Gene expression patterns define pathways correlated with loss of differentiation in lung adenocarcinomas

Chad Creighton[a,*], Samir Hanash[b], David Beer[c]

[a] *Bioinformatics Program, University of Michigan, Ann Arbor, MI 48109, USA*
[b] *Department of Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI 48109, USA*
[c] *Department of Surgery, University of Michigan, Ann Arbor, MI 48109, USA*

Received 8 January 2003; revised 16 February 2003; accepted 27 February 2003

First published online 18 March 2003

Edited by Robert B. Russell

Abstract An analysis of microarray data from 86 lung adenocarcinomas reveals hundreds of genes significantly correlated with tumor cell differentiation. A bioinformatics approach of linking these genes to public information from the Locuslink and KEGG databases yields evidence for a loss of tumor cell differentiation being associated with biological processes of cell division, protein degradation, pyrimidine and purine metabolism, oxidative phosphorylation, glyoxylate and dicarboxylate metabolism, folate biosynthesis, and glutamate metabolism. The increased expression of genes involved in these processes is consistent with increased proliferation and metabolism characteristics of poorly differentiated tumors. The complete results of this analysis are available at http://dot.ped.med.umich.edu:2000/pub/diff/index.htm.

Key words: Gene expression; Lung cancer; Differentiation; Microarray

1. Introduction

The loss of differentiation in cancer cells, a process by which the cells gradually lose the characteristics of their organ-specific tissue of origin, is a hallmark of the progression of the disease, and tumor differentiation status has prognostic significance for recurrence and death in lung cancer [1]. In a recent study by Beer et al. [2] that used global gene expression profiles to predict lung cancer patient survival, hierarchical clustering of profiles from lung tumors showed significant differences between the clusters with respect to differentiation status; as expected, lung tumors with similar differentiation clustered together. We have carried out a post-planned analysis of the Beer et al. dataset and have found numerous genes significantly correlated in terms of their expression with the level of tumor differentiation, most of them with negative correlations, i.e. the high expression of these genes being indicative of a loss of differentiation. While a list of hundreds of genes by itself may shed little light on the processes involved with a loss or gain of differentiation in cancer, we have also found, through linking the genes to public sources of gene annotation and biological pathway information, key biological features that appear significantly over-represented in the set of negatively correlated genes, providing valuable insight into and stimulating hypotheses surrounding the molecular biology of lung cancer.

2. Materials and methods

2.1. Data collection

We obtained the dataset of Affymetrix HuGeneFL gene chip profiles from 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors, that was generated by the previous study by Beer et al. [2]. The differentiation status of each tumor profiled had been established by a pathological assessment and consisted of a rating of ‘poor’, for poorly differentiated (of which there were 21 tumors out the 86 profiled), ‘moderate’, for moderately differentiated (41 tumors), or ‘well’, for well differentiated (23 tumors).

2.2. Identification of genes significantly correlated with tumor differentiation

We assigned each profiled tumor a number based on the differentiation status, −1 for ‘poor’, 0 for ‘moderate’ and 1 for ‘well’; one of the tumors had been rated as ‘moderate-to-well’ and received a value of 0.5. We then calculated the Pearson correlation coefficient between the intensity values for each probe set and the numerical values for differentiation. The significance of the magnitude of any given coefficient was estimated using 1000 random permutations of the profile differentiation labels to determine the likelihood distribution for the dataset. We calculated the number of genes with significance values less than 10^{-2}, less than 10^{-3}, and less than 10^{-4} for the magnitude of the correlation coefficient being as high as it was for the gene by chance.

2.3. Significantly enriched annotation terms and KEGG pathways among gene sets

For the set of genes showing the lowest negative correlations with tumor differentiation status, a search was made for ‘significantly enriched’ annotation terms, i.e. terms that appear disproportionately in the gene set to a significant extent. The same type of search was made as well for the set of genes showing the highest positive correlations with tumor differentiation. GO terms and other external annotation terms were obtained from the Locuslink database [3,4]. For a given annotation term appearing n times in one of our sets of k genes, where the term applied to a total of A genes out of the set of G unique genes profiled on the HuGeneFL chip, the probability P for the term occurring n or more times within a set of k genes randomly selected from the chip was calculated using the hypergeometric formula:

$$P=\sum_{i=n}^{\min[k,A]}\frac{\binom{A}{i}\binom{G-A}{k-i}}{\binom{G}{k}}$$

As each annotation term in the Locuslink database (on the order of 2000) was tested for our set of genes of interest, the true significance of a low P-value for an enriched term was estimated using 1000 Monte Carlo simulation tests, each test in which k genes were first

*Corresponding author. Fax: (1)-734-615 4637.
E-mail address: ccreight@umich.edu (C. Creighton).

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00259-X

randomly selected from the set of G genes, and P-values for the terms occurring within the k set of genes were then calculated in order to determine the likelihood distribution for the dataset. Although a simpler Bonferroni correction of multiplying a P-value by the total number of annotation terms tested might have been used to estimate its significance, the significance would then likely be greatly underestimated, as the terms are not equally likely to be assigned a given P-value (e.g. one term may be represented by 200 genes on the array chip, while another term may be represented by only two genes). Instead, for a P-value for a given term in a k set of genes calculated using the hypergeometric formula, we estimated the number of terms that would be expected to have a P-value lower than the given P-value in a k set of randomly selected genes, by first counting the number of occurrences within the set of 1000 simulation results in which any term received a lower P-value and then dividing that number by 1000.

In order to identify significantly enriched pathways within the gene sets, we carried out the same type of analysis described above using the complete set of 119 biological pathways catalogued for *Homo sapiens* in the KEGG encyclopedia [5]. Using the above hypergeometric formula, we calculated the probability P for n or more genes belonging to a given pathway out of a set of k genes randomly selected from the set of G genes profiled on the HuGeneFL chip, on which A genes belonging to the pathway were represented. We carried out 1000 Monte Carlo simulations as described above to determine the true significance of an enriched pathway.

3. Results

3.1. Genes showing significant positive or negative correlations with tumor differentiation

Table 1 shows the number of probe sets that could be considered significant at various significance levels for the magnitude of the coefficient of the correlation with tumor differentiation. Over 500 probe sets out of some 7029 on the HuGeneFL chip had a significance value less than 0.001. Fig. 1 shows a histogram of the correlation to differentiation for the probe sets, along with an averaging of histograms generated from 10 separate random permutations of the tumor differentiation labels in the actual dataset, illustrating that the tails of the distribution for the correlation coefficients from the actual dataset are much more spread out compared to the distribution tails for the permutated datasets. Roughly twice as many probe sets in the actual dataset show significant negative correlations than show positive correlations. The complete list of significant genes with respect to tumor differentiation is available at our web site [12].

For our subsequent analysis for enriched annotation terms and pathways, we selected the probe sets with an absolute correlation coefficient greater than 0.3, the choice of 0.3 as a cutoff being somewhat arbitrary, but which included 812 probe sets, each with a significance value less than 0.0043. From the 812 probe sets, we selected genes into two sets, one set of 462 genes with mRNA expression showing negative correlations with tumor differentiation, and the other of 286 genes with mRNA expression showing positive correlations with differentiation (the remaining probe sets either did not map to a known gene or referred to a gene that was profiled by two or more probe sets).

Table 1
The number of HuGeneFL probe sets found to be significantly correlated with tumor differentiation in the Beer et al. [2] dataset, with the approximate number of probe sets that may have a high absolute correlation value by chance (Expected)

Significance	Probe sets	Expected
10^{-4}	266	1
10^{-3}	536	7
10^{-2}	1050	71

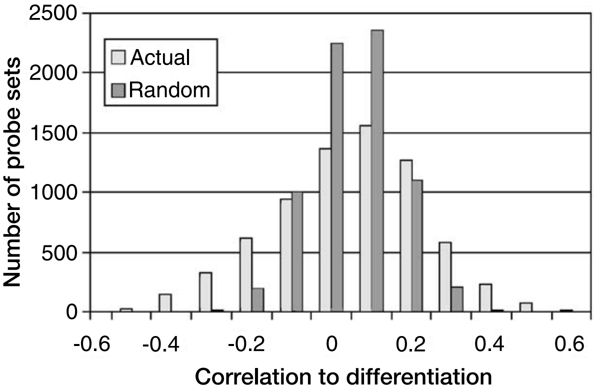

Fig. 1. Histogram of the correlation to differentiation for the probe sets in the lung tumor gene expression dataset from Beer et al. [2]. Also displayed is an averaging of histograms generated from 10 separate random permutations of the tumor differentiation labels in the original dataset.

3.2. Significantly enriched annotation terms for genes correlated with loss of tumor differentiation

We searched the set of genes positively correlated with differentiation and the negatively correlated set of genes separately for significantly enriched annotation terms (i.e. groups of functionally related genes). Table 2 displays the most significant terms found for the set of negatively correlated genes, which terms include *RNA-binding protein*, *26S proteasome*, *nuclear-cytoplasmic transport*, *DNA synthesis*, *small nuclear ribonucleoprotein*, *mitochondrial*, *ligase*, *protein folding*, *nuclear pore*, *RNA splicing*, *DNA replication factor*, *protein degradation,* and *protein synthesis*. If we consider the set of enriched terms listed in Table 2 to be significant over any other terms that could appear similarly enriched within a set of randomly selected genes, then we may expect on the order of 0.36 of the Table 2 terms to appear significant by chance.

While for the positively correlated genes, several enriched annotation terms of interest were found with a P-value less than 0.05, including *GTPase activating protein*, *gas exchange*, *hemoglobin*, *intercellular transport*, *cell-cell matrix adhesion*, *cell adhesion*, and *positive control of cell proliferation*, none of these terms could individually be considered truly significant, based on the Monte Carlo simulation test results, as multiple terms were tested (in a randomly selected set of 286 genes, on the order of one enriched term would be expected to have a P-value as low or lower than the lowest P-value that occurred in the actual 286 gene set). The complete set of enriched terms found for both sets of genes, including specifically which genes fall under which terms, is available at our web site [12].

3.3. Significantly enriched KEGG pathways for genes correlated with loss of tumor differentiation

Similar to our search for significant annotation terms, we searched both the positively correlated and the negatively correlated sets of genes separately for significantly enriched KEGG pathways. Table 3 displays the most significant pathways found for the set of negatively correlated genes, which pathways include *proteasome*, *cell cycle*, *pyrimidine metabolism*, *oxidative phosphorylation*, *glyoxylate* and *dicarboxylate*

Table 2
Significantly enriched annotation terms found for the set of 462 genes with significant negative correlations with tumor cell differentiation (with a coefficient value less than −0.3 and a significance value less than 0.005)

Category	Term	Total	Found	*P*-value	Expected
Biochemical function	RNA-binding protein	114	30	1.56E-07	0
Biochemical function	Proteasome subunit	30	14	1.58E-07	0
Cellular component	26S proteasome	32	14	4.27E-07	0
Cellular role	Nuclear-cytoplasmic transport	20	9	4.07E-05	0.008
Cellular role	DNA synthesis	51	15	5.24E-05	0.014
Molecular localization	RNA-associated	121	26	5.55E-05	0.014
Subcellular localization	Nuclear transport factor	4	4	8.36E-05	0.018
Subcellular localization	Nuclear	476	70	0.000106	0.031
Molecular function	Small nuclear ribonucleoprotein	10	6	0.000113	0.033
Subcellular localization	Mitochondrial	149	29	0.000139	0.036
Biochemical function	Ligase	33	11	0.000153	0.039
Biological process	Protein folding	24	9	0.000223	0.046
Cellular component	Nuclear pore	15	7	0.000232	0.048
Biological process	Ubiquitin-dependent protein degradation	12	6	0.000421	0.107
Cellular role	RNA splicing	37	11	0.000475	0.121
Cellular component	Spindle	9	5	0.000721	0.157
Molecular function	DNA replication factor	3	3	0.000877	0.189
Cellular role	Protein degradation	106	21	0.000904	0.36
Cellular role	Protein synthesis	128	24	0.000913	0.362

The number of occurrences for each term within the set of genes profiled on the HuGeneFL chip is listed, along with the number of genes found in the set of 462 that fall under the term. The *P*-value for each term was calculated using the hypergeometric formula, and the number of terms that could be expected to have a *P*-value less than the given *P*-value in a set of 462 randomly selected genes was calculated using simulation tests.

metabolism, *folate biosynthesis*, *glutamate metabolism*, *one carbon pool by folate*, *purine metabolism*, and *aminoacyl-tRNA biosynthesis*. If we consider these pathways to be significant over any other pathways that could appear similarly enriched within a set of randomly selected genes, then we may expect on the order of one of these pathways to appear significant by chance. As with the search for annotation terms, no truly significant pathways were found for the positively correlated genes. The complete set of enriched pathways found for both sets of genes, including which genes belong to which pathways, is available at our web site [12].

4. Discussion

The results of our analysis implicate on the order of hundreds of genes as being involved with a loss of differentiation in cancer, either directly or through processes indirectly associated with the progression of the disease. Furthermore, we found numerous classes of genes, defined by Locuslink annotation or KEGG pathway, that appear disproportionately within the set of genes showing the strongest negative correlations with tumor differentiation, to the extent that the expression of these genes is not likely to have resulted merely from an overall deregulation of gene transcription within tumor cells, but instead appears to be representative of a systematic up-regulation of several biological processes that can be linked to a loss of differentiation and an increase in cell proliferation in tumors. Conversely, if there had been no significant classes found associated with the negatively correlated genes, it might have been inferred that the genes as a whole have no particular biological features that distinguish them from any randomly selected set of genes. While DNA copy number has recently been shown to have a major direct role in the alteration of the transcriptome in cancer [6], the increased expression of genes associated here with the progression of the disease appears to represent more than the random amplification of individual genes irrespective of the processes in which they participate, though amplicons may include genes such as those that encode transcription factors, which can have systematic effects.

Table 3
Significantly enriched pathways found for the set of 462 genes with significant negative correlations with tumor cell differentiation

KEGG id	Category	Title	Total	Found	*P*-value	Expected
hsa03050	Sorting and degradation	Proteasome	24	14	4.83E-10	0.009
hsa04110	Cell growth and death	Cell cycle	55	16	4.9E-06	0.01
hsa00240	Nucleotide metabolism	Pyrimidine metabolism	50	15	6.38E-06	0.011
hsa00190	Energy metabolism	Oxidative phosphorylation	55	14	9.85E-05	0.016
hsa00630	Carbohydrate metabolism	Glyoxylate and dicarboxylate metabolism	4	3	0.002071	0.116
hsa00790	Metabolism of cofactors and vitamins	Folate biosynthesis	13	5	0.002707	0.19
hsa00251	Amino acid metabolism	Glutamate metabolism	19	6	0.003202	0.219
hsa00670	Metabolism of cofactors and vitamins	One carbon pool by folate	10	4	0.006332	0.42
hsa00230	Nucleotide metabolism	Purine metabolism	88	14	0.01184	0.889
hsa00970	Transcription	Aminoacyl-tRNA biosynthesis	18	5	0.012783	0.916

The number of genes that are profiled on the HuGeneFL chip for each pathway is listed, along with the number of genes found in the set of 462 that belong to the pathway. The *P*-value for each pathway was calculated using the hypergeometric formula, and the number of pathways that could be expected to have a *P*-value less than the given *P*-value in a set of 462 randomly selected genes was calculated using simulation tests.

A set of significantly enriched gene classes being associated with the genes showing the strongest positive correlations to tumor differentiation would be something of interest, as such classes might implicate one or more processes as acting counter to the processes related to a loss of differentiation. Though the most enriched classes found here for the positively correlated genes do make some sense from a biological standpoint (e.g. for classes related to cell adhesion, a less differentiated cell may be less connected to its adjacent cells), none of these classes could be considered strongly significant from a statistical standpoint, especially compared to the significance of the classes associated with the negatively correlated genes. This fact might suggest that there are no genetic programs that may become activated to keep the cell from becoming less and less differentiated, at least no known programs that could adequately be described in terms of either the Locuslink annotations or the KEGG pathways.

One model that could help explain there being no well-defined processes strongly associated with a gain of tumor differentiation is that the loss of transcription of individual genes normally expressed in the more differentiated cells may be indicative of the process of natural selection that can underlie the progression of the disease [7], as genes that offer no competitive advantage to a cancer cell may be selectively removed from the transcriptome in a random manner, as in the case of gene deletion. If a deleted gene happens to be a transcriptional repressor, then the net effect could be a systematic up-regulation of the set of genes that would normally be repressed, which may account in part for the observation made here of many more genes being negatively correlated to tumor differentiation than being positively correlated.

Taken together, the set of significantly enriched annotation terms found for genes negatively correlated to differentiation are indicative of an overall process of increased cell proliferation and increased protein synthesis and degradation. This finding helps confirm what should already be known, as cells that are continuously dividing will lose the characteristics of a differentiated cell, and as protein degradation plays a critical role in cell proliferation in cancer [8]. The observation of the cell cycle and proteasome being included in the set of significant pathways reinforces the conclusions drawn from the set of significant annotation terms. Other pathways that appear significant may be indicative of the increased energy requirements of proliferating cells (e.g. oxidative phosphorylation), but other significant pathways with links to cancer suggested in previous studies may merit further investigation. In particular, glutamate antagonists have been shown to limit tumor growth [9], and antimetabolites of purine and pyrimidine nucleotide metabolism can cause many types of cancer cells to terminally differentiate into mature, non-proliferating cells [10]. In addition, many studies suggest that the cellular level of folate, which has a central role in biological methylation and nucleotide synthesis, can modulate carcinogenesis [11].

Our study illustrates an exploratory approach of analyzing global gene expression data by looking for known classes of genes that appear over-represented within a set of genes of interest, which can provide clues as to the types of processes that underlie the expression of the genes and can stimulate new hypotheses from which a directed study to address a specific question may follow.

References

[1] Harpole Jr., D.H., Herndon 2nd, J.E., Wolfe, W.G., Iglehart, J.D. and Marks, J.R. (1995) Cancer Res. 55, 51–56.

[2] Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J.M., Iannettoni, M.D., Orringer, M.B. and Hanash, S. (2002) Nat. Med. 8, 816–824.

[3] Pruitt, K.D. and Maglott, D.R. (2001) Nucleic Acids Res. 29, 137–140.

[4] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Nat. Genet. 25, 25–29.

[5] Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002) Nucleic Acids Res. 30, 42–46.

[6] Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L. and Brown, P.O. (2002) Proc. Natl. Acad. Sci. USA 99, 12963–12968.

[7] Simpson, A.J. and Camargo, A.A. (1998) Semin. Cancer Biol. 8, 439–445.

[8] Naujokat, C. and Hoffmann, S. (2002) Lab. Invest. 82, 965–980.

[9] Rzeski, W., Ikonomidou, C. and Turski, L. (2002) Biochem. Pharmacol. 64, 1195–1200.

[10] Hatse, S., De Clercq, E. and Balzarini, J. (1999) Biochem. Pharmacol. 58, 539–555.

[11] Choi, S.W. and Mason, J.B. (2002) J. Nutr. 132, 2413S–2418S.

[12] http://dot.ped.med.umich.edu:2000/pub/diff/index.htm