Understanding Chronic Fatigue Syndrome (CFS) from CAMDA Data: A Systems Biology Approach

Vasyl Pihur, Susmita Datta, Somnath Datta

Department of Bioinformatics and Biostatistics,
University of Louisville, KY 40292, USA

susmita.datta@louisville.edu
somnath.datta@louisville.edu

@ CAMDA 2007, Valencia, December 13, 2007

Chronic Fatigue Syndrome (CFS):

- **CFS is a relatively rare, poorly understood, complex disorder that is characterized by severe and chronic physical and mental fatigue not attributable to other causes (diseases)**

- **Sometimes accompanied by other symptoms such as weak immune response, digestive problems and depression**

- **A great deal of effort has been put forth in recent years in collecting clinical, gene expression, genotypic and proteomic data by the Chronic Fatigue Syndrome Group at CDC in an attempt to find a genetic basis of CFS**

- **These data have been analyzed by numerous researchers (and research teams) in the last two years resulting in a special issue of the journal *Pharmacogenomics* and were also as part the CAMDA conference in 2006**

- **The type of success has been mixed and limited**

- **Our attempt in analyzing these data as part of this year's CAMDA competition takes a systems biology approach where we study groups of genes (called modules) obtained from gene-gene association/interaction networks since genes do not act alone, especially, for a complex disorder**

- **At the end, we identify eleven "interesting" genes which may play important roles in certain aspects of CFS or related symptoms**

- **The CDC Chronic Fatigue Syndrome Research Group provided challenge datasets consisting of clinical, microarray, proteomics, and SNP data that were used for both CAMDA 2006 and CAMDA 2007 competitions**

- **227 subjects filled self-administered questionnaires and had their blood drawn for lab analysis**

- **Clinical, Microarray (gene expression), SNP (genotypic) and SELDI-TOF (proteomic) data were collected**

- **We attempt to integrate all four types of data in our analysis although we mainly focus on the microarray data**

Microarray Data:

- **CAMDA 2006 microarray data consists of 177 arrays, 9 of which were repeated twice at different times during the study**

- **We discarded these 9 microarrays for multiplicity reasons and additional 5 arrays were excluded from this analysis due to the absence of clinical information on the subjects. Thus, we started our analysis with 163 arrays. Subtracted ARM (Artifact-removed) density column which is already adjusted for the background density was log-transformed to stabilize the variance**

Clinical Data

- **Clinical data contains extensive information on 227 subjects and can be linked to microarray and SNP data via the ABTID subject ID**

- **The two pieces of clinical data that we made use of were the *Intake Classific* variable which classifies patients into 5 categories and the *Cluster* variable provides information on the severity of the symptoms ("Worst", "Middle", "Least")**

- **ANOVA F-test for each probe was carried out to determine differentially expressed genes across the five groups. 286 probes were identified as differentially expressed (p-values < 0.01). Since we are not interested in determining the differentially expressed genes per se, multiplicity correction was not used. The reduced microarray data consisting of 286 probes and 163 samples (subjects) was used later for further analysis**

Network Analysis:

- **For checking consistency and added validation, we employ two computational network inference techniques**

- **The first method is based on the Partial Least Squares regression (PLS) : Pihur et al., 2007, under revision (also see Datta, Susmita, 2001)**

- **The second method is based on the Partial Correlations (PC) : Schäfer and Strimmer, 2005, *Bioinformatics***

The PLS method

- **Uses PLS based scores to measure strength of association/interaction**

- **Uses empirical Bayes technique to judge whether a score is significant or not**

- **Uses local fdr to control the error rate**

- **Nonparametric**

The PC method

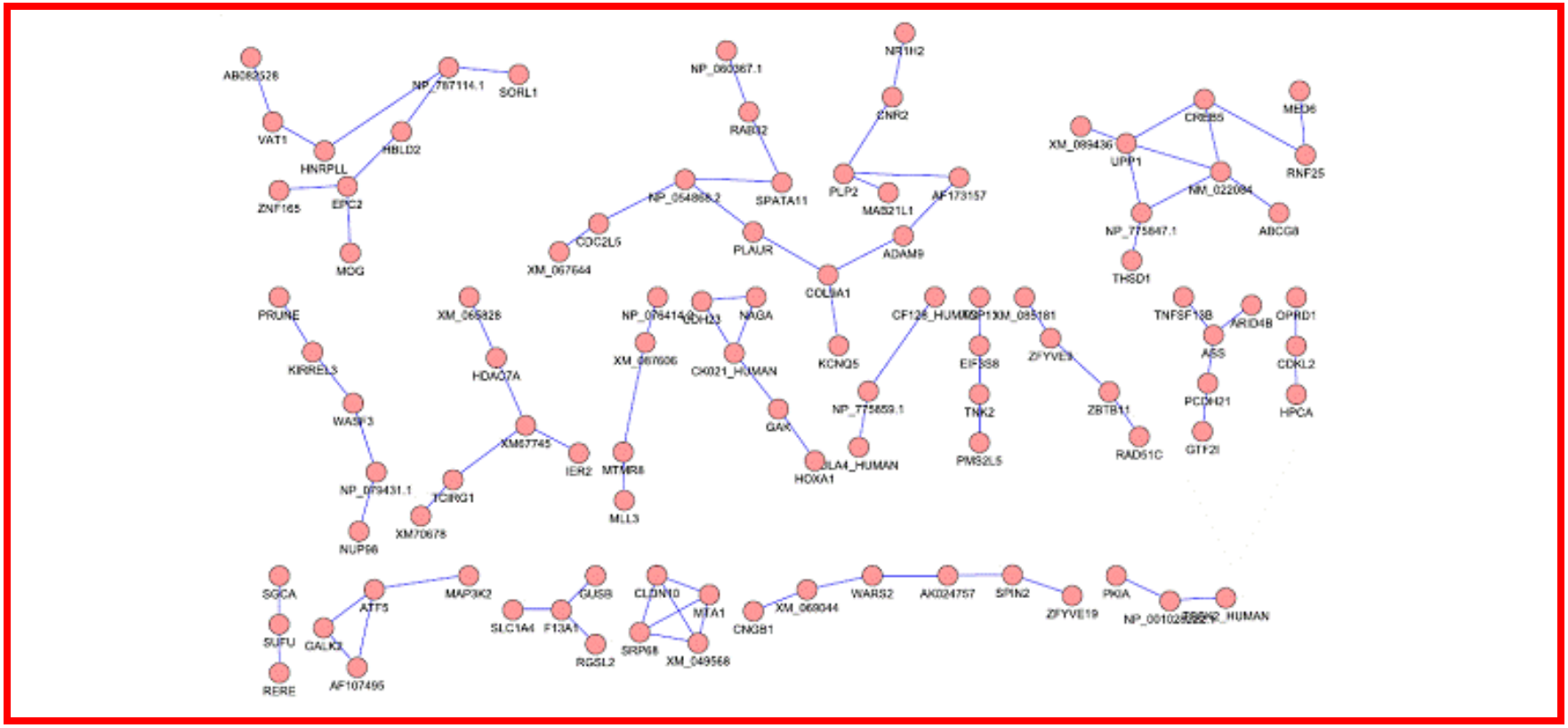
- **Uses partial correlations to measure strength of association/interaction**

- **Uses parametric models and approximate tests to judge whether a score is significant or not**

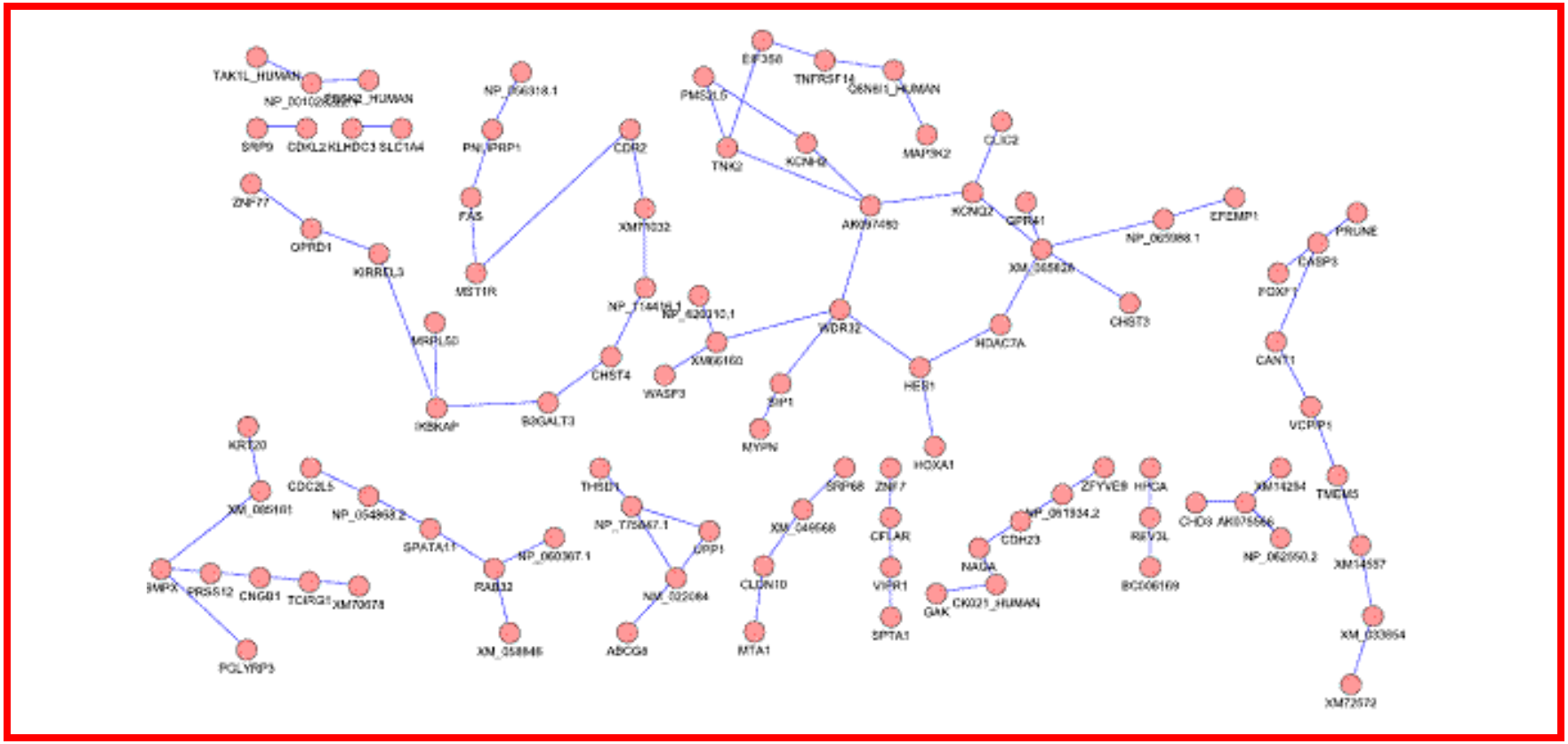
- **Uses local fdr to control the error rate**

- **Parametric**

Gene-Gene Network constructed using the PLS method



Gene-Gene Network constructed using the PC method



Summary tables

- **Gene association modules were defined to be clusters of 4 or more connected genes such that genes in two distinct components are not connected by an edge**

- **Percentages of each module's average association score when compared to the module with the largest average association score are reported**

- **Genes shown in red are the genes that appear in both tables**

Based on PLS

# of Genes	Average Scores(%)	Gene Symbols
4	100	MTA1, SRP68, XM_049568, CLDN10
9	88	CREB5, MED6, UPP1, NP_775847.1, ABCG RNF25, XM_089436, THSD1, NM_022084
5	85	HOXA1, NAGA, GAK, CK021_HUMAN, CDH
6	81	IER2, TCIRG1, XM67745, XM_06582 XM70678, HDAC7A
*5	79	WASF3, NUP98, PRUNE, NP_079431.1, KI REL3
6	78	ZFYVE19, AK024757, CNGB1, WARS2, SPIN XM_069044
5	77	PCDH21, ASS, GTF2I, ARID4B, TNFSF13B
9	77	AB082528, HNRPLL, HBLD2, ZNF165, MO SORL1, VAT1, EPC2, NP_787114.1
4	77	MTMR8, NP_076414.2, MLL3, XM_087606
4	76	ZFYVE9, RAD51C, XM_085181, ZBTB11
4	75	TNK2, EIF3S8, PMS2L5, TCP11
4	73	MAP3K2, ATF5, AF107495, GALK2
15	72	CDC2L5, PLP2, NR1H2, PLAUR, SPATA1 NP_060367.1, KCNQ5, COL9A1, AF17315 XM_067644, MAB21L1, CNR2, NP_054868. RAB32, ADAM9
4	18	SLC1A4, F13A1, RGSL2, GUSB

Based on PC

# of Genes	Average Scores(%)	Gene Symbols
4	100	SRP68, MTA1, XM_049568, CLDN10
5	85	ABCG8, NP_775847.1, UPP1, NM_022084, THSD1
9	85	CASP3, XM72572, TMEM5, XM14557, CANT1, XM_033654, FOXF1, VCPIP1, PRUNE
*24	84	CHST3, SIP1, TNK2, CLIC2, AK097480, NP_065988.1, XM_065828, EIF3S8, HES1, HOXA1, PMS2L5, KCNH2, XM66160, TNFRSF14, EFEMP1, KCNQ2, WASF3, Q8N8I1_HUMAN, MYPN, HDAC7A, WDR32, NP_620310.1, GPR41, MAP3K2
6	84	NP_060367.1, SPATA11, XM_058846, CDC2L5, RAB32, NP_054868.2
6	83	NAGA, CDH23, GAK, NP_061934.2, CK021_HUMAN, ZFYVE9
14	82	CHST4, CDR2, NP_114416.1, NP_056318.1, IKBKAP, KIRREL3, FAS, ZNF77, B3GALT3, MST1R, XM71032, PNLIPRP1, OPRD1, MRPL50
4	81	VIPR1, CFLAR, SPTA1, ZNF7
8	79	CNGB1, KRT20, TCIRG1, PGLYRP3, PRSS12, SMPX, XM_085181, XM70678
4	78	CHD3, AK075566, XM14294, NP_062550.2

Prediction of Symptom Severity

- **We investigate the ability of each module to predict the CFS severity level**

- **We fit a log-linear model for each gene module to regress the clinical variable *'Cluster'* on the set of expression profiles of genes included in the module**

- **The overall predictive ability of the CFS severity by a given module can be judged on the basis of the likelihood ratio test which compares the full model (all genes in a module included as covariates in the model) and the null model which includes no covariates**

- **Small p-values indicate that gene association modules are effective in predicting the symptom severity categories**

- **These should be interpreted as descriptive statistics rather than conclusive evidence from formal statistical tests since, among other things, these are not adjusted for multiple testing**

PLS results

# of Genes	Average Scores(%)	Gene Symbols	Severity p-value
4	100	MTA1, SRP68, XM_049568, CLDN10	0.7588
9	88	CREB5, MED6, UPP1, NP_775847.1, ABCG8, RNF25, XM_089436, THSD1, NM_022084	0.1645
5	85	HOXA1, NAGA, GAK, CK021_HUMAN, CDH23	0.4978
6	81	IER2, TCIRG1, XM67745, XM_065828, XM70678, HDAC7A	0.5051
*5	79	WASF3, NUP98, PRUNE, NP_079431.1, KIRREL3	**0.0154**
6	78	ZFYVE19, AK024757, CNGB1, WARS2, SPIN2, XM_069044	0.0081
5	77	PCDH21, ASS, GTF2I, ARID4B, TNFSF13B	0.3015
9	77	AB082528, HNRPLL, HBLD2, ZNF165, MOG, SORL1, VAT1, EPC2, NP_787114.1	0.0063
4	77	MTMR8, NP_076414.2, MLL3, XM_087606	0.0032
4	76	ZFYVE9, RAD51C, XM_085181, ZBTB11	0.2778
4	75	TNK2, EIF3S8, PMS2L5, TCP11	0.1286
4	73	MAP3K2, ATF5, AF107495, GALK2	0.0436
15	72	CDC2L5, PLP2, NR1H2, PLAUR, SPATA11, NP_060367.1, KCNQ5, COL9A1, AF173157, XM_067644, MAB21L1, CNR2, NP_054868.2, RAB32, ADAM9	0.0145
4	18	SLC1A4, F13A1, RGSL2, GUSB	0.0053

PC results

# of Genes	Average Scores(%)	Gene Symbols	Severity p-value
4	100	SRP68, MTA1, XM_049568, CLDN10	0.7588
5	85	ABCG8, NP_775847.1, UPP1, NM_022084, THSD1	0.0329
9	85	CASP3, XM72572, TMEM5, XM14557, CANT1, XM_033654, FOXF1, VCPIP1, PRUNE	0.1299
*24	84	CHST3, SIP1, TNK2, CLIC2, AK097480, NP_065988.1, XM_065828, EIF3S8, HES1, HOXA1, PMS2L5, KCNH2, XM66160, TNFRSF14, EFEMP1, KCNQ2, WASF3, Q8N8I1_HUMAN, MYPN, HDAC7A, WDR32, NP_620310.1, GPR41, MAP3K2	**0.0169**
6	84	NP_060367.1, SPATA11, XM_058846, CDC2L5, RAB32, NP_054868.2	0.0315
6	83	NAGA, CDH23, GAK, NP_061934.2, CK021_HUMAN, ZFYVE9	0.1886
14	82	CHST4, CDR2, NP_114416.1, NP_056318.1, IKBKAP, KIRREL3, FAS, ZNF77, B3GALT3, MST1R, XM71032, PNLIPRP1, OPRD1, MRPL50	1e-04
4	81	VIPR1, CFLAR, SPTA1, ZNF7	0.0105
8	79	CNGB1, KRT20, TCIRG1, PGLYRP3, PRSS12, SMPX, XM_085181, XM70678	0.0932
4	78	CHD3, AK075566, XM14294, NP_062550.2	0.0536

Integration of SNP data

- **Forty two Single nucleotide polymorphisms (SNP's) for 10 different genes were genotyped**

- **For the purposes of this analysis, we selected two SNP's, hCV245410 (on gene TPH2) and hCV7911132 (on gene SLC6A4), which were previously identified (Presson et al., 2006, CAMDA) to be associated with CFS severity**

- **Once again, we study whether the gene modules are predictive of the genotypes at these SNP locations by fitting log-linear models**

PLS results

# of Genes	Average Scores(%)	Gene Symbols	Severity p-value	hCV245410 p-value	hCV7911132 p-value
4	100	MTA1, SRP68, XM_049568, CLDN10	0.7588	0.3869	0.0328
9	88	CREB5, MED6, UPP1, NP_775847.1, ABCG8, RNF25, XM_089436, THSD1, NM_022084	0.1645	0.2970	0.1271
5	85	HOXA1, NAGA, GAK, CK021_HUMAN, CDH23	0.4978	0.2636	0.6640
6	81	IER2, TCIRG1, XM67745, XM_065828, XM70678, HDAC7A	0.5051	0.5825	0.1689
*5	79	WASF3, NUP98, PRUNE, NP_079431.1, KIR-REL3	**0.0154**	**0.0163**	0.1665
6	78	ZFYVE19, AK024757, CNGB1, WARS2, SPIN2, XM_069044	0.0081	0.7775	0.1203
5	77	PCDH21, ASS, GTF2I, ARID4B, TNFSF13B	0.3015	0.4987	0.0112
9	77	AB082528, HNRPLL, HBLD2, ZNF165, MOG, SORL1, VAT1, EPC2, NP_787114.1	0.0063	0.8304	0.1047
4	77	MTMR8, NP_076414.2, MLL3, XM_087606	0.0032	0.5670	0.2732
4	76	ZFYVE9, RAD51C, XM_085181, ZBTB11	0.2778	0.4110	9e-04
4	75	TNK2, EIF3S8, PMS2L5, TCP11	0.1286	0.6770	0.1270
4	73	MAP3K2, ATF5, AF107495, GALK2	0.0436	0.2459	0.1904
15	72	CDC2L5, PLP2, NR1H2, PLAUR, SPATA11, NP_060367.1, KCNQ5, COL9A1, AF173157, XM_067644, MAB21L1, CNR2, NP_054868.2, RAB32, ADAM9	0.0145	0.0960	0.2859
4	18	SLC1A4, F13A1, RGSL2, GUSB	0.0053	0.7451	0.4517

PC results

# of Genes	Average Scores(%)	Gene Symbols	Severity p-value	hCV245410 p-value	hCV7911132 p-value
4	100	SRP68, MTA1, XM_049568, CLDN10	0.7588	0.3869	0.0328
5	85	ABCG8, NP_775847.1, UPP1, NM_022084, THSD1	0.0329	0.2552	0.1428
9	85	CASP3, XM72572, TMEM5, XM14557, CANT1, XM_033654, FOXF1, VCPIP1, PRUNE	0.1299	0.5498	0.2732
*24	84	CHST3, SIP1, TNK2, CLIC2, AK097480, NP_065988.1, XM_065828, EIF3S8, HES1, HOXA1, PMS2L5, KCNH2, XM66160, TNFRSF14, EFEMP1, KCNQ2, WASF3, Q8N8I1_HUMAN, MYPN, HDAC7A, WDR32, NP_620310.1, GPR41, MAP3K2	**0.0169**	**0.0586**	0.6642
6	84	NP_060367.1, SPATA11, XM_058846, CDC2L5, RAB32, NP_054868.2	0.0315	0.3867	0.1895
6	83	NAGA, CDH23, GAK, NP_061934.2, CK021_HUMAN, ZFYVE9	0.1886	0.8259	0.08
14	82	CHST4, CDR2, NP_114416.1, NP_056318.1, IKBKAP, KIRREL3, FAS, ZNF77, B3GALT3, MST1R, XM71032, PNLIPRP1, OPRD1, MRPL50	1e-04	0.9611	0.4225
4	81	VIPR1, CFLAR, SPTA1, ZNF7	0.0105	0.5447	0.7448
8	79	CNGB1, KRT20, TCIRG1, PGLYRP3, PRSS12, SMPX, XM_085181, XM70678	0.0932	0.6724	0.4883
4	78	CHD3, AK075566, XM14294, NP_062550.2	0.0536	0.838	0.9018

Integration of proteomic data

- To further validate the two modules of genes that we selected one from the PLS and one from the PC networks, we analyzed the protein spectra available for 63 subjects in the study

- Serum was originally separated into six fractions of which we use the last 4 and then applied to 3 different SELDI surfaces and 2 laser settings giving us a total combination of 24 different experimental settings

- Spectra for 32 combined cases (CFS and ISF subjects) and 31 controls were compared for each combination of chip surface and fraction

Pre-processing

• We removed the first 4000 m/z values from our analysis which roughly corresponds to m/z values smaller than 1700 Da

• We divided the spectrum into the bins of size 10 and took the maximum intensity value in each bin. That reduced the data by a factor of 10, leaving 2650 m/z values in the data for further analysis

• We estimate the standard deviation for each m/z bin and take the median of these as a measure of noise standard deviation σ. Intensity values smaller than 3σ were considered to be pure noise. If this happened in all samples, the m/z value was removed from the analysis. Then the data was then log transformed

Selection of most discriminating features

• We have run a number of well regarded classifiers based on the class information we the hope of identifying the features with most classification ability; however this approach was abandoned since none of the classifiers produced desirable classification error rate when cross validation was used

• For each of the remaining m/z bin we perform a t-test to compare the case and the control samples and identified the discriminating features by the magnitude of the p-values

Prediction by the gene modules

• **We fitted regression models to predict the intensity values of the ten most discriminating features from the collection of expressions of the genes in the two modules (from PLS and PC, respectively) identified by our past analysis**

• **For a number of the settings the gene modules are predictive (small p-values) of the intensities of at least some of the top ten features**

- **The setting IMAC30, fraction 4, high laser shows particularly strong agreement with the PLS module**

p-values from the PLS module

IMAC30 F4	0.9358	0.0409	0.3409	0.1622	0.4425	0.0334	0.1191	0.0961	0.0640	0.0139

- **The setting H50, fraction 6, low laser shows particularly strong agreement with the PC module**

p-values from the PC module

H50 F6	0.0717	0.0406	0.3171	0.1058	0.5441	0.0272	0.0718	0.0577	0.0777	0.0947

Discussion of Results

- **Two gene association modules (indicated by asterisks) are of interest based on their predictive ability of symptom severity and of, at least, one of the SNP genotypes**

- **These gene modules exhibit some connection with the proteomic data as well**

- **Eleven genes that are in common between these two gene modules**

- **It is plausible that these genes are responsible for certain aspects of CFS or its symptoms**

- **The first gene on the list WASF3 (aka WAVE3) is thought to take part in the p38 MAPK regulatory pathway (Sossey-Alaoui et al., 2005). On the other hand, in recent animal model studies (Katafuchi et al., 2006), it has been demonstrated that regulation of brain cytokines through p38 MAPK pathway is involved in the in the central mechanisms of fatigue and therefore may play a role in the pathogenesis of the CFS**

- **The list also includes autoimmune response gene NUP98 and genes related to tumor activities (PRUNE, TNK2, HOXA1)**

- **Gene expression of HDAC7A has been shown to be correlated with unexplained fatigue in a past study (Whistler et al., 2006)**

- **The gene GPR41's role in autoimmune disorders including CFS has been hypothesised in Staines (2005)**

Gene	GO Process	Pathways	Description
WASF3	Cell Organization and Biogenesis, Metabolism	Adherens Junction	Actin-binding WH2
NUP98	Cell Organization and Biogenesis, Transport, DNA Replication	RAN regulation	
PRUNE	Energy production and conversion	Purine metabolism	Glycoside hydrolase, Phosphoesterase
KIRREL	Signal Transduction, Cell Adhesion		Integral to membrane, protein binding
TNK2	Cell Organization and Biogenesis, Signal Transduction, Protein amino acid phosphorylation	Regulation of CDC42 activity, Regulation of RAC1 activity	PAK-box/P21-Rho-binding, Protein kinase
EIF3S8	Protein Biosynthesis		Translation initiation factor activity
HOXA1	Transcription	p44/42 MAP kinase	Sequence-specific DNA binding
PMS2L5	DNA Repair		ATP binding, damaged DNA binding
HDAC7A	DNA Metabolism, Transcription		Histone deacetylase 7A
GPR41	Signal Transduction	p53/Bax pathway	G Protein-Coupled Receptor
MAP3K2	Protein amino acid phosphorylation	Mapk signaling, Gap Junction	Mitogen-activated protein kinase

REFERENCES

Datta, Susmita (2001). Gene Expression, 9, 257-264.

Katafuchi et al. (2006). Ann. N.Y. Acad. Sci., 1088, 230-237.

Pihur et al. (2007). Preprint

Presson et al. (2006). CAMDA 2006 Conference Paper.

Schäfer, J. and Strimmer, K. (2005). Bioinformatics, 21(6), 754-64.

Sossey-Alaoui et al. (2005). Experimental Cell Research, 308(1), 135-145.

Staines, D. (2005). Medical Hypothses, 65, 29-31.

Vernon, S. D. and Reeves, W. C. (2006). Pharmacogenomics, 7, 345-354.

Whistler et al. (2006). Future Medicine, 7(3) 395-405.

THANK YOU FOR YOUR ATTENTION!

THANK YOU FOR YOUR ATTENTION!