Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2012 ; 9(5): 1399–1409. doi:10.1109/TCBB.2012.65.

Detecting Phenotype-Specific Interactions between Biological Processes from Microarray Data and Annotations

Nadeem A. Ansari,

Microsoft Corp., 16095 Cleveland ST #218, Redmond, WA 98052.

nadeem.ansari@microsoft.com

Riyue Bao,

The Department of Biological Sciences, Wayne State University, 5047 Gullen Mall, Detroit, MI

48202. rybao@wayne.edu

Călin Voichița, and

The Department of Computer Science, Wayne State University, 5057 Woodward Ave, Detroit, MI

48202. calin@wayne.edu

Sorin Drăghici

The Department of Obstetrics & Gynecology, Wayne State University, 3750 Woodward Ave.,

Detroit, MI 48201, the Department of Clinical and Translational Science, Wayne State University,

Detroit, MI 48201, and the Department of Computer Science, Wayne State University, 5057

Woodward Ave., Detroit, MI 48202. sorin@wayne.edu

Abstract

High throughput technologies enable researchers to measure expression levels on a genomic scale. However, the correct and efficient biological interpretation of such voluminous data remains a challenging problem. Many tools have been developed for the analysis of GO terms that are over- or under-represented in a list of differentially expressed genes. However, a previously unexplored aspect is the identification of changes in the way various biological processes interact in a given condition with respect to a reference. Here, we present a novel approach that aims at identifying such interactions between biological processes that are significantly different in a given phenotype with respect to normal. The proposed technique uses vector-space representation, SVD-based dimensionality reduction, differential weighting, and bootstrapping to assess the significance of the interactions under the multiple and complex dependencies expected between the biological processes. We illustrate our approach on two real data sets involving breast and lung cancer. More than 88 percent of the interactions found by our approach were deemed to be correct by an extensive manual review of literature. An interesting subset of such interactions is discussed in detail and shown to have the potential to open new avenues for research in lung and breast cancer.

© 2012 IEEE

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-10-0249. Digital Object Identifier no. 10.1109/TCBB.2012.65.

APPENDIX

SUPPLEMENTAL MATERIALS

A complete list of results in Excel format is available as online supplemental material on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2012.65>. The file consists of nine worksheets. The first worksheet describes the columns of the remaining worksheets; next four worksheets contain the results for the four weighting schemes when applied to Van't Veer (2002) breast cancer data set; the last four worksheets contain results for four weighting schemes when applied to Beer (2002) lung cancer data set.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

Keywords

Phenotype-specific interactions; biological processes; microarrays; gene ontology; single value decomposition

1 INTRODUCTION

HIGH-THROUGHPUT technologies such as microarrays allow researchers to measure the expression of thousands of genes simultaneously [1], [2]. Researchers, however, still face the challenge of translating such expression measurements into a better understanding of the underlying biological phenomenon. Existing annotations and gene ontologies such as GO [3] are essential in this step. An approach based on the identification of the biological processes that are significantly over- or under-represented in the set of DE genes is widely used for this purpose [4], [5], [6]. As of 2005, there were at least 15 such tools focusing on this type of ontological analysis [7]. Since then, even more tools have been developed. A more recent review [8] included not fewer than 68 tools in this area. Efforts have been made to improve this approach in various ways such as by including the entire list of DE genes rather than a subset of DE genes (e.g., GSEA [9] and related), taking into consideration the structure of the GO DAG [10], [11], taking into consideration sets of GO terms at a time [12], etc. A previously unexplored aspect is the identification of those relationships between biological processes that are significantly perturbed in a given condition. Here, we will present a first attempt toward this goal.

The technique described here aims to identify nontrivial relationships between biological processes represented by GO terms that are statistically significant in a given condition. The primary goal of this technique is to suggest novel hypotheses and directions for further exploration of the condition under study. A nontrivial relationship is a relationship that cannot be inferred from the directed acyclic graph (DAG) structure of the GO. Any relationships that can be directly inferred from the GO structure (e.g., a relationship between a given GO term and any of its distant ancestors or descendants) are not considered specific to any given condition and hence, are discarded by our technique.

A relationship between two biological processes, as predicted by our method, can also: 1) reveal a biological interdependence or interaction among GO terms that cannot be inferred easily from existing literature and that may be specific to the given condition; 2) suggest novel gene networks or possible interconnections in a gene network formed by the genes annotated with predicted associated terms; and 3) be used, in further studies, to predict novel gene annotation when one of the functions in the relationship is already known to be annotated with a gene and the other is not. As such, these functional relationships have the potential to broaden the knowledge base of research and enhance our understanding of biological phenomenon under study.

2 METHODS

2.1 Data Representation

The set of all genes considered for an experiment, e.g., the list of all genes present on the microarray used in the experiment, along with their annotations form a reference data set R . Gene annotations are maintained in various public repositories such as the ones provided by the GO Consortium [3]. Here, we used Onto-Express [4] in order to obtain all biological processes in GO for the genes in our data set. We represent these genes and their functional annotations in a reference gene-function matrix GF^R with M rows and N columns. The rows and columns of this matrix represent the genes and biological processes, respectively. The

“unknown” biological process was discarded because it conveys no useful information. The element a_{ij} of this matrix is 1 if and only if gene i is involved in the biological process j .

$$GF_{M \times N}^R = \{a_{ij}\} = \begin{cases} 1 & \text{if gene } i \text{ is annotated with GO term } j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The construction of gene-function matrix (GF^R) is similar to the one described in our previous work [13] with the exception that here we do not propagate gene annotations to parent terms up in the GO graph. This matrix can also be written as

$$GF_{M \times N}^R = (f_1 f_2 \dots f_j \dots f_N), \quad (2)$$

where $f_j = (a_{1j} a_{2j} \dots a_{Mj})^T$ is a function vector that describes the biological process j .

In a given experiment, a subset of genes of interest is selected from the reference data set, R . Such a subset can be formed for instance, with the genes found to be differentially expressed (DE) between two phenotypes compared. The list of selected genes and their annotations form an experiment data set $E \subset R$. We use the genes and biological processes in E , to form a gene-function matrix GF^E that will describe the given experiment

$$GF_{m \times n}^E = \{a_{ij}\} = \begin{cases} 1 & \text{if DE gene } i \text{ is annotated with GO term } j \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In this formula, m is the number of selected genes of interest that are known to be involved in n biological processes. Since the set of genes in E is a subset selected from the reference gene set in R , $m < M$ and $n \leq N$.

Each of the matrices of biological function vectors, described above, represents a vector space model. In principle, a number of different matrix factorization techniques could be applied here in order to try to extract hidden relationships while also reducing the influence of the noise present in the data. Such methods include non-negative matrix factorization (NMF) [14], principal component analysis (PCA), spectral clustering, and singular value decomposition (SVD), among others. Depending on the various choices of the various parameters the above methods are related to various degrees [15], [16]. In this work, we have chosen to use SVD, mainly because SVD was proven to work well in the information retrieval (IR) context [17], [18], [19], as well as in our previous work on the semantic analysis of functional annotations of gene [20]. Even though we have chosen to use SVD here, other matrix factorization techniques may work equally well or better.

Each of the two gene-function matrices above will then be decomposed using a SVD.

2.2 Singular Value Decomposition and Dimensionality Reduction

The SVD of a $p \times q$ gene-function matrix GF , where $p \leq q$, is a factorization of the form [21]:

$$GF_{p \times q} = G_{p \times l} \times S_{l \times l} \times F_{l \times q}^T, \quad (4)$$

where G and F are the matrices of left and right singular vectors, and S is the $l \times l$ diagonal matrix of singular values. Here, l is the rank of GF (i.e., the number of linearly independent rows or columns), which is at most $\min(p, q)$. The matrix S also decouples the expression of each gene (function) vector from all other gene (function) vectors. Graphically, the SVD decomposition is represented in Fig. 1.

The diagonal values of matrix S have the property that $s_1 \geq s_2 \geq \dots \geq s_l$. In essence, each singular value captures the amount of variability in the data along the direction of the corresponding singular vector. Several of these singular values in the S matrix are small and can be ignored without a significant loss of the features characterizing the data. The choice of the threshold used for the variance has to be a good compromise between the need to eliminate the variance due to the noise while also striving to preserve most or all of the variance due to the underlying biological phenomenon. A value around 80 percent was shown to work well in GO-related data mining applications [20]. Hence, in this work we kept the first k largest singular values of the matrix S corresponding to 80 percent of the variance in the data. In other words, k is the smallest index for which

$$\frac{\sum_{i=1}^k |s_i|}{\sum_{i=1}^l |s_i|} \geq 0.80. \quad (5)$$

The other singular values from $k+1$ to l , together with the corresponding columns and rows in the matrices G and F^T were eliminated, as shown in Fig. 2.

This allows a projection of the data in a lower dimensional space without significant information loss. This reduced-dimensionality matrix will be

$$GF_k = G_k \times S_k \times F_k^T. \quad (6)$$

Note that this approximation captures the most important features of the data without losing any gene or function since the reduced gene-function matrix GF_k has the same size as that of the original gene function matrix. Here, the subscript k is used to denote the fact that these matrices are those corresponding to the first k singular values rather than to indicate the dimensionality of the matrix as used in (4) and before. This procedure is applied to both the original reference GF^R , and experiment GF^E matrices.

2.3 Covariances and Correlations between Biological Processes

For each of these two GF_k matrices, we can first center the data in their columns, and then calculate a covariance matrix between all biological processes

$$FF_k = \{\sigma_{ij}\} = GF_k^T \times GF_k, \quad (7)$$

as well as the corresponding correlation matrix

$$R = \{r_{ij}\} = \left\{ \frac{f_i^T \times f_j}{\sqrt{\sigma_{ii} \cdot \sigma_{jj}}} \right\} = \left\{ \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \cdot \sigma_{jj}}} \right\}, \quad (8)$$

where f_i and f_j are two arbitrary function vectors, σ_{ij} their covariance, σ_{ii} and σ_{jj} the corresponding variances, and r_{ij} their correlation coefficient.¹

In essence, this generic correlation matrix captures the correlations between biological processes in a given state described by an arbitrary gene function expression matrix. We will use this approach to calculate two specific correlation matrices. The first such matrix, denoted by R^E , is calculated from GF^E and captures the correlations between all biological

¹Here, R is a matrix of elements representing the correlation coefficients between the biological processes. However, when R is used as a superscript, it denotes the reference data set.

processes in the given condition. The second such matrix, denoted by R^R , is calculated from GF^R and encodes the structure found in the annotation.

Since the genes and functions represented by the GF^E matrix are subsets of the genes and functions represented by the GF^R matrix, we reconstruct the R^R matrix to retain only those biological processes that are present in R^E , and call it $\widehat{R^R}$. Alternatively, $\widehat{R^R}$ matrix can be obtained by computing the pair-wise correlation coefficients between function vectors in the reduced size $GF^R_{M \times n}$ matrix. The element $r_{ij}^R \in \widehat{R^R}$ represents the correlation between biological processes i and j in the reference condition, based solely upon the known annotations. In contrast, the element $r_{ij}^E \in R^E$ represents the correlation between the same biological processes in the condition under study, based on the differentially expressed genes known to be involved in biological processes i and j .

2.4 Identifying Significant Changes

We are interested to identify those pairs of biological processes whose correlation has significantly changed between the reference and the given experiment. Fisher's z -transform [22] can be used to map the correlations into variables that follow a normal distribution

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right). \quad (9)$$

Here, z is the Fisher transformed value of a correlation coefficient element r . The standard

error of this distribution is $\sqrt{\frac{1}{\eta - 3}}$, where η is the number of data points used to compute r .

We compute z_{ij}^R and z_{ij}^E values from the elements r_{ij}^R and r_{ij}^E , respectively, using (9). Here, we consider that the z_{ij}^R value obtained in the absence of any fold changes is the mean of the normal distribution followed by this particular correlation, between functions i and j , when random changes occur between their annotated genes. Our null hypothesis is that the z_{ij}^E observed in the given condition for a particular pair of biological processes is due to noise alone. The research hypothesis is that the interaction between the biological processes i and j is significantly changed in the given condition. The Z statistic to test this can be calculated as follows:

$$Z = \frac{z_{ij}^E - z_{ij}^R}{\sqrt{\frac{1}{m-3}}}. \quad (10)$$

Here, m is the number of data points (genes) used in the computation of r_{ij}^E . We used the False Discovery Rate (FDR) [23] method to correct the Z -statistic p-value for multiple comparisons.

In order to assess the significance of the correlation change between the experiment and reference, while taking into consideration the potentially complex dependencies between various biological processes, we used a bootstrap procedure. We performed the bootstrap by randomly selecting a set of m genes (same number as the experiment) from the reference and repeating the procedure described above with the new set as the experiment. We repeated this procedure 1,000 times and tracked the number of times a predicted functional association appeared with the same or more extreme Z value. Bootstrap runs with 10,000

iterations did not affect significantly the prediction, while the computation time was increased many folds. Hence, we considered the use of 1,000 iterations as sufficient for this application. The significance threshold used for the bootstrap p-value is set at 5 percent. We also discarded any pair of biological processes with absolute correlation coefficient value less than 0.20 since it is indicative of a weak correlation.

2.5 Weighting Schemes

Matrices GF^R and GF^E in (1) and (3), respectively, represent gene-function data in binary format. Previous work in IR shows that the weighted representation of information improves the quality of retrieval over a binary representation based upon the assumptions that the repeated terms in a document are well suited to describe the topic of the document, while infrequent terms across the document collection are suited to differentiate between documents. Similar assumptions may hold for genes and their annotations.

Since we have two gene-function matrices, we will use a two-letter notation separated by a hyphen to denote the weighting scheme used in each case. The first letter denotes the weight used for the GF^R matrix, and the second letter denotes the weight used for the GF^E matrix. When a scheme enhances an existing scheme by assigning additional weight to the elements of GF matrices, we prepend the enhancing scheme acronym to the enhanced scheme name. In the simplest version of our approach described above, we assigned 1-s to the DE genes regardless of their actual expression change. This was done for both the reference, as well as the experiment matrix. Hence, we will refer to this as 1 – 1.

However, the actual gene expression change can carry meaningful information. If genes involved in different biological processes change in a coherently coordinated way, this may indicate a couple or interaction between those biological processes. This fold-change can be captured and taken into consideration during the analysis by redefining our GF^E matrix as described below

$$GF_{m \times n}^E = \{a_{ij}\} = \begin{cases} e_i & \text{if DE gene } i \text{ is annotated with GO term } j \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Here, e_i is the normalized and log-transformed fold-change measured for gene g_i in the given condition. Normalization is performed by dividing each fold-change by the absolute maximum of all fold-changes. Since in the reference condition, there are no fold changes, we still use 1-s in the construction of the reference matrix GF^R , as described above. Hence, this weighting scheme will be referred to as 1 – e .

We define our third scheme by adapting a weighting scheme commonly used in information retrieval. In this case, we will assign a weight w_{ij} to each element of the gene-function matrix. This weight is computed as the product of two terms: gene-bias (gb_j) and inverse-annotation-bias (iab_i). These terms are defined as follows:

$$w_{ij} = gb_j \cdot iab_i, \quad (12)$$

$$gb_j = \frac{1}{\text{Number of genes annotated with } f_j}, \quad (13)$$

$$iab_i = -\ln \frac{\text{Number of annotations for } g_i}{\text{Total annotations}}. \quad (14)$$

The term gb_j is meant to remove the gene annotation bias in the GO. The GO DAG is defined in such a way that general, broader terms are associated with the nodes at the top of the DAG, while more specific terms are associated with nodes found in the lower part of the DAG. Clearly, very specific terms provide more information and tend to be associated with very few genes, while general terms are associated with a lot of genes but they carry less information about each such gene. The term gb_j tries to quantitatively remove this bias by a factor inversely proportional to the number of genes annotated with a given term.

Similarly, a gene annotated with fewer GO terms is either a good indicator of the distinctive biological functions it participates in, or has not been studied much so far. As annotation databases become more comprehensive, the former explanation will be more likely. The term iab_j exploits the former assumption and assigns higher weight to a gene with fewer annotations.

With this IR-inspired weighting scheme our matrices can be defined as follows:

$$GF_{M \times N}^R = \{a_{ij}\} = \begin{cases} w_{ij} & \text{if gene } i \text{ is annotated with GO term } j \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

where, w_{ij} is computed as described in (12)–(14). The matrix $GF_{m \times n}^E$ corresponding to the given experiment is obtained in a similar way. We will refer to this weighting scheme as $IR_1 - 1$.

In our fourth scheme, we use fold-changes of the differentially expressed genes for the GF^E matrix in the IR scheme and call it $IR_1 - e$. The elements of GF^E matrix for this scheme are defined as follows:

$$GF_{m \times n}^E = \{a_{ij}\} = \begin{cases} e_i \cdot w_{ij} & \text{if DE gene } i \text{ is annotated with GO term } j \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

The definition of GF^R remains the same as defined in (15).

3 RESULTS AND DISCUSSION

We illustrate our proposed technique on two data sets. A first such set included genes identified as relevant for the prediction of breast cancer outcome [24]. Veer et al. selected a set of 231 genes from an initial set of more than 24,000 (reference) genes, which were found to be differentially expressed and correlated to the clinical outcome of breast cancer. We translated these genes into Entrez gene identifiers (IDs) using Onto-Translate [25]. The mapping produced 19,292 reference and 147 differentially expressed Entrez Gene IDs. 13,201 of these initial genes were found by Onto-Express [4] to be annotated with 3,670 biological processes, out of which 267 biological processes were annotated with 124 differentially expressed genes. The reference gene function matrix GF^R thus constructed had a dimension of $13,201 \times 3,670$, and the condition specific experiment gene function matrix GF^E had a dimension of 124×267 .

After running each of the four weighting schemes described above, we selected only those associations where the biological processes involved were represented by at least two genes. Also, we ignored any relationships that could be inferred exclusively from the GO structure without the use of any DE genes (such as ancestor-descendant at any depth in the GO graph, etc.). We used a significance threshold of 0.05, FDR as the method for multiple comparison correction, and bootstrapping to establish significance in the presence of potential dependencies.

Our analysis identified a number of pairs of biological processes that interact in a statistically significant way in this condition. A subset of results obtained for this data set is shown in Table 1. The complete list of results can be found in the supplemental materials, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2012.65>.

One of the statistically significant interactions was reported to happen between proteolysis (GO:0006508) and positive regulation of apoptosis (GO:0043065). The ana-phase-promoting complex (APC) is known to regulate degradation of intracellular proteins. It has been recently shown that UV radiation induced human cell death by activating the proteolysis of CDH1, an APC activator, which resulted in the accumulation of cell cycle proteins and thus facilitated UV-triggered apoptosis [26]. In addition, caspase-3, a member of cysteinyl aspartate specific protease (caspase) family, was identified as a key mediator of mammalian cell apoptotic cascade through proteolytic activation of various cellular proteins. One of caspase-3 downstream targets is protein kinase C delta (PKCdelta), which is associated with dopaminergic cell death in Parkinson's disease [27]. These studies suggest that the two processes can be coupled in specific circumstances such as reaction to UV radiation and Parkinson's disease. Our analysis shows that these two processes are also linked in some way in breast cancer, which is consistent with previous studies in [28], [29] showing potential antitumor agents, such as quercetin and resveratrol, induced tumor-cell specific apoptosis in a caspase 3-dependent manner. Interestingly, in our analysis, the link between the two processes was not established by a substantial number of common genes (in this case there is only one common gene, MMP9, as shown in Table 1) but rather by the expression of the other four DE genes involved in proteolysis (PITRM1, RBP3, CTSL2, and PCSK6), which appeared to be correlated to the expression of the other three DE genes involved in positive regulation of apoptosis. It is also worth mentioning that no links can be established between these two processes from the GO structure alone.

Our analysis also reported a statistically significant interaction between DNA replication initiation (GO: 0006270) and transcription (GO:0006350), as well as between DNA replication initiation and regulation of transcription-DNA dependent (GO:0006355). In fact, the promoter region of mammalian aldolase B (AldoB) gene has been discovered to have a dual function. In differentiated hepatoma cells, it participates in the regulation of transcription as a normal promoter, while in dedifferentiated hepatoma cells, it is transcriptionally silenced and functions as an origin of DNA replication [30]. Genome-wide studies showed that the DNA replication initiation sites and transcription unit regions are coordinately mapped in the mouse genome, and the location and type of the promoters were shown to be associated with the efficiency and specificity of the replication origin (ORI) [31]. Based on this previous work, together with our findings reported here, we can suggest that transcription may have a regulatory role in the initiation of DNA replication in breast cancer, possibly through some promoter regions having a dual role, similar to that of the AldoB promoter in hepatoma. Follow-up work in this direction could start with the genes associated with these processes and also found to be DE in this condition: MCM6, CCNE2, KIAA1442, EZH2, SEC14L2, BTG2, KIAA1442, and HMGB3. In fact, in the genome-wide studies mentioned above, it was revealed that one of the DE genes included in our analysis, HMGB3, has a promoter that overlaps with one of their newly identified CpG islands-associated ORIs [31].

Another statistically significant interaction was reported between DNA-dependent regulation of transcription (GO:0006355) and DNA repair (GO:0006281). In higher eukaryotes, one essential mechanism to maintain the accuracy of DNA-dependent transcription is transcription-coupled repair (TCR), in which case DNA lesions located on the actively transcribed strand of expressed genes is preferentially repaired compared to the

non-transcribed strand. This mechanism suggests a direct correlation of transcriptional regulation and DNA repair mediated by RNA Polymerase II basal transcription factor TFIID, disruption of which causes multiple organ disorder and is usually fatal [32]. Our analysis suggests that the two processes are also associated in breast cancer, possibly through shared key regulators. In fact, the protein encoded by one of the predominant tumor suppressor genes, breast cancer 1 (BRCA1), has been found to directly participate in activation of transcriptional regulation complex and DNA repair machinery in response to DNA damage, linking the maintenance of chromosomal stability to tumor suppression, possibly through transcription-coupled DNA repair [33]. It would be noteworthy to investigate factors interacting with genes found to be DE in breast cancer and involved in those processes, such as MCM6, KIAA1442, EZH2, HMGB3, BTG2, and CCNE2, for the purpose of identifying regulator(s) of both processes as potential anticancer drug target(s).

While several of the significant interactions detected by our analysis involved biological processes that have one or more annotated gene(s) in common, some interactions were reported between processes that have no common genes whatsoever. Such interactions are particularly important since no connection between those biological processes can be established based on existing annotations alone. One example is that of DNA replication initiation (GO:0006270 with annotated genes MCM6 and CCNE2) and phosphoinositide (PI)-mediated signaling (GO:0048015 with annotated genes CDC28, AURKA, NDC80, RFC4). In the experiment GF matrix, half of the processes annotated with gene CCNE2 are also annotated with CDC28 and NDC80. Because these genes are involved, in general, in the same or closely related biological processes, we can infer that the other processes with no common gene may also hold yet unknown interactions. One such interaction would be between DNA replication initiation and phosphoinositide (PI)-mediated signaling. Furthermore, our literature review shows that PI-mediated signaling pathway plays a crucial role in a variety of cellular functions, including cell growth, proliferation, differentiation, survival, etc. One major signal transducer in this pathway is the proto-oncogene protein Phosphoinositide 3-kinase (PI3K), which interacts with multiple tumor progression and metastasis pathways, therefore serves as a good candidate drug target in cancer therapy [34]. In addition to its regulatory role in DNA transcription and translation, PI3K also plays an important part in the initiation of DNA synthesis through modulating stabilization of c-Myc oncogenic transcription factor, a key regulator involved in precise control of cell cycle progression [35]. Based on this previous work, together with our findings, our analysis suggests that PI signaling may contribute to breast cancer through DNA replication-related mechanisms, in which case deregulation of PI3K activity possibly induced by abnormal level of growth factors may lead to enhanced amplification of DNA materials, which facilitates cell proliferation in tumor progression. Follow-up work in this direction would be identification of putative interactions between PI-mediated signaling pathway and the two genes associated with DNA replication initiation and found to be DE in breast cancer, MCM6, and CCNE2.

The second data set consisted of 100 genes from Affymetrix Human Gene FL microarray identified as significantly different between good and poor prognosis in patients with lung adenocarcinoma [36]. The mapping of the microarray probes returned 5,541 reference and 96 differentially expressed Entrez Gene IDs. Five thousand and thirty seven of these initial genes were found by Onto-Express to be annotated with 2,908 biological processes, out of which 248 biological processes were annotated with 87 differentially expressed genes. The reference gene function matrix GF^R thus constructed had a dimension of $5,037 \times 2,908$, and the condition specific gene function matrix GF^E had a dimension of 87×248 .

While most of the interactions between biological processes in this data set involved generic processes occurring in almost all types of cells, such as DNA replication, transcription, and

apoptosis, interactions identified in Beer's data set involved very specific biological processes that occur in only a subset group of cells, including blood coagulation, inflammatory response, and cellular defense response. For instance, our analysis reported a statistically significant interaction between antiapoptosis (GO:0006916) and inflammatory response (GO:0006954). Inflammatory response is the organism's most important protective mechanism in response to infections and/or traumatic damages, removing the invading stimuli as well as initiating the healing process. Inflammation is known to be associated with various kinds of cancer, in which condition the balance between cell proliferation and cell death in normal stage is interrupted due to loss of cell cycle checkpoint control and/or suppression of the apoptosis machinery [37]. Many inflammatory factors have been found to contribute to tumor progression in liver, stomach, and/or intestine tissues, among which a key player is the nuclear factor kappa B (NF-kappaB). It gets activated during inflammation response and subsequently triggers expression of downstream antiapoptotic genes, further enhancing tumor cell survival [38]. Our analysis suggests that inflammation may play a similar role in promoting lung carcinogenesis, possibly through the release of signals that induce inhibition of apoptosis in cancer cells. Indeed, sulindac and other nonaspirin nonsteroidal antiinflammatory drugs (NSAIDs) that target inflammatory cytokines have been shown to successfully attenuate lung cancer development with induced cell apoptosis [39]. Follow-up work in this direction could start with the inflammation-associated factors RELA, BMP2, and TNFAIP6 that were found to be DE in lung cancer for their potential interactions with antiapoptosis pathways. In fact, Tumor Necrosis Factor Alpha Induced Protein 6 (TNFAIP6), also known as TSG-6, has been identified as a potential target of Hypoxia-Inducible Factor(HIF)-1alpha on its inhibition of human small cell lung cancer (SCLC) cell apoptosis [40].

Another example of change in correlation detected by our analysis is between interspecies interaction between organisms (GO:0044419) and negative regulation of apoptosis (GO:0043066). As one major type of interspecies interactions, infection happens when a foreign species (for instance, virus, or bacterium) invades a host organism (say, human) and interferes with its normal functioning, which may lead to severe diseases and cancer [41]. Parasitic bacteria species, such as *Toxoplasma gondii* (cause of Toxoplasmosis), were known to inhibit the apoptosis of infected host cells possibly by interfering with the caspase cascade, therefore increasing their survival opportunity [42]. Moreover, Hepatitis B (HBV) virus infection, the major cause of hepatocellular carcinoma (HCC), has been shown to activate overexpression of apoptosis inhibitors in the host cells [43]. These studies indicate a clear interaction between the two processes in the circumstances of Toxoplasmosis and liver cancer. Even more relevant literature exists linking viral infection with lung cancer and adenocarcinoma in particular [44], [45], which reported an important role of oncogenic viruses such as Human papillomaviruses (HPVs) and JC virus (JCV) in lung carcinogenesis. Our data indicate that the two processes may be coupled in some way in lung cancer, in which case virus-induced negative regulation of apoptosis in the host cells may be the mechanism through which cells invaded by the virus escape death and start immortalization. Consistent with this, significant down-regulation of apoptosis inducers, including tumor necrosis factor (TNF)-related apoptosis-inducing ligand (TRAIL), was observed in immortal Li-Fraumeni syndrome (LFS) cells compared to normal condition [46]. Of the four DE genes (FADD, KRT7, KRT19, and KRT18) annotated to be associated with interspecies interaction, FADD (Fas-Associated protein with Death Domain) initiates apoptosis through death-receptor signaling while the epithelial keratin protein KRT18 plays a role in cell resistance to drug-induced apoptosis [47], [48]. Therefore, interesting follow-up work would be to investigate their potential interaction with proteins produced by the lung cancer-related viruses. In fact, one major mechanism in which the adenoviruses (cause of respiratory infections) manage to establish a long term infection is to inhibit the host cell's apoptosis by blocking the interaction between FADD and death effector filaments, which is mediated by

the viral protein E1B [49], suggesting those interspecies interaction-related DE genes may serve as good candidate targets for anticancer therapy.

Our analysis also found a significant change of correlation between Wnt receptor signaling pathway, calcium modulating pathway (GO:0007223), and cell differentiation (GO:0030154). Different from the canonical Wnt signaling pathway which is dependent on the transcription factor beta-catenin, Wnt/calcium pathway is triggered by the Wnt5a class (Wnt5a, Wnt4, Wnt11) resulting in activation of two calcium-sensitive kinases, Calmodulin-dependent Protein Kinase II (CamKII), and Protein Kinase C (PKC) [50], [51]. In addition to its established role in embryonic dorsal-ventral patterning, there is also evidence that Wnt/calcium signaling participates in metastasis through modulation of cell proliferation and motility mediated by Wnt5a [52], [53]. More evidence exists linking Wnt/calcium pathway to lung cancer. For example, [54] reported significant expression level changes of multiple noncanonical Wnt signaling components including Wnt5a and Wnt11 in squamous cell lung cancer cells compared to normal cells. Our data suggest a regulatory role of Wnt/calcium pathway in lung cancer related to cell differentiation, possibly in a way that deregulation of Wnt/calcium signals lead to de-differentiation of fully developed lung cells, resulting in uncontrolled cell growth and eventually malignant stage. Consistent with this, the data of [55] support a role of Wnt11 in promoting cardiac differentiation through activation of Wnt/calcium signaling mediator PKC. In addition, together with Wnt5a, Wnt11 also regulates the differentiation pattern of myogenic cells in limb development [56]. Follow-up work in this direction could start with the two DE Wnt ligands in lung cancer (WNT10B, WNT1) that were annotated to be associated with Wnt/calcium signaling.

While most significantly interacting biological processes detected in our study have at least one gene in common, approximately 30 cases with distinct input gene sets are of particular interest. One example of that involves the blood coagulation (GO:0007596) and negative regulation of cell cycle (GO:0045786). One candidate mediator that possibly establishes the bridge between the two processes is the serine protease prothrombin. It is known to play a critical role in blood coagulation, in which case prothrombin is converted into its active form, thrombin, and promotes the formation of blood clots by converting soluble fibrinogen into insoluble fibrin [57]. Abnormally prolonged prothrombin time, an indication of reduced coagulation function, has been discovered to be associated with aggravated stage in cases of gastric and lung cancer [58], [59]. Moreover, data of [60] supported that prothrombin is involved in cell cycle regulation as a DNA synthesis inhibitor, possibly through an induction of cytoskeleton deorganization. This inhibition was only observed in hepatocytes but not in hepatoma cells likely due to different extracellular matrix (ECM) protein interactions in normal versus cancer condition. These previous studies clearly indicated a regulatory role of blood coagulation factor(s) in cell cycle progression. Our results suggest that a similar mechanism may exist in lung cancer, in which case the two processes interact in a way that dysregulation of blood coagulation mediators activated in clotting cascade causes interruptions in cell cycle control, resulting in unlimited cell proliferation and growth. Follow-up work in this direction could start with the genes identified to be DE in lung cancer and associated with these two processes, SERPINE1, ITGA2, BMP2, and INHA. In fact, one of the blood coagulation-associated DE gene SERPINE1 (serine protease inhibitor, clade E, member 1), has already been found to participate in cell proliferation promotion through inducing cell cycle reentry as part of the wound repair program in replicatively competent keratinocytes [61].

Another interesting interaction change was detected between the negative regulation of cell cycle (GO:0045786) and cell adhesion (GO:0007155). As part of the tumor microenvironment (stroma), reduced cell adhesion is often observed between tumor and normal cells [62]. Abnormal expression of cell adhesive molecules, such as integrins,

catenins, and E-cadherins, has been reported to interact with various intra- and/or extracellular signaling factors and promote tumor progression [63]. One example of that is B-cell Non-Hodgkin Lymphoma (NHL), in which case the lymphoma cells are maintained in a dormant stage when attached to bone marrow stromal cells, but reenter proliferation upon detachment and progress into metastasis thereafter. This is due to a cell adhesion-induced reversible cell cycle arrest in adherent lymphoma cells via up-regulation of cyclin-dependent kinase inhibitor (CDKI) p27, and the arrest is removed when dissociation happens [64]. Based on this previous work, together with our findings, we can suggest a regulatory role of cell adhesion in cell cycle progression in lung cancer development. Consistent with this, E-cadherin-catenin complex, the critical mediator of normal cell-cell contact, has been reported to be associated with lung carcinogenesis [65]. Moreover, other existing studies support E-cadherin as a growth inhibitor in mammary carcinoma cells through inducing cell cycle arrest, which is associated with elevated protein level of p27, the same factor that mediates the dormancy of lymphoma cells in NHL [66]. Interestingly, in the Genes-to-Systems Breast Cancer (G2SBC) Database [67], the expression profile of p27 is highly correlated to that of trophoblast glycoprotein (TPBG), one of the cell adhesion-associated DE genes in our study. Therefore, follow-up work in this direction could start with the five DE genes involved in cell-cell contact, TPBG, RND3, ITGA2, LAMB1, and TNFAIP6, for their possible interactions with cell cycle regulators in the cases of lung cancer. A list of significant interactions between biological processes discussed here are shown in Table 2. A complete list of results can be found in Table 3 (Additional file 1, available in the online supplemental material).

3.1 Related Work

Overrepresentation studies, in general, ignore the biological fact that most complex functions in cells are carried out through the concerted activities of many gene products. Current vocabularies such as that provided by GO are classified mostly based on individual gene functions as opposed to the functionality of a set of interacting genes. Antonov and Mewes proposed a method to find overrepresentation of complex functions annotated to a set of genes of interest, wherein a gene may belong simultaneously to two or more functional categories [12], [68]. They defined the complex functional categories as algebraic combination of base categories, such as union, intersection, or difference of the gene sets corresponding to the base categories. Nam et al. analyzed the differentially expressed gene sets using composite GO annotations using a parametric (Z-) test [69].

A recently developed tool, ToppCluster [70], finds association between functional categories by computing the significance scores from the p-values of overrepresented categories. Euclidean-based distance between significance scores is used to perform a 2D hierarchical clustering and to produce either a heatmap or a Cytoscape-based network, which is then used to find association between the functional categories or annotated genes. Another tool, GENECODIS [71], finds significant concurrent annotations in gene lists by finding the significance of co-occurrence of annotations with at least three genes. The hypergeometric and chi-squared tests are employed to compute the significance level of their prediction. Unlike ToppCluster, our proposed method does not rely on other overrepresentation techniques that are based on single annotations, and unlike GENECODIS, our technique can find significant functional categories with no common input gene. In addition, these studies find groups of functional categories associated with some list of differentially expressed genes, while the method described here finds biological process pairs, whose association has changed in the studied phenomenon.

The novelty of our approach can be demonstrated by the following example. Suppose, a DE gene g_a is annotated with a biological process A and, under the same biological condition, another DE gene g_b is annotated with a biological process B . The biological process A might

have been shown to interact in a specific condition, in another research, with the biological process *B*. However, current analysis tools may fail to provide such information due to the nature and/or structure of the ontology and the database. As such, the current methods and techniques, oblivious of any interaction between biological processes, may only provide the two GO categories as over- or underrepresented, individually or collectively as a complex category and not reveal further insight about the nature or strength of association among functions in the complex category.

4 CONCLUSIONS

In this paper, we presented a method that analyzes expression data obtained from a microarray experiment together with all existing gene annotations to find statistically significant interactions between specific biological processes involved in the given biological condition. Applying the proposed method to two independent data sets yielded a number of interesting statistically significant interactions specific to the given conditions. Most of these interactions are well supported by independent studies from the literature. For the purpose of evaluation, we manually surveyed the literature and assigned each predicted biological processes pair to one of the three categories: 1) supported by literature and considered trivial, 2) supported by literature and considered nontrivial, and 3) yet unknown. Predicted pairs assigned to categories 1 and 2 were considered correct or true positive, while the pairs assigned to category 3 were assumed to be false positive. More than 88 percent of the associations predicted by any of the four schemes were found to be correct and valid. A subset of such interactions was discussed in details and shown to have the potential to open a number of new avenues for research in lung and breast cancer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work has been partially supported by the following grants: NIH RO1 DK089167, NIH STTR R42GM087013, and NSF DBI-0965741 (to S.D.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

Biographies

Nadeem A. Ansari received the PhD degree in computer science from Wayne State University, Detroit. He is currently a software development engineer in Test 2 (SDet II) in Exchange Mobility at Microsoft. His current research interests are focused on data mining approaches and test automation for quality assurance.

Riyue Bao received the bachelor's degree in biological sciences from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and the PhD degree in biological sciences in 2012 from Wayne State University, Detroit. From 2004 to 2005, she worked as an assistant researcher in the School of Life Sciences, USTC, Hefei, China. Her research interests include molecular evolution, data mining, and bioinformatics.

C lin Voichi a received the BSc degree in computer science from "Babes-Bolyai" University, Cluj-Napoca, Romania, and the MSc degree in computer science from Wayne State in 2010. He is working toward the PhD degree in the Department of Computer Science, Wayne State University, Detroit. His research interests include systems biology, machine learning and cancer bio-marker discovery. He is a member of the IEEE.

Sorin Dr ghici received the PhD degree in computer science from the University of St. Andrews, United Kingdom. He currently holds the Robert J. Sokol MD Endowed chair in Systems Biology in the Department of Obstetrics and Gynecology, and is a professor in the Department of Clinical and Translational Science and the Department of Computer Science, as well as the head of the Intelligent Systems and Bioinformatics Laboratory at Wayne State University. He has published two books on microarray analysis entitled *Data Analysis Tools for Microarrays* (Chapman and Hall/CRC Press, 2003) and *Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition* (CRC Press, Mathematical & Computational Biology Series, 2012), eight book chapters, and more than 100 peer-reviewed journal and conference papers. He is a senior member of the IEEE.

References

1. DeRisi JL, Iyer VR, Brown PO. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*. 1997; 278:680–686. [PubMed: 9381177]

2. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum Information about a Microarray Experiment (MIAME)-Toward Standards for Microarray Data. *Nature Genetics*. Dec.2001 29(4):365–371. [PubMed: 11726920]
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*. 2000; 25:25–29. [PubMed: 10802651]
4. Khatri P, Draghici S, Ostermeier GC, Krawetz SA. Profiling Gene Expression Using Onto-Express. *Genomics*. Feb.2002 79(2):266–270. [PubMed: 11829497]
5. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global Functional Profiling of Gene Expression. *Genomics*. Feb.2003 81(2):98–104. [PubMed: 12620386]
6. Rhee YS, Wood V, Dolinski K, Draghici S. Use and Misuse of the Gene Ontology Annotations. *Nature Rev Genetics*. Jul; 2008 9(7):509–515. [PubMed: 18475267]
7. Khatri P, Draghici S. Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems. *Bioinformatics*. 2005; 21(18):3587–3595. [PubMed: 15994189]
8. Huang DW, Sherman BT, Lempicki RA. Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Research*. 2009; 37(1):1–13. [PubMed: 19033363]
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc Natl Academy of Sciences USA*. 2005; 102(43):15545–15550.
10. Alexa A, Rahnenfuhrer J, Lengauer T. Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure. *Bioinformatics*. 2006; 22(13):1600–1607. [PubMed: 16606683]
11. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved Detection of Overrepresentation of Gene-Ontology Annotations with Parent Child Analysis. *Bioinformatics*. 2007; 23(22):3024–3031. [PubMed: 17848398]
12. Antonov AV, Mewes HW. Complex Functionality of Gene Groups Identified from High-Throughput Data. *J Molecular Biology*. 2006; 363(1):289–296.
13. Khatri P, Done B, Rao A, Done A, Draghici S. A Semantic Analysis of the Annotations of the Human Genome. *Bioinformatics*. 2005; 21(16):3416–3421. [PubMed: 15955782]
14. Lee D, Seung H. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*. 1999; 401(6755):788–791. [PubMed: 10548103]
15. Ding C, He X, Simon H. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. *Proc SIAM Data Mining Conf*. 2005:606–610.
16. Gaussier E, Goutte C. Relation between Plsa and Nmf and Implications. *Proc 28th Ann Int'l ACM SIGIR Conf Research and Development in Information Retrieval*. 2005:601–602.
17. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by Latent Semantic Analysis. *J Am Soc for Information Science*. 1990; 41(6):391–407.
18. Berry MW, Dumais ST, O'Brien GW. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Rev*. 1995; 37(4):573–595.
19. Berry MW, Drmac Z, Jessup ER. Matrices, Vector Spaces, and Information Retrieval. *SIAM Rev*. 1999; 41(2):335–362.
20. Done B, Khatri P, Done A, Draghici S. Predicting Novel Human Gene Ontology Annotations Using Semantic Analysis. *IEEE/ACM Trans Computational Biology and Bioinformatics*. Jan-Mar; 2010 7(1):91–99.
21. Golub, G.; van Loan, CF. *Matrix Computations*. The Johns Hopkins Univ Press; 1983.
22. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*. May; 1915 10(4):507–521.

23. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Royal Statistical Soc Series B (Methodological)*. 1995; 57:289–300.
24. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart A, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen others AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*. Jan. 2002 415(6871):530–536. [PubMed: 11823860]
25. Khatri P, Desai V, Tarca AL, Sellamuthu S, Wildman DE, Romero R, Draghici S. New Onto-Tools: Promoter-Express, nsSNPCounter, and Onto-Translate. *Nucleic Acids Research*. 2006; 34:W626–W631. [PubMed: 16845086]
26. Liu W, Li W, Fujita T, Yang Q, Wan Y. Proteolysis of CDH1 Enhances Susceptibility to UV Radiation-induced Apoptosis. *Carcinogenesis*. 2008; 29:263–272. [PubMed: 18174259]
27. Kitazawa M, Anantharam V, Kanthasamy AG. Dieldrin Induces Apoptosis by Promoting Caspase-3-Dependent Proteolytic Cleavage of Protein Kinase Cdelta in Dopaminergic Cells: Relevance to Oxidative Stress and Dopaminergic Degeneration. *Neuroscience*. 2003; 119:945–964. [PubMed: 12831855]
28. Chien SY, Wu YC, Chung JG, Yang JS, Lu HF, Tsou MF, Wood WG, Kuo SJ, Chen DR. Quercetin-Induced Apoptosis Acts through Mitochondrial- and Caspase-3-Dependent Pathways in Human Breast Cancer MDA-MB-231 Cells. *Human Experimental Toxicology*. 2009; 28(8): 493–503. [PubMed: 19755441]
29. Alkhalaf M, El-Mowafy A, Renno W, Rachid O, Ali A, Al-Attyiah R. Resveratrol-Induced Apoptosis in Human Breast Cancer Cells Is Mediated Primarily through the Caspase-3-Dependent Pathway. *Archives of Medical Research*. 2008; 39(2):162–168. [PubMed: 18164959]
30. Miyagi S, Zhao YP, Saitoh Y, Tamai K, Tsutsumi KI. Replication of the Rat Aldolase B Locus Differs Between Aldolase B-Expressing and Non-Expressing Cells. *FEBS Letters*. 2001; 505:332–336. [PubMed: 11566199]
31. Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gomez M. Transcription Initiation Activity Sets Replication Origin Efficiency in Mammalian Cells. *PLoS Genetics*. 2009; 5:e1000446. (Epub). [PubMed: 19360092]
32. Sarker AH 4, Tsutakawa SE, Kostek S, Ng C, Shin DS, Peris M, Campeau E, Tainer JA, Nogales E, Cooper PK. Recognition of RNA Polymerase II and Transcription Bubbles by XPG, CSB, and TFIIH: Insights for Transcription-Coupled Repair and Cockayne Syndrome. *Molecular Cell*. 2005; 20(2):187–198. [PubMed: 16246722]
33. Venkitaraman AR. Functions of BRCA1 and BRCA2 in the Biological Response to DNA Damage. *J Cellular Sciences*. 2001; 114:3591–3598.
34. Fry MJ. Phosphoinositide 3-Kinase Signalling in Breast Cancer: How Big a Role Might It Play. *Breast Cancer Res*. 2001; 3:304–312. [PubMed: 11597319]
35. Kumar A, Marqués M, Carrera AC. Phosphoinositide 3-Kinase Activation in Late G1 Is Required for c-Myc Stabilization and S Phase Entry. *Molecular Cell Biology*. Dec.2006 26(23):9116–9125.
36. Beer DG, Kardia SL, Huang C, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. *Nature Medicine*. Jul; 2002 8(8):816–824.
37. Coussens LM, Werb Z. Inflammation and Cancer. *Nature*. 2002; 420:860–867. [PubMed: 12490959]
38. Pikarsky E, Porat RM, Stein I, Abramovitch R, Amit S, Kasem S, Gutkovich-Pyest E, Urieli-Shoval S, Galun E, Ben-Neriah Y. NF-kappaB Functions as a Tumour Promoter in Inflammation-Associated Cancer. *Nature*. 2004; 431:461–466. [PubMed: 15329734]
39. Berman KS, Verma UN, Harburg G, Minna JD, Cobb MH, Gaynor RB. Sulindac Enhances Tumor Necrosis Factor-Alpha-Mediated Apoptosis of Lung Cancer Cell Lines by Inhibition of Nuclear Factor-kappaB. *Clinical Cancer Research*. 2002; 8(2):354–360. [PubMed: 11839649]
40. Wan J, Ma J, Mei J, Shan G. The Effects of HIF-1alpha on Gene Expression Profiles of NCI-H446 Human Small Cell Lung Cancer Cells. *J Experimental Clinical Cancer Research*. 2009; 28:150.

41. Parsonnet J. Bacterial Infection as a Cause of Cancer. *Environmental Health Perspective*. 1995; 103:263–268.
42. Keller P, Schaumburg F, Fischer SF, Hacker G, Gross U, Luder CG. Direct Inhibition of Cytochrome C-Induced Caspase Activation in Vitro by Toxoplasma Gondii Reveals Novel Mechanisms of Interference with Host Cell Apoptosis. *FEMS Microbiology Letters*. 2006; 258(2): 312–319. [PubMed: 16640590]
43. Lu X, Lee M, Tran T, Block T. High Level Expression of Apoptosis Inhibitor in Hepatoma Cell Line Expressing Hepatitis B Virus. *Int'l J Medical Sciences*. 2005; 2(1):30–35.
44. Klein F, Amin Kotb WF, Petersen I. Incidence of Human Papilloma Virus in Lung Cancer. *Lung Cancer*. 2009; 65:13–18. [PubMed: 19019488]
45. Zheng H, Abdel Aziz HO, Nakanishi Y, Masuda S, Saito H, Tsuneyama K, Takano Y. Oncogenic Role of JC Virus in Lung Cancer. *J Pathology*. 2007; 212:306–315.
46. Kulaeva OI, Draghici S, Tang L, Kraniak JM, Land SJ, Tainsky MA. Epigenetic Silencing of Multiple Interferon Pathway Genes After Cellular Immortalization. *Oncogene*. Jun; 2003 22(26): 4118–4127. [PubMed: 12821946]
47. Chinnaiyan AM, O'Rourke K, Tewari M, Dixit VM. FADD, A Novel Death Domain-Containing Protein, Interacts with the Death Domain of Fas and Initiates Apoptosis. *Cell*. 1995; 81:505–512. [PubMed: 7538907]
48. Oshima RG. Apoptosis and Keratin Intermediate Filaments. *Cell Death and Differentiation*. 2002; 9(5):486–492. [PubMed: 11973607]
49. McNeas AL, Gooding LR. Adenoviral Inhibitors of Apoptotic Cell Death. *Virus Research*. 2002; 88:87–101. [PubMed: 12297329]
50. Du SJ, Purcell SM, Christian JL, McGrew LL, Moon RT. Identification of Distinct Classes and Functional Domains of Wnts through Expression of Wild-Type and Chimeric Proteins in Xenopus Embryos. *Molecular Cell Biology*. 1995; 15(5):2625–2634.
51. Pongracz J, Stockley R. Wnt Signalling in Lung Development and Diseases. *Respiratory Research*. 2006; 7(1):15. [PubMed: 16438732]
52. Liang H, Chen Q, Coles AH, Anderson SJ, Pihan G, Bradley A, Gerstein R, Jurecic R, Jones SN. Wnt5a Inhibits B Cell Proliferation and Functions as a Tumor Suppressor in Hematopoietic Tissue. *Cancer Cell*. 2003; 4:349–360. [PubMed: 14667502]
53. Weeraratna AT, Jiang Y, Hostetter G, Rosenblatt K, Duray P, Bittner M, Trent JM. Wnt5a Signaling Directly Affects Cell Motility and Invasion of Metastatic Melanoma. *Cancer Cell*. 2002; 1:279–288. [PubMed: 12086864]
54. Lee EHL, Chari R, Lam A, Ng RT, Yee J, English J, Evans KG, MacAulay C, Lam S, Lam WL. Disruption of the Non-Canonical Wnt Pathway in Lung Squamous Cell Carcinoma. *Clinical Medicine Oncology*. 2008; 2:169–179. [PubMed: 20401333]
55. Pandur P, Laschel M, Eisenberg LM, Kuhl M. Wnt-11 Activation of a Non-Canonical Wnt Signalling Pathway Is Required for Cardiogenesis. *Nature*. 2002; 418(6898):636–641. [PubMed: 12167861]
56. Anakwe K, Robson L, Hadley J, Buxton P, Church V, Allen S, Hartmann C, Harfe B, Nohno T, Brown AM, Evans DJ, Francis-West P. Wnt Signalling Regulates Myogenic Differentiation in the Developing Avian Wing. *Development*. Aug; 2003 130(15):3503–3514. [PubMed: 12810597]
57. Yegneswaran S, Mesters RM, Griffin JH. Identification of Distinct Sequences in Human Blood Coagulation Factor Xa and Prothrombin Essential for Substrate and Cofactor Recognition in the Prothrombinase Complex. *J Biological Chemistry*. 2003; 278:33312–33318.
58. Kwon HC, Oh SY, Lee S, Kim SH, Han JY, Koh RY, Kim MC, Kim HJ. Plasma Levels of Prothrombin Fragment F1+2, D-Dimer and Prothrombin Time Correlate with Clinical Stage and Lymph Node Metastasis in Operable Gastric Cancer Patients. *Japanese J Clinical Oncology*. 2008; 38(1):2–7.
59. Buccheri G, Ferrigno D, Ginardi C, Zuliani C. Haemostatic Abnormalities in Lung Cancer: Prognostic Implications. *European J Cancer*. 1997; 33:50–55. [PubMed: 9071899]
60. Carr BI, Kara S, Wanga M, Wang Z. Growth Inhibitory Actions of Prothrombin on Normal Hepatocytes: Influence of Matrix. *Cell Biology Int'l*. 2007; 31:929–938.

61. Qi L, Higgins SP, Lu Q, Samarakoon R, Wilkins-Port CE, Ye Q, Higgins CE, Staiano-Coico L, Higgin PJ. SERPINE1 (PAI-1) Is a Prominent Member of the Early G0 → G1 Transition 'Wound Repair' Transcriptome in p53 Mutant Human Keratinocytes. *J Investigative Dermatology*. 2008; 128:749–53.
62. Behrens J. The Role of Cell Adhesion Molecules in Cancer Invasion and Metastasis. *Breast Cancer Research and Treatment*. 1993; 24(3):175–84. [PubMed: 8435473]
63. Sung SY, Hsieh CL, Wu D, Chung LW, Johnstone PA. Tumor Microenvironment Promotes Cancer Progression, Metastasis, and Therapeutic Resistance. *Current Problems in Cancer*. 2007; 31:36–100. [PubMed: 17362788]
64. Lwin T, Hazlehurst LA, Dessureault S, Lai R, Bai W, Sotomayor E, Moscinski LC, Dalton WS, Tao J. Cell Adhesion Induces p27Kip1-Associated Cell-cycle Arrest through Down-Regulation of the SCFSkp2 Ubiquitin Ligase Pathway in Mantle-Cell and Other Non-Hodgkin B-Cell Lymphomas. *Blood*. 2007; 110(5):1631–1638. [PubMed: 17502456]
65. Bremnes RM, Veve R, Hirsch FR, Franklin WA. The E-cadherin Cell-cell Adhesion Complex and Lung Cancer Invasion, Metastasis, and Prognosis. *Lung Cancer*. 2002; 36:115–24. [PubMed: 11955645]
66. Croix BS, Sheehan C, Rak JW, Florenes VA, Slingerland JM, Kerbel RS. E-Cadherin-Dependent Growth Suppression Is Mediated by the Cyclin-Dependent Kinase Inhibitor p27(KIP1). *J Cell Biology*. 1998; 142(2):557–571.
67. Viti F, Mosca E, Merelli I, Calabria A, Alfieri R, Milanesi L. Ontological Enrichment of the Genes-to-Systems Breast Cancer Database. *Proc Third Int'l Conf Metadata and Semantic Research*. Oct.2009 46:171–182.
68. Antonov AV, Schmidt T, Wang Y, Mewes HW. ProfCom: A Web Tool for Profiling the Complex Functionality of Gene Groups Identified from High-Throughput Data. *Nucleic Acids Research*. 2008; 36(Web Server issue):W347–W351. [PubMed: 18460543]
69. Nam D, Kim S, Kim S, Yang S, Kim S, Chu I. ADGO: Analysis of Differentially Expressed Gene Sets Using Composite Go Annotation. *Bioinformatics*. 2006; 22(18):2249–53. [PubMed: 16837524]
70. Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. ToppCluster: A Multiple Gene List Feature Analyzer for Comparative Enrichment Clustering and Network-Based Dissection of Biological Systems. *Nucleic Acids Research*. 2010; 38:W96–W102. [PubMed: 20484371]
71. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: A Web-Based Tool for Finding Significant Concurrent Annotations in Gene Lists. *Genome Biology*. 2007; 8:R3. [PubMed: 17204154]

Fig. 1. Singular value decomposition of a gene-function matrix

Singular value decomposition of gene function matrix, GF into G , S , and F^T . G and F are matrices of left and right singular vectors and are orthonormal, i.e., $GG^T = I$, $F^T F = I$, where I is the identity matrix; S is a diagonal matrix of singular values s_i such that $s_1 \ s_2 \ \dots \ s_l$.

Fig. 2. Approximation of a gene-function matrix

GF_k is an approximation of original gene function matrix, GF , in a subspace with only k singular vectors. However, note that the dimensionality of GF_k remains the same, i.e., there is an element for every gene and every biological process.

Selected Pairs of Biological Processes that Interact in a Statistically Significant Way in the Breast Cancer Data Set

TABLE 1

Scheme	GO Term 1	GO Term 2	FDR p-Val	Bootstrap p-Val	Common Input Genes
1-e, IR_1-e	Proteolysis	Positive regulation of apoptosis	0.00010	0.029	MMP9
1-1	Transcription	DNA replication initiation	0.02601	0.043	MCM6
1-1	Regulation of transcription, DNA-dependent	DNA replication initiation	0.00880	0.007	MCM6
1-1	DNA repair	Regulation of transcription, DNA-dependent	0.03368	0.020	BTG2
IR_1-1	Vesicle-mediated transport	Transcription from RNA polymerase II promoter	0.00206	0.000	-
IR_1-1	DNA replication initiation	Phosphoinositide-mediated signaling	0.00001	0.000	-

TABLE 2
Selected Pairs of Biological Processes that Interact in a Statistically Significant Way in the Lung Cancer Data Set

Scheme	GO Term 1	GO Term 2	FDR p-Val	Bootstrap p-Val	Common Input Genes
<i>IR_1-1</i>	anti-apoptosis	Inflammatory response	.000000	.031	RELA
<i>1-1, IR_1-1-e</i>	Interspecies interaction between organisms	Negative regulation of apoptosis	.039476	.015	KRT18
<i>1-e, IR_1-1-e</i>	Wnt receptor signaling pathway, calcium modulating pathway	Cell differentiation	.000000	.026	WNT1
<i>IR_1-e</i>	Blood coagulation	Negative regulation of cell cycle	.046280	.020	-
<i>IR_1-e</i>	Negative regulation of cell cycle	Cell adhesion	.000000	.019	-