Revealing global regulatory perturbations across human cancers

Hani Goodarzi[#], Olivier Elemento[#$], and Saeed Tavazoie[*]

Inventory

Figure S1. This figure relates to Figure 1 in the main text and shows the schematics of our approach.

Figure S2. This figure relates to Figure 2 in the main text and contains supplemental results discussed in the main text regarding the associations discovered in Figure 2.

Figure S3. This figure relates to Figure 4 in the main text and shows the expression level of Sp1 and NF-Y and their target genes in more detail as opposed to the average values presented if Figure 4B.

Table S1. This table relates to Figure 5 in the main text and lists the tumor vs. normal gene expression datasets compiled for this study.

Figure S4. This figure relates to Figure 6 in the main text and contains three parts. Figure S4A shows the cancer regulatory map discussed in the main text. The conservation scores of these elements are shown in Figure S4B. Figure S4C is the complete association map from which Figure 6 was extracted as an exemplary subset.

Table S2. This table relates to Figure 6 in the main text and lists a number of significant associations between putative regulatory elements and pathways as reported in Figure 6.

Figure S5. This figure relates to Figure 7 in the main text and shows additional analyses performed on the gene expression profiles obtained from the TF knock-down and decoy vs. scrambles experiments.

Supplementary Materials

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.

Supplemental Figures

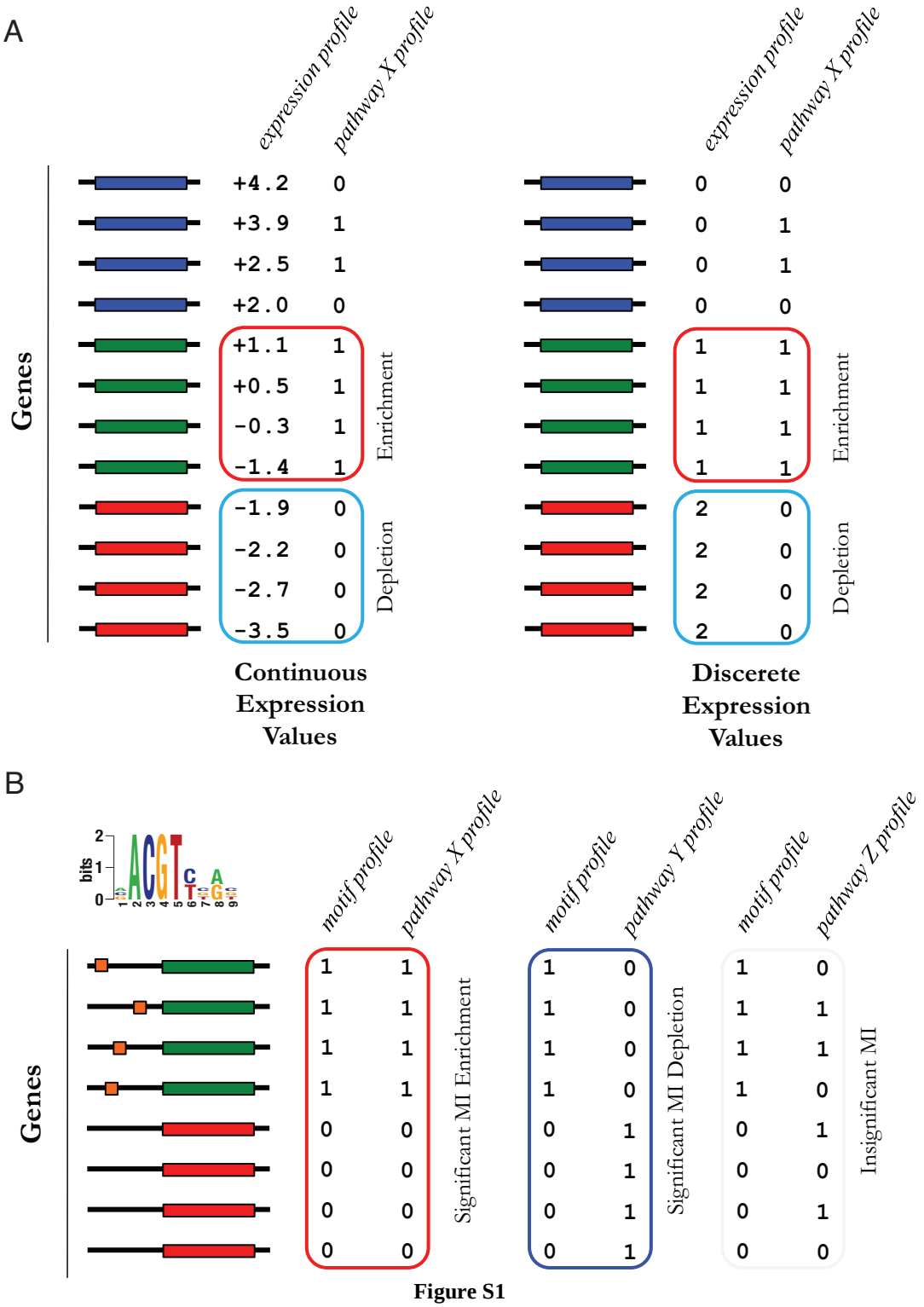


Figure S1

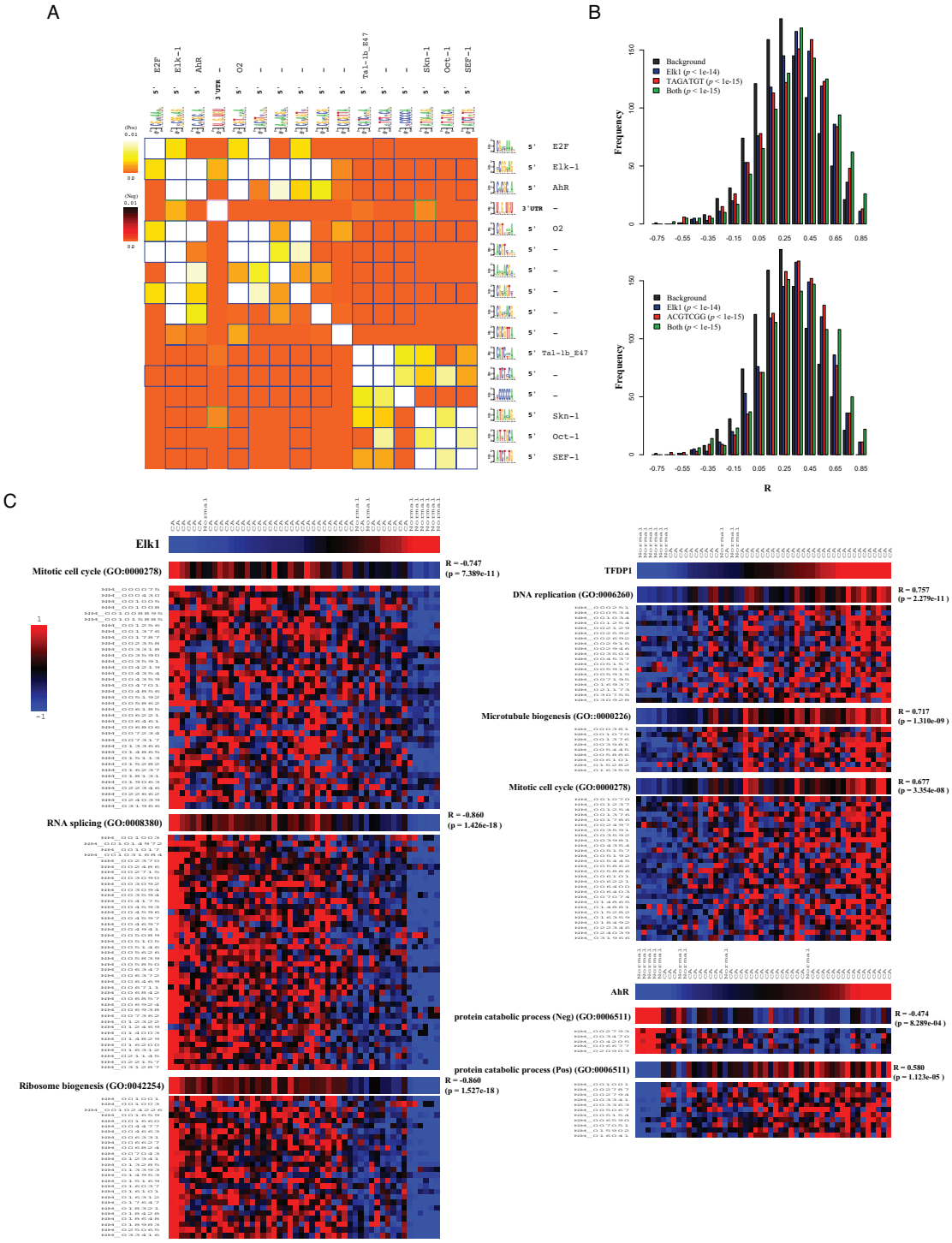

A
B
C
Background
Elk1 (p < 1e-14)
TAGATGT (p < 1e-15)
Both (p < 1e-15)
ACGTCGG (p < 1e-15)
Frequency
R
Elk1
Mitotic cell cycle (GO:0000278)
R = -0.747
(p = 7.389e-11)
RNA splicing (GO:0008380)
R = -0.860
(p = 1.426e-18)
Ribosome biogenesis (GO:0042254)
R = -0.860
(p = 1.527e-18)
TFDP1
DNA replication (GO:0006260)
R = 0.757
(p = 2.279e-11)
Microtubule biogenesis (GO::0000226)
R = 0.717
(p = 1.310e-09)
Mitotic cell cycle (GO:0000278)
R = 0.677
(p = 3.354e-08)
AhR
protein catabolic process (Neg) (GO:0006511)
R = -0.474
(p = 8.289e-04)
protein catabolic process (Pos) (GO:0006511)
R = 0.580
(p = 1.123e-05)

Figure S2

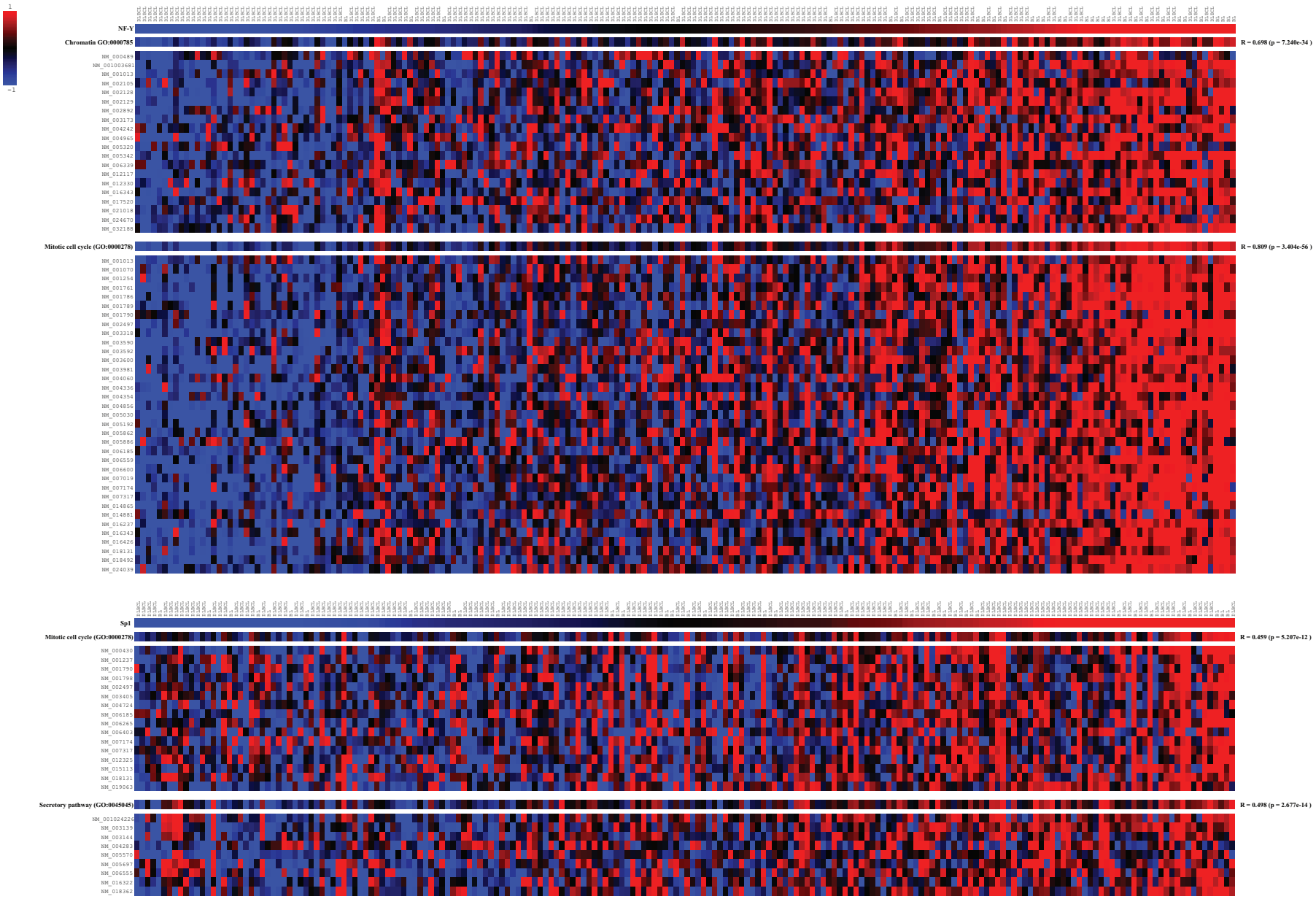

1
−1
NF-Y
Chromatin GO:0000785
R = 0.698 (p = 7.240e-34)
NM_000489
NM_001003681
NM_001013
NM_002105
NM_002128
NM_002129
NM_002892
NM_003173
NM_004242
NM_004965
NM_005320
NM_005342
NM_006339
NM_012117
NM_012330
NM_016343
NM_017520
NM_021018
NM_024670
NM_032188
Mitotic cell cycle (GO:0000278)
R = 0.809 (p = 3.404e-56)
NM_001013
NM_001070
NM_001254
NM_001761
NM_001786
NM_001789
NM_001790
NM_002497
NM_003318
NM_003590
NM_003592
NM_003600
NM_003981
NM_004060
NM_004336
NM_004354
NM_004856
NM_005030
NM_005192
NM_005862
NM_005886
NM_006185
NM_006559
NM_006600
NM_007019
NM_007174
NM_007317
NM_014865
NM_014881
NM_016237
NM_016343
NM_016426
NM_018131
NM_018492
NM_024039
Sp1
Mitotic cell cycle (GO:0000278)
R = 0.459 (p = 5.207e-12)
NM_000430
NM_001237
NM_001790
NM_001798
NM_002497
NM_003405
NM_004724
NM_006185
NM_006265
NM_006403
NM_007174
NM_007317
NM_012325
NM_015113
NM_018131
NM_019063
Secretory pathway (GO:0045045)
R = 0.498 (p = 2.677e-14)
NM_001024226
NM_003139
NM_003144
NM_004283
NM_005570
NM_005697
NM_006555
NM_016322
NM_018362

Figure S3

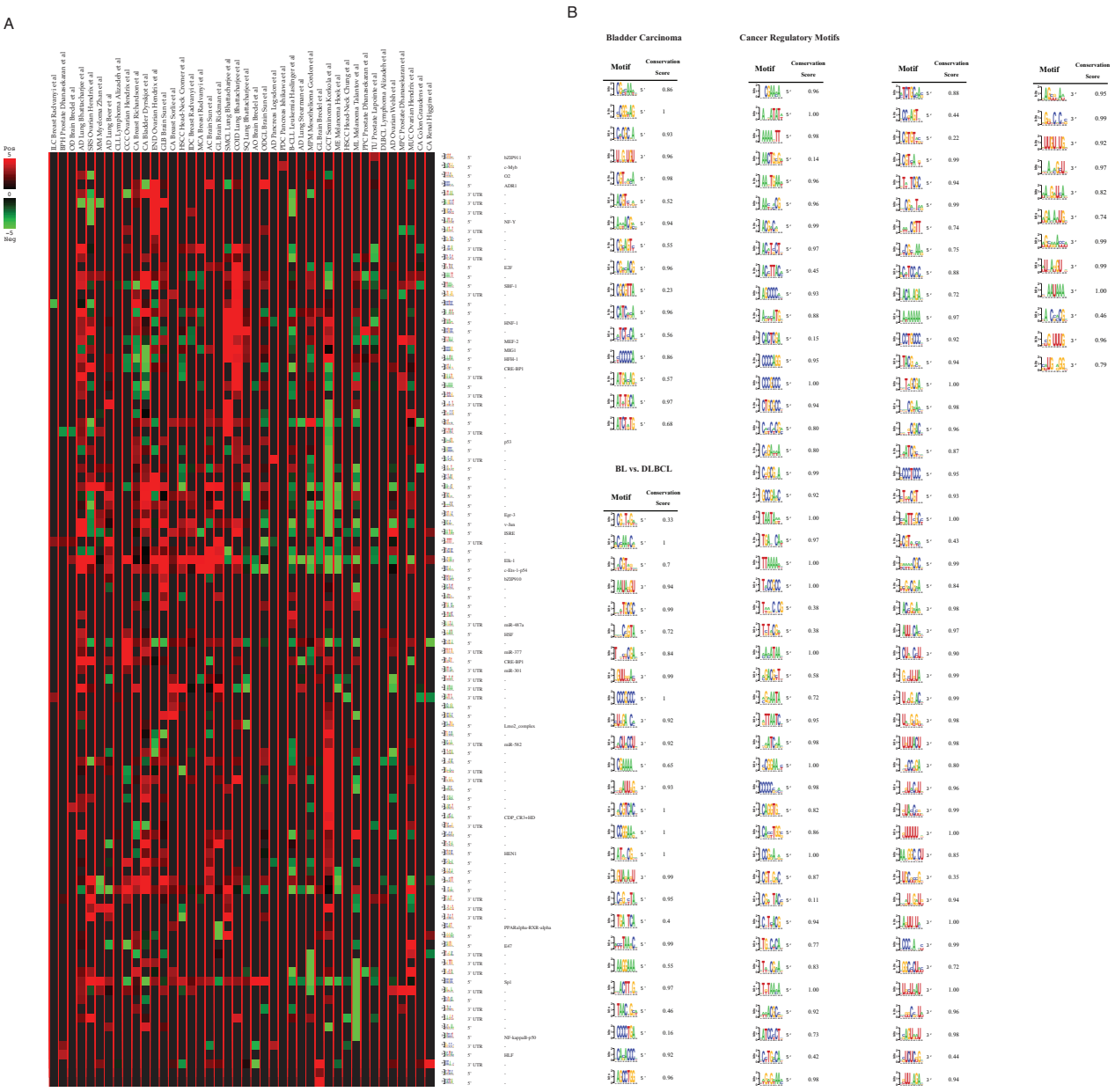

A
B
Pos
5
0
-5
Neg
Bladder Carcinoma
Motif
Conservation Score
Cancer Regulatory Motifs
Motif
Conservation Score
Motif
Conservation Score
Motif
Conservation Score
BL vs. DLBCL
Motif
Conservation Score

Figure S4 (A, B)

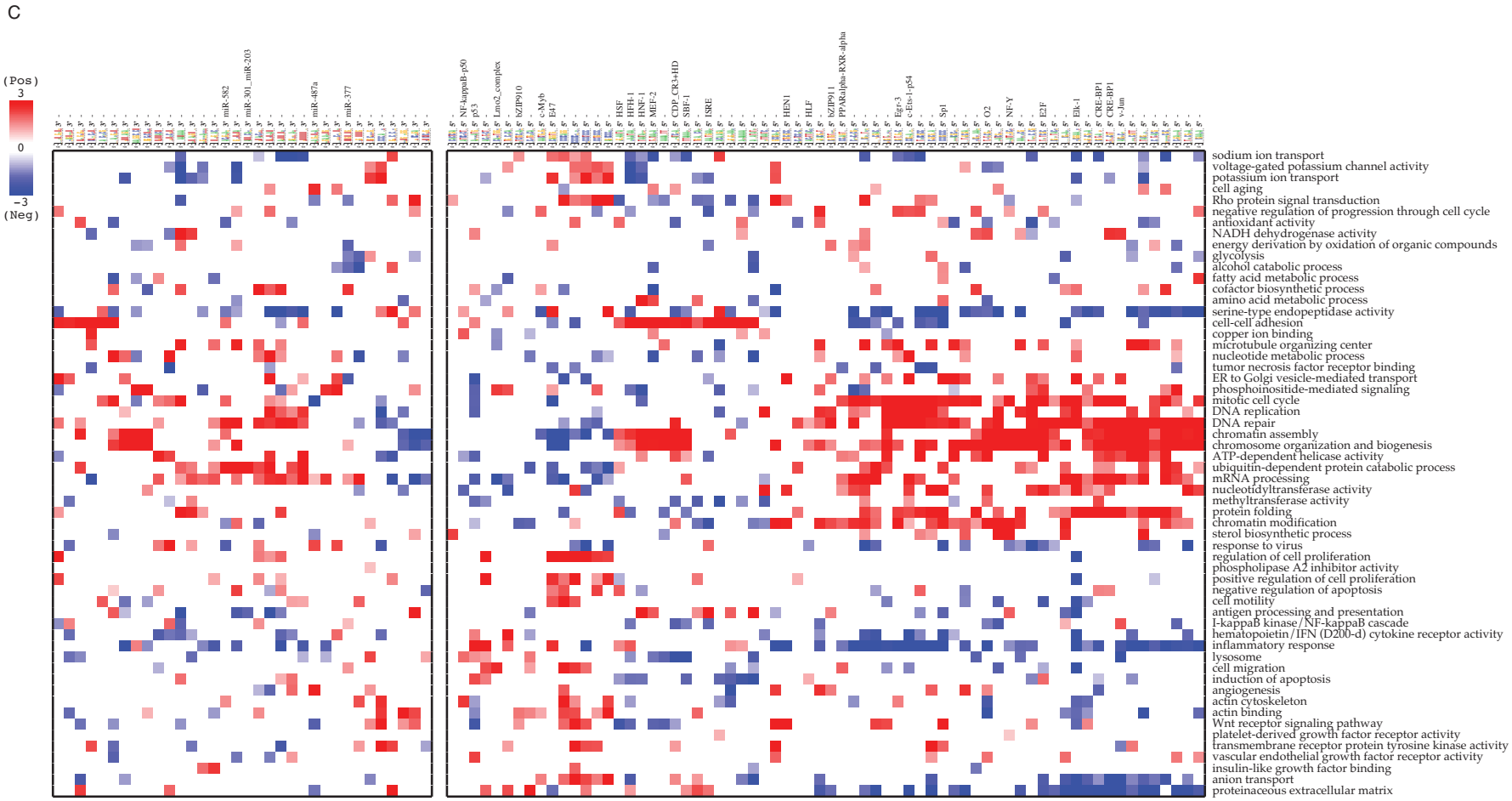
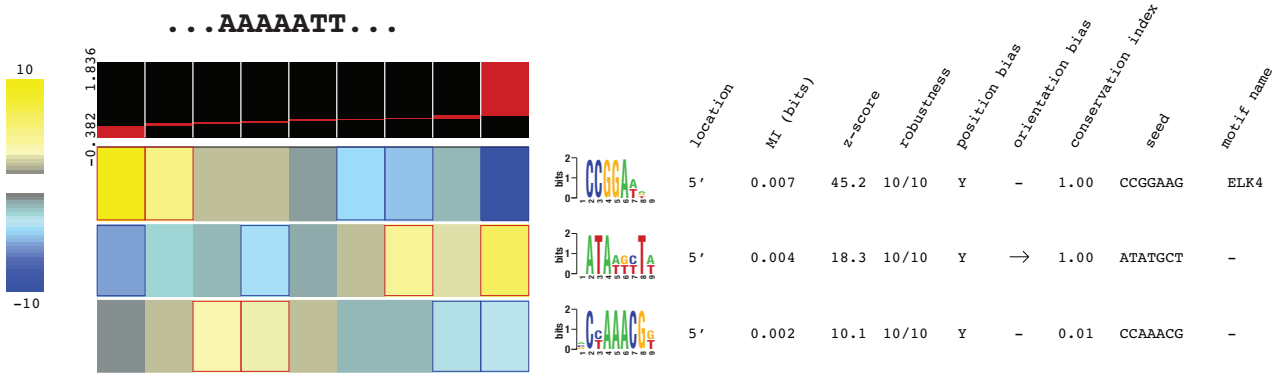

C
(Pos)
3
0
−3
(Neg)
miR-582
miR-301_miR-203
miR-487a
miR-377
NF-kappaB-p50
p53
Lmo2_complex
BZIP910
c-Myb
Ets1
CDP_CR3+HD
SBF-1
ISRE
HSF
HSF-1
HNF-1
MEF-2
HEN1
HLF
BZIP911
PPARalpha-RXR-alpha
Egr-3
c-Ets-1-p54
Sp1
C2
NF-Y
E2F
Elk-1
CRE-BP1
CRE-BP1
v-Jun
sodium ion transport
voltage-gated potassium channel activity
potassium ion transport
cell aging
Rho protein signal transduction
negative regulation of progression through cell cycle
antioxidant activity
NADH dehydrogenase activity
energy derivation by oxidation of organic compounds
glycolysis
alcohol catabolic process
fatty acid metabolic process
cofactor biosynthetic process
amino acid metabolic process
serine-type endopeptidase activity
cell-cell adhesion
copper ion binding
microtubule organizing center
nucleotide metabolic process
tumor necrosis factor receptor binding
ER to Golgi vesicle-mediated transport
phosphoinositide-mediated signaling
mitotic cell cycle
DNA replication
DNA repair
chromatin assembly
chromosome organization and biogenesis
ATP-dependent helicase activity
ubiquitin-dependent protein catabolic process
mRNA processing
nucleotidyltransferase activity
methyltransferase activity
protein folding
chromatin modification
sterol biosynthetic process
response to virus
regulation of cell proliferation
phospholipase A2 inhibitor activity
positive regulation of cell proliferation
negative regulation of apoptosis
cell motility
antigen processing and presentation
I-kappaB kinase/NF-kappaB cascade
hematopoietin/IFN (D200-d) cytokine receptor activity
inflammatory response
lysosome
cell migration
induction of apoptosis
angiogenesis
actin cytoskeleton
actin binding
Wnt receptor signaling pathway
platelet-derived growth factor receptor activity
transmembrane receptor protein tyrosine kinase activity
vascular endothelial growth factor receptor activity
insulin-like growth factor binding
anion transport
proteinaceous extracellular matrix

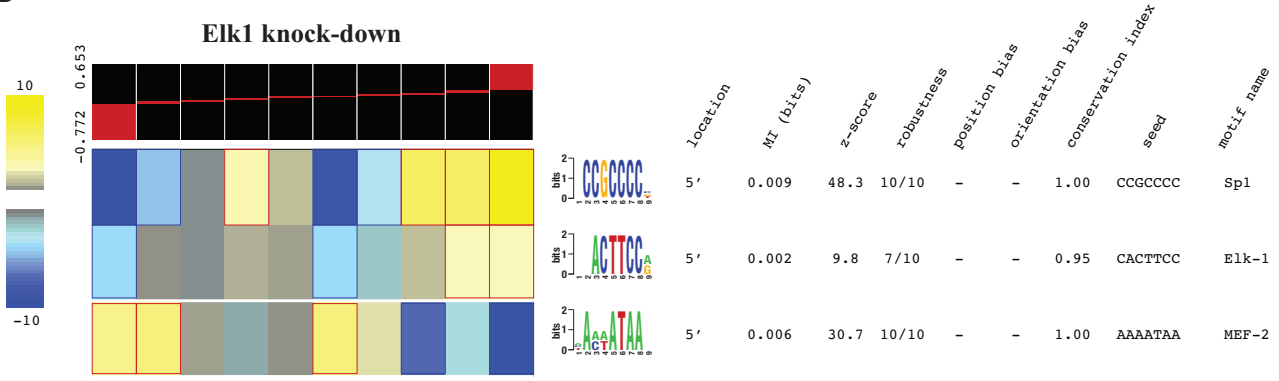
Figure S4 (C)

A



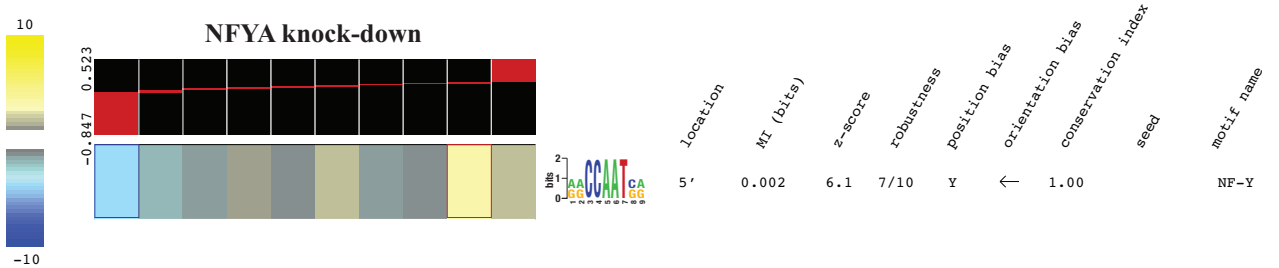
location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5′	0.007	45.2	10/10	Y	-	1.00	CCGGAAG	ELK4
5′	0.004	18.3	10/10	Y	→	1.00	ATATGCT	-
5′	0.002	10.1	10/10	Y	-	0.01	CCAAACG	-

B



location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5′	0.009	48.3	10/10	-	-	1.00	CCGCCCC	Sp1
5′	0.002	9.8	7/10	-	-	0.95	CACTTCC	Elk-1
5′	0.006	30.7	10/10	-	-	1.00	AAAATAA	MEF-2

C



location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5′	0.002	6.1	7/10	Y	←	1.00		NF-Y

Figure S5

Legends

Figure S1. iPAGE and PRMG Schematics, related to Figure 1. (A) Two exemplary expression profiles are shown: discrete (*e.g.* cluster indices from co-expression clustering) and continuous (*e.g.* log of fold change in expression level in the tumor sample compared to normal). Mutual information is then used to assess the level by which given pathway profiles are informative of these expression profiles. **(B)** Two exemplary expression profiles are shown: discrete (*e.g.* cluster indices from co-expression clustering) and continuous (*e.g.* log of fold change in expression level in the tumor sample compared to normal). Mutual information is then used to assess the level by which given pathway profiles are informative of these expression profiles.

Figure S2. Measuring co-regulation in the identified associations, related to Figure 2. (A) The regulatory interaction matrix for the bladder carcinoma *cis*-regulatory elements. **(B)** 1000 gene pairs are selected from bladder carcinoma dataset (Dyrskjot et al., 2004) and the distribution of their Pearson correlation coefficients is plotted. The background set includes all the genes in the dataset; whereas 'Elk1' is limited to the genes harboring Elk1 motifs in their upstream sequences. Similarly, TAGATGT plot represents the genes harboring [ACU]U[ACU]G[ACG]UGU (a novel 3' UTR element). We have also included this distribution for the simultaneous occurrence of these two motifs. Similarly, the distribution of R-values are shown for Elk1 and ACGTCGG (a upstream element [ACG]ACGT[CT][CGT][AG][CGT]) and their simultaneous presence. **(C)** Shown are the expression of Elk1, TFDP1 and AhR across the bladder carcinoma samples and correlations with target genes within their associated pathways. The expression of each pathway is calculated as the average normalized expression of the genes listed. A regression test is then used to calculate the correlation coefficients and their associated *p*-values.

Figure S3. The expression of key transcription factors and their targets in BL vs. DLBCL dataset, related to Figure 4. The normalized expression of NF-Y and Sp1 are shown across the BL vs. DLBCL samples. Similarly, the expression of the target genes in their associated pathways is also shown along with an average normalized expression for each pathway and its correlation with the upstream transcription factor.

Figure S4. Identifying the *cis*-regulatory elements that are informative of the expression patterns in various cancer datasets and their associations with known and putative downstream pathways, related to Figure 6. (A) Cancer regulatory map. The level of significance by which the genes harboring a given putative *cis*-regulatory element are up or down regulated is depicted here. This matrix is formatted to include only the known motifs and those that are significantly associated with more than 3 cancers. **(B) Network-level conservation scores.** This figure shows our discovered motifs and their network-level conservation scores with respect to the chicken genome (Elemento and Tavazoie, 2005). Values range from 0 to 1, with 1 being most conserved. **(C) The complete cancer pathway-regulatory interaction map.** Figure 6 in the main paper is a summarized version of this matrix.

Figure S5. FIRE analysis of experimentally tested associations, related to Figure 7. (A) The motifs that are most informative of the decoy AAAA[ATG]TT vs scrambled microarray experiment. **(B)** Knocking down Elk1 results in the upregulation of genes harboring Sp1, Elk1 and MEF-2 binding sites. **(C)** Knocking down NFYA results in upregulation of genes harboring the NF-Y binding site.

Table S1. The list, tissue and references of the cancer gene expression studies used to compile our initial dataset, related to Figure 5.

Table S2. A list of predictions based on the associations in the Cancer Pathway-Regulatory Interaction Map, related to Figure 6.

Supplemental Tables

Table S1

Tissue	Sample Name	Sample
Bladder	CA Bladder Dyrskjot et al	Carcinoma (Dyrskjot et al., 2004)
Brain	GBM Brain Liang et al	Glioblastoma Multiforme (Liang et al., 2005)
	OD Brain Bredel et al	Oligodendroglioma (Bredel et al., 2005)
	GL Brain Bredel et al	Glioblastoma (Bredel et al., 2005)
	AO Brain Bredel et al	Anaplastic Oligoastrocytoma (Bredel et al., 2005)
	GL Brain Rickman et al	Glioma (Rickman et al., 2001)
	ODGL Brain Sun et al	Oligodendroglioma (Sun et al., 2006)
	AC Brain Sun et al	Astrocytoma (Sun et al., 2006)
	GLB Brain Sun et al	Glioblastoma (Sun et al., 2006)
Breast	CA Breast Sorlie et al	Carcinoma (Sorlie et al., 2001)
	CA Breast Richardson et al	Carcinoma (Richardson et al., 2006)
	MCA Breast Radvanyi et al	Metastatic Breast Carcinoma (Radvanyi et al., 2005)
	ILC Breast Radvanyi et al	Invasive Lobular Carcinoma (Radvanyi et al., 2005)
	IDC Breast Radvanyi et al	Invasive Ductal Carcinoma (Radvanyi et al., 2005)
Colon	CA Colon Graudens et al	Carcinoma (Graudens et al., 2006)
Head-neck	HSCC Head-Neck Cromer et al	Head-Neck Squamous Cell Carcinoma(Cromer et al., 2004)
	HSCC Head-Neck Chung et al	Head-Neck Squamous Cell Carcinoma (Chung et al., 2004)
Leukemia	B-CLL Leukemia Haslinger et al	Chronic Lymphocytic Leukemia (Haslinger et al., 2004)
Lung	AD Lung Beer et al	Adenocarcinoma (Beer et al., 2002)
	AD Lung Bhattacharjee et al	Adenocarcinoma (Bhattacharjee et al., 2001)
	COID Lung Bhattacharjee et al	Carcinoid (Bhattacharjee et al., 2001)
	SQ Lung Bhattacharjee et al	Squamous Cell Lung Carcinoma (Bhattacharjee et al., 2001)
	SMCL Lung Bhattacharjee et al	Small Cell Lung Cancer (Bhattacharjee et al., 2001)
	AD Lung Stearman et al	Adenocarcinoma (Stearman et al., 2005)
Lymphoma	FL Lymphoma Alizadeh et al	Follicular Lymphoma (Alizadeh et al., 2000)
	DLBCL Lymphoma Alizadeh et al	Diffuse Large B-Cell Lymphoma (Alizadeh et al., 2000)
	CLL Lymphoma Alizadeh et al	Chronic Lymphocytic Leukemia (Alizadeh et al., 2000)
Melanoma	ML Melanoma Talantov et al	Cutaneous melanoma (Hoek et al., 2006)
	ME Melanoma Hoek et al	Melanoma (Talantov et al., 2005)
Mesothelioma	MPM Mesothelioma Gordon et al	Malignant Mesothelioma (Gordon et al., 2005)
Myeloma	MM Myeloma Zhan et al	Multiple Myeloma (Zhan et al., 2002)
Ovarian	AD Ovarian Welsh et al	Adenocarcinoma (Welsh et al., 2001)
	CCC Ovarian Hendrix et al	Clear Cell Carcinoma (Hendrix et al., 2006)
	MUC Ovarian Hendrix et al	Mucinous Adenocarcinoma (Hendrix et al., 2006)
	SRS Ovarian Hendrix et al	Serous Adenocarcinoma (Hendrix et al., 2006)
	END Ovarian Hendrix et al	Endometrioid Adenocarcinoma (Hendrix et al., 2006)
Pancreas	PDC Pancreas Ishikawa et al	Pancreatic Ductal Carcinoma (Ishikawa et al., 2005)
	AD Pancreas Logsdon et al	Adenocarcinoma (Logsdon et al., 2003)
Prostate	MPC Prostate Dhanasekaran et al	Metastatic Prostate Cancer (Dhanasekaran et al., 2001)
	PPC Prostate Dhanasekaran et al	Primary Prostate Cancer (Dhanasekaran et al., 2001)
	BPH Prostate Dhanasekaran et al	Benign Prostatic Hyperplasia (Dhanasekaran et al., 2001)
	TU Prostate Lapointe et al	Primary Tumor (Lapointe et al., 2004)
Renal	CA Renal Higgins et al	Carcinoma (Higgins et al., 2003)
	RCCC Renal Boer et al	Clear Cell Renal Cell Carcinoma (Boer et al., 2001)
	RCCC Renal Lenburg et al	Clear Cell Renal Cell Carcinoma (Lenburg et al., 2003)
Seminoma	GCT Seminoma Korkola et al	Germ Cell Tumor (Korkola et al., 2006)

Table S2

GO terms	Motifs	Significance
chromatin assembly	3′ 	p < 1e-86.7
DNA packaging	3′ 	p < 1e-32.5
chromosome organization and biogenesis	5′ 	p < 1e-11.1
ribonucleoprotein complex	5′ 	p < 1e-8.7
DNA packaging	5′ 	p < 1e-8.4
DNA repair	5′ 	p < 1e-7.7
mRNA processing	3′ 	p < 1e-7
cell-cell adhesion	3′ 	p < 1e-6.9
cell-cell adhesion	3′ 	p < 1e-6.7
ubiquitin-protein ligase activity	3′ 	p < 1e-6.4
protein-tyrosine kinase activity	3′ 	p < 1e-6.3
Golgi vesicle transport	5′ 	p < 1e-6.3
cytoskeletal protein binding	3′ 	p < 1e-6.2
GPI anchor binding	3′ 	p < 1e-6.2
DNA repair	3′ 	p < 1e-6
humoral immune response	5′ 	p < 1e-6
phosphoinositide-mediated signaling	3′ 	p < 1e-6
mitotic cell cycle	5′ 	p < 1e-5.9
mitosis	5′ 	p < 1e-5.5
response to wounding	5′ 	p < 1e-5.4
response to wounding	5′ 	p < 1e-5.3
cell-cell adhesion	5′ 	p < 1e-5.2
mRNA metabolic process	5′ 	p < 1e-5.1
small GTPase mediated signal transduction	5′ 	p < 1e-5
Wnt receptor signaling pathway	5′ 	p < 1e-4.8

Supplemental Procedures

In what follows, we provide a detailed description of the methods used in this study. The iPAGE software along with a minitutorial and supplemental results for this study (both experimental and computational) are available online at http://tavazoielab.princeton.edu/iPAGE/. Our approach, described here, involves the discovery of the informative *cis*-regulatory elements and cellular pathways from gene expression datasets; a subsequent analysis recovers the pathways that are likely regulated by the identified putative binding sites. A schematic of the FIRE/iPAGE framework is presented in Figure 1.

Pre-processing of input datasets

All cancer microarray datasets used in this study where 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 (i.e., urinary bladder cancer), each gene was associated with a continuous expression value based on the following equation:

(1) $v = s(1-p)$,

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 with maximal and minimal values of 1 and -1 respectively.

In the discrete case, genes were first clustered into $\sim\sqrt{N}$ groups (N is the total number of genes) based on their expression values in the normal and tumor samples, using the k-means unsupervised clustering approach. 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 a single background cluster. Subsequently, each gene was associated with the cluster index of the cluster to which it belongs.

FIRE: De novo discovery of informative regulatory elements

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

iPAGE: A detailed explanation of the algorithm

Expression profile

An *expression profile* is defined across N genes, where each gene is associated with a unique expression measure. Expression measures, discrete or continuous, can be obtained from a variety of gene-level measurements or analyses. For example, cluster indices from a partitioning process or the ranks obtained from sorting are discrete measures; whereas, results from a single microarray or any continuous-type statistic (e.g., p-values) are continuous values. In this study we have demonstrated this unifying capacity of iPAGE; e.g., in the bladder carcinoma we have used a continuous statistic derived from Student's t-test while in the BL vs DLBCL case we employed discrete indices obtained from clustering of gene expression values across all the samples. From here forward, we refer to these lists of input values as *expression profiles*. Schematized continuous and discrete expression profiles are shown in Figure S1A.

Pathway Profile

Each gene can be associated with a subset of *M* known pathways (e.g. from the Gene Ontology annotations). For each pathway, the *pathway profile* is defined as a binary vector with *N* elements, one for each gene. In this profile, "1" indicates that the gene belongs to the pathway and "0" indicates that it does not. A schematized pathway profile is shown in Figure S1A.

Quantizing continuous expression profiles

Although the concept of mutual information is defined for both discrete and continuous random variables, in practice, continuous data are discretized before calculating the mutual information (MI) values. Our quantization procedure is based on the maximum entropy principle (so as to make the least assumptions about the underlying data distribution), and involves using equally populated "expression bins". Thus, the discretization step only requires a single parameter, i.e., the number of genes in each bin. In the default iPAGE settings, the number of bins (N_e) is determined by:

(2) $$N_e \cdot N_m = N/50$$

where N_m is the number of bins in the pathway profile (here N_m=2). Although determining N_e values from Eqn. (2) allows a reliable calculation of mutual information (Slonim et al., 2005), other values can also be explored by the user. In this study, we used the continuous mode in one of the datasets and variations in the number of bins did not significantly change the results. Indeed, when we ran FIRE and iPAGE on the bladder carcinoma dataset with various numbers of bins (10, 50, 100 and 250), the identified seeds (k-mers) largely overlapped, with hypergeometric *p*-values always less than 1e-53 (down to 1e-281 in some comparisons). We made the same observation for the number of iPAGE-identified pathways, with hypergeometric *p*-values always less than 1e-20 (down to 1e-83).

Calculating the mutual information values

Given a *pathway profile* and an *expression profile* with *Ne* bins (or clusters), we create a table *C* of dimensions 2× *Ne*, in which $C(1, j)$ represents the number of genes that are contained in the j^{th} expression bin and are also present in the given pathway. $C(2, j)$, on the other hand, contains the number of genes that are in the j^{th} expression bin but are not assigned to the pathway. Given this table, we calculate the empirical mutual information as follows:

(3) $$I(\text{candidate pathway}; \text{expression}) = \sum_{i=1}^{2}\sum_{j=1}^{N_e} P(i, j)\log\frac{P(i, j)}{P(i)P(j)},$$

where $P(i, j) = C(i, j)/N$, $P(i) = \sum_{j=1}^{N_e} P(i, j)$ and $P(j) = \sum_{i=1}^{2} P(i, j)$.

Randomization-based statistical testing

To assess the statistical significance of the calculated MI values, we use a non-parametric randomization-based statistical test. Given *I* as the real MI value and keeping the pathway profile unaltered, the expression profile is shuffled 10,000 times and the corresponding MI values I_{random} are calculated. A pathway is accepted only if *I* is larger than (1-*max_p*) of the I_{random} values (*max_p* is set to 0.005 by default). This corresponds to a p-value < 0.005. In iPAGE, pathways are first sorted by information (from informative to non-informative). Starting from the most informative pathway, the statistical test described above is applied to each pathway, and pathways that pass the test are returned (provided they also pass the conditional information test described below). When *k* contiguous pathways in the sorted list do not pass the test, the procedure is stopped (*k* is set to 20 by default).

Removing redundantly informative pathways

Due to the hierarchical and nested nature of pathway annotations (e.g. Gene Ontology), many pathways display some level of redundancy, i.e., two pathways may be represented by very similar sets of genes (e.g.

GO:0006511, ubiquitin dependent protein catabolic process and GO:0019941, modification dependent protein catabolic process). To discover representative pathways and remove redundant ones, we require that each returned pathway be highly informative about the expression profile, but also bring a significant amount of new information compared to all other significantly informative pathways as calculated by conditional mutual information (Cover and Thomas, 2006). To achieve this, we require that each candidate pathway fulfills

$$\frac{I(\text{candidate pathway; expression} \mid \text{accepted pathways})}{I(\text{candidate pathway; accepted pathway})} > r \quad (4)$$

for all already accepted pathways, i.e., all pathways that have already passed the statistical and conditional information tests. An identical criterion was used in FIRE (Elemento et al., 2007). In iPAGE, *r* is set to 5 by default and only the pathways satisfying the above equation are presented in the graphical output; however, the list of all significant pathways is also created and stored as a text file.

Pathway over- and under-representation

Informative pathways are generally over-represented or under-represented in certain expression clusters/bins. To quantify the level of over- and under-representation, the hypergeometric distribution is used to calculate two distinct *p*-values:

$$p_{over}(X \geq x) = \sum_{i=x}^{N} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad \text{for over-representation} \quad (5)$$

and

$$p_{under}(X \leq x) = \sum_{i=0}^{x} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \quad \text{for under-representation,} \quad (6)$$

where *x* equals the number of genes in the given expression bin/cluster which are also assigned to the given pathway. *m* is the number of genes assigned to the pathway, *n* is the number of genes in the expression bin and *N* is the total number of genes. If $p_{over} < p_{under}$, we consider the pathway to be over-represented in the expression bin/cluster; otherwise, it is under-represented.

iPAGE graphical output

The over- and under-representation *p*-values described in the previous section are used to draw a heatmap, i.e., a graphical representation of pathway over- and under-representation across all expression bins/clusters. In this heatmap, the rows represent the significantly informative pathways and the columns are the expression bins/clusters. Colors indicate over- or under-representation levels. The red color-map indicates (in $\log_{10}$) the over-representation *p*-values; whereas, the blue color-map shows under-representation.

Additional iPAGE output files

In addition to the graphical heatmap, iPAGE generates files containing the actual log(*p*-values) for over- and under-representations, and the list of removed redundant pathways.

False Discovery Rate (FDR)

In order to measure the FDR of our method, we randomly shuffled the gene labels of the gene expression profile and counted the number of pathways discovered compared to the non-shuffled data. For the BL vs DLBCL dataset, in two random trials we found on average one significant pathway in comparison with 525

pathways discovered in the real dataset. In the continuous bladder carcinoma dataset, we found 224 significant pathways, whereas, in randomized expression profiles 4.5 significant pathways were deemed significant. Similar tests on other datastes puts the false discovery rate of iPAGE in the range of 0.001 to 0.02 with continuous datastes biased towards higher FDRs.

iPAGE command line
The basic command line syntax for iPAGE is :

```
perl page.pl --expfile=<inp> --species=<sp> --exptype=<type>
```

where <inp> indicates the input expression profile (a two-column tab-delimited text file with gene names in the first column and expression measures in the second), <sp> indicates the species, and <type> indicates whether the expression profile is discrete (e.g., cluster indices) or continuous (e.g., expression values obtained from a single microarray experiment). We have prepackaged pathway annotations for many species, ranging from bacteria to human.

For example, the following command line will run iPAGE on a continuous *E. coli* expression profile :

```
perl page.pl --expfile=./TEST/continuous.exp --species=human_go --exptype=continuous
```

iPAGE creates an expfile_PAGE directory where the results are saved to (./TEST/continuous.exp_PAGE in this case).

Pathway-Regulatory Interaction Map Generator (PRMG)

Motif definition
As described in (Elemento et al., 2007), regulatory elements (motifs) are defined as *regular expressions* and can only consist of the following characters: A, C, G, T, [AC], [AG], [AT], [CG], [CT], [GT], [ACG], [ACT], [AGT], [CGT], and N (equivalent to [ACGT]).

Motif profile
We look for motifs both in 5' upstream (DNA motifs) and in 3'UTR sequences (RNA motifs). Given a motif, the *motif profile* is defined as a binary vector with *N* elements, where for each gene, "+1" indicates the presence and "0" indicates the absence of the motif in the corresponding promoter (or 3'UTR). "1" indicates that at least one match of the regular expression is present in the sequences (see Figure S1B). For 5' sequences both strands are searched; whereas, in 3' UTR sequences only the transcribed strand is considered.

We used a generic definition of *active motif profile* (Elemento et al., 2007) to build the pathway-regulatory map; i.e., we only count the motif occurrences that are in expression cluster/bins in which the motif is over-represented. This approach filters out motif occurrences that are unlikely to be functional.

Pathway Profiles
For each pathway, the *pathway profile* is defined the same as in iPAGE.

Creating pathway-regulatory interaction maps
In the first step, we calculate the mutual information between the *motif profile* and *pathway profile* for each pair of motifs and pathways. We then assess the significance of these associations through 1,000 random shuffles of the motif profile and recalculating the MI values. By default, a category is accepted only if the real

MI is larger than 995 of the random values. The associations that pass this test are deemed significant and their under- or over-representation *p*-values are calculated using equations (5) and (6).

Graphical output
We build a matrix with motifs as columns and pathways as rows where the non-zero elements represent the –log10(*p*-value) in case of over-representations and log10(*p*-value) otherwise. This matrix is then visualized as a blue-red heatmap with red and blue elements representing positive and negative associations respectively. A schematic representation of this method is shown in Figure S1B.

PRMG command line

The PRMG script (prmg.pl) is part of the iPAGE package and is located in the PAGEvx.x directory; however, it also relies on FIRE outputs to run (FIRE is available at http://tavazoielab.princeton.edu/FIRE/):

```
export FIREDIR=/path/to/FIRE
export PAGEDIR=/path/to/iPAGE
perl prmg.pl --expfile=<inp> --species=<sp>
```

where <inp> indicates the input expression profile, <sp> indicates the species. The script does not work on *expfile* itself, but uses it to locate iPAGE and FIRE summary files (in expfile_PAGE and expfile_FIRE directories).

For example, the following command line will run PRMG on a continuous expression profile:

```
perl prmg.pl --expfile=./TEST/continuous.exp --species=human_go
```

The results are written to a motif_cat.cdt file and the graphical representations are created in the motif_cat.eps and motif_cat.pdf files. In order to run this program you also need to install the cluster 3.0 perl binder at http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm and define a global variable termed USRDIR pointing to the directory where this binder is installed:

```
export USRDIR=/path/to/cluster3
```

The Regulatory Network of Bladder Carcinoma
Our general meta-analysis of bladder cancer vs normal (Dyrskjot et al., 2004) reveals the most prominent signatures of a cancer state: in this case, a faster cell cycle and a repressed immune response through the regulatory effects of E2F and SEF1/E47 transcription factors respectively. We also identified a range of significant *cis*-regulatory elements, including a putative 3'UTR element, NUNGNUGU (seed UAGAUGU/TAGATGT) (Figure 2B, main text). Our approach also reveals that several of these motifs co-occur in promoters or 3'UTRs of the same genes (Figure S2A), thus suggesting possible cooperations between the regulatory factors that bind them. In order to provide additional evidence for these predicted cooperations, we used the approach described in Pilpel et al. (2001) to compare the extent to which two of the novel motifs (one DNA and one RNA) cooperate with the Elk-1 motif in co-regulating their target genes. First, we selected 1000 random pairs of genes and used Pearson correlation to calculate the correlation coefficient (R) between the expression levels of each pair across the bladder carcinoma dataset (Dyrskjot et al., 2004). We then repeated the same procedure, this time on the set of genes harboring an Elk1 motif in their upstream sequence. As it is shown in Figure S2B, the distribution of the resulting R-values from Elk1 target genes is shifted to the right compared to the random background values (*p*-value <1e-14). The genes harboring the DNA motif [ACG]ACGT[CT][CGT][AG][CGT] (seed ACGTCGG), show a distribution similar to that of Elk1 (*p*-value <1e-15) and focusing on the genes that harbor both of the motifs results in an even larger shift towards higher R-values (*p*-value <1e-15). The *p*-values reported in each case have been calculated using Mann-Whitney test,

comparing each distribution to that of the background. Repeating this procedure for [ACU]U[ACU]G[ACG]UGU (seed UAGAUGU/TAGATGT), resulted in comparable distributions (Figure S2B). These observations further highlight the biological relevance of the discovered novel motifs and provide additional support for the predicted motif interactions.

Building a regulatory map of cancer deregulation

We studied regulatory perturbations across many cancer types to capture both globally deregulated and more cancer-specific pathways. We compiled a compendium of 46 cancer versus normal gene expression microarray datasets (Table S1). We then processed the samples and used iPAGE to build a cancer pathway map (Figure 5 in the main text). We also used FIRE on our compendium to build a cancer regulatory map. In essence, sequence motifs whose associated genes show significant deregulation in the tumor samples are identified and compiled to form this regulatory map (Figure S4A). Apart from their independent occurrences in multiple datasets, most of these motifs also have high network-level conservation scores (Figure S4B). Figure S4B also includes the *cis*-regulatory elements identified in the bladder carcinoma and lymphoma datasets.

Subsequently, using an information-theoretical approach, we associated the discovered *cis*-regulatory elements with the deregulated pathways to build a cancer pathway-regulatory interaction map (Figure S4C). In Table S2, we list a number of novel and significant associations from this map, representing an unknown regulatory protein (or miRNA) potentially regulating the associated pathway through recognition of the corresponding sequence motif. In some cases, we have predicted novel associations for known transcription factors (or miRNAs).

Experimental validation of the discovered regulatory associations

Transfection of siRNAs targeting Elk1 and NFYA transcription factors

ON-Target*plus*™ (Dharcomon) set of siRNAs for each TF were transfected into MDA-MB-231 cells (growin in D10F medium) using Lipofectamine™ 2000 (invitrogen). 72 hours after transfection, RNA samples were extracted from the cells (mirVana™ miRNA Isolation Kit) and were subjected to cDNA synthesis (SuperScript® III RTS First-Strand cDNA Synthesis Kit from Invitrogen). mRNA knock-down in each sample was verified using SYBRE Green qPCR reactions (Universal ProbeLibrary Assay Design Center, Roche Applied Science). For each TF, we selected two of the successfully knocked-down transfections and extracted their total RNA along with mock-transfected cells as controls. We then differentially labeled the RNA samples with Cy3 and Cy5 dyes and hybridized them to Agilent human gene expression arrays (4×44k). The genes with significant discordant changes between the two biological replicates were filtered out and for the rest, the Cy3/Cy5 values were averaged and combined into a single dataset as log of ratios. The expression profiles are deposited in GEO (GSE18849) and are also available at http://tavazoielab.princeton.edu/iPAGE/.

Transfection of decoy and scrambled oligonucleotide sequences

For the validation experiments, we chose two of the genes implicated by FIRE to have a version of AAAA[ATG]TT (NM_000337 and NM_001024660). For each gene, we then synthesized a 19bp sequence containing the AAAA[ATG]TT motif. These sequences were also randomly shuffled to create scrambled sequences as controls. The resulting sequences were synthesized as double stranded oligonucleotides: Decoy1: caattGAAATTTTGagcaa, Scrambled1: gtTtATAcAcTaaaGaTGa, Decoy2: gctggAAAAAATTTaagac, Scrambled2: aagATTgctAgAAgAaATc. We then transfected these oligonucleotides into MDA-MB-231 cells grown in D10F medium at a concentration of 1 μM (TransIT®-Express Transfection Reagent). 72 hours post-transfection, we extracted RNA and differentially labeled the samples with Cy3 or Cy5 dyes. The samples were then hybridized to Agilent human gene expression arrays (4×44k). The Cy3/Cy5 ratios from the two sets were then averaged, filtered and combined in a single dataset

as log of ratios. In this step, we filtered out ~2000 genes that showed significantly discordant expression level changes in the two biological replicates. The expression profiles are deposited in GEO (GSE18844) and are also available at http://tavazoielab.princeton.edu/iPAGE/.

FIRE analysis of experimentally tested associations

As shown in Figure S5, we used FIRE to also analyze our gene expression profiles from both decoy vs. scrambled dataset and TF knock-down datasets. In the AAAA[ATG]TT dataset, in addition to ATA[AT][GT][CT]T[AT] (which resembles the reverse complement of AAAA[ATG]TT), we also discovered Elk4 and another novel motif (Figure S5A). The observed deregulation in the Elk4 downstream genes explains the up-regulation in the cell cycle genes, as this TF is a known modulator of mitosis. Similarly, we observed an up-regulation in the genes harboring the CCAAT motif in the NFYA knock-down dataset (Figure S5C).

Supplemental References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503-511.

Amin, J., Ananthan, J., and Voellmy, R. (1988). Key features of heat shock regulatory elements. Mol Cell Biol *8*, 3761-3769.

Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., *et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med *8*, 816-824.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., *et al.* (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A *98*, 13790-13795.

Boer, J.M., Huber, W.K., Sultmann, H., Wilmer, F., von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Fuzesi, L., *et al.* (2001). Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. Genome Res *11*, 1861-1870.

Bredel, M., Bredel, C., Juric, D., Harsh, G.R., Vogel, H., Recht, L.D., and Sikic, B.I. (2005). Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. Cancer Res *65*, 8679-8689.

Busche, S., Descot, A., Julien, S., Genth, H., and Posern, G. (2008). Epithelial cell-cell contacts regulate SRF-mediated transcription via Rac-actin-MAL signalling. J Cell Sci *121*, 1025-1035.

Chung, C.H., Parker, J.S., Karaca, G., Wu, J., Funkhouser, W.K., Moore, D., Butterfoss, D., Xiang, D., Zanation, A., Yin, X., *et al.* (2004). Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell *5*, 489-500.

Cromer, A., Carles, A., Millon, R., Ganguli, G., Chalmel, F., Lemaire, F., Young, J., Dembele, D., Thibault, C., Muller, D., *et al.* (2004). Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis. Oncogene *23*, 2484-2498.

Crosby, M.E., Jacobberger, J., Gupta, D., Macklis, R.M., and Almasan, A. (2007). E2F4 regulates a stable G2 arrest response to genotoxic stress in prostate carcinoma. Oncogene *26*, 1897-1909.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A., and Chinnaiyan, A.M. (2001). Delineation of prognostic biomarkers in prostate cancer. Nature *412*, 822-826.

Fukada, T., Ohtani, T., Yoshida, Y., Shirogane, T., Nishida, K., Nakajima, K., Hibi, M., and Hirano, T. (1998). STAT3 orchestrates contradictory signals in cytokine-induced G1 to S cell-cycle transition. Embo J *17*, 6670-6677.

Gordon, G.J., Rockwell, G.N., Jensen, R.V., Rheinwald, J.G., Glickman, J.N., Aronson, J.P., Pottorf, B.J., Nitz, M.D., Richards, W.G., Sugarbaker, D.J., *et al.* (2005). Identification of novel candidate

oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. Am J Pathol *166*, 1827-1840.

Graudens, E., Boulanger, V., Mollard, C., Mariage-Samson, R., Barlet, X., Gremy, G., Couillault, C., Lajemi, M., Piatier-Tonneau, D., Zaborski, P., *et al.* (2006). Deciphering cellular states of innate tumor drug responses. Genome Biol *7*, R19.

Haslinger, C., Schweifer, N., Stilgenbauer, S., Dohner, H., Lichter, P., Kraut, N., Stratowa, C., and Abseher, R. (2004). Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. J Clin Oncol *22*, 3937-3949.

Hendrix, N.D., Wu, R., Kuick, R., Schwartz, D.R., Fearon, E.R., and Cho, K.R. (2006). Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas. Cancer Res *66*, 1354-1362.

Higgins, J.P., Shinghal, R., Gill, H., Reese, J.H., Terris, M., Cohen, R.J., Fero, M., Pollack, J.R., van de Rijn, M., and Brooks, J.D. (2003). Gene expression patterns in renal cell carcinoma assessed by complementary DNA microarray. Am J Pathol *162*, 925-932.

Hoek, K.S., Schlegel, N.C., Brafford, P., Sucker, A., Ugurel, S., Kumar, R., Weber, B.L., Nathanson, K.L., Phillips, D.J., Herlyn, M., *et al.* (2006). Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. Pigment Cell Res *19*, 290-302.

Ishikawa, M., Yoshida, K., Yamashita, Y., Ota, J., Takada, S., Kisanuki, H., Koinuma, K., Choi, Y.L., Kaneda, R., Iwao, T., *et al.* (2005). Experimental trial for diagnosis of pancreatic ductal carcinoma based on gene expression profiles of pancreatic ductal cells. Cancer Sci *96*, 387-393.

Korkola, J.E., Houldsworth, J., Chadalavada, R.S., Olshen, A.B., Dobrzynski, D., Reuter, V.E., Bosl, G.J., and Chaganti, R.S. (2006). Down-regulation of stem cell genes, including those in a 200-kb gene cluster at 12p13.31, is associated with in vivo differentiation of human male germ cell tumors. Cancer Res *66*, 820-827.

Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., *et al.* (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A *101*, 811-816.

Lenburg, M.E., Liou, L.S., Gerry, N.P., Frampton, G.M., Cohen, H.T., and Christman, M.F. (2003). Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. BMC Cancer *3*, 31.

Liang, Y., Diehn, M., Watson, N., Bollen, A.W., Aldape, K.D., Nicholas, M.K., Lamborn, K.R., Berger, M.S., Botstein, D., Brown, P.O., *et al.* (2005). Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. Proc Natl Acad Sci U S A *102*, 5814-5819.

Logsdon, C.D., Simeone, D.M., Binkley, C., Arumugam, T., Greenson, J.K., Giordano, T.J., Misek, D.E., Kuick, R., and Hanash, S. (2003). Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. Cancer Res *63*, 2649-2657.

Pilpel, Y., Sudarsanam, P., and Church, G.M. (2001). Identifying regulatory networks by

combinatorial analysis of promoter elements. Nat Genet *29*, 153-159.

Radvanyi, L., Singh-Sandhu, D., Gallichan, S., Lovitt, C., Pedyczak, A., Mallo, G., Gish, K., Kwok, K., Hanna, W., Zubovits, J., *et al.* (2005). The gene associated with trichorhinophalangeal syndrome in humans is overexpressed in breast cancer. Proc Natl Acad Sci U S A *102*, 11005-11010.

Richardson, A.L., Wang, Z.C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J.D., Livingston, D.M., and Ganesan, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell *9*, 121-132.

Rickman, D.S., Bobek, M.P., Misek, D.E., Kuick, R., Blaivas, M., Kurnit, D.M., Taylor, J., and Hanash, S.M. (2001). Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res *61*, 6885-6891.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., *et al.* (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet *24*, 227-235.

Singh, V., and Aballay, A. (2006). Heat-shock transcription factor (HSF)-1 pathway required for Caenorhabditis elegans immunity. Proc Natl Acad Sci U S A *103*, 13092-13097.

Slonim, N., Atwal, G.S., Tkacik, G., and Bialek, W. (2005). Information-based clustering. Proc Natl Acad Sci U S A *102*, 18297-18302.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A *98*, 10869-10874.

Stearman, R.S., Dwyer-Nield, L., Zerbe, L., Blaine, S.A., Chan, Z., Bunn, P.A., Jr., Johnson, G.L., Hirsch, F.R., Merrick, D.T., Franklin, W.A., *et al.* (2005). Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model. Am J Pathol *167*, 1763-1775.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.

Sun, L., Hui, A.M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., *et al.* (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell *9*, 287-300.

Talantov, D., Mazumder, A., Yu, J.X., Briggs, T., Jiang, Y., Backus, J., Atkins, D., and Wang, Y. (2005). Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clin Cancer Res *11*, 7234-7242.

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampton, G.M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci U S A *98*, 1176-1181.

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M., *et al.* (2002). Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. Blood *99*, 1745-1757.