



For reprint orders, please contact:
reprints@futuremedicine.com

Gene expression correlates of unexplained fatigue

**Toni Whistler^{1†},
Renee Taylor²,
R Cameron Craddock¹,
Gordon Broderick³,
Nancy Klimas⁴ &
Elizabeth R Unger¹**

[†]Author for correspondence
¹Centers for Disease Control
and Prevention,

Viral Exanthems and
Herpesvirus Branch,
Atlanta, GA, 30333, USA
Tel.: +1 404 639 1305;
Fax: +1 404 639 3540;
E-mail: taw6@cdc.gov

²University of Illinois at
Chicago,
Department of Occupational
Therapy, Chicago,
IL, 60612, USA
Tel.: +1 312 996 3412;
Fax: +1 312 413 0256;

E-mail: rtaylor@uic.edu
³University 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

⁴Miami Veterans Affairs
Medical Center, Miami,
FL, 33125, USA
Tel.: +1 305 324 3267;
Fax: +1 305 324 3139;
E-mail: Nancy.klimas@
va.gov

Quantitative trait analysis (QTA) can be used to test whether the expression of a particular gene significantly correlates with some ordinal variable. To limit the number of false discoveries in the gene list, a multivariate permutation test can also be performed. The purpose of this study is to identify peripheral blood gene expression correlates of fatigue using quantitative trait analysis on gene expression data from 20,000 genes and fatigue traits measured using the multidimensional fatigue inventory (MFI). A total of 839 genes were statistically associated with fatigue measures. These mapped to biological pathways such as oxidative phosphorylation, gluconeogenesis, lipid metabolism, and several signal transduction pathways. However, more than 50% are not functionally annotated or associated with identified pathways. There is some overlap with genes implicated in other studies using differential gene expression. However, QTA allows detection of alterations that may not reach statistical significance in class comparison analyses, but which could contribute to disease pathophysiology. This study supports the use of phenotypic measures of chronic fatigue syndrome (CFS) and QTA as important for additional studies of this complex illness. Gene expression correlates of other phenotypic measures in the CFS Computational Challenge (C3) data set could be useful. Future studies of CFS should include as many precise measures of disease phenotype as is practical.

Differential gene expression analysis is one of the most widely used applications of microarray technology, whether it be for examining gene expression differences between predefined classes (class comparisons) or the classification of samples based on gene expression patterns (class predictions). An alternative approach to the data analysis of genome-wide expression data integrates measures of physiological traits with gene expression to identify genes and biological processes important in the resulting phenotype. The quantitative trait analysis (QTA) tool within Biometric Research Branch (BRB) ArrayTools [1] uses the Spearman correlation coefficient to test whether the expression of a particular gene correlates with a quantified trait. The program uses a multivariate permutation probability to control the number of false discoveries.

For complex diseases that are defined by case definition, rather than a diagnostic laboratory test, correlation of genetic changes with phenotypic measures of characteristics of the illness (i.e., endophenotypes) has been used to avoid errors in case ascertainment [2]. The approach may also be useful for gene expression data. The study of chronic fatigue syndrome (CFS) may benefit from endophenotype analysis, as there are ambiguities in the application of the case definition [3], in addition to uncertainties about different disease subgroups.

An ideal endophenotype should be reliably measurable both over time and by different observers. As fatigue is an internal state, measures rely on self-report and may therefore be limited in their precision. However, the multidimensional fatigue inventory (MFI) developed by Smet and colleagues [4] does evaluate multiple dimensions of fatigue. As unexplained fatigue is the central feature of CFS, we chose to use the CFS Computational Challenge (C3) data (Vernon and Reeves [5]) to explore the gene expression correlates of fatigue using QTA. Fatigue is a common symptom in many illnesses as well as in the general population. We started our analysis with this trait, despite the fact that fatigue measures are imprecise, as unexplained fatigue is the central feature of CFS. The C3 data includes the MFI to measure fatigue.

We found that QTA identifies 839 of the 19,760 microarray measured genes as showing a significant correlation with one or more MFI measures of fatigue. The fatigue-associated gene set implicates many basic cellular pathways and includes novel genes that are not functionally annotated or associated with identified pathways. There is some overlap with genes implicated in other studies using differential gene expression. However, QTA allows detection of alterations that may not reach statistical significance in class comparison

Keywords: CFS, fatigue, gene expression, MFI, quantitative trait analysis

future
medicine

analyses, but which could contribute to disease pathophysiology. This study supports the use of phenotypic measures of CFS and QTA as important for additional studies of this complex illness. Gene expression correlates of other phenotypic measures in the C3 data set could be useful. Future studies of CFS should include as many precise measures of disease phenotype as practical.

Methods

Data were derived from the population-based study of CFS described in the introductory paper of this issue [5] that was made available as the CFS computational challenge. 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 [6], 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 fatigue. Thus, our analysis focused on data from 112 female subjects. Using disease classification based on the CFS research case definition as measured by scores on the symptom inventory, MFI and Short Form (SF)-36 instruments [7], our study population included 40 CFS sufferers, 37 non-fatigued (NF) controls, and 35 individuals with unexplained fatigue with symptoms or severity short of the case definition (insufficient symptom fatigue [ISF]). One additional subject was excluded due to inadequate microarray data (see below); however, all 111 subjects with complete data were included in the QTA.

Quantitative measure of fatigue

We used the MFI [4], a 20-item self-report instrument that addresses fatigue as a multidimensional construct, covering the dimensions general fatigue, physical fatigue, mental fatigue, reduced motivation and reduced activity. There are four items to each dimension measured on a Likert scale of 1 to 5 (higher score indicates greater fatigue), resulting in scores for each dimension that range from 4 (least) to 20 (most). The MFI shows good internal consistency, inter- and intra-rater reliability [8]. This instrument was tested for its psychometric properties and found to have good internal consistency and construct validity in samples with CFS [9].

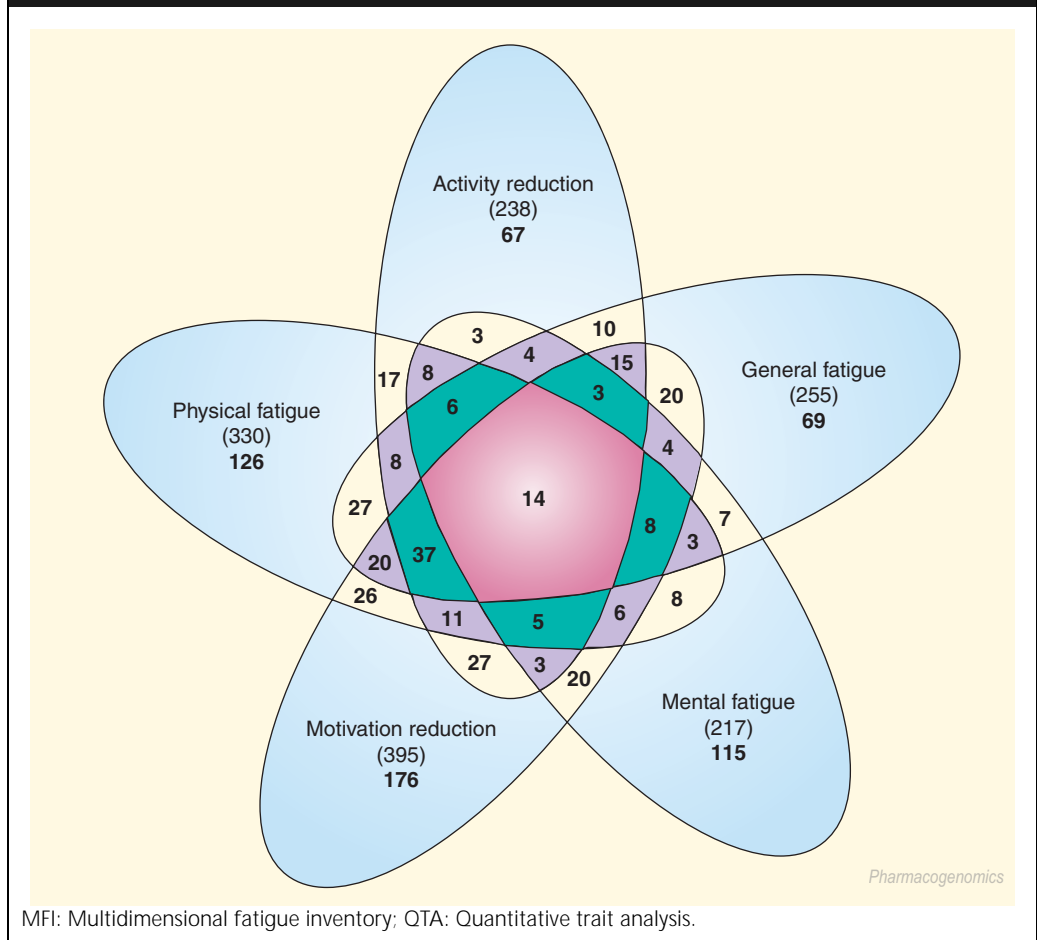
Microarray data

The collated, background-subtracted net intensity values for microarrays in the C3 data set were processed prior to use in QTA. The data were normalized with multiplicative scatter correction [10] followed by quantile normalization [11], 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. 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 gene features from the entire data set were filtered to remove those with little or no variability by an F-test. Technical error was estimated for each feature as the pooled variance between replicates, this value was used as the denominator of the F ratio. The numerator or the signal variance was calculated as the variance of a feature across all 111 microarrays. The F test was calculated for each feature and the resulting p-values were corrected for multiple testing using the Benjamini-Hochberg criterion [12] with a false discovery rate of 0.001. Features with an F test result that did not meet this criterion were removed from further analysis. This reduced the number of features from 19,760 to 15,136.

Quantitative trait analysis

We correlated the normalized log-transformed signal intensities of the 15,136 genes with the subject's MFI scores for general fatigue, physical fatigue, mental fatigue, motivation reduction and activity reduction using quantitative trait analysis performed with BRB Arraytools v3.30 beta_3a [1]. We computed a statistical significance level for each gene based on testing the hypothesis that the Spearman's correlation between gene expression and the fatigue parameter was zero. The Spearman coefficient was chosen over the Pearson correlation as the fatigue measures were ordinal. These p-values were then used in a multivariate permutation test [13,14] in which the fatigue scores were randomly permuted among arrays. We used the multivariate permutation test to provide 80% confidence that

Figure 1. Venn diagram showing relationship of the 873 microarray features associated with one or more MFI dimensions of fatigue by QTA.



the false discovery rate was less than 10%. The false discovery rate is the proportion of the list of genes claimed to be differentially expressed that are false positives. The multivariate permutation test is nonparametric and does not require the assumption of Gaussian distributions. Those genes with Spearman's coefficients of $p < 0.01$ were used as correlates of fatigue.

Gene set annotations

To translate the lists of fatigue-associated genes into a functional profile from which we may be able to offer some insight into the cellular mechanisms relevant to fatigue we used two analytical tools:

- Gene Ontology Tree Machine (GOTM [14,101]). This web-based platform annotates the gene symbol data using Gene Ontology (GO) hierarchies [102]. A hypergeometric test is used to compare the distribution of selected genes to those annotated in the whole human genome to

indicate GO terms with enriched gene numbers, and thus biological functions of potential importance.

- Pathway Miner [103] maps Genbank accession numbers onto the metabolic, cellular and regulatory process pathways from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [104], BioCarta [105] and GenMAPP [106] resources.

Results

The QTA identified 873 gene features whose expression significantly correlated to scores in one or more of the five dimensions of the MFI (Figure 1). The Spearman coefficients ranged from 0.47 to -0.39. Nearly two-thirds of these gene features (553/873, 63.3%) were associated with a single fatigue dimension. The other third were highly shared between dimensions; there were 14 genes associated with all five MFI dimensions (Table 1).

Table 1. Fourteen genes identified by QTA to be common to all five subscales of fatigue as measured by MFI.

Accession number	UniGene cluster	Gene name	Gene symbol	Locus link ID
AC002550	Hs.460217	Hypothetical protein MGC35048		
	Hs.546412	Esophageal cancer associated protein		
AF026692	Hs.105700	Secreted frizzled-related protein 4	<i>SFRP4</i>	6424
AF208843	Hs.4859	Cyclin L1	<i>CCNL1</i>	57018
AF286904	Hs.23270	Enhancer of polycomb homolog 2 (<i>Drosophila</i>)	<i>EPC2</i>	26122
AK024391	Hs.246112	Activating signal cointegrator 1 complex subunit 3-like 1	<i>ASCC3L1</i>	23020
AK056531	Hs.213087	Odz, odd Oz/ten-m homolog 4 (<i>Drosophila</i>)	<i>ODZ4</i>	26011
AK091309	Hs.437409	Transmembrane protein 20	<i>TMEM20</i>	159371
AP001748	Hs.431043	PBX/knotted 1 homeobox 1	<i>PKNOX1</i>	78999
	Hs.365116	U2(RNU2) small nuclear RNA auxiliary factor 1	<i>U2AF1</i>	91369
	Hs.184085	Crystallin, α A	<i>CRYAA</i>	29075
BC000207	Hs.209979	Leucine rich repeat and fibronectin type III domain containing 4	<i>LRFN4</i>	126014
BC005853	Hs.463426	Ankyrin repeat domain 40	<i>MGC15396</i>	
NM_014162	Hs.439352	HSPC072 protein	<i>HSPC072</i>	7450
NM_130771	Hs.347655	Osteoclast-associated receptor	<i>OSCAR</i>	
U90142	Hs.519635	Butyrophilin, subfamily 2, member A1	<i>BTN2A1</i>	
X04385	Hs.440848	Von Willebrand factor	<i>VWF</i>	

ID: Identification; MFI: Multidimensional fatigue inventory; QTA: Quantitative trait analysis.

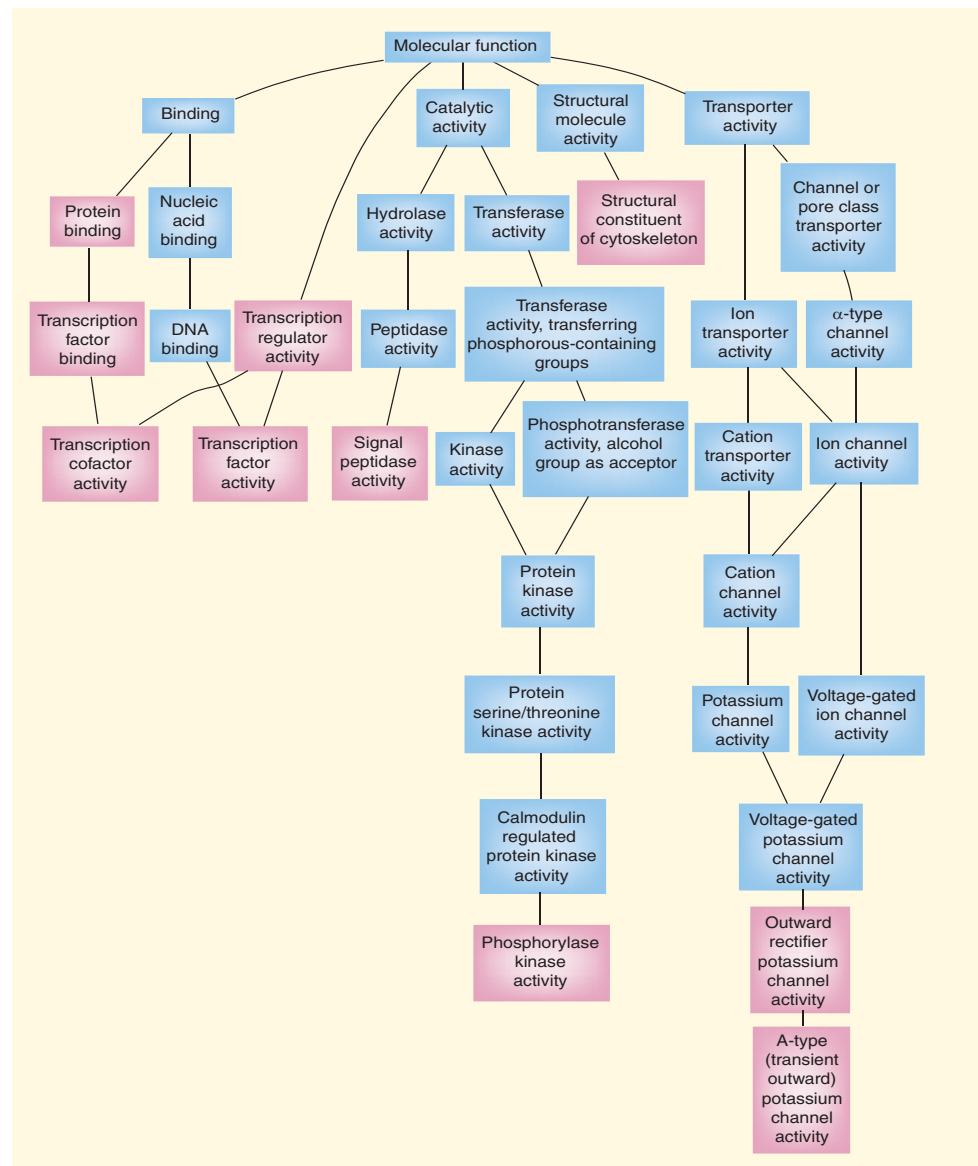
The 873 microarray features represented 839 nonoverlapping Genbank accession numbers and included 696 (83%) that mapped to a Human Genome Organization (HUGO) gene symbol, of which only 687 were unique. GO annotation and evaluation using GOTM is restricted to the 687 with gene symbols. GO annotation was found for 460, 447 and 404 gene symbols in the biological process, molecular function and cellular component categories, respectively (Figures 2, 3 and 4). The fatigue genes were significantly enriched in 44 biological process (Figure 2), ten molecular function (Figure 3) and 11 cellular component (Figure 4) GO categories. To verify that this enrichment did not result because of genes included in the microarray, we repeated the analysis against the microarray gene list using Web-based Gene Set Analysis Toolkit (WebGestalt) [107] and found the same ontologies were being enriched (results not shown).

The largest group of enriched biological process GO categories encompassed 89 gene symbols and was related to development, particularly morphogenesis and organ development, predominately muscle (Figure 2). The second group within enriched biological process GO encompassed 51 gene symbols and related

to cell organization and biogenesis. The third group related to metabolism and encompassed 272 gene symbols. Several metabolic processes are enriched, including glucan, polysaccharide, energy reserve and lipid metabolism (32 gene symbols). Within the molecular function, GO categories enriched in fatigue genes (Figure 3), the largest number of gene symbols (134) were categorized as protein binding, the majority in transcription factor binding (22) and transcriptional regulator activity (66). The enriched cellular component GO categories (Figure 4) include eukaryotic translation initiation factor 4F complex and several pertaining to endocytic vesicles and microtubules.

Pathway Miner was able to map slightly more than a quarter of fatigue-associated genes to pathways (215 of the 839 unique gene accessions; 25.6%; Table 2). Almost 40% of the fatigue genes annotated in Pathway Miner were mapped in metabolic pathways (82; 38.1%). Table 3 shows the pathways where four or more genes were mapped. The metabolism pathways identified by Pathway Miner show considerable overlap with those highlighted by GOTM. For example, the six enriched subcategories of GOTM polysaccharide metabolism (16 genes in

Figure 3. Directed acyclic graph view of enriched Gene Ontology categories as determined for the fatigue gene set using a hypergeometric test for each of the ontologies – molecular function.



Categories in the pink boxes represent significantly enriched gene numbers ($p < 0.01$), while those in blue boxes are nonenriched parents.

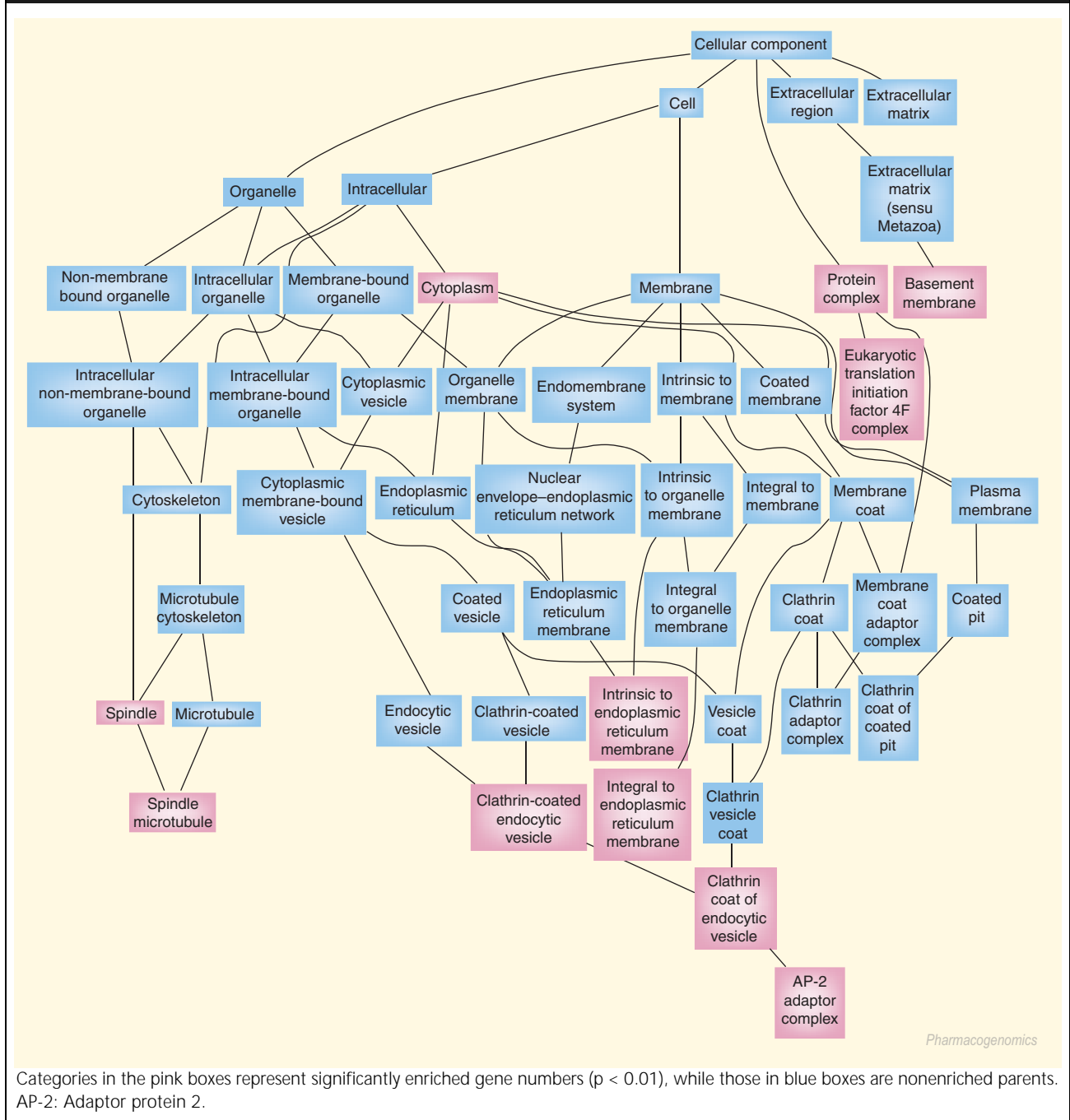
total) overlap with several Pathway Miner-identified metabolic pathways associated with glycolysis and gluconeogenesis, and oxidative phosphorylation (27 genes in total).

Pathway Miner annotation of the fatigue-associated genes implicated six different signaling pathways (Table 3):

- Janus kinase-signal transducer and activator of transcription (Jak-STAT) (seven genes, KEGG)
- Toll-like receptor (TLR) (four genes, KEGG)

- Target of rapamycin (mTOR) (five genes, BioCarta)
- G-protein signaling (four genes, GenMAPP); three other G-protein coupled receptor (GPCR)-associated pathways were identified in GenMAPP and BioCarta accounting for a total of 13 unique gene symbols
- Wnt signaling identified in both KEGG (ten genes) and GenMAPP (seven genes) sources with 12 unique genes

Figure 4. Directed acyclic graph view of enriched Gene Ontology categories as determined for the fatigue gene set using a hypergeometric test for each of the ontologies – cellular component.



- Three associated mitogen-activated protein kinase (MAPK) pathways including 16 non-overlapping gene products: MAPK signaling (BioCarta four genes; KEGG 13 genes), the extracellular signal-regulated kinases (ERK) 1 and 2 (BioCarta three genes), and the p38 MAPK signaling pathways (BioCarta three genes). A total of 16 nonoverlapping gene products were associated with MAPK signaling.

Discussion

Using QTA to correlate microarray results with MFI measures of fatigue, we identified 839 fatigue-associated genes (873 microarray features). These genes represent approximately 4% of the total number measured (873 of 19,760 array features). While a relatively high proportion of genes were associated with fatigue, the Spearman correlation coefficients were modest, 0.47 to -0.39.

Table 2. Summary of the Pathway Miner mapping of the 215 unique fatigue-associated genes.

Source, type	Number of pathways	Total
BioCarta		
C	190	98
M	4	4
GenMAPP		
C	19	39
M	10	22
KEGG		
C	24	73
M	61	72

*Pathways are designated as cellular & regulatory processes (C) or metabolism (M).
GenMAPP: Gene Map Annotator and Pathway Profiler; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Almost two-thirds of the genes were associated with only one fatigue dimension, but the remaining third were highly shared between dimensions. A total of 14 genes (Table 1) were associated with all five MFI dimensions of fatigue.

We used GOTM and Pathway Miner to map the fatigue-associated genes to functions and pathways. However, only 83% of the Genbank accession numbers (696/839) were linked with a gene product (HUGO gene symbol). Less than 70% of those with gene symbols had GO designations, and even fewer (26% of those submitted) could be assigned to known pathways. Overall, 50–75% of the genes correlating with MFI scores have unknown function and cannot be assigned to pathways. The lack of complete gene annotation limits the functional interpretation of the data. To maximize the number of annotated genes we submitted the entire fatigue-associated gene set to GOTM and Pathway Miner, and did not analyze the five dimensions of fatigue separately. All fatigue dimensions, except mental fatigue, shared more than half of their associated gene features with one or more other dimensions.

Despite limitations in annotation, the fatigue-associated genes mapped to several basic cellular functions including metabolism, transcriptional regulation and cell signaling pathways. Cell signaling pathways have extensive interaction and regulate nearly all biological processes, for example, metabolism, cell growth, proliferation and apoptosis, so it is not surprising to find that six major signaling pathways were associated with measures of fatigue. The GO annotation showed 66 gene products associated with transcription regulator activity, 22 of

which were transcription factor proteins. It is not surprising that polysaccharide metabolism (glycolysis and gluconeogenesis) and oxidative phosphorylation, the principal source of high-energy molecules in every cell, are pathways to which several fatigue-associated genes mapped. Clearly this analysis indicates many highly interconnected basic cell processes are altered or perturbed with fatigue. However, the pathogenesis of fatigue is not elucidated.

Prior studies of peripheral blood gene expression in CFS have relied on differential gene expression to determine genes associated with illness. The QTA approach avoids strict disease classification, and relies on correlations of genes with measures of disease-associated phenotypes. In this analysis we focused on fatigue. Interestingly, three fatigue-associated proteins in the mTOR pathway (eukaryote initiation factors, protein components essential for translation through a protein complex involved in mRNA cap recognition) were previously identified as differentially expressed in CFS [16,17]. Furthermore, nine of 16 genes validated as differentially expressed in CFS by Kaushik and colleagues [17] were associated with measures of fatigue by QTA. These included proteins in the adenosine triphosphate (ATP)-binding cassette subfamily, eukaryote transcription initiation factor complex, mitochondrial ribosomal proteins, anaphase promoting complex subunits and programmed cell death proteins. The overlap in genes associated with CFS by differential gene expression and with fatigue by QTA supports the utility of these complementary methods of analysis. In addition, it reinforces the central nature of fatigue to CFS. As QTA allows integration of other biological measures with gene expression, genes involved in specific phenotypes can be identified, extending the usefulness of microarray technology beyond class comparison.

There are, however, several limitations to this approach to data integration. QTA assumes that coordinated gene regulation is the same for all individuals in the population, independent of disease status, and that fatigue is a spectrum in the population. This may not be correct; there may be a disease-dependent association of gene expression with fatigue. This raises the question whether explained fatigue, for example that resulting from radio- or chemotherapy, or as a response to physical exercise, would have the same gene associations. Expression patterns common to fatigue in diverse disease associations may suggest novel approaches to therapeutic interventions. Further

Table 3. Results of mapping fatigue-associated genes to pathways using Pathway Miner.

Pathway name	No. genes in pathway	Type	Genes in the pathway
HIV-1 Nef: negative effector of Fas and TNF	5	C	<i>BIRC2, PSEN1, TRAF2, PTK2, CDC2L2</i>
mTOR signaling pathway	5	C	<i>EIF4G3, EIF4G2, EIF4E, RPS6, PTEN</i>
Aggrin in postsynaptic differentiation	6	C	<i>ITGB1, PTK2, LAMA4, ARHGEF6, DMD, RAPSIN</i>
Nuclear receptors in lipid metabolism and toxicity	4	C	<i>CYP4A11, ABCD2, CYP1A2, CYP2E1</i>
Mechanism of gene regulation by peroxisome proliferators via PPAR α	4	C	<i>STAT5B, NR2F1, EHHADH, SRA1</i>
Regulation of eIF4e and p70 S6 kinase	4	C	<i>EIF4G3, EIF4G2, EIF4E, PTEN</i>
β -arrestins in GPCR desensitization	4	C	<i>GNAS, ARRB1, DNMI1, AP2A1</i>
MAPKinase signaling pathway	4	C	<i>MAP3K13, TRAF2, IKBKB, MAP3K7</i>
Hs_GPCRs_class_A_rhodopsin-like	6	C	<i>NTSR1, FPRL1, GPR74, GPR21, IL8RA</i>
Hs_Wnt_signaling	7	C	<i>WNT5A, CCND3, WNT3, PPP2R5C, FZD3, PAFAH1B1, SFRP4</i>
Hs_glycogen_metabolism	6	M	<i>GYS1, PPP2R1B, PHKG2, PPP2R5C, GYG, PHKA1</i>
Hs_glycolysis_and_gluconeogenesis	4	M	<i>GAPD, PFKM, BMP4, G6PC</i>
Hs_peptide_GPCRs	4	C	<i>NTSR1, FPRL1, IL8RA</i>
Hs_G_protein_signaling	4	C	<i>PLCB3, PERQ1, GNAS, PDE4D</i>
Wnt signaling pathway	10	C	<i>WNT5A, PSEN1, PLCB3, CCND3, PPP2R1B, WNT3, FZD3, MAP3K7, SFRP4, PLCE1</i>
Lysine degradation	8	M	<i>DOT1L, HSD17B12, NAT6, SET -P7, EHHADH, BBOX1, GCDH, signal peptidase</i>
Tryptophan metabolism	8	M	<i>HRMT1L6, ASMT, HRMT1L1, EHHADH, CYP1A2, GCDH, CYP2E1</i>
MAPK signaling pathway	13	C	<i>DUSP14, DUSP6, HSPB1, CACNG6, FGF13, MAP3K13, TRAF2, MAPT, ARRB1, IKBKB, MAP3K7, YWHAZ, FLNB</i>
Glycerolipid metabolism	6	M	<i>HSD17B12, NAT6, AGPAT1, PAFAH1B1, YWHAZ, CHAT</i>
Complement and coagulation cascades	7	C	<i>MASP2, C6, IF, C1QA, VWF, SERPINA5, TFPI</i>
Fatty acid metabolism	7	M	<i>CYP4A11, EHHADH, CYP1A2, GCDH, CYP2E1, HADHB</i>
Cell cycle	7	C	<i>CDK4, CCNB3, CCND3, PLK1, MAD1L1, CCNE1, HDAC7A</i>
Jak-STAT signaling pathway	7	C	<i>STAT5B, CCND3, PIK3C3, IL2, CSF3R, IL6ST, PIAS2</i>
Starch and sucrose metabolism	5	M	<i>GYS1, GAK, G6PC, U5 snRNP, ERCC2</i>
Neurodegenerative disorders	5	C	<i>HD, PSEN1, GAPD, MAPT, APOE</i>
Benzoate degradation via CoA ligation	5	M	<i>HSD17B12, GAK, NAT6, EHHADH, GCDH</i>
Butanoate metabolism	5	M	<i>HMGCL, OXCT1, HSD17B12, NAT6, EHHADH</i>
Alzheimer's disease	5	C	<i>PSEN1, GAPD, MAPT, SERPINA3, APOE</i>
Pyrimidine metabolism	4	M	<i>AK3L1, POLR2C, UMPS, DCTD</i>
Tyrosine metabolism	4	M	<i>HRMT1L6, NAT6, HRMT1L1, TPO</i>
Galactose metabolism	4	M	<i>GALE, HSD17B12, PFKM, G6PC</i>
Toll-like receptor signaling pathway	4	C	<i>PIK3C3, IKBKB, MAP3K7, IFNAR2</i>
Cholera – infection	4	C	<i>PLCB3, GNAS, ATP6V1C1, ATP6V1D</i>
Oxidative phosphorylation	4	M	<i>COX7A2, ATP6V1C1, ATP6V1D</i>
Glycolysis/gluconeogenesis	4	M	<i>BPGM, GAPD, PFKM, G6PC</i>
Purine metabolism	5	M	<i>POLR2C, HPRT1, PDE4D, DGUOK, testicular soluble adenylyl cyclase</i>
Valine, leucine and isoleucine degradation	5	M	<i>EHHADH, OXCT1, HADHB, NAT6, HMGCL</i>
Apoptosis	4	C	<i>BIRC2, TRAF2, PIK3C3, TNFSF10</i>

Pathways with four or more genes are listed.

C: Cellular and regulatory processes; M: Metabolism.

studies of additional fatigued groups will be required to address this issue. Another limitation is the measure of fatigue itself. While MFI is one of the most widely used measures of fatigue in diverse clinical settings, it is a self-reported instrument. Errors of measurement cannot be determined, and fatigue remains difficult to rigorously and precisely measure.

Outlook

The identification of the function of all genes that contribute to specific biological processes and complex traits is one of the major challenges in the postgenomic era. We believe the next few years will see a vast improvement in the annotation of

databases that will include inferred and experimentally investigated gene functions, high-quality annotation of alternative splicing, pseudogenes and promoter regions. This will allow for very comprehensive understanding of the pathophysiology of disease. More of a systems biology approach to research will evolve, incorporating information from many sources. This will have a major impact on the way we discover and develop new medicines to treat chronic diseases.

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.

Highlights

- Using quantitative trait analysis (QTA), we integrated genome-wide expression data with measures of fatigue from the multidimensional fatigue inventory (a validated, self-report questionnaire) to explore the biological correlates of unexplained fatigue from a continuum of subjects with chronic fatigue syndrome (CFS) or CFS-like illness.
- Approximately 4% of the genes studied correlated with one or more dimensions of fatigue measured in the multidimensional fatigue inventory (MFI) (839 genes). We translated the list of fatigue-associated genes into functional profiles using Gene Ontology (GO) or pathway annotations. Approximately 50% of the data were mapped to a GO category, and only about 25% of the data mapped to a pathway.
- Biological areas associated with fatigue in this data set indicated impaired energy metabolism. Several signal transduction pathways were correlated with fatigue and these regulate a broad range of biological processes, for example, metabolism, the growth and the proliferation of a cell and apoptosis.
- There is some overlap with genes implicated in other studies using differential gene expression. However, QTA allows detection of alterations that may not reach statistical significance in class comparison analyses, but which could contribute to disease pathophysiology. This study supports the use of phenotypic measures of CFS and QTA as important for additional studies of this complex illness.
- One of several areas of research that need to be expanded upon from this study is whether the expression patterns associated with fatigue from this study of CFS subjects are common to fatigue in other diseases.
- There are several issues that urgently need to be addressed with regard to the annotation of gene expression data. Microarray probe selection for the commercially available arrays used in this study relied on earlier genome and transcriptome annotation, which is significantly different from current knowledge. Therefore, probe set information provided by the manufacturers of microarrays are no longer consistent with gene and transcript models in major public databases. This has a great impact on the analysis and interpretation of the data.
- The accurate and increased annotation of biological data, such as gene functions will provide a vital resource for future study. Currently this is a stumbling block to gene expression analysis. In this study less than 50% of gene functions were known.

Bibliography

1. Simon RM, Lam A: *BRB-ArrayTools User Guide*. Version 3.2 (2004).
2. Skuse, DH: Endophenotypes and child psychiatry. *Br. J. Psychiatry* 178, 395–396; 395–396 (2001).
3. 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).
4. Smets EM, Garssen B, Bonke B *et al.*: The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. *J. Psychosom. Res.* 39, 315–325 (1995).
5. Vernon SD, Reeves WC: The challenge of integrating disparate high-content data: epidemiological, clinical and laboratory data collected during an in-hospital study of chronic fatigue syndrome. *Pharmacogenomics* 7(3), 345–354 (2006).
6. Rinn JL, Snyder M: Sexual dimorphism in mammalian gene expression. *Trends Genet.* 21, 298–305 (2005).
7. 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).
8. Meek PM, Nail LM, Barsevick A *et al.*: Psychometric testing of fatigue instruments for use with cancer patients. *Nurs. Res.* 49, 181–190 (2000).
9. Wagner D, Nisenbaum R, Heim C *et al.*: Psychometric properties of the CDC symptom inventory for assessment of chronic fatigue syndrome. *Popul. Health Metr.* 3, 1–8 (2005).
10. Geladi P, MacDougall D, Martens H: Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* 39, 491–500 (1985).
11. Bolstad BM, Irizarry RA, Astrand M *et al.*: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193 (2003).
12. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. [Ser. B]* 57, 27–39 (1995).

13. Simon R, Korn EL, McShane LM, Radmacher MD, Wright G, Zhao Y: *Design and Analysis of DNA Microarray Investigations*. Springer Verlag, Heidelberg, Germany (2003).
 14. Korn EL, Troendle JF, McShane LMLM *et al.*: Controlling the number of false discoveries: application to high-dimensional genomic data . *J. Stat. Plan. Inference* 124, 379–398 (2005).
 15. Zhang B, Schmoyer D, Kirov S *et al.*: GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 5, 16 (2004).
 16. Whistler T, Unger ER, Nisenbaum R *et al.*: Integration of gene expression, clinical, and epidemiologic data to characterize chronic fatigue syndrome. *J. Transl. Med.* 1, 10 (2003).
 17. Kaushik N, Fear D, Richards SC *et al.*: Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome. *J. Clin. Pathol.* 58, 826–832 (2005).
- Websites
101. Gene ontology Tree machine (GOTM) is a web-based platform for interpreting microarray data or other interesting gene sets using Gene Ontology. <http://genereg.ornl.gov/gotm>
 102. The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism www.geneontology.org
 103. Online resource for easy access to collective and integrated information from various public biological resources. www.biorag.org
 104. Kyoto Encyclopedia of Genes and Genomes www.genome.jp/kegg
 105. Resource for studying pathways and examining how genes interact through dynamic graphical models. www.biocarta.com
 106. GenMAPP is a free computer application designed to visualize gene expression and other genomic data on maps representing biological pathways and groupings of genes. www.genmapp.org
 107. A web-based platform for interpreting microarray data or other interesting gene sets using Gene Ontology. <http://genereg.ornl.gov/webgestalt/>