The Pharmacogenomics Journal (2003) 3, 308–311

www.nature.com/tpj

CLINICAL IMPLICATION

Making sense of microarray data to classify cancer

S Hanash[1] and C Creighton[2]

[1]*Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA;* [2]*Bioinformatics Program, Ann Arbor, MI, USA*

The Pharmacogenomics Journal (2003) **3**, 308–311. doi:10.1038/sj.tpj.6500209
Published online 4 November 2003

Profiling gene expression using DNA arrays has had a tremendous impact on biomedical research. From a disease investigation point of view, applications of DNA microarrays include uncovering unsuspected associations between genes and specific clinical features of disease, resulting in novel, molecular-based disease classifications. Cancer is a case in point. Most published studies of cancers using DNA microarrays have either examined a pathologically homogeneous set of tumors to identify clinically relevant subtypes, for example, responders *vs* nonresponders, or pathologically distinct subtypes of cancer of the same lineage, for example, high-stage *vs* low-stage tumors to identify molecular correlates, or tumors of different lineages to identify molecular signatures for each lineage. A study of cutaneous T-cell lymphoma by Kari *et al*,[1] published recently, typifies both what one hopes to gain from disease investigations using DNA microarrays and the limitations of such studies.

Primary cutaneous lymphomas are a heterogeneous group of lymphomas of T- or B-cell origin that represent a relatively common type of lymphoma and their incidence appears to be increasing. The two predominant subtypes of cutaneous T-cell lymphomas are mycosis fungoides, a mostly indolent variety, and its leukemic counterpart the Sezary syndrome, an aggressive variety characterized by skin involvement, lymphadenopathy and circulating atypical lymphocytes, the so-called Sezary cells. Kari *et al* used cDNA microarrays to study gene expression patterns in peripheral blood mononuclear cells from patients with the leukemic form of cutaneous T-cell lymphoma. The goal of the study was to identify markers that may be useful for diagnosis or prognosis, or that might provide new targets for treating this disease. The approach was to uncover gene expression differences between cells from 18 patients with high Sezary cell counts and an appropriate (Th2-skewed) cell fraction from nine normal controls. The differences in gene expression observed reflected many of the observed characteristics of the disease. Overexpressed genes in disease samples included some genes required for Th2 differentiation characteristic of Sezary cells. The analysis, however, did not uncover changes consistent with the hypothesis of defective apoptotic pathways in this disease. An important objective of the study was to identify markers for cutaneous T-cell lymphoma given the paucity of such markers. A member of the plastin gene family and a chemokine (CX3CR1) inappropriately expressed represented such potential novel markers. Two genes found to have a high predictive power to classify patients and controls were STAT4 and GTPase RhoB. These two genes alone accurately classified the high Sezary cell patients and controls. A signature profile with 10 genes was uncovered that identified a class of patients who succumb to the disease early, irrespective of their tumor burden. The study therefore uncovered a wealth of findings that shed some light on the biology of this disease and uncovered markers that may have a practical utility.

The DNA microarray studies described above and others in the literature indeed point to the great utility of DNA microarrays for uncovering patterns of gene expression that are clinically informative. Have the data been thoroughly analyzed? There is no shortage of analytical tools for uncovering patterns in microarray data. An important challenge for microarray analysis is to understand at a mechanistic level the significance of associations observed between subsets of genes and clinical features of disease. Another challenge is to identify the smallest but most informative sets of genes associated with specific clinical features, which then could be interrogated using technologies available in clinical laboratories, as appears to have been accomplished in this study. Another challenge is to determine how well RNA levels of predictive genes correlate with protein levels. A lack of correlation may imply that the predictive property of the gene(s) is independent of gene function.

To increase the effectiveness of DNA microarray analysis, global gene expression data may be combined with external data sources, such as gene annotation, in order to associate the expression patterns of a set of genes with the biological processes that they may represent. A welcome trend of data sharing allows others to analyze previously published microarray data and to combine multiple data sets. For illustration, we examined the data set published by Kari *et al* to see what we could uncover. In our analysis, we relied on the Gene Ontology (GO) annotation. The Gene Ontology Consortium[2] has defined a controlled vocabulary for describing genes in terms of their molecular function,

participation in biological processes and cellular locations. The GO annotations are making possible the high-throughput analyses of gene expression in terms of functional gene class associations, which otherwise would require laborious and somewhat subjective manual literature searches.

Using the data set from Kari *et al*, we searched a set of 122 genes found overexpressed in patients with high blood tumor burden, or Sezary cell count, compared to healthy controls ($P<0.01$, fold change >1.5), for significantly enriched (over-represented) GO terms, as described elsewhere.[3] We made the same search for a set of 280 genes found underexpressed in patients with high Sezary cell count ($P<0.01$, fold change <0.67). Our premise is that annotation terms that are shared by a significant number of genes within a large gene set may provide clues as to the processes driving the coordinate expression of the genes as a whole. Numerous enriched terms were found for the set of 280 underexpressed genes with $P<0.001$, including *class II major histocompatibility complex antigen* (five genes represented), *cytokine-binding activity* (six), *mitochondrion* (26), *electron transporter activity* (12) and *nucleotide metabolism* (four); these enriched terms could suggest a downregulation in CTCL of processes related to the immune response and mitochondrial function. Terms found enriched for the set of 122 overexpressed genes with $P<0.05$ include cell adhesion (nine genes represented) and cell cycle arrest (three).

The enriched GO terms listed above represent only a fraction of the genes significantly expressed in CTCL, and additional gene-to-process associations, not currently described in the biomedical literature or public annotation sources, may be inferred from data mining of large expression profile data sets. Our premise in this case is that genes that are coordinately expressed participate in closely related biological processes.[4] For a given gene, a GO term may be associated if the gene is correlated in expression with a significant number of other genes that share the given GO term annotation. We examined the expression patterns of 60 genes highly underexpressed in the Kari *et al* data set for patients with high Sezary cell count ($P<0.01$, fold change <0.33) that were also represented in a large independent data set of leukemia expression profiles from Armstrong *et al*.[5] For each of the 60 genes, the set of genes with significant positive correlations ($P<0.01$) with the given gene in the Armstrong data set was searched for significantly enriched GO terms ($P<0.0001$). In this way, 1963 gene-to-term associations, involving all 60 genes, were found. We performed two simulation tests to assess the number of random gene-to-term associations that could exist in the Armstrong data set, in one test permuting the expression values and

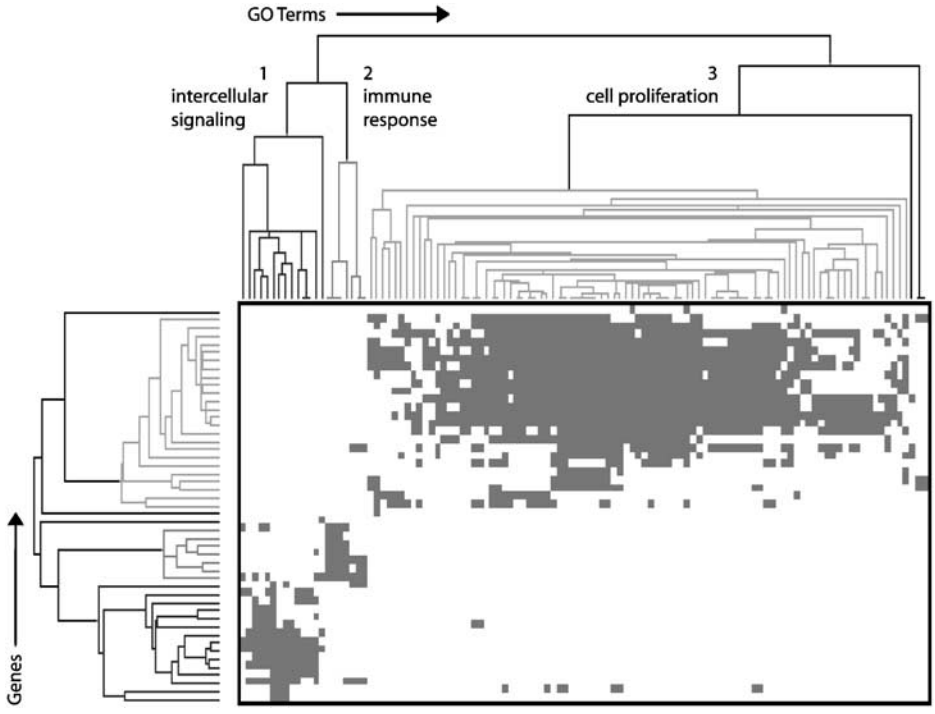

Figure 1 Hierarchical clustering of associations of GO terms for genes found underexpressed in patients with high Sezary cell count ($P<0.01$, fold change <0.33). For each gene-to-term association represented here, the given gene was found positively correlated in expression with a significant number of other genes that share the given GO term annotation. The rows in the matrix diagram represent genes; the columns represent terms. An entry in the matrix indicates that the corresponding gene-to-term association was found in the leukemia profile data set from Armstrong *et al* with $P<0.0001$. Three major clusters are highlighted corresponding to terms related to (1) intercellular signaling, (2) the immune response, and (3) cell proliferation. Table 1 lists the genes that fall under each cluster.

Table 1 GO term associations from Figure 1 for genes underexpressed in patients with high Sezary cell counts

Gene	*Gene product description*
Cluster 1—integral to plasma membrane; receptor activity; signal transducer activity; cell surface receptor-linked signal transduction; cell motility; G-protein-coupled receptor protein signaling pathway; cell–cell signaling; development; organogenesis; morphogenesis; extracellular	
CCL2	Small inducible cytokine A2
CD8B1	CD8 antigen, beta polypeptide 1 (p37)
CTSL	Cathepsin L
GPNMB	Glycoprotein (transmembrane) nmb
IL1R1	Interleukin 1 receptor, type I
ITGB4	Integrin, beta 4
MAL	Mal, T-cell differentiation protein
MAOA	Monoamine oxidase A
ME1	Malic enzyme 1, NADP(+)-dependent, cytosolic
PLAU	Plasminogen activator, urokinase
STAT4	Signal transducer and activator of transcription 4
TNFAIP6	Tumor necrosis factor, alpha-induced protein 6
Cluster 2—immune response; response to biotic stimulus; defense response; vacuole; lytic vacuole; lysosome	
CCL4	Small inducible cytokine A4 (homologous to mouse Mip-1b)
CYP1B1	Cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile)
FCER2	Fc fragment of IgE, low-affinity II, receptor for (CD23A)
GZMK	Granzyme K (serine protease, granzyme 3; tryptase II)
IL4R	Interleukin 4 receptor
MMP9	Matrix metalloproteinase 9 (gelatinase B, 92 kDa gelatinase, 92 kDa type IV collagenase)
TIMP1	Tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)
Cluster 3—DNA repair; DNA replication; nucleolus; cell cycle; cell proliferation; mitosis; mRNA processing; mRNA splicing; ubiquitin-dependent protein catabolism; 26S proteasome; spliceosome complex; translation initiation factor activity; mitochondrion; oxidative phosphorylation; tricarboxylic acid cycle; cytochrome c oxidase activity	
AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
AP1B1	Adaptor-related protein complex 1, beta 1 subunit
ATOX1	ATX1 (antioxidant protein 1, yeast) homolog 1
ATP5G3	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9) isoform 3
CD164	CD164 antigen, sialomucin
CDC2	Cell division cycle 2, G1–S and G2–M
DRG1	Developmentally regulated GTP-binding protein 1
HADH2	Hydroxyacyl-coenzyme A dehydrogenase, type II
HLA-DQB1	Major histocompatibility complex, class II, DQ beta 1
LDHA	Lactate dehydrogenase A
LMNB2	Lamin B2
NDUFS1	NADH dehydrogenase (ubiquinone) Fe-S protein 1 (75 kDa) (NADH-coenzyme Q reductase)
OXCT	3-oxoacid CoA transferase
PCNA	Proliferating cell nuclear antigen
RUNX1	Runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)
SATB1	Special AT-rich sequence-binding protein 1 (binds to nuclear matrix/scaffold-associating DNA's)
SLC25A11	Solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11
SPINT2	Serine protease inhibitor, Kunitz type, 2
TOP2A	Topoisomerase (DNA) II alpha (170 kDa)
TXNRD1	Thioredoxin reductase 1
UBE2C	Ubiquitin carrier protein E2-C
VDAC1	Voltage-dependent anion channel 1
YWHAZ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
ZNF204	Zinc-finger protein 204

in another test permuting the annotation assignments. Neither search with these randomized data sets yielded more than 15 associations, indicating that most of the actual associations found were not the result of chance.

We used average linkage hierarchical clustering[4] to obtain a global view of the gene-to-term associations mined from the Armstrong leukemia expression data set. Figure 1 shows the resulting cluster diagram (with 113 GO terms that were associated with at least five genes being represented and with 49 genes that were associated with at least one of these terms). Genes are represented in the rows of the matrix diagram, and

terms are represented in the columns. An entry in the diagram indicates that the given gene (underexpressed in patients with high Sezary count compared to healthy controls) was coexpressed with a significant number of genes that share the given annotation term. GO terms that are closely related to each other by the biological concepts that they represent were found to cluster together. The clustering diagram defines three distinct major clusters of genes and terms related to intercellular signaling (labeled as Cluster '1' in the figure), the immune response (labeled Cluster '2'), and cell proliferation (labeled Cluster '3'). Table 1 lists the genes that fall under each cluster, with example terms.

Our GO term clustering analysis indicates that many of the genes underexpressed in CTCL may be associated with processes of cell proliferation, the immune response or intercellular signaling, which suggests a hypothesis that the pathogenesis of CTCL involves a downregulation of these processes. CTCL is characterized by the accumulation of malignant cells with a low proliferative index, which appears consistent with the observation made here of numerous genes associated with proliferation being underexpressed in CTCL. The observed underexpression in CTCL of numerous genes involved in the immune response, including several genes encoding for the class II major histocompatibility antigen complex, might be construed as contradicting one hypothesis that CTCL may be a malignancy of T cells stimulated to proliferate against its own tumor antigens.[6] There has been much speculation that CTCL cells are defective in their apoptotic pathways, and that the disease is linked to an accumulation rather than a true proliferation of T cells.[7] Underexpressed genes in CTCL thought to mediate apoptosis, including STAT4, CTSL (cathepsin L), IL1R1 (interleukin 1 receptor, type I) and TNFAIP6 (tumor necrosis factor, alpha-induced protein 6), are associated here with intercellular signaling-related terms.

However perfected DNA microarrays and their analytical tools become for disease profiling, they will not eliminate a pressing need for other types of profiling technologies that go beyond measuring RNA levels, particularly for disease-related investigations. DNA microarrays have limited utility for the analysis of biological fluids and for uncovering directly in the fluid, assayable biomarkers. There is a need to assay protein levels and activity. Numerous alterations may occur in proteins that are not reflected in changes at the RNA level, providing a compelling rationale for additional, direct analysis of gene expression at the protein level. The next challenge is to integrate RNA data with protein data.

DUALITY OF INTEREST

None declared

Correspondence should be sent to:
SM Hanash, Department of Pediatrics, University of Michigan, 1150 W. Medical Center Drive, MSRBI, Room A520, Ann Arbor, MI 48109, USA.
Tel: + 734 763 9311
Fax: + 734 647 8148
E-mail: shanash@umich.edu

REFERENCES

1 Kari L, Loboda A, Nebozhyn M, Rook AH, Vonderheid EC, Nichols C *et al.* Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. *J Exp Med* 2003; **197**: 1477–1488.
2 Ashburner M, Ball CA, Blake JA, Botstein D, Cherry JM *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
3 Creighton C, Kuick R, Misek DE, Rickman DS, Brichory FM, Rouillard JM *et al.* Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice. *Genome Biol* 2003; **4**: R46.
4 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
5 Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 2002; **30**: 41–47.
6 Edelson RL. Cutaneous T cell lymphoma: the helping hand of dendritic cells. *Ann NY Acad Sci* 2001; **941**: 1–11.
7 de Arruda MV, Watson S, Lin C-S, Leavitt J, Matsudaira P. Fimbrin is a homologue of the cytoplasmic phosphoprotein plastin and has domains homologous with calmodulin and actin gelation proteins. *J Cell Biol* 1990; **11**: 1069–1079.