BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1122–1129
doi:10.1093/bioinformatics/bts112

Gene expression

Advance Access publication March 8, 2012

Deriving transcriptional programs and functional processes from gene expression databases

Jeffrey T. Chang*

Department of Integrative Biology and Pharmacology, The University of Texas Health Science Center in Houston, Houston, TX 77030, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: A system-wide approach to revealing the underlying molecular state of a cell is a long-standing biological challenge. Developed over the last decade, gene expression profiles possess the characteristics of such an assay. They have the capacity to reveal both underlying molecular events as well as broader phenotypes such as clinical outcomes. To interpret these profiles, many gene sets have been developed that characterize biological processes. However, the full potential of these gene sets has not yet been achieved. Since the advent of gene expression databases, many have posited that they can reveal properties of activities that are not evident from individual datasets, analogous to how the expression of a single gene generally cannot reveal the activation of a biological process.

Results: To address this issue, we have developed a high-throughput method to mine gene expression databases for the regulation of gene sets. Given a set of genes, we scored it against each gene expression dataset by looking for enrichment of co-regulated genes relative to an empirical null distribution. After validating the method, we applied it to address two biological problems. First, we deciphered the E2F transcriptional network. We confirmed that true transcriptional targets exhibit a distinct regulatory profile across a database. Second, we leveraged the patterns of regulation across a database of gene sets to produce an automatically generated catalog of biological processes. These demonstrations revealed the power of a global analysis of the data contained within gene expression databases, and the potential for using them to address biological questions.

Contact: jeffrey.t.chang@uth.tmc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 8, 2011; revised on February 16, 2012; accepted on March 4, 2012

1 INTRODUCTION

The state of a cell is reflected in its transcriptional profile. Both molecular events (such as the activation of a cell surface receptor, loss of function of a tumor suppressor or initiation of the cell cycle) and functional attributes (sensitivity to therapeutics or propensity to metastasize) are evident in gene expression (Bild *et al*., 2006; Lamb *et al*., 2006; Miller *et al*., 2005; Perou *et al*., 2000; Whitfield *et al*., 2002; Zhang *et al*., 2009). They can be measured on a DNA microarray and recognized by computational algorithms. Because such a diverse range of molecular activities can be identified from the same gene expression profiles (Hughes *et al*., 2000; Lamb *et al*., 2006), we argue that gene expression contains the complete (or nearly complete) compendium of the underlying cellular activities. In other words, a gene expression profile embodies the state of a cell.

Currently, a major challenge is to interpret gene expression profiles. A common approach is to construct narratives based on the functions of the genes showing changes in gene expression. This is made difficult, however, by a limited understanding of what the genes do, compounded by pleiotropy in gene function. One solution to these problems is to consider that biological processes are brought about by the activation of many genes. Thus, taking into account the co-ordinated activities of sets of genes yields insight into processes that are not evident from the expression of single genes. To capitalize on this idea, many have generated gene expression signatures that evince a biological activity. Although signatures can take different forms, the vast majority consists simply of lists of genes that are involved in biological processes. This type of signature is also called a gene set, and MSigDB is the largest database (Liberzon *et al*., 2011).

To predict the activation of signatures from gene expression profiles, many computational algorithms have been developed (Barbie *et al*., 2009; Liu *et al*., 2008; Spang *et al*., 2002; Subramanian *et al*., 2005; Tavazoie *et al*., 1999). While these algorithms operate on single datasets, many have anticipated methods that can tap into gene expression databases ever since their development a decade ago (Bassett *et al*., 1999; Hunter *et al*., 2001). The expectation was that such resources contain rich biological phenotypes that could be mined for value. To do this, a wealth of approaches have been developed that can search gene expression databases based on associated metadata, but there are fewer options for searching the gene expression profiles themselves (Praz *et al*., 2004; Rhodes *et al*., 2007; Yu *et al*., 2009; Zhu *et al*., 2008). To date, there have been two systems. The first, SPELL, was developed to predict Gene Ontology (GO) terms for yeast genes based on patterns seen across a gene expression database (Ashburner *et al*., 2000; Hibbs *et al*., 2007). Although it does score the datasets, the focus is on the annotation of single genes, and it does not provide a statistical significance. A more recently developed algorithm does provide the significance, but requires as input a weighted list of genes and thus cannot be applied to gene sets, which are unweighted (Engreitz *et al*., 2010b). Therefore, there is not yet an ability to fully capitalize on the data within gene expression databases.

To address the need for a method to mine gene expression databases, we have developed a novel approach to identify the

*To whom correspondence should be addressed.

datasets in which a given gene set is regulated. At its heart is a *coherence*-based algorithm based on the straightforward notion that if a gene set is tightly regulated in a dataset, as evidenced by high correlation among its genes, then it must play an important role in that condition (Banerjee and Zhang, 2003; Beer and Tavazoie, 2004; Pilpel *et al*., 2001; Singh *et al*., 2007). Applying this, algorithm across a gene expression database then generates a *regulatory profile* that reveals the cellular conditions in which a specific biological process is tightly regulated. To accomplish this, we extended the coherence algorithm to work at large scale by developing a novel approach to estimate the significance of a coherence score. Then, we evaluated the accuracy of the resulting method on two tasks in which the outcome is known: identifying regulation of oncogenic pathways, and predicting the tissue specificity of gene sets. Finally, we applied it to address two unsolved biological questions: unraveling the combinatorial regulatory network of the E2F transcription factors, and developing an unbiased catalog of biological processes. These experiments demonstrated the utility of the concept of a regulatory profile and the power of leveraging gene expression databases in addressing a range of biological questions.

2 METHODS

We have developed a method to quantify the evidence for the regulation of a gene set in microarray gene expression dataset (Fig. 1A). This method was an extension of a previously described coherence score that measured whether a set of genes in a gene set was significantly co-regulated in a dataset (Pilpel *et al*., 2001). Briefly, to calculate this score given a gene set and a microarray gene expression dataset, we selected the genes from the gene set and calculated the pairwise Pearson correlations of their expression profiles (Fig. 1B). Next, we discarded the insignificant correlations that could be obtained from at least 95% of random pairs of genes. The coherence score was the maximum number of genes with significant correlations. To obtain a *P*-value for the score, we generated a null distribution by repeating this procedure with sets of randomly selected genes (Fig. 1C). This background distribution was conditioned upon the size of the gene set, and empirically, we found that this distribution also varied across datasets (data not shown). Because it was not computationally practical to sample the null distribution exactly for all gene set sizes across a large gene expression database, we developed a novel strategy where we computed the null distribution for a limited number of sizes of gene sets and then estimated the distributions for the remaining sizes (Fig. 2A). We did this by modeling the background distribution using Parzen windows (Fig. 2B) and then interpolating the distributions for arbitrary sizes of gene sets (Fig. 2C). See the Supplementary Materials for a detailed description of this algorithm.

3 RESULTS

3.1 Matching gene sets to datasets

To evaluate the accuracy of the coherence score, we applied it to nine microarray datasets containing gene sets of cell cycle signaling pathways generated from human mammary epithelial cells (Bild *et al*., 2006; Gatza *et al*., 2010). Each dataset consisted of samples from ∼10 replicates of cells where an oncogenic pathway was activated, and ∼10 replicates where the pathway was inactivated. From each dataset, we derived a set of the genes whose expression changed in response to pathway activity. Then, we tested whether the coherence score could recognize regulation of each gene set in the proper dataset.

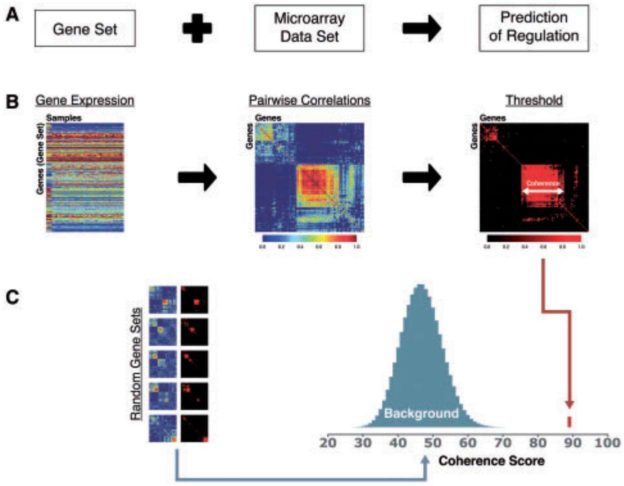


Fig. 1. (**A**) We predict whether a set of genes shows evidence of transcriptional regulation in a microarray gene expression dataset. (**B**) To do this, we start first with the gene expression profiles of the genes in the gene set across the samples in the dataset. Shown are the genes from gene set BILD_E2F3_ONCOGENIC_SIGNATURE from MSigDB v3.0 in dataset GSE23402 from GEO. Next, we calculate the Pearson correlations of each pair of genes in the gene set. Then, we apply a threshold to remove all non-significant correlations. We finally derive a *coherence score* based on the number of genes that exhibit significant correlation. (**C**) We can calculate the statistical significance of the coherence score from the scores that are obtained by gene sets comprised of randomly selected genes.

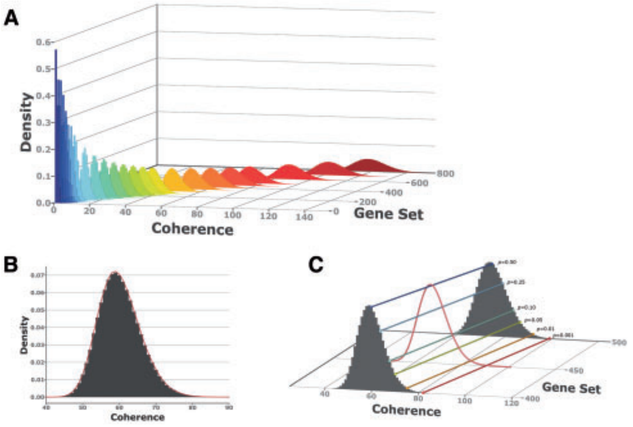


Fig. 2. (**A**) The background distribution of coherence scores varies with the size of the gene set. This plot shows histograms of the density (*Y*-axis) of the coherence scores (*X*-axis) from randomly generated gene sets of various sizes (*Z*-axis) sampled from dataset GSE15026. As the size of the gene set increases, the expected size of the coherence score also increases. These comprise the null distribution of the coherence score for gene sets of specific sizes. (**B**) The density of the coherence score for a specific size of gene set (400 genes shown here) can be modeled with Parzen windows (shown in the red line). (**C**) For gene sets of sizes that have not previously been sampled, we can estimate its null distribution using linear interpolation from those that were empirically determined. Here, the null distribution for a gene set of size 450 (shown as a red line) is interpolated from those of sizes 400 and 500 (shown as black bars). Coherence scores with equivalent *P*-values across the three sizes of gene sets are shown.

Table 1. This table shows the coherence scores of gene sets indicating activation of pathways across the gene expression datasets

	Gene Sets	AKT1	CTNNB1	E2F1	E2F3	MYC	TP63	PIK3CA	HRAS	SRC	E2F1/2
Datasets	**AKT**	**91**	17	102	48	50	76	88	71	13	21
	CTNNB1	51	**34**	45	30	20	51	43	51	26	12
	E2F1	89	16	**161**	134	53	73	106	75	14	**24**
	E2F3	69	28	141	**161**	62	50	74	68	22	18
	MYC	62	27	45	28	**82**	47	45	67	24	14
	TP63	89	14	121	79	42	**82**	99	66	15	23
	PIK3CA	85	19	124	77	57	79	**121**	74	14	21
	HRAS	59	15	57	37	60	61	63	**153**	15	10
	SRC	61	33	59	35	34	62	49	61	**33**	15
	Total genes	210	72	205	226	182	204	217	234	54	27

Each column represents a gene set and the rows contain of the datasets. The last row indicates the number of unique genes in each the gene set. For each gene set, the dataset with the highest coherence score is indicated by bold values.

For the evaluation, we calculated the coherence scores for each gene set across each dataset (Table 1). As a control, we selected a biologically distant E2F gene set generated from genes induced by E2F1 or E2F2 in mouse embryo fibroblasts (Ishida *et al*., 2001). The data showed that the gene set for each pathway obtained the highest coherence score in the corresponding dataset, indicating that the coherence measure could correctly identify the dataset with the most significant regulation of the gene set. To show that the coherence scores were statistically significant, we generated the null distribution for each gene set from 100 000 random samples and found that the correct dataset in each case was significant with $P < 0.00001$ (Supplementary Table S1). Furthermore, the relative magnitudes of the scores recapitulated known biological connections among the pathways. For instance, while the E2F1 gene set scored highest in the E2F1 dataset, the second highest score was obtained in the E2F3 dataset. Conversely, the E2F3 gene set was second highest in the E2F1 dataset. The close relation of the coherence scores likely reflects the fact that these proteins can transcriptionally activate many of the same targets (Trimarchi and Lees, 2002). However, there were known distinctions in their functions as well. E2F1 has a unique ability to induce apoptosis in a manner connected to PI3K/Akt signaling that is not been seen in E2F3 (Hallstrom *et al*., 2008; Kowalik *et al*., 1995). Mirroring the biochemical dichotomy, the E2F1 gene set scored relatively high in the PI3K and Akt datasets, whereas the E2F3 gene set did not. In the analysis of the mouse E2F1/2 gene set, it scored highest against the E2F1 dataset and fifth against E2F3, behind P63, PI3K and Akt, suggesting that the conditions used to generate this gene set may have triggered an apoptotic response. Taken together, these analyses demonstrated that the coherence score could accurately quantify regulation among closely linked pathways.

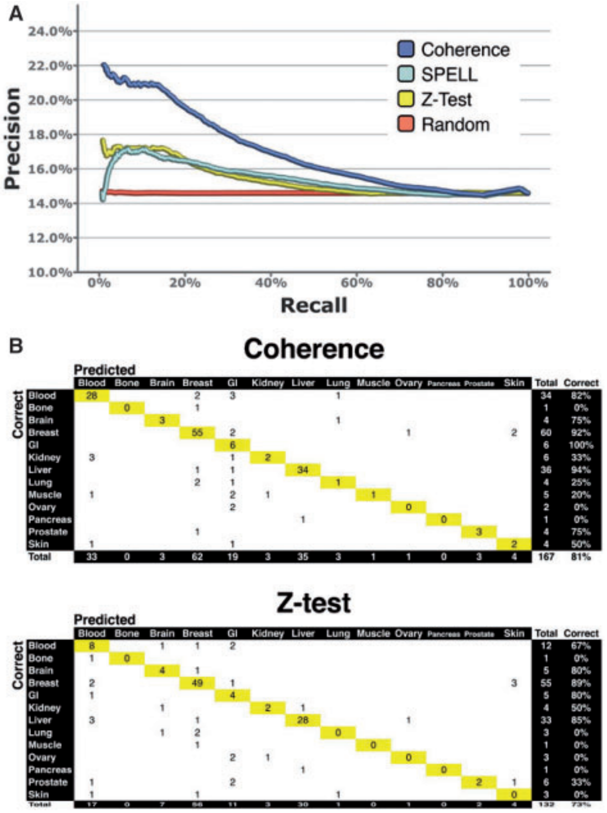
3.2 Predicting tissue specificity of gene sets

As a further evaluation of the ability of the algorithm to quantify regulation, we tested whether it could correctly distinguish the tissue of origin of a gene set. We leveraged the two facts that (i) many gene sets were generated within a specific tissue context, and (ii) many gene expression datasets contained samples collected from a single type of tissue. Because gene sets could contain imprints of tissue-specific transcriptional programs, we reasoned that they should score higher in datasets from the same tissue than those from others. This was a difficult test case because the biological processes embodied by the gene sets were convoluted with the signal from the tissue type.

For this experiment, we collected 406 gene sets from MSigDB and 918 datasets from the Gene Expression Omnibus (GEO) gene expression database covering 13 different tissues (Barrett *et al*., 2011). We then applied our algorithm and scored the significance of all pairwise associations of the gene sets and datasets. As a comparison, we implemented the Z-test and SPELL algorithms, as well as a random scoring function as a baseline (Hibbs *et al*., 2007). For the 4 algorithms, we ranked all 372 708 associations by score and calculated the recall and precision curves, counting an association correct if the gene set and dataset were generated from the same tissue. While all algorithms performed better than random, the coherence score obtained significantly higher precision for the same levels of recall as compared with the others (Fig. 3A).

Although this experiment demonstrated an ability to identify tissue-specific signals in the gene sets, a limitation of the approach stems from the fact that many biological processes generated similar transcriptional profiles across tissues, and in this case, a gene set would obtain a high regulation score in multiple tissues. To control for this, we collated all scores for each gene set and scored the significance of their association with each tissue type. So for each gene set, we counted the number of datasets of each tissue type that was significantly associated with it (false discovery rate; FDR < 0.01), and calculated the statistical significance of the association using a Fisher's exact test (Benjamini and Hochberg, 1995). We then predicted the tissue that obtained the highest classification. If no datasets reached the statistical cutoff, then no prediction was made. While the previous test generated 372 708 predictions, this one produced only 406, one for each gene set.

Comparing the coherence and Z-test algorithms, we found that the coherence score could predict the tissue for 167 out of 406 gene sets (the remaining did not achieve statistical significance), of which 81% were correct (Fig. 3B, Supplementary Table S2). In comparison, the Z-test made fewer predictions, 132, with an accuracy of 73%. Thus, the coherence score obtained a higher accuracy with a larger number of predictions. Because the SPELL algorithm did not produce P-values, we could not determine a statistical cutoff for determining association. Nevertheless, we tried a range of *ad-hoc* cutoffs and found that it could achieve 172 predictions with an accuracy of 60% (data not shown).



	Blood	Bone	Brain	Breast	GI	Kidney	Liver	Lung	Muscle	Ovary	Pancreas	Prostate	Skin	Total	Correct
Blood	28			2	3			1						34	82%
Bone		0		1										1	0%
Brain			3					1						4	75%
Breast				55	2					1			2	60	92%
GI					6									6	100%
Kidney	3				1	2								6	33%
Liver				1	1		34							36	94%
Lung				2	1			1						4	25%
Muscle	1				2	1			1					5	20%
Ovary					2					0				2	0%
Pancreas							1				0			1	0%
Prostate				1								3		4	75%
Skin	1				1								2	4	50%
Total	33	0	3	62	19	3	35	3	1	1	0	3	4	167	81%

	Blood	Bone	Brain	Breast	GI	Kidney	Liver	Lung	Muscle	Ovary	Pancreas	Prostate	Skin	Total	Correct
Blood	8		1	1	2									12	67%
Bone	1	0												1	0%
Brain			4	1										5	80%
Breast	2			49	1								3	55	89%
GI	1				4									5	80%
Kidney			1			2	1							4	50%
Liver	3			1			28			1				33	85%
Lung			1	2				0						3	0%
Muscle				1					0					1	0%
Ovary					2	1				0				3	0%
Pancreas							1				0			1	0%
Prostate	1				2							2	1	6	33%
Skin	1			1				1					0	3	0%
Total	17	0	7	56	11	3	30	1	0	1	0	2	4	132	73%

Fig. 3. (**A**) This plot quantifies the ability of the coherence, SPELL, Z-test, and a random baseline algorithm to identify gene sets and datasets of the same tissue type. The recall (X-axis) and precision (Y-axis) curves are shown for each algorithm. (**B**) For each gene set, we group together the scores across all datasets and generate a single tissue type prediction for that gene set. These matrices show the correct tissue type on the vertical access and the predicted tissue type on the horizontal. The numbers in the matrices indicate the number of gene sets predicted. The gene sets on the diagonal are predicted correctly, while those off the diagonal are incorrect. The final two columns of the matrices show the total number of gene sets predicted and the percent of predictions that are correct for each tissue type. Prediction matrices are provided for the coherence and Z-test algorithms.

Overall, these results demonstrated that the coherence measure was best able to detect the regulation of tissue-specific transcriptional programs in the gene sets. Furthermore, it also suggested that while the gene sets embodied biological processes, they could also be influenced by the context of the cell type, a dimension that had not previously been systematically investigated.

3.3 Deciphering transcriptional regulatory programs

Having observed that our algorithm could accurately detect regulation of gene sets, we then applied it to decipher transcriptional regulatory programs. While the principle that combinations of transcription factors regulate the expression of genes is now well understood, knowledge of the functioning of specific transcription factors and co-factors remains limited (Fedorova and Zink, 2008; Ravasi *et al.*, 2010; Stanojevic *et al.*, 1991). One common approach to identifying co-factors is based on over-representation of transcription factor binding sites in sequences from promoters or ChIP experiments (Berger *et al.*, 2008; Harbison *et al.*, 2004; Johnson *et al.*, 2007). A limitation is that sequences do not provide evidence of functional activation. To address this, we applied our algorithm to identify transcriptional co-factors based on evidence of regulation across a large gene expression database.

We chose as our model of combinatorial regulation the E2F family of transcription factors. E2F regulates the cell cycle transition from G_1 to S phase and is also involved in apoptosis, development and other functions (Dimova and Dyson, 2005; Nevins, 1998). Functionally, it sits at the nexus of many processes that drive the cancer phenotype and is, thus, an important area of research. E2F activates target genes dependent upon the presence of transcriptional co-activators. While some co-factors have been established, they do not yet explain the regulation of the diversity of activities and it is believed that others are required to maintain proper E2F function.

To identify potential E2F co-factors, we paired it with each other transcription factor from the TRANSFAC and JASPAR databases, a total of 634 possible pairs (Bryne *et al.*, 2008; Matys *et al.*, 2006). For each pair, we collected the genes whose promoters contained binding sites for both transcription factors at a significance of $P < 0.0001$ within 500 bp (Fig. 4A). Those genes comprised a gene set for the predicted regulatory target of E2F and its co-factor. We scored each of the gene sets against the 2641 datasets in our GEO database (Supplementary Table S3) and obtained their regulatory profiles. In each profile, we counted the number of datasets that exhibited significant regulation (FDR < 0.01; Supplementary Table S3). Finally, we plotted the regulatory profiles of each co-factor that was significant in at least 10 datasets across 2 principal components (Fig. 4B). These results revealed that the co-factors that exhibited the broadest regulation corresponded to known co-factors, such as NFY, SP1, TFE3 (an EBOX binding protein) and YY1 (Giangrande *et al.*, 2003; Karlseder *et al.*, 1996; Schlisio *et al.*, 2002; Zhu *et al.*, 2004). As a control, we also identified E2F itself. One known co-factor that we did not recognize was MYB (Zhu *et al.*, 2004). A major difference between MYB and the other co-factors is that it regulates events at the G_2/M transition of the cell cycle while the others operate at G_1/S. This suggests that the approach outlined here may be sensitive to the regulatory profile of the co-factors being profiled. Nevertheless, the algorithm recovered known E2F co-factors, and the remaining ones represented experimentally testable predictions that were supported by evidence from both promoter sequence and transcriptional regulation. Using this analysis, we could predict the identity of the co-factors, but could not distinguish which E2F is the partner because they all interact with the same consensus promoter sequence. Given these results, this initial demonstration shows the power of leveraging large gene expression databases for decoding combinatorial transcriptional regulatory programs and suggests a framework for deciphering the complete transcriptional regulatory network across all genes.

3.4 An unbiased catalog of biological processes

A long-standing problem is how to identify the underlying processes that drive the phenotype of a cell. While gene expression profiles, which show the activation of individual genes, is one approach, individual genes provide only limited insight into the processes that are driven by the actions of multiple ones. The solution, then, is to examine sets of genes for enrichment of GO annotations

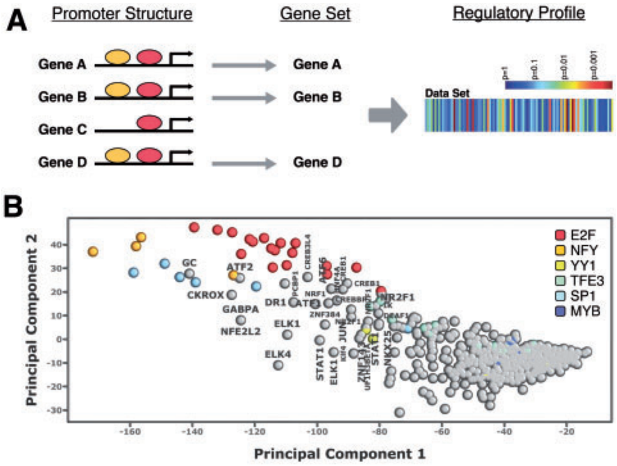


Fig. 4. (**A**) To score the regulatory potential of a pair of transcription factors, we first identify the genes in the genome that contain binding sites for E2F as well as a potential co-factor. These genes comprise the gene set for the co-factor. Then, we score the regulation of this gene set across our gene expression database, generating a profile showing the breadth of the regulation of this gene set across the datasets. (**B**) We plot the regulatory profiles of each transcription co-factor onto the first two principal components. Each circle represents a position weight matrix (PWM) for a co-factor. Some co-factors are represented by multiple PWMs, and appear more than once in the plot. E2F co-factors that have previously been identified experimentally are colored according to the legend.

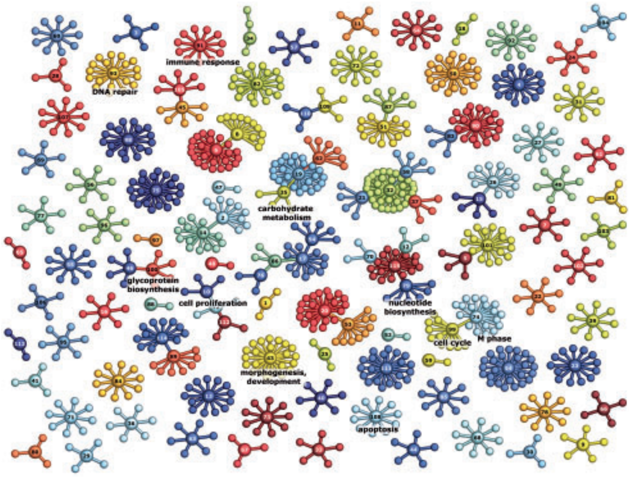


Fig. 5. This plot shows distinct transcriptional regulatory profiles. Each sphere represents a gene set. They are grouped into processes according to similarity of the regulation exhibited across a gene expression database. Each distinct process is represented by an exemplar, as determined by the affinity propagation algorithm. The exemplars are shown with a larger diameter and are labeled with unique identifiers. There is no significance to the coordinates of the layout. Each process is shown in a different color, chosen randomly and GO terms are labeled for selected ones.

(Huang da *et al.*, 2009a). However, a limitation of this approach is the dependence upon the comprehensiveness of GO, which is created based on manual curation of biological functions. To investigate this issue, we sought to develop a catalog of unbiased biological processes (UBPs).

To define the UBPs, our solution was to leverage gene sets. While in principle, MSigDB already provides gene sets that can be used to describe biological processes, the reality is that many distinct gene sets can describe the same one (Ein-Dor *et al.*, 2005). Thus, we used MSigDB as a starting point and refined it based on their regulatory profiles. We did this by computing a profile for each of the 6769 gene sets against our GEO database and discarding the ones that did not achieve significant regulation (see Section 2). Then, to identify groups of gene sets sharing similar transcriptional profiles, we applied affinity propagation (Frey and Dueck, 2007). This yielded a collection of 115 distinct UBPs made from 995 gene sets (Fig. 5). Each UBP was associated with between 2 and 32 gene sets (median 7), and 11 and 4431 genes (median 593).

To compare the UBPs with GO, we annotated each one using GATHER (Chang and Nevins, 2006). First, we noted that many of the GO functions were significant across multiple UBPs (Supplementary Table S4). For example, the annotations *immune response* and *metabolism* were each significantly associated with 16 programs, with no overlap between them. Indeed, 32% of all UBPs were involved in either cell cycle: metabolic or immune response activities, demonstrating their breadth and complexity of regulation. However, these general processes could be broken down into more specific functions. For instance, UBP 99 consists of gene sets more generally associated with *cell cycle*, and is connected to UBP 74, which is more specifically associated with *M phase* and related cytoskeletal events (Fig. 5). Similarly, metabolic activities could be broken down into *glycoprotein biosynthesis* (UBP 100), *nucleotide biosynthesis* (UBP 79), *carbohydrate metabolism* (UBP 35) and others. This demonstrated that these processes corresponded to well-defined functions and also made distinctions among them.

Overall, we saw a broad concordance between our processes and GO terms (Supplementary Table S5). As examples, UBP 91 contained gene sets for chemokines, cytokines, IL4 and inflammatory response, a clear correspondence to its GO term for *immune response* (FDR = 10^{-225}). Similarly, UBP 93 included gene sets for base excision repair, double strand break repair and homologous recombination; and was annotated with the GO term for *DNA repair* (FDR = 10^{-225}). Finally, UBP 108, UV response, apoptosis and programmed cell death, had the GO term for *apoptosis* (FDR = 10^{-225}).

However, a significant number of UBPs did not relate directly to GO functions. One possibility is that they were artifacts that did not correspond to any physiological events. To determine whether this was the case, we examined the gene sets for these UBPs and found several reasons why they could not be annotated with GO. For one, the gene sets for UBP 43 suggested a regulatory program associated with differentiation of breast cancer subtypes, e.g. *FARMER_BREAST_CANCER_BASAL_VS-_LUMINAL*. The assigned GO terms, *morphogenesis* (FDR = 10^{-12}) or *development* (FDR = 10^{-10}), supported the biology driving the subtypes, but did not include one specifically associated to the breast subtypes in the process. While GO included terms describing breast differentiation, they referred only to events in normal development. Even though this UBP was among the largest processes (comprised of 23 individual gene sets), indicating a great interest in this phenotype, there was not a GO annotation available for describing it. As another example, UBP 10 was comprised of gene sets for transcriptional targets of MYC, a master regulator of the cell cycle. But, it was assigned a more generic GO annotation of

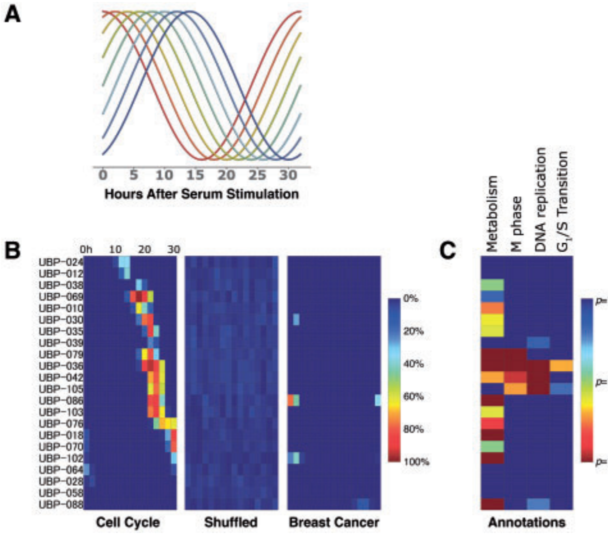


Fig. 6. (**A**) We generated 16 sinusoidal waves that peaked every 2 h across a 32 h cell cycle dataset. The first eight waves are shown here. (**B**) These heatmaps show the percentage of simulations in which a UBP (on the rows) are significantly associated with a sine wave that peaks at a specific time (on the columns). The UBPs are manually ordered so that those that peak earlier are shown closer to the top. The three heatmaps show the results when the procedure is applied to a cell cycle dataset, the dataset shuffled to remove temporal patterns and a breast cancer dataset. (**C**) This heatmap shows the association between the genes in each UBP with GO annotations.

cell proliferation (FDR = 10^{-17}). To test whether the differences in the GO terms were due to limitations in the annotation tool, we re-ran the analyses using DAVID and saw no notable differences (data not shown) (Huang da *et al*., 2009b). Thus, in these cases, the GO annotations supported the process described by the gene sets, but the ontology itself lacked precise terms to describe the underlying process.

To evaluate (i) whether the UBPs were biologically meaningful, and (ii) whether the GO annotations accurately described the function of the UBPs, we scored the UBPs against a cell cycle dataset measuring the expression profiles of synchronized human cells for one cell cycle after serum stimulation (Bar-Joseph *et al*., 2008). Since it has previously been shown that cycling genes can be modeled by sine waves (Whitfield *et al*., 2002), we generated a series of waves that peaked every 2 h, and used Significance Analysis of Microarrays (SAM) to score whether the gene sets from the UBPs exhibit sinusoidal oscillation (Fig. 6A; Efron and Tibshirani, 2007). We used two negative controls: we shuffled with disrupted temporal order and a breast cancer dataset that should contain no sinusoidal expression patterns (Sircoulomb *et al*., 2010). From this experiment, we find a set of UBPs that are expressed at different times across the cell cycle, and not in the negative controls (Fig. 6B). Examining the GO annotations associated with them, we find that *Metabolism* is associated with UBPs across all time points, while *DNA replication* is limited to UBPs that peak ~20–26 h (Fig. 6C), which corresponds closely with the accumulation of 4N DNA content in the published flow cytometry data. This demonstrates that the UBPs are biologically meaningful and that their activity can be detected in gene expression data.

Although GO is by far the most popular vocabulary for describing biological processes, as attested by the number of tools developed to use those annotations, our analysis shows that it is limited in its ability to describe the entirety of biological processes. The reason for this is due to both the enormity of the task of describing all processes of interest and another is because many processes are not yet studied or understood. Therefore, we propose that an unbiased approach to catalog observed biological processes results in a more precise and empirical measure of independently regulated biological processes. We do not claim that the UBPs will replace GO, but it may supplement it with an independent catalog of biological process.

To our knowledge, there are two prior efforts to automatically derive biological processes from databases of gene expression profiles. One identified network motifs that are common across multiple human datasets, resulting in 143 400 functionally homogeneous *modules* (Huang *et al*., 2007). The second performed a matrix decomposition using independent component analysis across human gene expression data (Engreitz *et al*., 2010a). Using this approach, they found 423 statistically independent components that could be linked to functional annotations. Including UBP, there are now three distinct automatically generated databases of biological process: one that represents process as a gene set, one where processes are modeled as weighted gene signatures and one where processes are networks of co-expressed genes. While it is likely that these catalogs would also be incomplete, they nevertheless provide practical frameworks that can be refined as algorithms and databases are updated.

4 CONCLUSIONS

Signatures or gene sets form the dominant strategy behind current efforts to decipher the meaning behind cellular transcriptional profiles. While a gene expression profile can be thought of as a high-throughput method to measure the expression levels of genes, signatures provide a means to interpret the larger scale processes that are brought about by the concerted efforts of groups of genes. The ultimate goal is the capacity to derive an understanding of the state of a cell based on its transcriptional profile. To get closer to this goal, we have investigated the novel concept of a regulatory profile that revealed the regulation of gene set across a gene expression database, and we have developed a method to derive this profile. Analogous to how a gene expression profile reveals information not evident in single genes, a regulatory profile reveals patterns by tapping into the power and richness of biological phenotypes across a gene expression database.

While the use of the coherence measure to quantitate the regulation of a gene set was not novel, our contribution was the mechanics to compute it so that it could be scaled to score gene sets across gene expression databases. However, there were several limitations of our approach. First, because we used an empirical null distribution, the sensitivity of the approach was limited by the amount of sampling done. This leads to a fundamental consequence of this approach, which is that, while we can distinguish the relative regulation of a gene set across multiple datasets, the complementary analysis is difficult. Given a single dataset, it can be hard to differentiate the activation across multiple gene sets, since they may all be very significant, as can be seen in Supplementary Table S1. Furthermore, the sampling required would grow linearly with the size of the gene expression database. Finally, our scoring method did not distinguish between genes that were up-regulated and those that were down-regulated.

Nevertheless, applying regulatory profiles, we provided demonstrations of how they may be used to solve critical biological problems. In the first analysis, we used the profiles to recapitulate the known portion of the E2F transcriptional regulatory program, as well as to predict several other previously unreported co-factors. For the second, we leveraged the profiles to generate an unbiased catalog of 115 distinct biological processes. Because these were generated in an unbiased method, we propose that this represents a more robust foundation to decipher the processes underlying a cellular state. In sum, these experiments have revealed the value that can be derived by tapping into the power of large gene expression databases.

ACKNOWLEDGEMENTS

I thank Joseph Nevins, Sayan Mukherjee and the Nevins lab for helpful discussions.

Funding: J.T.C. is supported by NIH R00LM009837 and grant R1006 from the Cancer Prevention & Research Institute of Texas.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Bar-Joseph,Z. *et al.* (2008) Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl Acad. Sci. USA*, **105**, 955–960.

Barbie,D.A. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.

Bassett,D.E., Jr. *et al.* (1999) Gene expression informatics—it's all in your mine. *Nat. Genet.*, **21**, 51–55.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.Roy. Stat. Soc.*, **57**, 289–300.

Berger,M.F. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.

Bild,A.H. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

Chang,J.T. and Nevins,J.R. (2006) GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics*, **22**, 2926–2933.

Dimova,D.K. and Dyson,N.J. (2005) The E2F transcriptional network: old acquaintances with new faces. *Oncogene*, **24**, 2810–2826.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Ein-Dor,L. *et al.* (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Engreitz,J.M. *et al.* (2010a) Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.*, **43**, 932–944.

Engreitz,J.M. *et al.* (2010b) Content-based microarray search using differential expression profiles. *BMC Bioinformatics*, **11**, 603.

Fedorova,E. and Zink,D. (2008) Nuclear architecture and gene regulation. *Biochim. Biophys. Acta.*, **1783**, 2174–2184.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gatza,M.L. *et al.* (2010) A pathway-based classification of human breast cancer. *Proc. Natl Acad. Sci. USA*, **107**, 6994–6999.

Giangrande,P.H. *et al.* (2003) Identification of E-box factor TFE3 as a functional partner for the E2F3 transcription factor. *Mol. Cell. Biol.*, **23**, 3707–3720.

Hallstrom,T.C. *et al.*(2008) An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. *Cancer Cell*, **13**, 11–22.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hibbs,M.A. *et al.* (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.

Huang,Y. *et al.* (2007) Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics*, **23**, i222–i229.

Huang da,W. *et al.* (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Huang da,W. *et al.* (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Hunter,L. *et al.* (2001) GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, **17** (Suppl. 1), S115–S122.

Ishida,S. *et al.* (2001) Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol. Cell. Biol.*, **21**, 4684–4699.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Karlseder,J. *et al.* (1996) Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F. *Mol. Cell. Biol.*, **16**, 1659–1667.

Kowalik,T.F. *et al.* (1995) E2F1 overexpression in quiescent fibroblasts leads to induction of cellular DNA synthesis and apoptosis. *J. Virol.*, **69**, 2491–2500.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Liu,Z. *et al.* (2008) Singular value decomposition-based regression identifies activation of endogenous signaling pathways in vivo. *Genome Biol.*, **9**, R180.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Miller,L.D. *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.

Nevins,J.R. (1998) Toward an understanding of the functional complexity of the E2F and retinoblastoma families. *Cell Growth Differ.*, **9**, 585–593.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Praz,V. *et al.* (2004) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res.*, **32**, D542–D547.

Ravasi,T. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.

Rhodes,D.R. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.

Schlisio,S. *et al.* (2002) Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. *EMBO J.*, **21**, 5775–5786.

Singh,L.N. *et al.* (2007) TREMOR—a tool for retrieving transcriptional modules by incorporating motif covariance. *Nucleic Acids Res.*, **35**, 7360–7371.

Sircoulomb,F. *et al.* (2010) Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer*, **10**, 539.

Spang,R. *et al.* (2002) Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol.*, **2**, 369–381.

Stanojevic,D. *et al.* (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science*, **254**, 1385–1387.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Trimarchi,J.M. and Lees,J.A. (2002) Sibling rivalry in the E2F family. *Nat. Rev. Mol. Cell Biol.*, **3**, 11–20.

Whitfield,M.L. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Yu,Y. *et al.* (2009) GEOGLE: context mining tool for the correlation between gene expression and the phenotypic distinction. *BMC Bioinformatics*, **10**, 264.

Zhang,X.H. *et al.* (2009) Latent bone metastasis in breast cancer tied to Src-dependent survival signals. *Cancer Cell*, **16**, 67–78.

Zhu,W. *et al.* (2004) E2Fs link the control of G1/S and G2/M transcription. *EMBO J.*, **23**, 4615–4626.

Zhu,Y. *et al.* (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**, 2798–2800.