

Running Title: Cellular Pathways in Cancer

Can we identify of Cellular Pathways Implicated in Cancer using Gene Expression Data?

Nigam Shah¹, Jorge Lepre², Yuhai Tu² and Gustavo Stolovitzky^{2*}

1 Pennsylvania State University, University Park, PA 16802, USA

2 IBM Computational Biology Center, T.J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA

Keywords: gene expression; signalling pathways; functional genomics; human cancer.

Abbreviations: SOI: set of interest.

**Corresponding Author:* gustavo@us.ibm.com, (914) 945-1292 (off); (914)945-4104 (FAX)

Abstract

The cancer state of a cell is characterized by alterations of important cellular processes such as cell proliferation, apoptosis, DNA-damage repair, etc. Some of these alterations involve modifications of the expression of genes that participate in the pathways responsible for these processes. From this simple observation it follows that the expression of genes associated with cancer related pathways should exhibit differences between the normal and the cancerous states. We explore various means to find those differences. Interestingly, these differences can only be identified when groups of genes, as opposed to isolated genes, are considered. Instead of using all the individual genes on a DNA array in comparing cohorts of cancer patients with control subjects, our analysis searches for signals within small subsets of genes that are associated with specific pathways, thereby substantially increasing the sensitivity of the analysis, while preserving the biological information contained in the pathways. We analyze 6 different pathways (p53, Ras, Brca, DNA damage repair, NF κ b and β -catenin) and 4 different types of cancer: colon, pancreas, prostate and kidney. Our results are found to be mostly consistent with existing knowledge of the involvement of these pathways in different cancers. Our analysis constitutes proof of principle that it may be possible to predict the involvement of a particular pathway in cancer or other diseases by using gene expression data. Such method would be particularly useful for the types of diseases where biology is poorly understood.

Introduction

The use of DNA microarrays to measure genome-wide RNA expression levels has become an established research methodology in genomics.^{[1], [2], [3]} Within the biomedical sciences, cancer research has benefited enormously from the use of high-throughput gene expression studies. One of the main aims of high throughput gene-expression in cancer research is to identify genes and gene expression profiles that show a consistent change between cancer and normal states. Much effort^[4-17] has been devoted to develop data analysis techniques to carry out this research program. Indeed, many methods have been successfully used to isolate genes that seem to be differentially expressed in the transformed cells. However, even though these results are useful for staging, prognosis and diagnosis purposes, they shed little if any light on the underlying biology^[16] due to the inherent noise in the technology^[18] and the large number of genes which sometimes are irrelevant for the task at hand. Without prior knowledge, it is extremely difficult to assert whether these signals are artifacts or truly involved in the causation of the cancer. The power of the microarray technology, the ability to measure expression of thousands of genes at the same time, thus becomes a potential major hindrance.

In this paper we propose an approach to bridge in part the gap between the end result of a gene expression study and our ability to transform these results into understanding of the underlying biology. In particular, we want to determine whether a given pathway is involved in a given type of cancer by investigating the gene expression profiles of only those genes that are associated with the specific pathway. By focusing on a relatively small set of genes collected by using by prior biological knowledge of a given pathway, our analysis is more sensitive because we exclude the majority of the irrelevant genes on the chip.

Our approach is composed of three major steps. The starting step is a careful selection of the genes that are known to be associated with the pathways that are crucial in the regulation of important cellular processes such as cell proliferation, apoptosis, DNA-damage repair, etc. Next, using only the genes that are associated with a specific pathway, we develop three different methods to find signatures in the microarray data in terms of the degree of differential expression between cancer and control samples. Finally, to assess the statistical relevance of our findings in the previous step, we create *pseudo-pathways* by randomly picking the same number of genes as in the real pathway from the thousands of genes available from the same microarray. The statistical relevance is determined by the likelihood of finding the same signatures in the pseudo-pathways.

We applied our method to gene expression data for four different cancers (colon, pancreas, prostate and kidney) and for six different pathways (p53, Ras, Brca, DNA damage repair, NF κ b and β -catenin). When we find significant signatures for a given pathway in a particular type of cancer, they are usually supported by the existing biological knowledge. Our approach, therefore, could be useful as a first pass tool to determine whether a particular pathway may be involved in the causation or underlying biology of those cancers (or diseases in general) where the biology is poorly understood.

Material and Methods

Expression data. Two previously reported sets of microarray data were used in this study. The first set was described in^[5]. This data set contains 218 tumor and 90 normal tissue samples

covering 14 different tumor types. One tumor type (melanoma) does not have control samples, the CNS data is from heterogeneous tumor types and 8 tumor types have a very unbalanced number of tumor *versus* normal tissues. We decided to include in our study only those datasets that had roughly equal number of tumor and normal samples. *That left us with four tumor types: Colon (11 cancer, 11 control), Pancreas (11 cancer, 10 control), Prostate (10 cancer, 9 control) and Kidney (11 cancer, 13 control), which met our data requirements of homogeneity in the tumor type and balance between tumor and normal sets.* The hybridizations^[5] were done using the Hu6800 and Hu35KsubA GeneChips from Affymetrix. The second set of microarray data used in this paper pertains to colon cancer data with known p53 mutation status. This data, first reported in^[19], contains 18 cancer and 18 control samples. We used only the adenocarcinoma data, which was generated using the Hu6500 GeneChip.

Pathway information. We considered six pathways: p53, Ras, NF κ b, Brca, β -catenin and DNA-repair pathway, whose derangement could be related to causation of cancer. Note here that the term ‘pathway’ is used in its most general sense, e.g., the ‘p53 pathway’ is not a real biological signaling pathway in the strict sense. However because of the importance of p53 and its associated genes in the cancer process, it is beneficial to treat the group of genes associated with it as a “pathway”. For each of these pathways, we compile a list of *associated genes* by gathering a substantial amount of biological information from a variety of sources, including journal articles and various databases. The information sources are listed in Table 1. We refer to each one of these lists as a ‘set of interest’ (SOI). In our study the criteria for inclusion of a gene in an SOI for a pathway was kept loose to include as many genes as possible for a pathway to enhance the chance of detecting a signal. This is necessary because the number of genes in SOI is limited by the list of genes probed by a given microarray experiment. For example, in the search for the p53 associated genes, the literature yielded 188 implicated genes, of which only 60 could be mapped to one of the GeneChips used in the study. (Genes like PTEN, had no associated probe in the array, and thus were excluded.) The GeneChips can have more than one probe per gene, e.g. JUN had 3 probes associated with it, and hence the total number of gene probes for the p53 pathway was 105. Table 2 details the number of gene probes for each SOI (diagonal) as well as the number of probes that are shared by more than one SOI. The percentages in Table 2 refer to the percentage of overlap in relation to the smallest SOI. For the p53 pathway, we also analyzed another data set (Notterman et al., 2001) that used a different GeneChip, the HU6500. Therefore we needed to compile a second SOI (with 47 probes) for this chip corresponding to the p53 pathway genes.

P/A filter. Each gene in an Affymetrix GeneChip is assigned a status or *call* of either Present (P) or Absent (A) or Marginal (M). When the call for a gene is A, the corresponding numerical value of the average difference is very unreliable. In our analyses we only accept a gene for further analysis if there is a minimum fraction of subjects for whom that gene’s call was either P or M. Only genes whose percentage of absent calls is less than 70% of the total data set (counting both the case and control sets) are further processed. The filtered out genes are not further considered in the signature detection methods.

Logarithmic transformations. In some of the signature detection methods described below, the logarithm to the base 10 of the average differences was used. Because average differences can be negative (the Affymetrix MAS 4.0 software suite was used to process the data after

hybridization), we set the data to take the value 1 when the actual average difference was less or equal to 1.

Gene Expression Data Analyses. Expression values were used as reported in the source papers without any data preprocessing. For each cancer set analyzed, we have two groups of subjects: the phenotype or case set (denoted below by C1), and the control set (denoted below by C2). Below we describe three signature detection methods. Each method consists of two parts: feature selection and statistical evaluation. The latter will allow us to assign a p -value to the discovered signature.

Feature Selection I: Signal to noise method. This analysis is a univariate treatment of gene expression in which each gene is examined independently to check whether it is differentially expressed between classes C1 and C2. This analysis addresses the question of whether the difference between the mean expression level in C1 and C2 is large when compared with the average variability of expression around the mean in each class. In other words, we estimate the extent of overlap between the distribution of expression values of a gene in C1 with its distribution in C2. A simple measure of this overlap, which we call the signal to noise ratio s , is given by the formula

$$s = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2},$$

Where μ_1 is the mean logarithmic expression of the gene in cancer tissue (class C1), μ_2 is the mean logarithmic expression of the gene in normal tissue (class C2), σ_1 and σ_2 are the standard deviations of the logarithmically transformed expression values in C1 and C2 respectively.

Feature Selection II: Pair-wise correlation method. This analysis is a bivariate treatment of gene expression in which pairs of genes are considered to check whether their expression across both classes C1 and C2 is considerably correlated. The correlation r between gene X and gene Y is given by the formula

$$r = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}},$$

Where x_i and y_i are respectively the gene expression values of genes X and Y in the i -th sample, and i spans all the samples in both C1 and C2. \bar{x} and \bar{y} are the mean of the gene expression value for genes X and Y across all the samples in both C1 and C2. We used the absolute value of r in our subsequent calculations. The rationale for this measure is that the correlation r between genes X and Y will be larger if there is a differential behavior in gene expression between cancer patients and controls subjects for both X and Y.

Feature Selection III: Pattern discovery method: the Genes@Work algorithm. We used the gene-expression pattern discovery algorithm Genes@Work^[15] (available for download from <http://www.research.ibm.com/FunGen/index.html>) as our basic tool to perform multivariate gene selection. A Genes@Work pattern consists of a subset of genes that express differentially in a subset of the phenotype set with respect to the control set. Differential expression is determined as follows: for each gene g , a gene expression probability density $p_g(e)$ is computed empirically

from the control set, where e is the expression level. The algorithm discovers *all* the subgroups of patients within the phenotype group for which there is a subset of genes with the following property: for every gene g in this subset of genes the integral of $p_g(e)$ over the expression range in the subgroup of patients is less than a user-defined threshold δ . In other words, each gene in a pattern expresses at a similar level in the patient subset, the level of similarity given by δ . The number of patients that compose a pattern is called its support. The patterns reported by Genes@Work are *maximal* in the number of genes and in support: we cannot add another gene to a pattern without decreasing its support, and we cannot increase the support of a pattern without decreasing the number of genes participating in it. The statistical significance of a Genes@Work pattern is given by its p -value, defined as the probability of observing one or more similar patterns (i.e., a pattern with the same number of genes and support) in a set of random patients whose genes g 's are distributed according to the probability density function $p_g(e)$ (the null hypothesis^[15]). The input parameters for Genes@Work are the minimum number of genes and the support of a pattern, the degree of similarity of gene expression for each gene, δ , and the maximum p -value allowed. Given a set of input parameters for Genes@Work, it is possible that no pattern is found. The protocol we used for exploring the parameter space is described next. We start to look for a pattern supported by at least half the patients, one third of the genes in the SOI and $\delta=0.001$. If no pattern is found δ is doubled, until we reach a delta of 0.128, each time looking for presence of patterns. If no patterns are found during this process we decrease the number of genes that are required to be in the pattern by 1 and restart the process with $\delta=0.001$ (the 'supported by half the patients' condition remains the same). This process is iterated until we find a pattern in the cancer data set supported by half the patients and maximum number of genes. There has to be a minimum of 2 genes in a pattern. If we do not find a pattern of at least 2 genes we declare that no patterns are found.

Statistical significance of the feature selection results. In order to assess the statistical significance of finding K features (genes, pairs of genes or patterns) for a particular SOI in a particular cancer, we generate M gene sets, which we call *pseudo-pathway*. Each pseudo-pathway contains as many genes as genes are in our SOI. The genes in the pseudo-pathway are picked randomly from the original set of genes in the GeneChip used in the assay. In this way, the genes within the pseudo-pathways do not carry any biological signature, nor are they associated with a unique biological process as the genes in an SOI do. The same feature selection method that produced K features in our SOI in a particular cancer type is now applied to the M pseudo-pathways and the same cancer, yielding a probability $P(n)$ of finding n features when the feature selection step is applied to the pseudo-pathways. (M was set to be 1,000 for the signal to noise and pair-wise correlation methods; M was 100 for the pattern discovery method.) We can then to assign a p -value for our K features in the original SOI, according to the formula:

$$p - \text{value} = \sum_{n \geq K} P(n)$$

An SOI is implicated in a cancer by a particular feature selection method, if the p -value for the corresponding SOI and that cancer is less than or equal to a threshold, $P_0=0.05$.

Results

We compiled sets of interest (SOI) corresponding to six pathways of importance in cancer research^[20] and cell biology in general. These are the p53 pathway, involved in DNA damage checkpoint and regulation of apoptosis^[21]; the RAS pathway, involved in cell proliferation,

differentiation and cell morphology^[22]; the BRCA pathway, formed by genes associated with the BRCA1 and BRCA2 oncogenes and implicated in DNA repair, recombination and transcription^[23]; the NF κ b pathway, associated with regulation of cell differentiation and apoptosis^[24]; the β -catenin pathway, involved in cell proliferation, cell cycle progression and cell fate in development^[25]^[26]; and the DNA-repair pathway, which controls the repair of environmental damage to DNA^[27].

For each of the compiled SOI, and each of the cancers studied we looked for signatures in the data using the three feature selection schemes described in Materials and Methods.

Results of the signal to noise based feature selection scheme. For each SOI and for each particular cancer dataset, we count the number of genes K that pass the P/A-filter and whose signal to noise ratio is above threshold values 1 (We also considered threshold values of 2 and 3, with similar results –data not shown). We estimated the p -value of the number of features found as described in Material and Methods. The results are shown in Table 3. The entries equal to 1 in Table 3 indicate that no gene was found in the corresponding SOI with a signal to noise higher than the corresponding threshold. For all the thresholds, most of the p -values are quite large. This means that the number of genes in the SOI with large signal to noise values is comparable with what was expected from a random assortment of genes without any association with each other. This is possible as the involvement of a particular gene in a given cancer may not occur in all cases of such cancer and certainly not at all stages of the pathogenesis. Therefore there may not be a clear separation of expression values between the phenotype (C1) and the control (C2) cells for individual genes, since C1 contains samples from different patients at different stages of the cancer

Results of the pair-wise correlation based feature selection scheme. For each SOI and for each cancer, we apply the P/A-filter, and for the N genes that pass, we calculate the pair-wise correlation between all pairs, $\binom{N}{2}$ and count the number of pairs K , whose correlation absolute value is bigger than 0.9. (We followed the same procedure at correlation cutoffs 0.85, 0.9 and 0.95 before deciding on 0.9 –data not shown). We estimated the p -value of the K features as described in Material and Methods. The results are shown in Table 4. We see that there are a considerable number of statistically significant signatures. p -values less than 0.001 correspond to cases when none of the pseudo-pathways exhibited any pairs with correlation larger than 0.9. Notably, all the SOIs had a signature for two types of cancer, except for β -catenin for which no correlated pair above 0.9 was found. Upon closer inspection of the β -catenin SOI, we found that only a small number of the 30 probes passed the P/A filter, leaving only a few probes, thus reducing the actual ability of the method to find any signature, even if there is one, and constitutes one of the potential sources of false negatives in this study.

Results of the pattern-discovery based feature selection scheme. For each SOI and each cancer, we first apply the P/A-filter and then we search for the largest ‘pattern’ of gene expression profiles supported by at least half of the patients. The signature in this method is defined on the basis of similarity of gene expression within the cancer panel, and consistency in change in expression level between normal and cancer panels. Once a pattern is found in the SOI (see Materials and Methods for choice of parameters for the algorithm Genes@Work) we generate 100 pseudo-pathways and using the parameters at which we found the patterns, we search in

these random sets for presence or absence of patterns of similar size. The fraction of pseudo-pathways that have a pattern of similar or larger size than that in our SOI constitutes an estimate of the p -value for the signature in the SOI. Table 5 shows the results of the pattern discovery method.

Comparison between the methods. A cursory look at the results from the three methods suggests that multivariate methods are better than single gene based methods to find signatures in a collection of genes associated with a pathway. Of the signatures obtained using the multivariate methods, out of the 24 possible conditions in which a signature could be found, we found one in only 12 conditions. Ten of those conditions correspond to signatures detected using the pair-wise correlation, whereas 7 were discovered using the pattern discovery method. The two methods yielded a statistically significant signature simultaneously in 5 conditions. This shows that the two methods, while coinciding in about 50% of their predictions, are complementary in that they can pick different signatures. Furthermore, where they coincide the genes participating in the signature tends to be the same. Table 6 shows this coincidence in some detail, e.g. out of the 10 genes associated with patterns in the p53 pathway in pancreas, 9 are part of the 23 genes that formed a correlated pair above threshold according to the correlation method. (The list of genes forming the signatures is available from the authors upon request.)

Comparison between in-silico identification and experimentally validated involvement of pathways in cancer. In order to find out whether the statistically significant signatures we discovered make biological sense, we conducted a thorough search of the literature for involvement of each of the six pathways in the four cancers. We identified studies in which the 'key' member of the pathway (or its SOI) is examined experimentally in a particular cancer. If the 'key' member is affected in some manner (in most cases by a mutation), then it is reasonable to expect that some of the genes associated with the 'key' member could show changes in expression as the key members are usually transcription factors (e.g. p53, nfkb) themselves or are proteins that affect transcription indirectly (e.g. Ras). Therefore, if one of the key members of a pathway was found to be abnormal (for example, a point mutation) in a particular type of cancer, it was considered that the pathway maybe involved in that particular type of cancer. These direct biological evidences are compared with the statistically significant signatures discovered in the corresponding SOI, the results are as summarized in Table 7. The details of this search and comparison can be requested from the authors. Overall the agreement is very encouraging. If we ignore the signatures found in the Ras and β -catenin pathway (for reasons discussed later) then there are 12 instances where we expected a signature, 3 where we do not expect a signature, and 1 where there are conflicting reports. Taking the union of the signatures reported by the pair-wise correlation method and the Genes@work method we are able to detect 8 of the 12 instances (i.e., 67% of true positives.) Alternatively, we do not report a signature where we do not expect to find one (Brca in kidney cancer, and NF κ b in prostate and colon cancer). The pair wise correlation method *does* predict an involvement in the controversial case of NF κ b in kidney cancer. In this case our analysis tends to support the notion that NF κ b is involved in some way in the development of renal cell carcinoma^[28]. In such a scenario the signature that we find using the pair-wise correlation method and the genes participating in that signature become interesting study targets.

In connection with NF κ b in pancreas it is interesting to note that all the 5 genes (PRG1, FBP1, THBS2, VIM, FBN1) Genes@work identified as forming a pattern in the NF κ b pathway have NF κ b binding sites in their promoter^[29-32]. These 5 genes, which are all up regulated in the patients that support the pattern, are also present in the pair-wise correlation signature (Table 6). This would support the notion that NF κ b induction is in some way involved in survival of pancreatic cancer cells (and such induction would cause the up regulation of the 5 pattern genes), in agreement with the experimental finding that NF κ b induction is involved in protecting pancreatic cancer cells from anti-CD59 and TRAIL mediated apoptosis^[33]. Only a few of the signaling cytosolic proteins of the nfkb pathway have representatives in the microarrays analyzed. They are NF κ b, TNF- α , TNFR and Ubiquitin Conjugation Enzyme. We did not detect these gene products in the feature selection procedures, presumably because their activity in this pathway is regulated at the protein level and not at the mRNA level. Still, those 5 genes that participate in the signature, which are downstream of NF κ b, are detected by both of our multivariate methods. This suggests some differential activity of the NF κ b pathway in the cancer cells when compared with normal.

One important false negative in Table 7 is p53 in colon cancer. In the colon cohort used to produce Table 3-7, the p53 status of the patients from which the data is collected is not known. We redid the analysis with colon cancer data where the p53 status is known. We used the Notterman *et al.*'s data set (18 cases and 18 controls). Of the 18 cancer patients, 12 had a p53 mutation. Using this data set, we *did* find a signature using the Genes@Work method.

Thus, if we take into account the results found using Notterman *et al.*'s data, we are able to find statistically significant signatures in 9 out of 12 cases where we expect to find signatures (75% of true positives). In 5 out of 9 instances the pair-wise correlation and the Genes@work method agree, and in all these instances there is extensive overlap in the genes that form respective signatures.

It was mentioned earlier that the results for the Ras and β -catenin pathways fall in a gray zone because there are too few genes (only 30) from our compiled lists that were mapped onto the microarray data. The number of participating genes in these SOIs is further reduced when we apply the P/A filter (for example, only 7 genes pass the P/A filter in colon data for the β -catenin pathway and 13 pass in pancreas data for Ras pathway). So the signatures found using these pathways may not be as robust as one would expect from a fuller set of genes characterizing the pathway.

Summary and Discussion

The advent of microarray technology, allows researchers to conduct more descriptive research based on these high throughput data. Though microarray experiments provide us with much information about the system under study, they often leave us with more questions than answers. The work presented in this paper is an attempt to leverage prior biological knowledge in answering the simple, yet important question: Does a certain pathway count among the processes that show derangement in cancer?

Looking for collective behavior between multiple genes increases the sensitivity of our analysis, as shown in our study, where only the multivariate methods are successful in detecting

statistically significant signals. On the other hand, it is highly desirable to focus on a relatively small set of genes because it reduces the noise caused by thousands of irrelevant genes probed by the microarray. Leveraging our analysis with prior biological knowledge, i.e., concentrating on the subset of genes that are known to be associated with a particular pathway seems to balance these two seemingly contradictory aspects of the analysis. However, the information contained in the rest of genes participating in the array is not discarded; instead they are used as a control to evaluate the statistical significance of signatures found within the pathway genes. Of the 4 (reliable) pathways and 4 cancers analyzed, a literature search showed 12 instances of involvement of pathways in cancer. Our method found a statistically significant signature in 9 of those 12, showing a reasonably low false negative rate of about 25%. In the case of the Brca pathway in renal cancer, where the involvement is believed to be minimal, and NF κ b pathway in colon and prostate cancer, where no difference in NF κ b transcript between cancer and control has been reported (though activation of the protein was reported) our method correctly predicts no involvement of the corresponding pathway. So we get three true negatives out of three.

Like other classification problems, the false positive and false negative errors in our method depend on data quality and the significance threshold in the p-value. In general, our method is more prone to false negative errors than false positive errors for a couple of reasons: (1) Our feature selection schemes are incapable in capturing certain signatures e.g. if the relationship between two genes is highly non-linear, in which case a more adequate measure of relatedness would be the mutual information. (2) The list of genes representing a pathway can be incomplete, either because previous literature had not identified some genes as participants in a pathway, and/or that some pathway-participating genes do not have a corresponding probe in the gene expression array. Indeed the latter is the reason that we do not put too much emphasis in the results pertaining to the Ras and β -catenin pathways. (3) Stratification of the cohort of cancer patients can also hinder the correct prediction. We saw that when analyzing the colon cancer cohort from Ramaswamy *et al.*, the p53 pathway did not show a signature. However, when we analyzed the p53 pathway in the subgroup with p53-mutated patients in Notterman's *et al.*'s colon cancer data, we did find an involvement. (4) Finally, the involvement of the pathway in the specific cancer maybe only at the protein level, which will certainly escape the detection of any transcription based analysis.

It is worth noting that in the case of renal cell carcinoma, there are conflicting reports about the involvement of the NF κ b pathway in the response of cell lines to TNF-related apoptosis-induced ligand (TRAIL). Our method does find a signature, thus giving some support to the study that postulates involvement of NF κ b in kidney cancer. This is the kind of potential application that we envision for our method: in cancers where there is an unclear pathogenesis, and for which the biological knowledge is scant, (e.g., melanoma), the proposed methodologies can help identify which pathways, and which specific genes, are implicated in the cancerous transformation.

Our method could be extended to include other criteria for grouping genes. For example, genes can be grouped if they belong to the same protein complex, if they contain the same promoter region, etc. If upon analysis of these sets of genes we find a signature, then it might be surmised that such complex, or promoter is differentially active in the particular cancer. We believe that this method and its possible extensions have great potential for functional genomics research.

Acknowledgements. We benefited from fruitful discussions with J. Jeremy Rice. NS acknowledges support from an IBM summer fellowship.

Tables:

Table 1: Sources of Pathway gene information

	P53	Ras	Brca	Nfkb	β -catenin	Dna-repair
Key Paper	11571296	0655059	11832208	10602468	12040179	11181991
Data-base	DIP		BCGD			
Product-catalogue	p53 Gene Array (TranSignal™)			NFkB Gene Array (TranSignal™ & GEMArray™)		

Sources of information from where the genes participating in the six pathways listed in the first row were collected. The 'Key Paper' row contains the reference number for the paper.

Table 2: Number of probes in pathways lists, including intersections

	p53	Ras	Brca	Nfkb	β -	DNA-
p53	105					
Ras	0	30				
Brca	8	0	57			
Nfkb	8 (9%)	0	2 (3%)	89		
β-catenin	0	1 (3%)	1 (3%)	0	30	
DNA-rep	4 (8%)	0	8	0	0	49

Table 3: p-values for signature in Signal-to-Noise method

Signal to noise threshold = 1, 2.

	p53	Ras	Brca	Nfkb	β -catenin	Dna-repair
Colon	0.91, 0.22	0.99, 1	0.77, 0.37	0.95, 0.19	0.99, 1	0.99, 1
Pancreas	0.06, 0.18	0.44, 1	0.82, 1	0.75, 1	0.69, 1	0.89, 1
Prostate	0.28, 0.01	0.43, 1	0.63, 1	1, 1	1, 1	1, 1
Kidney	0.46, 0.20	1, 1	0.85, 1	0.96, 1	1, 1	0.80, 1

Table 4: p-values for signature in pair-wise correlation method

Pair wise correlation simulations at 0.90 cutoff

	p53	Ras	Brca	Nfkb	β -catenin	Dna-repair
Colon	0.708	0.001	1	0.496	1	1
Pancreas	<0.001	0.058	0.045	0.002	1	0.007
Prostate	0.001	<0.001	0.025	1	1	<0.001
Kidney	0.268	1	0.975	0.018	1	1

Cells with p-value less than or equal to 0.05 are colored gray

Table 5: p-values for presence of 'pattern' in cancer using the Genes@work method

	p53	Ras	Brca	Nfkb	β -catenin	DNA-repair
Colon	0.81	0.65	0.02	0.36	0.29	0.65
Pancreas	0.01	0.33	0.01	0.01	0.14	0.53
Prostate	0.11	0.05	0.19	0.21	0.05	0.01
Kidney	0.08	0.1	0.06	0.11	1	1

Cells with p-values less than or equal to 0.05 are colored gray

Table 6: Common genes between the two methods

	p53	Ras	Brca	Nfkb	β-catenin	DNA-rep
Colon						
Pancreas	10/23/9		5/7/4	5/9/5		
Prostate		4/10/2				2/12/1

Genes in Pattern/Genes from Correlated pairs/Common genes.

Table 7: Involvement of various signaling pathways in the cancers under study and overlap between our methods and known information


	p53	Ras	Brca	Nfkb	β-catenin	DNA-rep
Colon	X(42%)	X(40%)	X(3.5%)		X(80%)	X(15%)
Pancreas	X(40%)	X(90%)	X	x		X(66%)
Prostate	X(15%)		X		x	x
Kidney	X(15%)	#	#	?	x (protein)	x

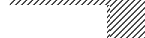
'X' = Involvement reported with an estimate of cases that have a mutation in the pathway. The parentheses contain the % of cases that were reported to have the mutation in the study.


'x' = Involvement reported without an estimate of the % of cases that have a mutation.

'#' = Involvement is believed to be minimal.

'?' = Contradictory evidence.

 Significant p-value in Genes@work method

 Significant p-value in Genes@work method only in Notterman et al. data

 Significant p-value in pair-wise corr

References

1. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*. Science, 1995. **270**(5235): p. 467-70.
2. Shalon, D., S.J. Smith, and P.O. Brown, *A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization*. Genome Res, 1996. **6**(7): p. 639-45.
3. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays*. Nat Biotechnol, 1996. **14**(13): p. 1675-80.
4. Yeang, C.H., et al., *Molecular classification of multiple tumor types*. Bioinformatics, 2001. **17 Suppl 1**: p. S316-22.
5. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures*. Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15149-54.
6. Klein, U., et al., *Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells*. J Exp Med, 2001. **194**(11): p. 1625-38.
7. Macgregor, P.F. and J.A. Squire, *Application of microarrays to the analysis of gene expression in cancer*. Clin Chem, 2002. **48**(8): p. 1170-7.
8. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns*. Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
9. Sherlock, G., *Analysis of large-scale gene expression data*. Curr Opin Immunol, 2000. **12**(2): p. 201-5.
10. Hoffmann, R., T. Seidl, and M. Dugas, *Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis*. Genome Biol, 2002. **3**(7): p. RESEARCH0033.

11. Holter, N.S., et al., *Fundamental patterns underlying gene expression profiles: simplicity from complexity*. Proc Natl Acad Sci U S A, 2000. **97**(15): p. 8409-14.
12. Raychaudhuri, S., J.M. Stuart, and R.B. Altman, *Principal components analysis to summarize microarray experiments: application to sporulation time series*. Pac Symp Biocomput, 2000: p. 455-66.
13. Brown, M.P., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines*. Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.
14. Mendez, M.A., et al., *Discriminant analysis to evaluate clustering of gene expression data*. FEBS Lett, 2002. **522**(1-3): p. 24-8.
15. Califano, A., G. Stolovitzky, and Y. Tu, *Analysis of gene expression microarrays for phenotype classification*. Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 75-85.
16. Kaminski, N. and N. Friedman, *Practical approaches to analyzing results of microarray experiments*. Am J Respir Cell Mol Biol, 2002. **27**(2): p. 125-32.
17. Szabo, A., et al., *Variable selection and pattern recognition with gene expression data generated by the microarray technology*. Math Biosci, 2002. **176**(1): p. 71-98.
18. Tu, Y., G.A. Stolovitzky, and U. Klein, *Quantitative noise analysis for gene expression microarray experiments*. Proc Natl Acad Sci U S A, 2002. **In press**.
19. Notterman, D.A., et al., *Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays*. Cancer Res, 2001. **61**(7): p. 3124-30.
20. Hahn, W.C. and R.A. Weinberg, *A subway map of cancer pathways*. 2002.
21. Vogelstein, B., D. Lane, and A.J. Levine, *Surfing the p53 network*. Nature, 2000. **408**(6810): p. 307-10.
22. Yamamoto, T., S. Taya, and K. Kaibuchi, *Ras-induced transformation and signaling pathway*. J Biochem (Tokyo), 1999. **126**(5): p. 799-803.
23. Venkitaraman, A.R., *Cancer susceptibility and the functions of BRCA1 and BRCA2*. Cell, 2002. **108**(2): p. 171-82.
24. Rayet, B. and C. Gelinias, *Aberrant rel/nfkb genes and activity in human cancer*. Oncogene, 1999. **18**(49): p. 6938-47.
25. Wong, N.A. and M. Pignatelli, *Beta-catenin--a linchpin in colorectal carcinogenesis?* Am J Pathol, 2002. **160**(2): p. 389-401.
26. Moon, R.T., et al., *The promise and perils of Wnt signaling through beta-catenin*. Science, 2002. **296**(5573): p. 1644-6.
27. Wood, R.D., et al., *Human DNA repair genes*. Science, 2001. **291**(5507): p. 1284-9.
28. Oya, M., et al., *Constitutive activation of nuclear factor-kappaB prevents TRAIL-induced apoptosis in renal cancer cells*. Oncogene, 2001. **20**(29): p. 3888-96.
29. Schafer, H., et al., *The promoter of human p22/PACAP response gene 1 (PRG1) contains functional binding sites for the p53 tumor suppressor and for NFkappaB*. FEBS Lett, 1998. **436**(2): p. 139-43.
30. Herzog, B., et al., *Characterization of the human liver fructose-1,6-bisphosphatase gene promoter*. Biochem J, 2000. **351 Pt 2**: p. 385-92.
31. Adolph, K.W., D.J. Liska, and P. Bornstein, *Analysis of the promoter and transcription start sites of the human thrombospondin 2 gene (THBS2)*. Gene, 1997. **193**(1): p. 5-11.
32. Chen, J.H., et al., *PEA3 transactivates vimentin promoter in mammary epithelial and tumor cells*. Oncogene, 1996. **13**(8): p. 1667-75.
33. Trauzold, A., et al., *CD95 and TRAIL receptor-mediated activation of protein kinase C and NF-kappaB contributes to apoptosis resistance in ductal pancreatic adenocarcinoma cells*. Oncogene, 2001. **20**(31): p. 4258-69.