# Gene expression profile of empirically delineated classes of unexplained chronic fatigue

*Liran Carmel[1], Sol Efroni[2], Peter D White[3], Eric Aslakson[4], Ute Vollmer-Conna[5] & Mangalathu S Rajeevan[4†]*

*[†]Author for correspondence*
*[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA*
*[2]National Cancer Institute Center for Bioinformatics, National Institutes of Health, Bethesda, Maryland, USA*
*[3]University of London, Department of Psychological Medicine, Barts, London and Queen Mary School of Medicine and Dentistry, London, UK*
*[4]Centers for Disease Control and Prevention, 1600 Clifton Road, MSG 41, Atlanta, GA 30333, USA*
*Tel.: +1 404 639 2931; Fax: +1 404 639 3540; E-mail: mor4@cdc.gov*
*[5]University of New South Wales, School of Psychiatry, Sydney, Australia*

**Objectives:** To identify the underlying gene expression profiles of unexplained chronic fatigue subjects classified into five or six class solutions by principal component (PCA) and latent class analyses (LCA). **Methods:** Microarray expression data were available for 15,315 genes and 111 female subjects enrolled from a population-based study on chronic fatigue syndrome. Algorithms were developed to assign gene scores and threshold values that signified the contribution of each gene to discriminate the multiclasses in each LCA solution. Unsupervised dimensionality reduction was first used to remove noise or otherwise uninformative gene combinations, followed by supervised dimensionality reduction to isolate gene combinations that best separate the classes. **Results:** The authors' gene score and threshold algorithms identified 32 and 26 genes capable of discriminating the five and six multiclass solutions, respectively. Pair-wise comparisons suggested that some genes (zinc finger protein 350 [*ZNF350*], solute carrier family 1, member 6 [*SLC1A6*], F-box protein 7 [*FBX07*] and vacuole 14 protein homolog [*VAC14*]) distinguished most classes of fatigued subjects from healthy subjects, whereas others (patched homolog 2 [*PTCH2*] and T-cell leukemia/lymphoma [*TCL1A*]) differentiated specific fatigue classes. **Conclusion:** A computational approach was developed for general use to identify discriminatory genes in any multiclass problem. Using this approach, differences in gene expression were found to discriminate some classes of unexplained chronic fatigue, particularly one termed interoception.

Keywords: chronic fatigue syndrome, Fisher quotient and discriminatory genes, gene expression and gene scores, interoception, latent class analysis, principal component analysis

Chronic fatigue syndrome (CFS) is a diagnosis of uncertain nosology and etiology [1–3]. There have been many studies of etiology and pathophysiology; however, consistent findings have been rare [3]. This may be due to the fact that CFS is heterogeneous, thus hindering any attempt to find biological markers. Several studies attempting to define CFS using symptoms and other clinical measures have produced evidence of heterogeneity [4–7], but no study has used biological measures to define the different putative endophenotypes that may make up CFS. Three small case–control studies of gene transcription in CFS have been published [8–10]. All three found differences in gene expression between CFS defined cases and healthy controls, but not one gene expression was common to two of the three studies. This discrepancy may be explained by the likely heterogeneity of CFS.

CFS case ascertainment is complicated by the fact that CFS is defined somewhat arbitrarily, although consensually [101]. This restrictive definition, useful for research though it is, omits two-thirds of people who have chronic medically unexplained fatigue [11]. In an attempt to refute or confirm the heterogeneity of CFS and chronic unexplained fatigue in general, principal components analysis (PCA) and latent class analysis (LCA) were used with clinical and biological measures from 159 female subjects of the Wichita (KS, USA) study [12].

More formally, a classification scheme is a division of a set of subjects into disjoint and not necessarily exhaustive classes. In other words, each subject is associated with, at most, one class. In this regard, a known subject is a subject that is associated with one particular class, while an unknown subject is a subject whose associated class is unknown. In studying diseases, a two-class classification scheme is typically used, with the two classes being healthy and ill. In the aforementioned study [12] the authors achieved a multiclass classification scheme consisting, in addition to a healthy class, of a collection of four or five different fatigue-related syndromes.

Given microarray data, it is relatively easy to detect genes whose expression level significantly differs between two classes, and a diverse repertoire of biomarker detection technique had been proposed [13–15]. The multiclass case poses a more difficult challenge, as typically no single gene can be solely used to discriminate between the classes. Recently several algorithms, predominantly based on machine learning techniques,

10.2217/14622416.7.3.375  ISSN 1462-2416

had been proposed for the multiclass problem [16–19]. Here, we would like to suggest a simple and intuitive technique, based on simple linear transformations of the data. This computational approach was further used to provide an external test of gene-expression-based validity for the multiple classes of unexplained chronic fatigue identified by PCA and LCA [12].

## Materials & methods

### *Subjects & classification of unexplained chronic fatigue*

This study adhered to human experimentation guidelines of the Helsinki Declaration and was approved by the Centers for Disease Control and Prevention (CDC) Institutional Review Board. All subjects were volunteers who gave informed consent.

Recruitment of subjects with medically unexplained chronic fatigue and matched controls was described by Vernon and Reeves [20]. Heterogeneity in female subjects from this group of subjects with fatiguing illness was assessed by Vollmer-Conna and colleagues [12] by PCA and LCA. Male subjects were excluded from the analysis since differentiation of distinct factors was strongly affected by normal male attributes (e.g., high testosterone and hemoglobin concentration) and there were insufficient male subjects to be adequately classified [12]. LCA resulted in two statistically coherent and interpretable classification schemes, a five-class solution (LCA5) and a six-class solution (LCA6) with clinical significance [12].

Among the classes in LCA5, Class 1 subjects were called 'obese hypnoea'; Class 2 captured well subjects; Class 3 subjects were 'obese hypnoea and stressed'; Class 4 were primarily 'interoceptive' and Class 5 included interoception–depressed subjects. Among the classes in LCA6 solution, Class 1 (obese hypnoea) contained obese subjects with prominent postexertional fatigue, sleep hypnoea and objective sleep disturbance; Class 2 (well) consisted of subjects who, although obese, were characterized by few symptoms, low depression scores and good objective sleep; Class 3 (obese hypnoea and stressed) captured individuals that were obese and had sleep hypnoea and a physiological stress response; Class 4 (interoception) differentiated a group with a lower body mass index and less depression, with symptoms of muscle pain and subjective sleep complaints but no objective sleep problems; Classes 5 and 6 were similar in that they captured less obese, but highly symptomatic and depressed individuals with prominent postexertional fatigue. In contrast to Class 5 (interoception–depression), subjects in Class 6 (multisymptomatic depressed, stressed and postmenopausal) were additionally defined by a lack of sex hormones and low heart rate variability during sleep, objective sleep disturbance and low cortisol. Detailed clinical characteristics and demographics of subjects assigned to different classes in the LCA5 and LCA6 solutions are given separately in this issue [12].

### *Microarray experiments*

Details on the collection and processing of peripheral blood mononuclear cells, total RNA extraction, cDNA synthesis, and microarray hybridization are provided by Vernon and Reeves [20]. Details on the quality control, technical replicates and normalization of the microarray data set of 15,315 genes used in this analysis were described by Whistler and colleagues [21]. The normalized expression levels of each gene across all the subjects were centered, that is brought to zero-mean. The present analysis was carried out with expression data from 111 out of 159 female subjects, only because microarray data from the remaining 48 subjects did not pass the quality control. The proportions of subjects assigned to different classes in the LCA5 and LCA6 solutions are shown in Figure 1 (individual subject assignment to classes is available upon request).

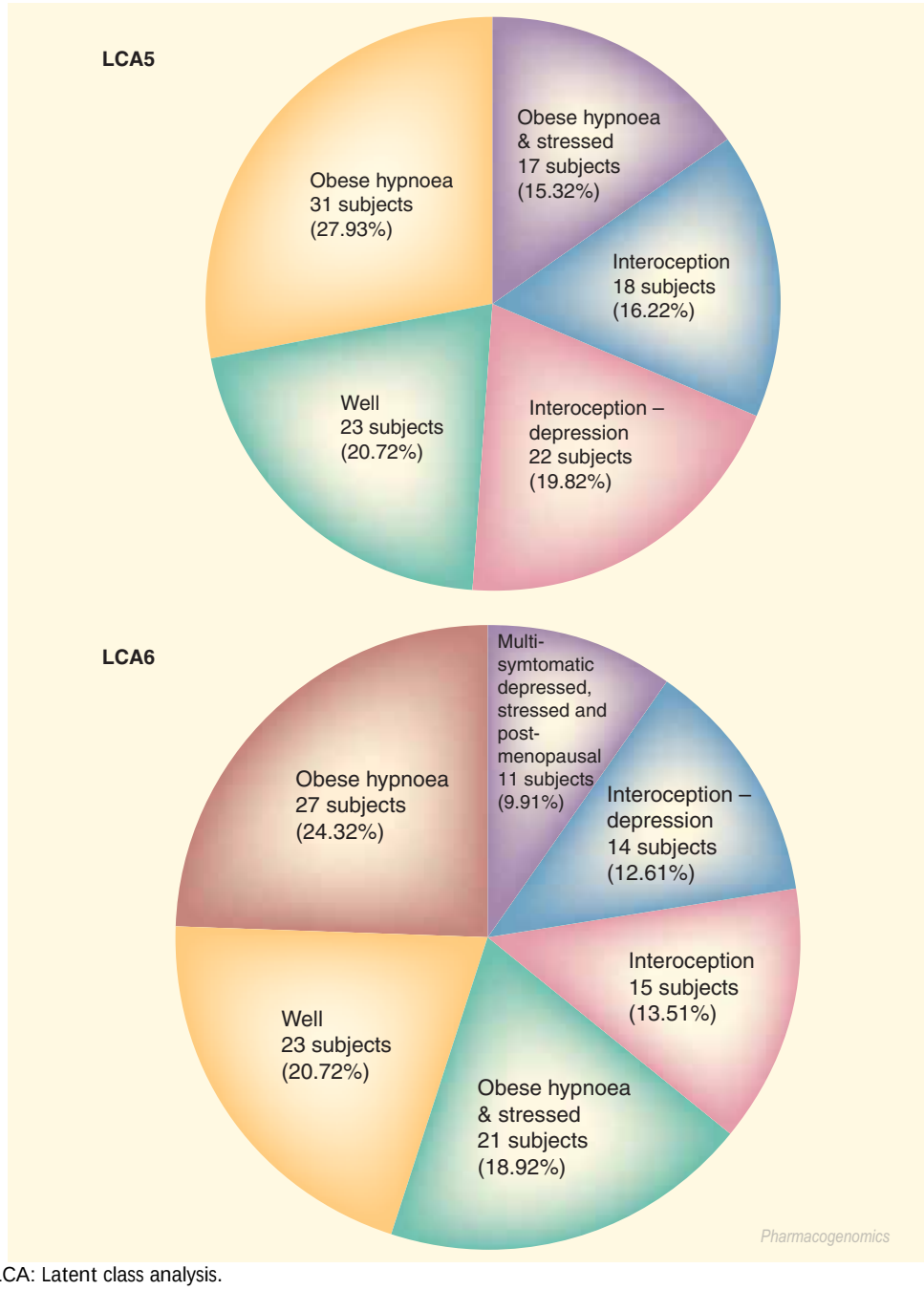
### *Data analysis*

#### *Unsupervised dimensionality reduction*

Each subject can be viewed as if it is described by a list of 15,315 variables, each variable being the expression level of a particular gene. Consequently, the 15,315-by-111 raw data matrix can be viewed as a collection of 111 subjects in the 15,315-dimensional gene space. In such situations, when the number of variables exceeds the number of subjects the correlation structure cannot be fully revealed. Indeed, it is well known that $N$ points can always be accurately embedded in $(N-1)$-dimensional space, suggesting that the 111 subjects can be described by a mere 110 variables (each is a linear combination of the original 15,315 variables).

We have used PCA to find such an orthonormal set of new variables, which are nothing but the first 110 principal components (PCs). An additional advantage of using PCA to reduce dimensionality is that the new variables are sorted by their importance in explaining the variability in the data. Actually, it is a good practice

**Figure 1. Distribution of 111 female subjects in each class in the LCA5 and LCA6 schemes.**



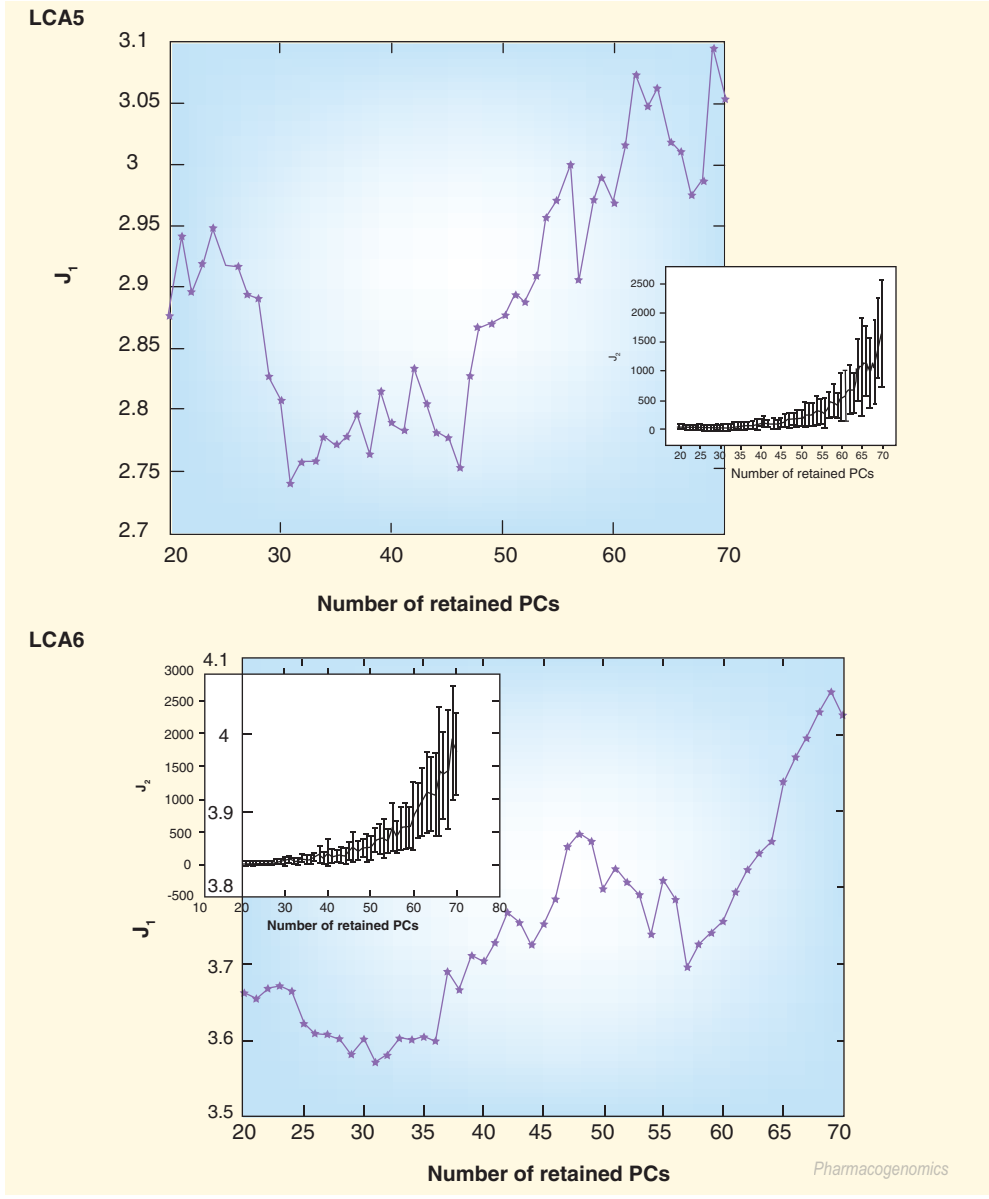

LCA: Latent class analysis.

to keep even lesser PCs, as removal of PCs with low eigenvalues removes noise while hardly affecting the structure of the data. Formally, let $X$ be the original 15,315-by-111 matrix of centered raw data, and let $U$ be the 15,315-by-$r$ matrix of the first $r$ PCs, $0<r\leq 110$. The projection of the data onto the PC space is $X' = U^T \cdot X$, which is of dimensions $r$-by-111. For the time being, $r$ is an unspecified parameter, but we shall later discuss how it can be determined in a data-driven fashion.

### *Supervised dimensionality reduction*

The $r$ PCs can be viewed as $r$ 'composite genes'. We can next ask which of these composite genes has a distinctly different expression pattern

**Figure 2. The $J_1$ measure for LCA5 (top) and LCA6 (bottom). The main figure depicts the change in $J_1$ as a function of the number of retained PCs. The small figure shows the $J_2$ values for the same data.**




$J_1$ measures the compactness of classes in the reduced dimensional space, whereas $J_2$ measures the amount of overfitting in the supervised dimensionality reduction.
LCA: Latent class analysis; PC: Principal component.

across the different classes. To this end, let the subjects be divided into $g$ classes, with the $i$th class containing $n_i$ subjects. Let $m_i$ and $S_i$ be the mean and the covariance matrix of the $i$th class, respectively. The average within-class covariance is defined as: $S_W = (1/n)\sum_{i=1}^{g} n_i S_i$

where $n = \sum_{i=1}^{g} n_i$ is the total number of subjects [22]. The between-class covariance is defined as:

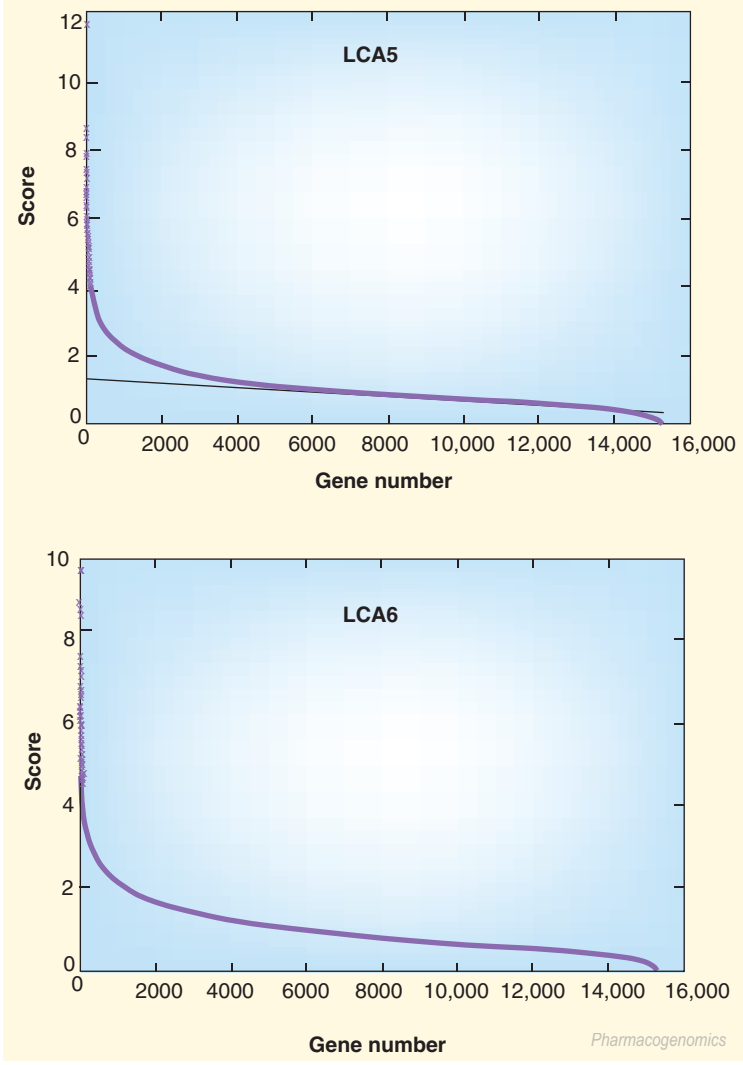
$$S_B = (1/n)\sum_{i=1}^{g} n_i (m_i - m)(m_i - m)^T$$

where $m$ is the mean of the entire data [22]. Using the well-studied Fisher quotient [22], we further reduce the dimensionality of the data by means of a linear transformation matrix $V$ of dimensions $r$-by-($g$-1). This matrix is the solution of:

$$V = \arg\max_v \frac{tr(v^T S_B v)}{tr(v^T S_W v)}$$

subject to the constraint $v^T S_W v = I$. The rational in this maximization is to find those variables for

**Figure 3. The sorted list of gene scores for LCA5 and LCA6.**




LCA: Latent class analysis.

which the classes are as compact as possible, while at the same time the classes are as separated from each other as possible. The matrix $V$ can be viewed as a set of $g-1$ new composite variables, called discriminant vectors (DVs). Just like the PCs, the DVs are sorted by their importance in separating the different classes. The transformed data is now $X'' = V^T X'$, which is a $(g-1)$-by-111 matrix. Note that in our case $g-1$ is 4 or 5, thus giving rise to an enormous reduction in dimensionality, from an original list of 15,315 genes to a compact set of 4–5 composite genes.

### *Gene scores*

We would like to infer from the composition of the DVs and the PCs on the relative contribution of the original genes to the separation between the classes. This is achieved in two steps. First, we assign each PC a score. The score of the $j$th PC is:

$$t_j = \frac{1}{M}\sum_{k=1}^{g-1} \mu_k v_{jk}^2 .$$

Here, $\mu_k$ is the value of the Fisher quotient for the $k$th DV, $M=\Sigma_{k=1}^{g-1}\mu_k$ is the total sum of the Fisher quotient values, and $v_{jk}$ is the $jk$th element in the matrix $V$, corresponding to the contribution of the $j$th PC to the $k$th DV.

Next, the original genes are assigned with scores. The score of the $i$th gene is:

$$s_i = N \cdot \sum_{j=1}^{r} t_j u_{ij}^2$$

where N is the total number of genes (15,315) and $u_{ij}$ is the $ij$th element in the matrix $U$, corresponding to the contribution of the $i$th gene to the $j$th PC. Each score is positive, and the sum of all scores is $N$. Obviously, the higher the score of a gene, the higher its contribution to the separation between classes.

### *The optimal number of PCs*

It is not straightforward to devise a criterion for the optimal number of PCs to retain, $r$. The rational is that taking too many PCs (overestimating $r$) results in accounting for too much noise in the analysis. On the other hand, too few PCs (underestimating $r$) results in using only a portion of the relevant information, obtaining poor separation between the classes.

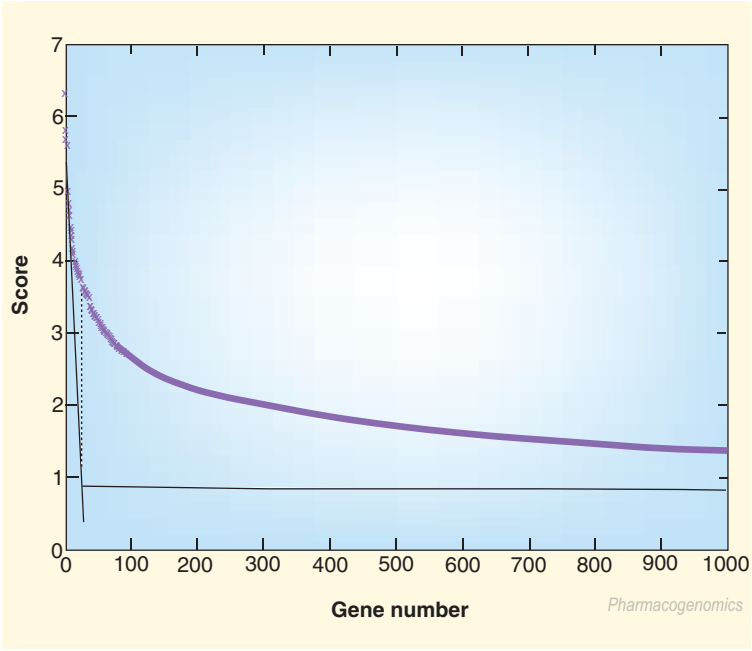
To formulate this intuition, the Mahalanobis distance [22] between each sample and all the classes was computed. Explicitly, the distance between the $i$th sample and the $k$th class is

$$d_{ik} = (x_i'' - \mu_k)^T \cdot S_W^{-1} \cdot (x_i'' - \mu_k)$$

where $x_i''$ is the $(g-1)$-dimensional representation of the $i$th subject. Notice that the average covariance matrix is taken as if it is the true covariance of each of the classes. For each sample, the minimal Mahalanobis distance (corresponding to its Mahalanobis distance from the closest class) is taken, and then averaged across all samples. For a given $r$, this results in a measure, denoted $J_1$, for the compactness of the classes in the $(g-1)$-dimensional space.

Another risk of keeping too many PCs is that of overfitting. To reduce this risk we define another measure, denoted $J_2$, for the amount of overfitting. To this end, we use the concept of cross-validation and assume that a certain fraction of the subjects, $f$, have unknown classification. The known samples are used to maximize the Fisher quotient, and then we compute the Mahalanobis distance between each unknown

**Figure 4. The intersection between the two asymptotes for the LCA6 classification scheme defines a short-list of size 26.**




LCA: Latent class analysis.

subject and all the classes. Analogously to $J_1$, $J_2$ is obtained by taking for each subject the minimum of these distances, and then averaging over all the subjects. The subjects that are assumed to be unknowns are chosen at random. To account for these random fluctuations, $J_2$ is computed several times for any given $r$. The values are used to determine the average value of $J_2$, as well its confidence interval.

## Results

The values of the $J_1$ measure for the LCA5 and LCA6 classification are shown in Figure 2. As expected, the curve is concave as too many PCs or too few PCs give rise to elevated $J_1$ values. We define the optimal number of PCs, $r$, as the minimum of this curve. Interestingly, $r = 31$ for both classification schemes. We also computed the $J_2$ measure (Figure 2) for the two classification schemes, using 20 repetitions for each value of $r$. Visual inspection of the result clearly reveals that at around $r = 31$ we are at reduced risk of overfitting.

Using $r = 31$, we have computed the score of each gene for either of the classification schemes (Figure 3; a complete list of all genes and scores is available upon request). The two score lists have a Spearman (rank) correlation of 0.96, suggesting a striking consistency between LCA5 and LCA6 solutions.

Taking advantage of the fact that the sorted scores are asymptotically linear toward the edges, we can define, for mere convenience, a short-list of discriminatory genes by intersecting the two asymptotes (Figure 4). For the LCA5 and LCA6, this gives a list of 26 and 32 genes, respectively (Table 1), with 19 genes in common between these solutions. These 19 common genes are to be taken as the best candidates for further investigation. Obviously, this is a rather arbitrary cut-off, and there is no limitation to the size of the list that one might want to investigate.

We tested the robustness of the choice of the number of PCs to retain, $r$, by repeating the entire analysis using different values of $r$. Analysis is robust with r values ranging from 25–35, as indicated by the high Spearman correlation (0.97–0.98) and by the high percentage (89%) of genes overlapping within the short list of genes (Table 2).

Further pairwise analysis was conducted using the set of discriminant genes for the LCA5 and LCA6 solutions (Table 3). This identified a total of 17 genes differentially expressed over twofold between the well subjects and subjects with symptoms in both LCA solutions. Among the differentially expressed genes on pairwise comparison, 13 were upregulated, and four were either up- or down-regulated depending on the comparison. Nine genes were differentially expressed in one or more pairwise comparison of both LCA5 and LCA6 class solutions, whereas seven genes were specific to LCA6 solution. Regardless of the LCA classification scheme, solute carrier family 1, member 6 (*SLC1A6*) was upregulated (three- to 30-fold) in all fatigued subjects compared with well subjects. Zinc finger protein 350 (*ZNF350*; three- to 92-fold) and F-box protein 7 (*FBX07*; four- to 28-fold) were also upregulated between all fatigued groups and well subjects except for the fatigued subjects in Class 4 of LCA6 solution. On the other hand, vacuole 14 protein homolog (*VAC14*) was down regulated (25- to 50-fold) in all fatigued subjects except for fatigued subjects in Class 4 of LCA6 solution. Class 4 in both LCA classification schemes was also unique with respect to the upregulation of patched homolog 2 (*PTCH2*; six- to 55-fold). Several of the differentially expressed genes between fatigued and well subjects were common for Classes 1, 3 and 5 of LCA5 solution. A more or less similar trend was seen with LCA6 solution also, although more genes were retained as differentially expressed in the pairwise comparison as opposed to LCA5 solution.

**Table 1. Genes discriminating multiple classes of fatigued subjects in LCA solutions.**

| LCA5* | | LCA6 | |
|---|---|---|---|
| **GenBank accession number** | **Gene score** | **GenBank accession number** | **Gene score** |
| **U38996** | 11.66 | U38996 | 9.71 |
| **AF087651** | 8.63 | AL135786 | 8.91 |
| **NM_004423** | 8.38 | AF087651 | 8.71 |
| **AK027244** | 7.93 | BC009921 | 8.60 |
| **AL135786** | 7.84 | NM_004423 | 7.34 |
| **BC009921** | 7.39 | AF233225 | 7.23 |
| **AK027172** | 7.31 | AF225419 | 7.14 |
| **AF225419** | 6.92 | NM_018052 | 6.84 |
| **AK024660** | 6.79 | BC020172 | 6.70 |
| **AF233225** | 6.71 | AK027244 | 6.63 |
| **AB029013** | 6.67 | U59494 | 6.41 |
| BC021294 | 6.62 | NM_012421 | 6.38 |
| **AF031588** | 6.43 | NM_000570 | 6.33 |
| BC001673 | 6.34 | AK027172 | 6.29 |
| L15702 | 6.33 | AB029013 | 6.14 |
| AF213465 | 6.11 | NM_014337 | 6.10 |
| **NM_012421** | 6.11 | NM_006593 | 6.06 |
| BC012375 | 6.11 | AL031316 | 5.99 |
| AF152493 | 6.11 | AF032119 | 5.96 |
| AF490768 | 6.10 | AF031588 | 5.89 |
| **NM_018052** | 6.01 | AK024660 | 5.87 |
| AB023230 | 5.94 | NM_002502 | 5.84 |
| **NM_014337** | 5.92 | NM_004526 | 5.81 |
| **BC028721** | 5.86 | BC009891 | 5.74 |
| **NM_007028** | 5.76 | NM_007028 | 5.73 |
| AF090988 | 5.75 | BCO28721 | 5.58 |
| BC029579 | 5.73 | | |
| AK023347 | 5.67 | | |
| **BC020172** | 5.67 | | |
| **NM_000570** | 5.59 | | |
| NM_021833 | 5.53 | | |
| NM_032493 | 5.52 | | |

**Genes in bold in LCA5 solution are also present in LCA6 solution.*
*LCA: Latent class analysis.*

## Discussion

Unexplained chronic fatigue remains a significant public health problem because of potential heterogeneity in the syndrome and consequent complexities with its diagnosis. Recently, gene expression profiling studies using microarray and differential display polymerase chain reaction (PCR) technologies were undertaken to identify biomarkers of CFS with subjects identified based on the 1994 CDC case definition [8–10,23], being compared with healthy controls. A standard case–control, two-class study design was used in these studies reporting deregulation in multiple pathways involved in immune function, cell-cycle control, nervous system function and viral infection [9]. No single abnormal gene expression was replicated in any of these studies. This involvement of genes from different disparate pathways may have resulted from the heterogeneity of subjects with fatigue defined homogeneously in these previous studies. Unlike earlier reports, we studied gene expression changes in fatigued subjects who were classified empirically into four or five discrete subgroups by Vollmer-Conna and colleagues [12] by PCA and LCA.

**Table 2. Robustness of multiclass discriminant gene analysis for LCA6 solution.**

| Number of PCA retained | Spearman correlation† | Size of short list | Overlap between short list |
|---|---|---|---|
| 20 | 0.82 | 27 | 15 (57.7%) |
| 25 | 0.97 | 33 | 23 (88.5%) |
| 35 | 0.98 | 29 | 23 (88.5%) |
| 40 | 0.88 | 41 | 20 (76.9%) |

*†Spearman correlation is computed between the new scores and the original ones computed with r = 31.*
*LCA: Latent class analysis; PCA: Principal component analysis.*

In the era of large-scale microarray experiments, identifying biomarkers for diseases has become a target of intensive research. Normally, these biomarkers are those genes whose expression level show significantly different behavior in healthy subjects compared with ill ones. In this study, the computational problem is somewhat more involved, as we target also the problem of discriminating multiple putative conditions. We suggest a novel way to measure each gene's contribution to the discrimination between all the classes [16–19]. This approach adopts ideas from the field of linear dimensionality reduction, and uses two consecutive linear transformations of the data. This algorithm is fairly general, and can be applied to any microarray data linked to a multiclass classification scheme.

Based on our algorithm, we were able to focus on 32 genes discriminating between the LCA5 classes and on 26 genes discriminating between the LCA6 classes, with 19 genes common among a total of 39 distinct genes between the solutions.

**Table 3. Pairwise expression pattern of genes discriminating LCA5 and LCA6 solutions of unexplained chronic fatigue.**

| GenBank accession number | Gene symbol | Expression pattern‡ | Fold difference* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LCA5 | | | | LCA6 | | | | |
| | | | Class 1 | Class 3 | Class 4 | Class 5 | Class 1 | Class 3 | Class 4 | Class 5 | Class 6 |
| NM_018052 | *VAC14* | ↓↑ | 0.02 | 0.03 | | 0.03 | 0.04 | 0.02 | 2.43 | 0.04 | 0.02 |
| NM_000570 | *FCGR3A* | ↑↓ | 2.94 | 2.43 | | 2.90 | 3.04 | 2.90 | 0.03 | 2.25 | 2.21 |
| NM_007028 | *TRIM31* | ↑ | | 4.50 | | 2.18 | | 2.18 | | 3.04 | 4.75 |
| NM_012421 | *RLF* | ↑ | | 4.81 | | 2.32 | | 2.32 | | 2.11 | 4.54 |
| L15702 | *BF* | ↑ | 5.12 | 5.28 | 7.93 | 9.10 | | | | | |
| BC028721 | *SLC1A6* | ↑ | 9.55 | 14.41 | 29.46 | 30.68 | 9.55 | 28.43 | 20.78 | 29.93 | 3.69 |
| AF233225 | *FBX07* | ↑ | 19.55 | 26.48 | 4.16 | 28.35 | 19.55 | 28.35 | | 26.80 | 27.75 |
| BC009921 | *ZNF350* | ↑ | 52.26 | 74.23 | 3.35 | 91.59 | 52.26 | 83.02 | | 64.83 | 82.48 |
| AB029013 | *WHSC1* | ↑↓ | | | 2.32 | 3.30 | | 2.82 | 2.28 | 2.44 | 0.06 |
| AF087651 | *PTCH2* | ↑ | | | 6.74 | | | | 55.96 | | |
| NM_006593 | *TBR1* | ↑↓ | | | | | 3.32 | 3.31 | 0.05 | 3.76 | 4.78 |
| U59494 | *THPO* | ↑ | | | | | 4.83 | 5.73 | 5.19 | 5.07 | |
| AF032119 | *CASK* | ↑ | | | | | 32.97 | 32.47 | | 21.57 | 42.03 |
| AF225419 | *HSCARG* | ↑ | | | | | | 2.17 | | | |
| BC009891 | *TCL1A* | ↑ | | | | | | | 7.47 | | |
| AL031316 | Not available | ↑ | | | | | | | | 2.03 | 2.94 |
| NM_002502 | *NFKB2* | ↑ | | | | | | | | 2.07 | 2.34 |

**Fold-differences (≥ twofold) with respect to expression in Class 2 of each solution. Fold differences were determined from median normalized expression values for each class. ‡↑ indicates up regulation and ↓ indicates down regulation.*
*BF: Human complement factor B; CASK: Calcium/calmodulin-dependent serine protein kinase; FBX07: F-box protein 7; FCGR3A: Fc fragment of IgG, low affinity IIIa, receptor (CD16a); HSCARG: Hypothetical protein LOC57407; NFKB2: Nuclear Factor κB2; PTCH2: Patched homolog 2; RLF: Rearranged L-myc fusion; SLC1A6: Solute carrier family 1 (high affinity aspartate/glutamate transporter), member 6; TBR1: T-box, brain, 1; TCL1A: T-cell leukemia/lymphoma 1A; THPO: Thrombopoietin; TRIM31: Tripartite motif-containing 31; VAC14: Vacuole 14 protein homolog; WHSC1: Wolf-Hirschhorn syndrome candidate 1; ZNF350: Zinc finger protein 350.*

**Table 4. Functions of genes with differential expression in pairwise comparisons of various classes of unexplained chronic fatigue subjects with healthy subjects in LCA solution.**

| GenBank accession number | Gene symbol | Gene name | Function |
|---|---|---|---|
| NM_018052 | *VAC14* | Vac14 homolog | Signal transduction, receptor activity |
| NM_000570 | *FCGR3A* | Fc fragment of IgG, low affinity IIIa, receptor | Immune response, receptor activity |
| NM_007028 | *TRIM31* | Tripartite motif-containing 31 | Protein ubiquitination |
| NM_012421 | *RLF* | Rearranged L-myc fusion sequence | Zinc ion binding, catalytic activity |
| L15702 | *BF* | Human compliment factor B | A component of alternative pathway of compliment activation. Immune function |
| BC028721 | *SLC1A6* | Solute carrier family 1 (high affinity aspartate/glutamate transporter) member 6 | Glutamate/aspartate transporter, synaptic transmission |
| AF233225 | *FBX07* | F-box protein 7 | Ubiquitin-dependent protein catabolism |
| BC009921 | *ZNF350* | Zinc finger protein 350 | Regulation of transcription |
| AB029013 | *WHSC1* | Wolf-Hirschhorn syndrome candidate | DNA binding, protein ubiquitination. |
| AF087651 | *PTCH2* | Patched homolog 2 (*Drosophila*) | Member of hedgehog signaling pathway, implicated in tumorigenesis. |
| NM_006593 | *TBR1* | T-box, brain, 1 | Transcription factor, brain development |
| U59494 | *THPO* | Thrombopoietin | Humoral growth factor for megakaryocyte proliferation and maturation |
| AF032119 | *CASK* | Calcium/calmodulin-dependent serine protein kinase | Cell adhesion, protein tyrosine kinase activity |
| AF225419 | *HSCARG* | HSCARG Protein | Transcriptional repressor activity |
| BC009891 | *TCL1A* | T-cell leukemia/lymphoma 1A | A protooncogene of T-cell malignancies |
| AL031316 | Not available | Not available | Not available |
| NM_002502 | *NFKB2* | Nuclear factor of κ light polypeptide gene enhancer in B-cells | Transcription factor, linked to inflammatory conditions of several diseases |

*HSCARG: Hypothetical protein LOC57407; IgG: Immunoglobulin G; LCA: Latent class analysis.*

Pairwise comparison within each LCA classification scheme identified a total of 17 genes differentially expressed between various classes of fatigued subjects and healthy Class 2 subjects. Broadly, these genes are implicated with immune function (Fc fragment of immunoglobulin G, low affinity IIIa, receptor [*FCGR3A*] and human complement factor B [*BF*]), transcription (*ZNF350*, T-box, brain, 1 [*TBR1*], hypothetical protein LOC57407 [*HSCARG*], nuclear factor κB2 [*NFKB2*]), ubiquitination (tripartite motif-containing 31 [*TRIM31*], rearranged L-myc fusion [*RLF*], *FBX07*, and Wolf-Hirschhorn syndrome candidate 1 [*WHSC1*]), signal transduction (*VAC14* and calcium/calmodulin-dependent serine protein kinase [*CASK*]) and transporter (*SLC1A6*) (Table 4). Among these differentially expressed genes, *SLC1A6*, *ZNF350* and *FBX07* are particularly interesting, since they were upregulated in all, or most, classes of fatigued subjects

compared with healthy subjects. *ZNF350* is reported to be associated with breast cancer 1 (*BRCA1*) for its transcriptional regulation of DNA damage-inducible genes, such as *GADD45* that functions in cell-cycle arrest [24]. *SLC1A6* (high affinity aspartate/glutamate transporter) is a member of the excitotory amino acid transporter of the CNS. These transporters maintain extracellular glutamate concentrations below excitotoxic levels, and limit the activation of glutamate receptors [25]. *FBX07* is a component of E3 ubiquitin protein ligases which function in phosphorylation-dependent ubiquitination [26]. *VAC14* (Vac14 homolog, *Saccharomyces cerevisiae*) was downregulated in most of the fatigued subjects, except those in Class 4 of the LCA6 solution in which it showed marginal upregulation. Reduced levels of *VAC14*, a novel positive regulator of PIKfyve, may render cells susceptible to developing cytoplasmic vacuoles [27], a hypothesis that needs further testing to detect abnormal cell morphology in CFS subjects.

Although a few genes clearly distinguished healthy subjects from all fatigued subjects, several of the differentially expressed genes were common for Classes 1, 3 and 5 of LCA5 solution with similar trend with LCA6 solution. However, Class 4 subjects, characterized as interoceptive, and thus likely to be associated with hypersensitivity in CNS processes, appeared to stand out from the rest of the fatiguing subjects in this pairwise comparison. *PTCH2*, a member of the hedgehog signaling pathway and implicated in tumorigenesis [28], and *TCL1A*, protooncogene of T-cell malignancies [29], were upregulated in Class 4 interoceptive subjects only in both LCA solutions. The functional role of patched homologs in the adult brain remains to be elucidated, but it is interesting to note its upregulation in interoceptive subjects considering the role of hedgehog signaling in establishing morphogenetic gradients during brain growth [30]. Fatigued subjects classified as interceptive were also different in terms of downregulation of *TBR1*, and for lack of differential expression of *CASK*, while these genes were upregulated in subjects belonging to other fatigued classes in LCA6 solution. *TBR1* is a neuron-specific T-box transcription factor, and in complex with *CASK*, targets the expression of several genes including *NMDAR* subunits 2b and subunit 1 [31]. Deregulation of *TBR-1/CASK* complex can thus adversely affect the neuronal activity and function in fatigued subjects, and possibly differently in interoceptive subjects.

Complement activation products as markers of CFS were proposed based on the observation of increased complement split product C4a in CFS subjects in response to exercise challenge [32]. We observed upregulation of *BF* in all fatigued subjects in the LCA5 solution, but differential expression of this gene was not apparent in the LCA6 solution. Likewise, differential expression of a number of genes (*TBR1*, thrombopoietin [*THPO*], *CASK*, *HSCARG*, *TCL1A*, and *NFKB2*) was unique to the LCA6 solution only. This suggests distinct profiles of gene expression associated with different LCA solutions, although the underlying reasons for this difference are not obvious from these results.

Many of the genes with interesting differences between LCA classes have no known association with any illnesses with or without fatigue as a symptom. The expression profile of these genes requires further experimental validation in terms of their specific association to one or more classes of unexplained chronic fatigue subjects derived from different populations.

## Outlook

In 5 or 10 years, we will have replicated or refined the heterogeneity of CFS using gene expression as an external validator. We will also have checked each of the genes in the short-list individually for their role in CFS pathophysiology, thus being able to validate the analytical technique that has been described here for the identification of these genes. Moreover, the same analytical technique will be used on different genomic datasets of multiple classes, further testing its range of applicability.

### Disclaimer

*The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency.*

### Acknowledgments

*The authors would like to thank Supraja Narasimhan for her bioinformatics support.*

**Highlights**

- Given a multiclass classification scheme and microarray data, the genes were quantitatively ranked by their contribution to the separation between the classes.
- Two consecutive linear transformations were used. First, principal components analysis (PCA) was applied and the last principal components (PCs) were removed to reduce noise. Second, the Fisher quotient was applied to obtain the most discriminatory combinations of genes.
- A technique was developed to estimate the number of PCs that should be removed.
- Gene expression supports the differentiation of healthy subjects from unexplained chronic fatigue.
- Gene expression provides partial support for discrimination between chronic unexplained fatigued subjects, which supports the use of syndrome heterogeneity in experimental designs of chronic fatigue syndrome (CFS) etiology and pathophysiological studies.
- Genes identified in this study as discriminating multiple classes of chronic fatigue are promising candidates for further studies on CFS biomarkers and pathophysiology.

## Bibliography


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

1. Lloyd AR: Chronic fatigue and chronic fatigue syndrome: shifting boundaries and attributions. *Am. J. Med.* 105, S7–S10 (1998).
2. Wessely SC, Hotopf M, Sharpe M: *Chronic Fatigue and Its Syndromes*. Oxford University Press, Oxford, UK (1998).
3. Afari N, Buchwald D: Chronic fatigue syndrome: a review. *Am. J. Psychiatry* 160, 221–236 (2003).
4. Hickie I, Lloyd A, Hadzi-Pavlovic D, Parker G, Bird K, Wakefield D: Can the chronic fatigue syndrome be defined by distinct clinical features? *Psychol. Med.* 25, 925–935 (1995).

• First empirical attempt to delineate chronic fatigue syndrome (CFS).

5. Wilson A, Hickie I, Hadzi-Pavlovic D *et al.*: What is chronic fatigue syndrome? Heterogeneity within an international multicentre study. *Aust. NZ J. Psychiatry* 35, 520–527 (2001).

• International replication of delineation of CFS.

6. Sullivan PF, Smith W, Buchwald D: Latent class analysis of symptoms associated with chronic fatigue syndrome and fibromyalgia. *Psychol. Med.* 32, 881–888 (2002).
7. Jason LA, Taylor RR, Kennedy CL *et al.*: Chronic fatigue syndrome: symptom subtypes in a community based sample. *Women Health* 37, 1–13 (2003).
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).

• First study of genomic expression of CFS.

9. 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).

• Genomic expression of CFS.

10. Powell R, Ren J, Lewith G *et al.*: Identification of novel expressed sequences, upregulated in the leucocytes of chronic fatigue syndrome patients. *Clin. Exp. Allerg.* 33, 1450–1456 (2003).
11. Darbishire L, Ridsdale L, Seed PT: Distinguishing patients with chronic fatigue from those with chronic fatigue syndrome: a diagnostic study in UK primary care. *Br. J. Gen. Pract.* 53, 441–445 (2003).
12. Vollmer-Conna U, Aslakson E, White PD: An empirical delineation of the heterogeneity of chronic unexplained fatigue. *Pharmacogenomics* 7(3), 355–364 (2006).

• First study of the heterogeneity of endophenotypes of CFS upon which this paper is based.

13. Liu H, Motoda H: *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA (1998).
14. Ross DT, Scherf U, Eisen MB *et al.*: Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* 24, 208–209 (2000).
15. Tan Y, Shi L, Hussain SM *et al.*: Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. *Bioinformatics* 22, 77–87 (2006).
16. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D: Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 21, 3896–3904 (2005).
17. Yeung KY, Bumgarner RE, Raftery AE: Bayesian model averaging: development of an improved multiclass, gene selection and classification tool for microarray data. *Bioinformatics* 21, 2394–2402 (2005).
18. Tsai CA, Lee TC, Ho IC, Yang UC, Chen CH, Chen JJ: Multiclass clustering and prediction in the analysis of microarray data. *Math Biosci.* 193, 79–100 (2005).
19. Tan Y, Shi L, Tong W, Wang C: Multiclass cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Res.* 33, 56–65 (2005).
20. 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).

• Basic description of the sample used in this study.

21. Whistler T, Taylor R, Craddock RC, Broderick G, Klimas N, Unger ER: Gene expression correlates of unexplained fatigue. *Pharmacogenomics* 7(3), 395–405 (2006).

• Describe quality control and normalization of microarray data used in this report.

22. Webb A: *Statistical Pattern Recognition*. 2nd edition. John Wiley & Sons, Chichester, UK (2002).
23. Steinau M, Unger ER, Vernon SD, Jones JF, Rajeevan MS: Differential-display PCR of peripheral blood for biomarker discovery in chronic fatigue syndrome. *J. Mol. Med.* 82, 750–755 (2004).
24. Tan W, Zheng L, Lee WH, Boyer TG: Functional dissection of transcription factor ZBRK1 reveals zinc fingers with dual roles in DNA-binding and BRCA1-dependent transcriptional repression. *J. Biol. Chem.* 279, 6576–6587 (2004).
25. Fairman WA, Vandenberg RJ, Arriza JL, Kavanaugh MP, Amara SG: An excitatory amino-acid transporter with properties of a ligand-gated chloride channel. *Nature* 375, 599–603 (1995).
26. Hsu JM, Lee YC, Yu CT, Huang CY: Fbx7 functions in the SCF complex regulating Cdk1-cyclin B-phosphorylated hepatoma

upregulated protein (HURP) proteolysis by a proline-rich region. *J. Biol. Chem.* 279, 32592–32602 (2004).

27. Sbrissa D, Ikonomov OC, Strakova J *et al.*: A mammalian ortholog of *Saccharomyces cerevisiae* Vac14 that associates with and upregulates PIKfyve phosphoinositide 5-kinase activity. *Mol. Cell. Biol.* 24, 10437–10447 (2004).
28. Rahnama F, Toftgard R, Zaphiropoulos PG: Distinct roles of *PTCH2* splice variants in Hedgehog signaling. *Biochem. J.* 378, 325–334 (2004).
29. Pekarsky Y, Zanesi N, Aqeilan R, Croce CM: Tcl1 as a model for lymphomagenesis. *Hematol. Oncol. Clin. North. Am.* 18, 863–879 (2004)
30. Tannahill D, Harris LW, Keynes R: Role of morphogens in brain growth. *J. Neurobiol.* 64, 367–375 (2005)
31. Wang TF, Ding CN, Wang GS *et al.*: Identification of Tbr-1/CASK complex target genes in neurons. *J. Neurochem.* 91, 1483–1492 (2004).
32. Sorensen B, Streib JE, Strand M *et al.*: Complement activation in a model of chronic fatigue syndrome. *J. Allergy. Clin. Immunol*. 112, 397–403 (2003).

•• First finding of immune abnormalities related to exercise in CFS.

Website

101. Reeves WC, Lloyd A, Vernon SD *et al.*: Identification of ambiguities in the 1994 chronic fatigue syndrome research case definition and recommendations for resolution. *BMC Health Serv. Res.* 3, 25 (2003). http://www.biomedcentral.com/1472–6963/3/25.

• Attempt to better define CFS for research purposes.