$ subunit (*CACNB4*), a gene involved in calcium channel activity, as well as differences in membrane-bound transcription factor peptidase, site 1 (*MBTPS1*) regulation of lipid synthesis.

We also used PLS to align the clinical feature space to the symptom target space to identify variation in clinical measures that correlate with illness. Clinical variables with the most significant and strongest contribution to the model are listed in Table 3. Many of these measures can be related to allostatic load, and its impact on vagal tone and variables describing heart rate variability (HRV) are prominent. A marker of cumulative wear and tear, a decline in HRV has been

attributed to a decrease in efferent vagal tone and reduced β -adrenergic responsiveness. HRV will also be influenced by aldosterone antagonists. For example, potassium will acutely improve cardiac vagal control [27]. Potassium levels have also been related to immune response [28]. In a recent study [29], over 50% of CFS patients were found to have an abnormal whole-body potassium level that associated with CD19⁺ and CD5⁺ cell counts.

It should be noted that the variable with the single highest contribution to defining the two-component clinical feature space is the number of medications being taken by the subject which target CNS function. Indeed, the results of a recent analysis based on classical univariate statistics also show the use of medication, including drugs targeting CNS function, to be more prevalent in CFS patients than in nonfatigued control subjects [30]. This only emphasizes the problem that medications bring to the analysis of CFS. Of course the difficulty of subjecting patients to a controlled wash-out period is only exacerbated in a population-based study such as this one. As a result, we

Highlights

- Principal components analysis (PCA) is an unsupervised approach to evaluate natural structure in the data and is particularly useful for analyzing gene expression as co-regulated genes become, in essence, repeated measures of the same underlying phenomenon resulting in a denoising or averaging effect.
- We constructed a symptom space from 59 variables known to contribute to the multiple dimensions of illness that make up chronic fatigue syndrome (CFS). These symptom constructs had a single common trend representing roughly a quarter of the overall variation and resulting in a distribution that correlated with disease.
- Unrotated PCA of gene expression did not capture illness effectively. However, using projection to latent structures (PLS) to rotate the gene expression features to the symptom space allows the relatively subtle proportion of gene expression variation that does correlate with illness (approximately 10%) to be identified.
- The model structure used here captures coordinated changes in a symptom space that includes all dimensions of CFS. This is quite different from addressing the elements of a single measurement tool such as multidimensional fatigue inventory (MFI) individually or relying on an *a priori* classification of subjects required in analyses of differential gene expression
- The model identifies 39 genes with the highest statistical significance and strength of contribution. Only a fraction of these can be functionally annotated at this time. Nonetheless, they are good candidates for further evaluation. Within the annotated group, the functions of oxidative stress response, ion transport and immune response appear prominently.
- The definition of the clinical feature space is dominated by variables describing heart rate variability (HRV) during sleep, a marker of cumulative wear and tear. Blood potassium and free T4 also figure prominently.
- The variable with the single highest contribution to defining the two-component clinical feature space is the number of medications being taken by the subject that target CNS function. This emphasizes the problem that medications bring to the analysis of CFS. We cannot determine if the medications are contributing to the phenotype, or whether they are a marker for illness (i.e., those who are ill are more likely to take medications).
- Though strong common patterns existed between most symptom constructs, Cambridge Neuropsychological Test Automated Battery (CANTAB) cognitive response stood apart and was not well described in either the two-feature gene space or clinical space.

cannot determine with certainty if the medications are contributing to the phenotype, or whether they are a coincidental marker for illness (i.e., those who are ill are more likely to take medications). However, while the use of CNS medications may be more prevalent in CFS patients, the prior study cited above [30] failed to find a correlation between CNS medication use and change in fatigue status. It could therefore be suggested that increased medication use in CFS patients is directed at symptom relief, and that these medications in fact do little to alter the underlying cause and mechanisms of fatigue. This argument is further supported by the significant alignment of

influential genes identified in this work with the ontology categories identified independently in another study [9], where a 7-day drug wash-out period was enforced.

Finally, when subjects are plotted in the 2D space created by the first two features of the gene expression and clinical variables, no distinct groupings appear. Though subjects with the same disease status are grouped together, as shown by the empirical probability distributions of the scatter of nonfatigued, ISF and CFS (Figure 3), these groups overlap. This suggests that for the subjects and measures included, unexplained fatiguing illness represents a spectrum. Differences between CFS and ISF could therefore be of disease severity, rather than type of disease.

This analysis identifies clinical and gene expression variables that merit further study. The main strength of this approach is that using the symptom space as a target allows relatively subtle disease relationships to be identified. Refining measures of CFS could improve the utility of the model. While we included CANTAB measures in the symptom space, they stood apart from the common patterns that existed between most symptom constructs. CANTAB cognitive response was not well described by the two-feature models, and it is possible that the models may be improved by omitting CANTAB from the symptom space. At the very least, this may lead to refinements in the definition of the symptom space that might better reflect all dimensions of CFS.

Outlook

Empirical models, be they simple polynomial functions, multivariate projections, or their nonlinear counterparts, support vector machines (SVMs) and artificial neural networks (ANNs), are all universal approximators of one type or another. These methods use simple functions and their weighted sums or products to approximate the aggregated effects of basic physico-chemical processes which underlie, and drive a system. This general form is used for two reasons:

- In many instances the exact mechanistic model is unknown
- The limited data would not support an unequivocal fit of this sophisticated model even if it were known

A more direct representation of nature, exact first-principles models, predict new outcomes more reliably than empirical models that

perform best over limited ranges. Nonetheless, these generic data-driven pathway discovery tools have, and will continue to assist in identifying many promising candidate gene regulatory networks. However, the robust predictive capabilities required for hypothesis testing will only be achieved through the development of detailed mechanistic models that capture the spatial and temporal dimensions of disease.

As the availability of raw computational power continues to improve, the principal obstacle to the development of these detailed models is our inability to measure anything more than coarse and incomplete snapshots of physiological and biochemical processes. Much has been said about the life sciences being data-rich, but in reality a modern oil refinery has several orders of magnitude more data recorded at sampling frequencies currently unimaginable in medicine. As a result, the pioneering developments enabling predictive computational medicine must first be achieved through contributions from advanced physics in

the area of instrument technology. This has already begun with quantum tagging allowing us to track the position and chemical state of individual protein molecules within a living cell. Measurements such as these will make it possible to estimate the diffusion properties of biomolecules and the dynamics of chemical transitions under actual *in vivo* conditions. Simple artificial organisms constructed with basic subcellular building blocks (BioBlocks-MIT) will also contribute to our understanding and enable hypothesis testing of cellular processes with molecular resolution. It is by examining and characterizing life's synergistic systems from the ground up that we can hope to better understand the staggering complexity that emerges from a whole that is truly greater than the sum of its parts.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the funding agency.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Evengard B, Evengard B, Schacterle RS, Komaroff AL: Chronic fatigue syndrome: new insights and old ignorance. *J. Intern. Med.* 246, 455–469 (1999).
2. Lyall M, Peakman M, Wessely S: A systematic review and critical evaluation of the immunology of chronic fatigue syndrome. *J. Psychosom. Res.* 55, 79–90 (2003).
3. Papanicolaou DA, Amsterdam JD, Levine S *et al.*: Neuroendocrine aspects of chronic fatigue syndrome. *Neuroimmunomodulation* 11, 65–74 (2004).
4. Moss-Morris R, Petrie KJ: Discriminating between chronic fatigue syndrome and depression: a cognitive analysis. *Psychol. Med.* 31, 469–479 (2001).
5. Pheley AM, Melby D, Schenck C, Mandel J, Peterson JK: Can we predict recovery in chronic fatigue syndrome? *Minn. Med.* 82, 52–56 (1999).
6. Reyes M, Dobbin JG, Nisenbaum R, Subedar N, Randall B, Reeves WC: Chronic fatigue syndrome progression and self-defined recovery: evidence from CDC surveillance system. *J. Chronic Fatigue Syndr.* 5, 17–27 (1999).
7. Vernon SD, Reeves WC: The challenge of integrating disparate high-content data: epidemiologic, clinical, and laboratory data collected during an in-hospital study of chronic fatigue syndrome. *Pharmacogenomics* 7(3), 345–354 (2006).
8. Vernon SD, Unger ER, Dimulescu IM, Rajeevan M, Reeves WC: Utility of the blood for gene expression profiling and biomarker discovery in chronic fatigue syndrome. *Dis. Markers* 18, 193–199 (2002).
- **This is one of the pioneering studies that demonstrated the successful use of gene expression measurements from peripheral blood mononuclear cells (PBMCs) in the study of chronic fatigue syndrome (CFS). This study also points to immune system activation in CFS patients, as do the results herein.**
9. Whistler T, Jones JF, Unger ER, Vernon SD: Exercise responsive genes measured in peripheral blood of women with chronic fatigue syndrome and matched control subjects. *BMC Physiol.* 5, 5 (2005).
- **This work again demonstrates the use of gene expression in PBMCs using a larger array of probes. More importantly, study results show a set of gene ontology categories that align well with gene functional groups found to distinguish CFS from nonfatigued subjects in this work.**
10. Reeves WC, Wagner D, Nisenbaum R *et al.*: Chronic fatigue syndrome – a clinically empirical approach to its definition and study. *BMC Med.* 3, 19 (2005).
11. Wagner D, Nisenbaum R, Heim C, Jones JF, Unger ER, Reeves WC: Psychometric properties of the CDC Symptom Inventory for assessment of chronic fatigue syndrome. *Popul. Health Metr.* 3, 8 (2005).
12. Martens H, Naes T: Multiplicative signal correction (MSC). In: *Multivariate Calibration*. John Wiley and Sons Ltd, New York, NY, USA, 345–351 (1991).
13. Geladi P, MacDougall D, Martens H: Linearization and scatter correction for near-infrared reflectance spectra of meat. *Appl. Spectroscopy* 39, 491–500 (1985).
14. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193 (2003).
15. Park T, Yi S-G, Lee SY, Lee JK: Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *BioTechniques* 38, 463–471 (2005).
16. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57(1), 289–300 (1995).
17. Wold H: Estimation of principal components and related models by iterative least squares. In: *Multivariate Analysis* Krishnaiah PR, (Ed.), Academic Press, New York, NY, USA, 391–420 (1966).
18. Efron B, Gong G: A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.* 37, 36–48 (1983).
19. Krzanowski WJ: Cross-validation in principal component analysis. *Biometrics* 43, 575–584 (1987).

20. Wold S, Trygg J, Berglund A, Antti H: Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* 58, 131–149 (2001).
21. Whistler T, Craddock RC, Taylor R, Broderick G, Klimas N, Unger ER: Gene expression correlates of fatigue. *Pharmacogenomics* 7(3), 395–405 (2006).
22. Jackson JE: *A user's guide to principal components*. John Wiley and Sons Ltd., New York, NY, USA (1991).
23. Kaushik N, Fear D, Richards SCM *et al.*: Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome. *J. Clin. Pathol.* 58, 826–832 (2005).
- **In this work a larger array is used than in [8] and [9]. Furthermore, polymerase chain reaction (PCR) is used to validate microarray results. Once again the functional categories identified align well with results from the present study.**
24. Kennedy G, Spence VA, McLaren M, Hill A, Underwood C, Belch JFF: Oxidative stress levels are raised in chronic fatigue syndrome and are associated with clinical symptoms. *Free Radic. Biol. Med.* 39, 584–589 (2005).
- **This presents new and convincing evidence linking CFS to excessive free radical generation. A very thorough review of literature is used to discuss the repercussions and possible causes of this free radical generation. These include possible ties to the gene functions identified in this study such as immune system activation and ion channel activity.**
25. Fantone JC, Ward PA: Polymorphonuclear leukocyte-mediated cell and tissue injury: oxygen metabolites and their relations to human disease. *Hum. Pathol.* 16, 973–978 (1985).
26. Elliott S J, Koliwad SK: Oxidant stress and endothelial membrane transport. *Free Radic. Biol. Med.* 49, 649–658 (1995).
27. Fletcher J, Buch AN, Routledge HC, Chowdhary S, Coote JH, Townend JN: Acute aldosterone antagonism improves cardiac vagal control in humans. *J. Am. Coll. Cardiol.* 43(7), 1270–1275 (2004).
28. Levite M, Cahalon L, Peretz A *et al.*: Extracellular K⁺ and opening of voltage-gated potassium channels activate T-cell integrin function: physical and functional association between Kv1.3 channels and β 1 integrins. *J. Exp. Med.* 191, 1167–1176 (2000).
29. Nijs J, Demanet C, McGregor NR *et al.*: Monitoring a hypothetical channelopathy in chronic fatigue syndrome: preliminary observations. *J. Chronic Fatigue Syndr.* 11(1), 117–133 (2003).
30. Jones JF, Nisenbaum R, Reeves WC: Medication use by persons with chronic fatigue syndrome: results of a randomized telephone survey in Wichita, Kansas. *Health Qual. Life Outcomes* 1, 74 (2003).
- **This work addresses the quintessential challenge that medication use presents to CFS studies. Of particular interest is the fact that study results fail to find a correlation between CFS remission and medication use. This supports the hypothesis that the medications in question do not significantly alter CFS pathophysiology.**
- Websites
101. BioRag (Bioresource for array genes), Bioinformatics group at Arizona Cancer Center. www.biorag.org