Revealing global regulatory perturbations across human cancers

Hani Goodarzi#, Olivier Elemento#, and Saeed Tavazoie*

Department of Molecular Biology & Lewis-Sigler Institute for Integrative Genomics
Princeton University, Princeton, NJ 08544.

*To whom correspondence should be addressed. Email: tavazoie@genomics.princeton.edu
#These authors contributed equally.

Keywords: Cancer, regulatory networks, signaling pathways, cancer pathway map, cancer regulatory map, PAGE, FIRE.
Summary

Perturbations to diverse regulatory pathways are essential steps along the path to neoplastic transformation. The discovery of these pathways in their native cellular context remains a fundamental challenge for cancer biology. We show that such pathway perturbations, and the cis-regulatory elements through which they operate, can be efficiently extracted from microarray expression studies. Our approach relies on information-theoretic analysis of expression levels, pathways, and local genomic sequences. Analysis across a diverse set of human cancers reveals the vast majority of previously known cancer signaling pathways. In addition, through de novo motif discovery, we associate these pathways with transcription-factor binding sites and miRNA targeting sequences, including those of E2F, NF-Y, p53, and let-7. Strikingly, the majority of the perturbations, strongly associated with novel cis-regulatory elements, fall outside of known cancer pathways. Our study provides a systems-level dissection of regulatory perturbations in cancer—an essential component of a rational strategy for therapeutic intervention and drug-target discovery.
Introduction

Precise molecular definition of pathologic states is an essential component of a rational approach to understanding and treating disease. This is especially true in cancer, where many complex cellular pathways contribute to the initiation and maintenance of the transformation process. Throughout the last decade, microarrays have been widely used for discovering significantly deregulated genes in the tumor samples in order to identify diagnostically and prognostically relevant “molecular signatures” (Rhodes et al., 2004). However, it is becoming increasingly clear that tumor state heterogeneity can often be more accurately described by the behavior of functionally coherent and coordinately regulated sets of genes. Thus, molecular signatures are moving towards pathway-level definitions (Segal et al., 2004; Subramanian et al., 2005). In fact, neoplastic transformation relies on deregulation of diverse oncogenic and tumor-suppressor pathways (Watters and Roberts, 2006), including stimulation of cell growth and proliferation and inhibition of cell cycle arrest and apoptosis (Adjei and Hidalgo, 2005). Global deregulation of these pathways is typically achieved through somatic mutations in key signaling molecules (e.g. Imai et al., 1998), transcription factors (e.g. Gallie, 1994), and post-transcriptional regulators such as microRNAs (e.g. Tavazoie et al., 2008; Wu et al., 2008). Systematic identification of these deregulated pathways and their underlying mutations is a crucial first step in developing a rational strategy for cancer therapy.

In this study, we present a coherent conceptual framework for systematic identification of deregulated pathways and the local regulatory elements through which these changes are orchestrated. Initially, we use a mutual information-based approach (PAGE for Pathway Analysis of Gene Expression) to discover the deregulated pathways across large-scale cancer gene expression measurements. Subsequently, we seek to identify the key regulators (transcription factors, RNA-binding proteins and miRNAs) whose perturbations are the direct or indirect cause of the observed changes in transcript levels. In many cases, due to the extensive regulation at the post-translational level, direct
identification of these factors is difficult to achieve from gene expression data alone. We argue that regulator identification is best achieved through the discovery of cis-regulatory elements that underlie the observed expression changes. For this task, we employ FIRE, a robust framework for finding cis-regulatory modules that are informative of the differential gene expression patterns (Elemento et al., 2007). The identity of these regulatory elements are then inferred by comparison to eukaryotic transcription factor binding sites in JASPAR (Sandelin et al., 2004) or TRANSFAC (Matys et al., 2006). Upon finding the motifs whose patterns of presence or absence across all considered regulatory regions (promoters or 3’UTRs) are most informative, we use another information-theoretic approach to associate the discovered sequence motifs with the deregulated cellular pathways. We show that this latter analysis essentially reveals the pathways that are regulated through the discovered putative binding sites by their associated regulatory proteins or RNAs.

Upon applying this generic computational approach to a large number of cancer gene expression datasets, we show that it successfully recapitulates the role of many known transcription factors and miRNAs in cancer. This comprehensive analysis of cancer gene expression represents a crucial first step towards achieving a systems-level understanding of regulatory perturbations in cancer. Our discovery of a large repertoire of significant cis-regulatory elements, many of which are RNA binding sites, and their associations with key cellular pathways, highlights our limited current understanding of cancer pathway perturbations.

**Results**

**Discovering deregulated pathways: description of approach and application to bladder cancer**

Urinary bladder cancer is the fourth most frequent cancer among men and it ranks as the fifth most common malignancy in the United States. Recently, gene expression signatures of bladder cancer have
been obtained through whole-genome microarrays (Dyrskjot et al., 2004). Using the same microarray dataset, we sought to discover the pathways that show significant differential expression in tumor samples compared to normal controls.

Several methods have been developed to perform this type of analysis. For example, the classic method based on the hypergeometric distribution (Tavazoie et al., 1999) relies on the statistical evaluation of the overlap between genes in a particular group, and genes associated with a given pathway. Since the gene expression data is often partitioned into many groups, and each group has to be tested independently for overlap with many pathways, this typically involves a large number of statistical tests, which leads to lower sensitivity upon correction for multiple hypothesis testing. On the other hand, T-profiler (Boorsma et al., 2005)—a more recent method that uses $t$-tests to discover deregulations—relies on the assumption that the expression values are normally distributed; also, similar to GSEA (Subramanian et al., 2005) it can only process continuous expression values (e.g., log-ratios from a single microarray experiment), meaning co-expression clusters can not be used as inputs. This is an important limitation, since using co-expression clusters can significantly improve the signal to noise ratio, by taking into account gene expression behavior across many conditions and/or perturbations (Tavazoie et al., 1999).

In this study, we describe a general information-theoretic approach for discovering deregulated pathways from genome-wide gene expression measurements without the limitations described above. Our framework, termed PAGE, operates on both continuous and discrete expression variables and uses mutual information (Cover and Thomas, 2006) to quantify the correlation between a given expression variable and pathway membership. The latter is obtained from publicly available pathway annotations, e.g., Gene Ontology (Ashburner et al., 2000) or MSigDB (Subramanian et al., 2005). Our approach uses non-parametric statistical tests to determine whether a pathway is significantly informative about
the observed expression measurements (i.e. whether the information is higher than expected by chance). The simultaneous contribution of pathway enrichments and depletions (i.e. when a pathway is under-represented in a cluster of co-expressed genes) to the mutual information value increases the overall sensitivity and specificity and results in very low false-discovery rates (see Table S1). Moreover, our approach includes a principled way to deal with pathway description redundancy and nesting, a feature of most functional annotations (e.g., GO) that can lead to tedious result interpretation. This is achieved in PAGE by using the notion of conditional information; briefly, PAGE only returns pathways that provide a significant amount of new information about the expression measurements given all other significantly informative pathways (see Supplemental Procedures for details). Finally, the automatic output of PAGE consists of a graphical heat-map where enrichment and depletion of each pathway, in co-expression clusters or expression bins, are shown (see Supplemental Procedures).

In the analysis of the bladder cancer gene expression dataset, we processed the samples so that each gene is assigned a real number, ranging from -1 to 1, representing the extent to which the gene is deregulated in the tumor samples compared to the controls from healthy individuals (zero indicates no change in expression; see Methods for details). We then used PAGE to search for the pathways (“Biological processes” categories in the Gene Ontology annotations) that are most informative about the observed gene expression changes. As shown in Figure 1, we found 16 non-redundant pathways with significant expression deregulation; including a global up-regulation of the “mitosis”, “DNA replication” and “oxidative phosphorylation” pathways, which are explained by the high cell proliferation rate in bladder tumors (Dyrskjot et al., 2004) and the elevated metabolic activity required to sustain it (Arora and Pedersen, 1988). On the other hand, our results indicate a general down-regulation of the “lymphocyte activation”, “immune response” and “cell adhesion” pathways. This implies a suppressed immune response in the tumors—which can reportedly be overcome by IL2
treatment (Velotti et al., 1991)—in addition to a higher probability of metastasis due to deregulation in cell adhesion components (Cooper and Pienta, 2000).

**De novo discovery of DNA and RNA cis-regulatory elements**

Upon discovering the pathways with significant perturbation (both up-regulated and down-regulated) in the tumor state compared to normal, we applied FIRE to identify the *cis*-regulatory elements that are informative about the observed gene expression changes (Elemento et al., 2007). In the final step of our analysis, we then evaluated whether the independently discovered pathways and regulatory sequences are mutually informative. This analysis enables us to associate regulators with their target genes, and to reconstruct the local regulatory networks responsible for cancer-related deregulation. Here (as in Elemento et al., 2007), we defined target genes as the set of genes whose promoters (or 3’UTRs for RNA elements) contain the corresponding sequence motif and belong to an expression cluster/bin where the same motif is significantly over-represented.

For the Dyrskjot et al. (2004) urinary bladder cancer dataset, we identified 16 informative upstream sequence motifs (including known binding sites for E2F, Elk-1, AhR, SEF-1 and E47) and a single 3’-UTR element (see Figure 2A). Approximately two thirds of these motifs are associated with genes that are up-regulated in the tumor state; whereas, the remaining third are enriched in down-regulated genes. Our analysis suggests that the Elk-1 transcription factor, a member of ETS family of ternary complex factors (TCF) and a target of the MAP kinase pathway, plays a central role in bladder cancer. The UNGNUGU element, a novel 3’ UTR motif, essentially shows a similar pattern of occurrence, suggesting a potential collaboration with Elk-1 (Figure 2A). This motif does not match any of the known miRNA target sites; it may be targeted by a yet to be discovered miRNA or by an RNA-binding regulatory protein (e.g. Wang et al., 2002).
Upon the discovery of known and putative regulatory elements, we systematically evaluated the mutual information between these elements and PAGE-discovered deregulated pathways. In a heat-map built from the resulting information values (Figure 2B; we call this representation pathway-regulatory map), we observed that Elk-1 binding sites are positively associated with several up-regulated pathways, namely “mitosis”, “DNA replication” and “protein degradation”, and negatively associated with several down-regulated ones (e.g. “lymphocyte activation” and “cell adhesion”). It has been previously shown that Elk-1 promotes cell cycle progression through transcriptional activation of \textit{c-fos} (Millar et al., 1982). The binding site for E2F also showed a significant association with “DNA replication” and “mitosis” (Figure 2B). Indeed, E2F is a known regulator of DNA replication and mitotic events (Ishida et al., 2001). We also predicted a potential association between AhR transcription factor and “ubiquitin-dependent protein degradation”. Prior evidence for this interaction exists in the literature: in breast cancer cells, it has been shown that the aryl hydrocarbon receptor (AhR) down-regulates estrogen receptor $\alpha$ through activation of the proteasome complex (Wormke et al., 2000). Furthermore, genes involved in “lymphocyte activation” and “humoral immune response” that showed a large decrease in average expression across the tumor samples are positively associated with the binding sites for the SEF-1 and E47 transcription factors. SEF-1 (SL3-3 enhancer factor 1) is a known activator of the T cell antigen receptor associated CD3 polypeptide and plays an important role in T cell specific transcription (Hallberg et al., 1992). Similarly, E47—a basic helix-turn-helix protein arising from alternative splicing of E2A’s primary transcript—is considered essential for defining leukocyte lineages (Shen and Kadesch, 1995).

Our analysis of bladder cancer microarray expression data recapitulates many previously known signaling pathway perturbations. In case of E2F and Elk-1 (whose binding sites we identified above), we speculate that mutations in their upstream signaling proteins (e.g. Rb and Erk2 respectively) result in higher activities of these transcription factors, which in turn translate into a faster cell cycle. On the
other hand, it has been noted that a suppressed immune response, in this case possibly originating from decreased E47 and SEF-1 activities, aids tumor survival (Zhang and Kaufman, 2006). Strikingly, half the regulatory elements we identified do not correspond to known transcription factor binding or miRNA targeting sites, but nonetheless are highly informative of regulatory perturbations in this dataset. The pathway-regulatory map (Fig. 2C) is a powerful starting point for exploring the biological role of these elements and their connections to known and novel signaling pathways.

**Comparative analysis of cancer sub-types (BL vs. DLBCL)**

Burkitt’s Lymphoma (BL) and Diffuse Large B-cell Lymphoma (DLBCL) are phenotypically similar with blurred distinctions; however, their treatment regimens differ drastically (Frost et al., 2004). In what follows, we employ our framework in order to capture a comprehensive comparative understanding of pathway perturbations between BL and DLBCL, and of the associated cis-regulatory elements.

We applied PAGE and FIRE to a microarray analysis of 36 BL and 166 DLBCL samples (Hummel et al., 2006). Based on their expression values across all the samples, we grouped the genes into 110 co-expression clusters (using the k-means clustering algorithm) with each gene uniquely assigned to an index representing a distinct cluster. This clustering process captures the intra-cancer gene expression heterogeneity, which is usually veiled when averaging expression values across multiple samples of the same tumor type. Using PAGE, we discovered a total of 51 significantly informative and non-redundant pathways. The representative pathways that are associated with the clusters showing differential average expression between BL and DLBCL samples are shown in Figure 3A (see Figure S8A for the complete list). We noted a marked increase in the expression of cell cycle-related genes (e.g. “mitotic cell cycle” and “DNA replication”) and a decrease in expression of immune response genes (e.g. “lymphocyte activation”) in BL samples compared to DLBCL.
Our results show that several of the highly informative cell cycle pathways are over-represented in co-expression clusters 17 and 6 (a subset of these are shown in Figure 3A). As indicated by their average expression, genes in these clusters tend to be expressed at a higher level in BL samples (Figure 3A). Along with cell cycle related genes, protein metabolism pathways such as “protein folding” and “protein catabolic process” are also identified as highly informative (Figure S8A). These are most highly associated with cluster 70 with a higher expression in BL samples (Figure S8A). Moreover, a number of pathways related to immune response, e.g. “cytokine receptor activity”, “lymphocyte activation” and “antigen processing”, are also significantly deregulated (Figure 3A and Figure S8A). In most cases, these pathways are associated with the clusters showing lower expression in BL compared to DLBCL (e.g. cluster 8 for “antigen processing”, cluster 41 and 58 for “lymphocyte activation”, cluster 35 and 72 for “inflammatory responses”, etc.). The higher expression of lymphocyte-specific pathways in DLBCL has been previously shown by employing immunohistochemical analysis, and revealed the overabundance of B cell activated markers (Gormley et al., 2005).

Application of FIRE to the same dataset revealed a total of 127 cis-regulatory elements (98 5’ upstream motifs and 29 3’ UTR elements) including many known transcription factor binding sites: E2F, ELK4, NF-Y, NF-AT, MYB and a microRNA target site for let-7 (Figure 3B and Figure S8B). The let-7 miRNA, whose target genes show significant up-regulation in BL samples, is a known regulator of cell proliferation, and let-7 mutations have been observed in human lung cancers (Johnson et al., 2007). As presented in the pathway-regulatory map, genes with a FIRE-identified NF-Y binding site are significantly associated with “mitotic cell cycle” (Figure 3C). NF-Y is a trimeric transcription factor with H2A/H2B-like subunits and is a known activator of G1-S cyclins with an established role in tumorigenesis through cyclin B2 over-expression (Jordan et al., 1999). Interestingly, we discovered a significant proximal co-occurrence of NF-Y and Sp1 binding sites in cluster 17 (Figure 4A). We
compared these co-occurrences with genes randomly selected from clusters 75 and 47. Both clusters 75 (enriched only in Sp1 motif) and 47 (enriched only in NF-Y motif) show negligible differential expression between BL and DLBCL samples ($t$-test $p$-value of 0.5 and 0.3 respectively); however, cluster 17 shows a highly significant average up-regulation in BL samples ($p$-value < $10^{-10}$), suggesting a functional interaction between NF-Y and Sp1 (Figure 4B). One of the shared targets of these two transcription factors with known over-expression in BL is A-myb (Facchinetti et al., 2000) whose binding site (TAACNG reported here as v-Myb) is also captured in our motif discovery step (Figure 3B). The average expression of the genes in cluster 17, across all the samples in the dataset, is shown in Figure 4B. The observed correlation between NF-Y mRNA expression and the genes in cluster 17 (R=0.73; $t$-test $p$-value = 9.65e-35) further supports the direct role of NF-Y in the regulation of the genes in this cluster.

In addition to NF-Y and Sp1 targets, the pathway-regulatory map reveals many known and novel associations (Figure 3C). For example, the prominent role of AP-1 proteins in different lymphomas (Vasanwala et al., 2002) is reflected in our results where the AP-1 motif (TGANTCA) shows a significant association with two pathways: “lymphocyte activation” and “cytokine receptor activity”. AP-1 is an important factor for leukocyte activation and differentiation in the immune system (Foletta et al., 1998). Besides AP-1, the pathway-regulatory map also highlights the known role of NF-AT in lymphocyte activation (Fisher et al., 2006).

Moreover, the pathway-regulatory map shows that E2F target genes are associated with mitotic cell cycle and RNA polymerase activity (Figure 3C). While the positive role of E2F in the regulation of cell proliferation is known (Ishida et al., 2001), E2F has not before been associated with Burkitt’s lymphoma. Alongside these transcription factors with a positive effect on cellular proliferation, we identified other regulators with potential key roles in defining the biological differences between BL
and DLBCL. For example, our results indicate that the binding site for the human X-box binding protein-1 (XBP1), a transcription factor that participates in the unfolded protein response (Calfon et al., 2002), is associated with “unfolded protein binding”. The latter pathway shows a significant up-regulation in BL samples (Figure 3B and C). Although this association has not been made before in the context of BL, it is known that sustaining the activation of the unfolded protein response (UPR) is important for tumor cells due to its cytoprotective action against the cytotoxic conditions (e.g. hypoxia and nutrient deprivation) that typically accompany the tumor state (Ma and Hendershot, 2004).

**Global analysis of pathway perturbations across cancers**

Our success in revealing regulatory perturbations in individual cancer sub-types motivated us to conduct a more comprehensive meta-analysis of perturbations across diverse human cancers. Our goal was to identify both generic and cancer-type specific deregulations and to reveal the \textit{cis}-regulatory sequences underlying these changes. To this end, we compiled data from 46 microarray studies of cancer versus normal tissues (see Table S2). In order to capture intra-cancer variation within samples, we employed the same pre-processing step as in the BL versus DLBCL analysis above; we first clustered the genes based on their expression across normal and tumor samples and then combined the clusters with low average differences into a single ‘background’ cluster (see Methods for details). We then used PAGE to find the pathways that best explain the resulting co-expression clusters. We combined the results obtained from all cancer datasets into a cancer pathway heatmap. This map reflects whether the informative pathways tend to be up or down-regulated in each cancer type (Figure 5; also see Figure S11 for the complete set of 269 pathways).

Some of these pathways are deregulated in many cancers, a subset of which includes well-known core cancer pathways while the others, to the best of our knowledge, have not been previously associated with the tumor state. As expected, our results show that pathways responsible for growth and
proliferation are consistently up-regulated in tumor samples as compared to normal controls. This includes mitotic cell cycle, DNA replication and chromatin assembly genes (Figure 5). Consistent with a short doubling time, metabolic pathways like those of “glycolysis” and “organic compounds oxidation” are up-regulated in many tumors (Arora and Pedersen, 1988); whereas, stress responses that lead to cell cycle arrest such as “negative regulation of progression through cell cycle” are often down-regulated. Among the signal transduction pathways, the expression of “NF-κB pathway” components is significantly increased in many cancers, implying the broad oncogenic role of this signaling pathway. NF-κB, a member of Rel family, regulates the immune and inflammatory response genes and functions as a negative regulator of apoptosis (Baeuerle and Baltimore, 1996). It is therefore a known promoter of cell growth, and increased levels of NF-κB have been reported in many solid and hematopoietic primary tumors and tumor cell lines (for review see Rayet and Gelinas, 1999).

Our results also suggest an important role for ion transport pathways in oncogenesis and/or tumor maintenance. For example, sodium and potassium transport activities are deregulated in many types of cancer (Figure 5). Active avoidance of sodium transport has been previously reported for some specific tumors (e.g. Morgan et al., 1986). It has also been shown that Na\(^+\)/K\(^+\) pump blockade decreases doxorubicin-mediated cytotoxicity and DNA strand breakage in human tumor cells (Lawrence and Davis, 1990). We also noted a general increase in the expression of the genes encoding anion transporters (especially phosphate transporters) in most of the tumor cells compared to their corresponding normal samples. The cytoplasmic Pi concentration has been suggested to play a critical role in metabolic control in animal cells; a measurable decline in cytoplasmic Pi is accompanied by a decrease in glycolytic or respiratory rates (Geck and Bereiter-Hahn, 1991). The degree to which limited Pi uptake restricts glycolysis, respiration, or cell growth in normal or malignant tissues has been studied extensively (e.g. Wehrle and Pedersen, 1982).
Alongside pathways showing deregulation across many cancers, we have also identified informative pathways that are only associated with a single or a small number of tumor classes. For example, “TNF receptor binding” pathway is most prominent in serous ovarian cancer. The molecular mechanisms of tumor survival in this cancer are not well-understood; however, a recent study has reported that the over-expression of tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) is correlated with prolonged survival in advanced ovarian cancers (Lancaster et al., 2003). VEGF receptor activity is another pathway that we find to be up-regulated primarily in renal cancers in our study. Interestingly, inhibitors of VEGF receptor have been widely considered as potential treatment for this type of cancer (Duncan et al., 2008).

The systematic PAGE analysis of cancer datasets described above allowed us to compare the corresponding cancers based on their pathway-level perturbations. In order to do so, we used hierarchical clustering to cluster the cancers based on their informative pathways (listed in Figure 5). The clustering results and the correlation matrix are shown in Figure S12. The hierarchical clustering tree in Figure S12 shows that independent cancer datasets of similar types are also similar in terms of their informative (and therefore deregulated) pathways. For example, in Figure S12, all non-small cell lung datasets are grouped together in one cluster (Beer et al., 2002; Bhattacharjee et al., 2001; Stearman et al., 2005) distinguishing them from the single small cell carcinoma (SMCL) dataset (Bhattacharjee et al., 2001). From a broader point of view, most of the cancers are clustered into three classes with the non-small cell lung carcinomas in one, breast, brain and ovarian cancers in another, and the rest in the third cluster. Two samples in particular show a highly negative correlation with other cancers: a melanoma dataset (Hoek et al., 2006) and a chronic lymphocytic leukemia (CML) dataset (Alizadeh et al., 2000). This anti-correlation largely results from a lower expression of cell cycle related genes in both of these tumors, compared to all other cancers (Figure 5). As Alizadeh et al. (2000) noted, CML is a slowly progressing disease with a low proliferation rate. Similarly, despite
being highly metastatic, a cohort of melanoma samples in this dataset has been shown to have a low proliferation rate (Hoek et al., 2006) thus explaining the lower expression of mitotic genes in this cancer.

While previous studies have also shown that pathway-level comparisons are more consistent and reproducible than gene-based comparisons (Subramanian et al., 2005), these studies have not revealed the mechanisms by which the observed perturbations may come about. In order to map the regulatory network onto these deregulated pathways, we systematically searched for informative cis-regulatory elements in each cancer gene expression dataset. Combining the resulting motifs into a non-redundant list, we ultimately generated a “cancer regulatory map” in which the up-regulation or down-regulation of the genes downstream (or upstream for 3’ UTR elements) of each motif is captured across all cancers (see Figure S13).

Subsequent systematic assessment of the relationship between the deregulated pathways and informative motifs suggests potential key roles for previously known regulatory elements as well as for many novel motifs (see cancer pathway-regulatory map in Figure S14). In Figure 6, we show a subset of motifs versus pathway associations from Figure S14. Our approach successfully assigns p53 and E2F to “induction of apoptosis” (Stiewe, 2007), Jun, Elk-1 and E2F to “mitotic cell cycle” (Gurzov et al., 2008; Ishida et al., 2001; Millar et al., 1982) and HSF to “protein folding” (Mosser et al., 1993). We hypothesize that the observed deregulations at the transcription level root from perturbations in the upstream signaling pathways leading to the activation or inactivation of key regulators. For example, E2F controls p53 via the activation of p19ARF, and in the tumor state E2F is typically uncoupled from apoptosis through mutations in ARF (Tsai et al., 2002). Similarly, the IFN-stimulated response element (ISRE) is associated with “antigen processing and presentation”. Displaying a broad range of
functions, interferon β increases gene expression at the transcriptional level through binding of factors to the ISRE upstream of interferon-inducible genes, such as HLA class I (Lefebvre et al., 2001).

In another case, genes harboring the MEF-2 motif show significant changes in expression level (Figure S13). These perturbations are, by and large, comparable across similar cancers, e.g. MEF-2 target genes tend to be up-regulated in most lung cancer samples (Figure S13). MEF-2, which in our study is associated with “cell-cell adhesion” (Figure 6), is a known regulator of αT-catenin promoter (Vanpoucke et al., 2004). The loss of α-catenin expression, a cadherin-associated protein, results in the disruption of cell-cell adhesion and is associated with an increase in tumor malignancy (Shimoyama et al., 1992).

Our cancer pathway-regulatory map in Figure S14 reveals many highly significant associations between known transcription factors and pathways whose roles in cancer have not been previously established. For example, in our results, the presence of a binding site for Lmo2 complex is associated with “cell migration” (Figure 6). Lmo2 complex is known to be activated through translocations in acute leukemia (Boehm et al., 1991).

In addition to promoter motifs, there are also a large number of predicted 3’ UTR elements associated with key pathways. Recent studies have highlighted the role of miRNAs in tumorigenesis and metastasis (Kumar et al., 2007; Tavazoie et al., 2008). Among them is miR-203 which was found to be down-regulated in metastatic cells from breast-cancer tumors (see Figure 1a in Tavazoie et al., 2008). Interestingly, our approach associates this miRNA with “ubiquitin-dependent protein catabolic process”. We have also assigned three known miRNA target sites, miR-377, miR487a and miR-582, to “cofactor biosynthesis”, “angiogenesis” and “protein degradation”, respectively (Figure 6).
We also discovered many novel motifs that are significantly associated with informative pathways (Figure S14; also see Table S3 for a list of predictions with their associated $p$-values). For example, AA[CT]N[AC]CG is a putative upstream binding element which our analysis associates with genes functioning in “chromatin assembly” (Figure 6). The genes with the NTACGN[AC]N binding sites, on the other hand, tend to function in “DNA repair”, “mRNA processing” and “protein folding” (Figure 6). Besides these upstream elements, we also report many novel associations with predicted RNA motifs. For example, GN[CU]U[GU]UA is associated with “DNA repair”, GGC[CU]CU[AU] with “chromatin assembly and AANGGCNCU with “PI3-K signaling” (Figure 6). Our discovery of a large number of RNA motifs suggests a key role for yet unknown miRNAs or RNA binding regulatory proteins in cancer.

**Discussion and Conclusion**

The identification of regulatory pathways whose perturbations are causal to the initiation and maintenance of the tumor state is one of the major challenges in cancer biology. In this study, we have introduced a novel approach, based on information theory, for simultaneous extraction of perturbed cellular pathways and their underlying regulatory programs from cancer versus normal gene expression datasets. Our results clearly show a general over-expression of mitotic pathways and down-regulation of immune response pathways in tumors compared to normal tissues; however, we did not detect any other “universal” tumor pathway signature. The diversity of the perturbed pathways, and their association with specific cancers, as represented in the cancer pathway map, highlights the broad heterogeneity underlying the tumor cellular state.

Despite this heterogeneity, pathway-level analysis of cancer gene expression can be employed for classification purposes in the sense that the tumors with similar phenotypes (in terms of which pathways are deregulated and to what extent) can be identified. Indeed, gene expression profiling has
proven helpful in cancer-type classification and prognosis prediction (Beer et al., 2002; Bhattacharjee et al., 2001), yet results from independent cancer studies often disagree (partly due to the high level of heterogeneity between tumor samples that are grouped together in the same class). However, despite the poor overlap between the signature genes identified in each study, our results show that the perturbed underlying pathways are similar. For example in two independent lung adenocarcinoma gene expression datasets (Beer et al., 2002; Bhattacharjee et al., 2001), we found 171 and 168 deregulated pathways with 129 pathways in common. Thus, while the signature genes from these two studies poorly overlap (12 out of 100 top genes with $p$-value of 0.01), a pathway-level comparison using PAGE shows highly similar perturbations at the pathway level ($p$-value $< 10^{-250}$). These results agree with previous studies that used alternative approaches for identifying perturbed pathways (Efroni et al., 2007; Subramanian et al., 2005).

In addition to uncovering deregulated pathways, we have employed a systematic cis-regulatory element discovery strategy in order to identify the regulators (transcription factors, miRNAs or RNA-binding proteins) and more generally the regulatory modules through which the perturbations in the cellular pathways come about. The signaling pathways that control the activity of these regulators are often known, and our approach uncovers a significant fraction of them. Our approach predicts the involvement of certain known transcriptional or post-transcriptional regulators in cancer-associated pathways, thus revealing putative oncogenes and tumor suppressors, and yielding valuable drug target candidates.

On average, however, only 10% (~32/292) of the discovered motifs correspond to previously known binding sites. This highlights both the complexity, and our relatively primitive understanding of the tumor state. For example, we have discovered 11 novel motifs with significant positive associations with DNA repair ($p$-value $< 10^{-3}$). Besides, many regulatory elements are highly informative about
groups of coordinately regulated genes in cancer versus normal tissues but are not associated with any known pathways. We hypothesize that these putative regulatory elements predict novel cancer-associated pathways. Identifying the corresponding signaling pathways will be an important direction for further research. Similarly, only a minority of the 3’UTR elements we discovered (~10%) match known miRNA targeting sites. These findings point to a largely unexplored role for post-transcriptional regulation (involving both miRNAs and RNA-binding proteins) in cancer.

We have introduced a principled framework for revealing regulatory perturbations in cancer. The implementation of this framework within powerful and user-friendly tools should allow rapid and comprehensive analysis of cancer expression data by experts and non-experts alike. As a final note, we stress that although our analyses here have been focused on gene expression perturbations in cancer, our framework is general in concept and can be utilized to study regulatory perturbations across other human diseases.

**Methods**

**Pre-processing of input datasets**

All cancer microarray datasets used in this study were downloaded from GEO (http://www.ncbi.nlm.nih.gov/projects/geo/). Each cancer versus normal dataset was converted into continuous or discrete gene expression profiles, as follows.

In the continuous case (e.g. urinary bladder cancer), each gene is associated with a continuous expression value based on the following equation:

\[ v = s(1 - p), \]
where \( p \) is a \( p \)-value calculated by performing a student’s \( t \)-test between the cancer samples and the normal controls. \( s \) is the sign of the difference between the average values in these two sets. Thus, \( v \) indicates the extent to which a gene is up-regulated or down-regulated in the cancer state.

In the discrete case, genes were first clustered into \( \sqrt{N} \) groups (\( N \) is the total number of genes), using the \( k \)-means unsupervised clustering approach, and based on their expression values in the normal and tumor samples. Then the clusters whose average expressions did not differ between the normal and cancer samples (nominal \( p \)-value from \( t \)-test > 0.05, where the \( t \)-test is performed on the expression profiles in each cluster) were combined into one background cluster. Subsequently, each gene was associated with the cluster index of the cluster to which it belongs.

**PAGE: Pathway Analysis of Gene Expression**

PAGE can process both continuous and discrete expression data, since the concept of mutual information (MI) is applicable to any type of random variable (Cover and Thomas, 2006). However, in practice, continuous expression profiles must be quantized in order to compute mutual information values. In PAGE, we quantize continuous expression data into equally populated bins (as described in Elemento et al., 2007).

PAGE takes as input an expression profile with \( N_e \) possible values (\( N_e \) equals the number of co-expression clusters or expression bins). In this profile, each gene is associated with a unique expression value. For each pathway, PAGE then creates a binary pathway profile, specifying whether each gene belongs to the pathway or not (see Supplemental Procedures.). The mutual information between the pathway profile and the expression profile is calculated using the following formula:
We use a randomization-based statistical test to determine the significance of each $MI$ value. Briefly, we shuffle the expression profile a large number of times (while keeping the pathway profile the same), calculate corresponding $MI$ values with the pathway profile, and calculate the fraction of shuffled $MI$ values that are greater than the real one (see Supplemental Procedures for details). In the default PAGE settings, we require this fraction to be less than 0.005.

To avoid testing all pathways, we apply the randomization-based test described above to pathways sorted by decreasing $MI$ values (i.e., starting with the most informative pathway). We stop testing when a specified number of consecutive pathways (set to 20 by default) do not pass the test. We then output all pathways with statistically significant $MI$ values.

Prior to testing a pathway, we evaluate whether a statistically significant pathway that is very similar to the one that is being evaluated has already been found. In order to do so, we require that the candidate pathway be highly informative about the expression, but also bring a significant amount of new information compared to all previously accepted pathways, while at the same time being minimally informative about these pathways. This is implemented in PAGE using a criterion based on the conditional mutual information (Cover and Thomas, 2006; Elemento et al., 2007), in which we require that:

$$\frac{I(\text{candidate pathway}; \text{expression} | \text{accepted pathways})}{I(\text{candidate pathway}; \text{accepted pathway})} > r$$

where $I(\text{candidate pathway}; \text{expression} | \text{accepted pathways})$ is the information that the pathway carries about the expression, given the accepted pathway profile. In the default PAGE settings (similar to FIRE), $r$ is set to 5.
We then determine the level with which the significantly informative pathways are over-represented or under-represented in each expression bin (or cluster). To do so, we use the hypergeometric distribution to calculate two distinct $p$-values for over and under-representations: $p_{\text{over}} = P(X \geq x)$ and $p_{\text{under}} = P(X \leq x)$ (where $x$ equals the number of genes in the given expression bin). If $p_{\text{over}} < p_{\text{under}}$ we consider the pathway to be over-represented in the expression bin/cluster; otherwise, it is under-represented. We use these $p$-values to draw a heatmap, in which rows represent significant pathways and columns correspond to expression bin/clusters (see. Supplemental Procedures for more details). In this heatmap, red entries correspond to pathway over-representations, while blue entries correspond to under-representations.

**FIRE: De novo discovery of informative regulatory elements**

FIRE was used with default settings, as described in Elemento et al, 2007.

**Pathway-Regulatory Maps: Associating regulatory elements with their target pathways**

In order to associate the FIRE cis-regulatory elements with the PAGE pathways they may control, we calculate pairwise MI values between all (pathway,motif) pairs. As described above, we first create pathway profiles which indicate whether each gene belongs to the evaluated pathway or not. Motif profiles are created in the same way, indicating whether each gene contains at least one copy of the evaluated motif in its promoter region (or 3’UTR for RNA motifs). The (pathway,motif) pairs that pass the randomization-based statistical test for significant $MI$ values are accepted. For these pairs, the sign and significance of the associations are determined using the hypergeometric distribution, as described above for PAGE. The hypergeometric $p$-values are then used to draw heatmaps, where columns correspond to cis-regulatory elements and rows correspond to pathways. In these heatmaps (which we call pathway-regulatory maps), red entries correspond to positive associations; such associations
suggest that the regulator that binds to the corresponding regulatory element regulates some aspects of the pathway. Blue entries indicate negative associations, the extent of which highlights the importance of the absence of a regulator for appropriate transcriptional control.
References


Figure Legends

Figure 1. Perturbed pathways in bladder cancer. This figure shows the informative pathways discovered by PAGE and their enrichment patterns across all expression bins. In this heatmap, rows correspond to pathways, and columns to consecutive expression bins. Red entries indicate over-representation of pathway genes in a given expression bin. Over- and under-representation are measured using hypergeometric \( p \)-values (log-transformed) as described in Methods. The top panel shows the range of relative expression values in each bin. Bins to the left contain genes with lower expression in cancer samples whereas the ones to the right contain genes with higher expression.

Figure 2. De novo discovery of cis-regulatory elements that are informative of pathway perturbations in bladder cancer (A) The enrichment patterns of the putative cis-regulatory elements discovered by FIRE, across the expression bins. In this heatmap, rows correspond to the discovered motifs and columns to expression bins. Yellow entries in the heatmap indicate motif over-representation (measured by negative log-transformed hypergeometric \( p \)-values) in a particular bin, while blue entries indicate under-representation (measured by log-transformed \( p \)-values). Some of the properties of each sequence motif are also shown, e.g. \( MI \) values, position biases, orientation biases, etc. (B) The resulting pathway-regulatory map showing the putative associations between regulatory elements and pathways. In this representation, rows correspond to informative PAGE pathways and columns to informative FIRE motifs. Red entries in this heatmap correspond to a positive association where the genes belonging to a pathway are also enriched in a given motif (measured using hypergeometric \( p \)-values). Blue entries correspond to significant motif depletions in the upstream sequences (or 3’UTRs) of genes in a given pathway.

Figure 3. Differential pathway perturbations between Burkitt’s lymphoma and diffuse large B-cell lymphoma. (A) Deregulated pathways as determined by PAGE and their enrichment patterns across BL/DLBCL co-expression clusters. In this representation, columns represent co-expression clusters while rows correspond to informative pathways. The red-blue entries in the heatmap show the enrichment/depletion of the members of each pathway across the co-expression clusters, measured by hypergeometric \( p \)-values (log-
transformed). The top panel shows the normalized average expression of each gene cluster in BL and DLBCL samples. (B) A subset of putative *cis*-regulatory elements discovered by FIRE in BL vs. DLBCL co-expression clusters. Similar to Figure 2A, rows correspond to motifs and columns to co-expression clusters. Yellow entries indicate motif over-representation (measured by negative log-transformed hypergeometric *p*-values) in a particular cluster, while blue entries indicate under-representation (measured by log-transformed *p*-values). (C) The pathway-regulatory map reveals the potential roles of the identified regulatory elements in bringing about the observed pathway perturbations. Red entries in this heatmap correspond to a positive association where the genes belonging to a pathway are also enriched in a given motif. Blue entries correspond to significant motif depletions in the upstream sequences (or 3’UTRs) of genes in a given pathway.

**Figure 4. Cis-regulatory element interactions and combinatorial regulation** (A) The FIRE regulatory interaction matrix for the *cis*-regulatory elements discovered in the BL vs. DLBCL dataset (Figure 3B), and an accompanying motif map showing co-localization of Sp1 and NF-Y sites. In the FIRE interaction matrix, lighter colors (white and yellow) correspond to significant motif co-occurrences. Darker colors correspond to joint-motif avoidance. ‘+’ signs indicate that two motifs tend to co-localize on the DNA or RNA sequences. The NF-Y and Sp1 binding sites show a significant proximal co-occurrence and co-localization in the promoters of their target genes. This co-localization is illustrated by a FIRE motif map, which shows where these two binding sites occur in the promoter sequences of genes in cluster 17, in comparison with genes randomly selected from clusters 75 and 47. (B) The average expression profile of genes in co-expression cluster 17, across all BL and DLBCL samples, shows a high correlation with NF-Y mRNA expression. The average expression profiles of the genes in clusters 75 and 47, although enriched in Sp1 and NF-Y putative sites respectively, are not correlated with BL vs. DLBCL classification.

**Figure 5. Cancer pathway map.** This figure shows the 58 non-redundant PAGE-discovered pathways with significant patterns of deregulations across 46 cancer vs. normal samples. Each entry in this heatmap represents the most significant over-representation of a given pathway across all non-background co-expression clusters for a given cancer. Over-representation is measured using log-transformed hypergeometric *p*-values. The colors
indicate whether the genes in a given pathway are up-regulated (red) or down-regulated (green) in the tumor samples. The pathways discussed in the text are highlighted in yellow.

**Figure 6. Cancer pathway-regulatory map.** This figure shows a subset of the *cis*-regulatory motif-pathway associations from the cancer pathway-regulatory map in Figure S14. As in Figure 2B, red entries represent positive associations between pathways and regulatory elements, i.e. cases where the genes belonging to a pathway are also enriched in a given motif. Blue entries represent motif depletion in the upstream sequences (or 3’UTRs) of genes in a given pathway. Associations are measured using log-transformed hypergeometric *p*-values.
Figures
ribosome biogenesis and assembly, GO:0042254  
ER to Golgi vesicle-mediated transport, GO:0006888  
tRNA aminoacylation, GO:0043039  
DNA replication, GO:0006260  
RNA splicing, GO:0008380  
mitotic cell cycle, GO:0000278  
ubiquitin-dependent protein catabolic process, GO:0006511  
microtubule biogenesis, GO:0000226  
oxidative phosphorylation, GO:0006119  
positive regulation of apoptosis, GO:0043065  
cAMP-mediated signaling, GO:0019933  
humoral immune response, GO:0006959  
potassium ion transport, GO:0006813  
homophilic cell adhesion, GO:0007156  
sodium ion transport, GO:0006814  
lymphocyte activation, GO:0046649

Figure 1
(A) Down-regulation

<table>
<thead>
<tr>
<th>Position</th>
<th>Orientation</th>
<th>Seed</th>
<th>Motif Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'</td>
<td>0.005</td>
<td>11.7</td>
<td>GCCAAA</td>
</tr>
<tr>
<td>5'</td>
<td>0.026</td>
<td>78.1</td>
<td>CCGGAAG</td>
</tr>
<tr>
<td>5'</td>
<td>0.010</td>
<td>26.1</td>
<td>CTCGCGA</td>
</tr>
<tr>
<td>3' UTR</td>
<td>0.004</td>
<td>8.4</td>
<td>UAGAUGU</td>
</tr>
<tr>
<td>5'</td>
<td>0.008</td>
<td>21.4</td>
<td>CGTCGCA</td>
</tr>
<tr>
<td>5'</td>
<td>0.007</td>
<td>18.5</td>
<td>ACGTGGG</td>
</tr>
<tr>
<td>5'</td>
<td>0.005</td>
<td>11.9</td>
<td>AGAACGC</td>
</tr>
<tr>
<td>5'</td>
<td>0.008</td>
<td>22.4</td>
<td>CGCAGTC</td>
</tr>
<tr>
<td>5'</td>
<td>0.009</td>
<td>21.4</td>
<td>CGTCACG</td>
</tr>
<tr>
<td>5'</td>
<td>0.004</td>
<td>8.8</td>
<td>CGCGTTA</td>
</tr>
<tr>
<td>5'</td>
<td>0.005</td>
<td>14.3</td>
<td>CATCTCA</td>
</tr>
<tr>
<td>5'</td>
<td>0.005</td>
<td>13.7</td>
<td>ATCTCCA</td>
</tr>
<tr>
<td>5'</td>
<td>0.005</td>
<td>10.6</td>
<td>TCCCCCA</td>
</tr>
<tr>
<td>5'</td>
<td>0.005</td>
<td>13.1</td>
<td>ATGAGAG</td>
</tr>
<tr>
<td>5'</td>
<td>0.004</td>
<td>10.5</td>
<td>ATTTGCA</td>
</tr>
<tr>
<td>5'</td>
<td>0.004</td>
<td>8.8</td>
<td>ATCTGTG</td>
</tr>
</tbody>
</table>

(B) 

- potassium ion transport GO:0006813
- sodium ion transport GO:0006814
- lymphocyte activation GO:0046649
- humoral immune response GO:0006959
- homophilic cell adhesion GO:0007156
- positive regulation of apoptosis GO:0043065
- oxidative phosphorylation GO:0006119
- microtubule biogenesis GO:000226
- mitotic cell cycle GO:0000278
- DNA replication GO:0006260
- ribosome biogenesis and assembly GO:0042254
- Ub-dependent protein catabolic process GO:0006511
- RNA splicing GO:0008380
- tRNA aminoacylation GO:0043039
- ER to Golgi vesicle-mediated transport GO:0006888

Figure 2
Figure 3
Figure 6