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 Procedures

In what follows, we provide a detailed description of the methods used in this study. The iPAGE software along with a minitutorial and other results from 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 and Figure S12.

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 S13.

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 S13.

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

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

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

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:

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

and

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

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. We have tabulated the results in Table S1, for the BL vs DLBCL and bladder cancer expression datasets (continuous and clustered profiles).

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 S14). 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 S14.

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
```

Experimental validation of the discovered regulatory associations

Transfection of siRNAs targeting Elk1 and NFYA transcription factors

ON-Target*plus*^{TM} (Dharcomon) set of siRNAs for each TF were transfected into MDA-MB-231 cells (growin in D10F medium) using Lipofectamine^{TM} 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 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 profile is available at http://tavazoielab.princeton.edu/iPAGE/.

Supplemental Results

iPAGE and Gene-set Enrichment Analysis

As mentioned in the main text, a variety of powerful approaches have been developed for gene-set enrichment analysis (e.g., Sinha et al., 2008; Subramanian et al., 2005). These computational methods rely on different statistical tests to assess the non-random distribution of pathway-membership across the expression values. For examples, Sinha et al. (2008) use the hypergeometric distribution to calculate the overlap between gene-expression clusters and pathway-memberships. Subramanian et al. (2005), on the other hand, use a rank-based approach to discover non-random patterns. These approaches are well-defined, powerful and tested; however, in a setting where a large number of statistical tests are performed, sensitivity is reduced mainly because of multiple-testing corrections.

In iPAGE, we employ mutual information to tackle this problem. Although iPAGE is only one component in our broader approach, it provides major advantages over the previous methodologies:

1. In comparison with the hypergeometric distribution, using a mutual information-based method notably decreases the number of statistical tests necessary (by the number of clusters/bins), resulting in a higher sensitivity at the same false discovery rate.
2. Mutual information also enables iPAGE to analyze both discrete and continuous inputs making it a universal approach for analyzing any type of data without the need for any upstream analysis. This is particularly important in this study where we analyze different whole-genome datasets.
3. iPAGE is also capable of discovering non-monotonic associations, a category that is masked in rank-based methods. In these associations, different components of a pathway may show opposite expression patterns relative to each other. For example, a number of metabolic pathways include both biosynthetic and catabolic genes and an increase in the final product typically requires the simultaneous down-regulation of the catabolic genes and up-regulation of the biosynthetic genes.
4. In iPAGE, we use conditional information to substantially reduce the redundancy, which in turn results in a concise and manageable graphical output.
5. We have been extra careful in limiting the computational expenses and our package is relatively fast. Also, we have attempted to make iPAGE as user-friendly as possible. In addition to the complete command-line version, we have developed Graphical User Interfaces for both Mac OS X and windows users for rapid and every-day use of the application by non-experts.

The Regulatory Network of Bladder Carcinoma

The general method for meta-analysis of bladder cancer vs normal dataset used in the main text (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 S1A), 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 S1B, 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 S1B). These observations further highlight the biological relevance of the discovered novel motifs and provide additional support for the predicted motif interactions.

While the continuous analysis of bladder carcinoma signatures apparently captures a global picture of deregulations, it does not take into account intra-cancer variations and may hide pathways that are deregulated in different subsets of samples. In order to increase our sensitivity in capturing a broader range of pathway deregulations, we first clustered the genes based on their expression across normal and tumor samples and then combined the clusters with low average differences between normal and tumor into a single background cluster. The informative pathways that iPAGE discovered from this discrete input data were generally similar to the pathways obtained when analyzing the continuous profile, as described in the main text. However, iPAGE also identified NF-κB, PI3-K and Rho signaling pathways as globally up-regulated in the tumor samples (Figure S15). Genes involved in "Negative regulation of apoptosis" also show a substantial increase in their expression compared to normal samples (Figure S15).

The identification of additional pathways without an increase in FDR (Table S1) suggests that using a cluster-based approach, which takes into account the intra-cancer variability, may increase sensitivity and reveal additional pathways. Interestingly, the benefits of using this approach become even more apparent when we search for *cis*-regulatory elements. We identified 128 significantly informative motifs, a summarized version of which is presented in Figure S16. The sequence motifs discovered in the continuous analysis are included among the new set of motifs (e.g. E2F and SEF-1); however, this analysis reveals many new known and novel regulatory elements including p53, Elk-1, Sp1, NF-Y and CRE-BP1 (Figure S16).

Among the known regulatory elements, Elk-1, Sp1, NF-Y and E2F show a significant association with "DNA replication" compared to E2F alone in the continuous method. E47 and SEF-1 are associated with the "immune response" pathways (Figure S17). In addition to these, we also captured the role of STAT3 transcription factor in cell cycle regulation. The signal transducers and activators of transcription (STATs) are a family of proteins with key roles in cellular differentiation and proliferation. Among these, STAT3 is known to be involved in coordinating G1-S transition (Fukada et al., 1998).

In the pathway-regulatory interaction map (Figure S17), up-regulation in the "protein folding" genes is associated with HSF (Heat Shock Factor). In higher eukaryotes, HSF binds the heat shock element (HSE) located in the promoters of heat shock genes to activate their transcription as part of the unfolded protein response (Amin et al., 1988). In addition to its apparent role in protein folding, HSF is also believed to participate in general immune response (Singh and Aballay, 2006) potentially explaining its association with this pathway in our analysis (Figure S17).

Besides the abovementioned transcription factors, we should also highlight the role of SRF in regulating the "cell junction" pathway. Assembly and disassembly of cell-cell junctions comprise a number of key events during physiological and pathological processes. The role of serum receptor factor (SRF) as a potential regulator of this process has been previously established by *in vivo* observation of the deficiencies in mutant backgrounds (Busche et al., 2008). As highlighted in our results (Figure S17), deregulation in the activity of this transcription factor may be crucial for tumor metastasis (*i.e.* benign to malignant transition).

A substantial fraction of *cis*-regulatory elements discovered in this study correspond to the binding sites of known transcription factors. The expression of these transcription factors can be viewed in the context of this dataset or other independent whole-genome datasets. This would enable us to compare the presence or absence of a given transcription factor to the expression of its putative downstream genes. Such analysis can act as a validation step for both the discovered motifs and the functions associated with them through the "pathway-regulatory interaction map". For example, here, we have focused on two transcription factors from the bladder carcinoma dataset: Elk1 and TFDP1 (a protein that binds E2F and promotes its DNA binding affinity; Crosby et al., 2007). Elk1, which was associated with mitosis, RNA splicing and ribosome biogenesis in our study (Figure 2C, main text), shows a significant anti-correlation with these genes, suggesting that it functions as a repressor (Figure S2). Similarly, TFDP1, which is a partner in E2F complexes, is significantly correlated with the expression of the genes in the E2F associated functions, namely DNA replication, microtubule biogenesis and mitotic cell cycle (Figure S3). We have also included the same analysis for AhR and its predicted association with Ub-dependent protein catabolic process (Figure S4).

In addition to GO pathways, we also used the MSigDB c2 gene-sets to re-analyze the bladder cancer dataset in its continuous format (as presented in the main text). The iPAGE results are presented in Figure S18. Up-regulation of many cancer-related genes and cell-cycle pathways along with a down-regulation of immune response pathways (e.g. cytokine pathway, IL and TNF mediated inflammatory responses) match the results presented based on the GO pathways in the main text. We have also included the pathway-regulatory interaction map obtained from these gene-sets (Figure S19).

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

Due to space limitations, for this analysis, we only included a limited number of deregulated pathways in the figures of the main paper. Here, however, we have presented the iPAGE and FIRE outputs with the complete list of the deregulated pathways and informative sequence motifs (Figure S5A and B). Some of the matrices presented here and in the next section are too large to be legibly fit in a page; however, the figures are vector graphics and can be enlarged when viewed electronically.

Similar to the Elk1 and TFDP1 analysis for the bladder carcinoma dataset, in Figure S6, we have included the expression of NF-Y (as the average expression of NFYB and NFYC genes) and SP1 in comparison with the expression of their associated downstream pathways. NF-Y, associated with cell cycle and chromatin biogenesis, is highly correlated with the genes in these pathways (*p*-value < 1e-30). Sp1 shows a similar correlation with cell cycle and secretory pathways (*p*-value < 1e-12).

Building a regulatory map of cancer deregulation

Next, 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 (see Table S2). We then processed the samples and used iPAGE to build a cancer pathway map (Figure S7 and Figure 5 in the main text). This systematic iPAGE analysis of cancer datasets 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, main text). The clustering results and the correlation matrix are shown in Figure S20. The hierarchical clustering tree in Figure S20 shows that independent cancer datasets of similar types are also similar in terms of their informative (and therefore deregulated) pathways. For example, in this figure, 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). Among all the cancers, two samples in particular show a highly negative correlation with others: 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 (Figure 5, main text). 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.

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 S8). Apart from their independent occurrences in multiple datasets, most of these motifs also have high network-level conservation scores (Figure S9). Figure S9 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 S10). In Table S3, 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). In Figure S11, we have used the NCI-60 gene expression panel to test some of these novel associations. For example, we have predicted Lmo2-complex as a direct regulator of PI3K signaling and cell migration. The members of these pathways that harbor an Lmo2 binding site are highly correlated with the expression of this transcription factor (Figure S11). We observed similar correlations between TCF3 (E47 complex) and the inflammatory response pathway and miR-203 and protein degradation pathway (Figure S11).

Here, we have also included the cancer pathway map built from MSigDB c2 gene-sets in Figure S21. The procedure is the same as above, with the exception that we have used MDigDB c2 gene-sets instead of GO terms (biological processes).

The discovered motifs are informative of gene expression patterns across independent datasets

We used the human gene-expression atlas (Su et al., 2004) and NCI-60 gene expression panel (Ross et al., 2000) to further test the functional relevance of our discovered elements. We clustered the genes in these two datasets into 70 clusters using the *k*-means approach, and then used FIRE in non-discovery mode (a mode in which the motif discovery is skipped and all the steps are performed on a set of input motifs) to test whether our 104 identified motifs show a significant non-random pattern across these datasets. As shown in Figure S22 and Figure S23, 74 of these motifs have significant mutual information values with the expression clusters further strengthening the evidence for the role of these elements in regulating gene expression.

FIRE analysis of experimentally tested associations

As shown in Figure S24, 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 S24B). 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 S24C).

Legends

Figure S1. Measuring co-regulation in the downstream genes of novel motifs. (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.

Figure S2. Elk1 expression across the bladder carcinoma samples and its comparison with target genes within the 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. TFDP1 (E2F associated protein) expression and its comparison with E2F associated pathways across the bladder carcinoma samples.

Figure S4. The correlation of transcription factor AhR with Ub-dependent protein catabolic process.

Figure S5. The complete iPAGE (A) and FIRE (B) outputs for the BL vs DLBCL dataset. While Figure 2A and B contain a summarized version of these results, here we have included the complete outputs.

Figure S6. The normalized expression of NF-Y and Sp1 across the BL vs. DLBCL samples along with the expression of their target genes.

Figure S7. The complete cancer pathway map without redundancy removal.

Figure S8. 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.

Figure S9. 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.

Figure S10. The complete cancer pathway-regulatory interaction map.

Figure S11. The expression level of Lmo2, TCF3 and miR-203 modules across the NCI-60 panel.

Figure S12. Conceptual schematic of our computational framework. We start from a gene expression dataset and use iPAGE and FIRE to discover the deregulated pathways and *cis*-regulatory elements that are informative of gene expression patterns. We then use pathway-regulatory interaction map (PRM) analysis to functionally annotate the identified motifs through associating them with their potential downstream pathways.

Figure S13. iPAGE schematic. 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 S14. Pathway-Regulatory Interaction Map Generator. The discovered informative pathways (from iPAGE) and putative *cis*-regulatory elements (from FIRE) are combined in this approach to associate regulatory proteins with their target pathways.

Figure S15. Bladder carcinoma versus normal samples (discrete mode). The informative pathways discovered by iPAGE that are deregulated in tumor samples compared to normal controls. The top panel shows the normalized average expression of each cluster in Bladder carcinoma and normal samples.

Figure S16. Discovering putative *cis*-regulatory elements driving bladder cancer deregulation. We used FIRE to identify putative DNA/RNA motifs whose downstream/upstream genes show significant deregulation in the tumor samples compared to normal controls. The top panel shows the normalized average expression of each cluster in Bladder carcinoma and normal samples.

Figure S17. The pathway-regulatory interaction map of bladder cancer. The discovered *cis*-regulatory elements (Figure S16) are positively or negatively associated with the significantly informative pathways (Figure S15).

Figure S18. iPAGE output for the bladder cancer dataset (continuous) using MSigDB c2 gene-sets.

Figure S19. The pathway-regulatory interaction map of bladder cancer using the MSigDB gene sets. The gene-sets discovered in Figure S18 are associated with *cis*-regulatory elements in Figure 2A (main text).

Figure S20. Correlation matrix calculated from our cancer pathway map. All the cancer samples are clustered based on their deregulations across different pathways. Each element in this matrix represents a pair-wise correlation value. On the right, the tree representing the hierarchical clustering is presented.

Figure S21. The complete cancer pathway map for MSigDB c2 gene-sets (without redundancy removal).

Figure S22. The discovered *cis*-regulatory elements in the cancer regulatory map are informative of gene expression clusters in the NCI-60 dataset.

Figure S23. The discovered *cis*-regulatory elements are informative of gene expression clusters in the human gene expression atlas.

Figure S24. FIRE analysis of experimentally tested associations. (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 number of pathways discovered by iPAGE in 3 datasets studied in the main text in comparison with the number of pathways obtained from the same but randomly shuffled datasets.

Table S2. The list, tissue and references of the cancer gene expression studies used to compile our initial dataset.

Table S3. A list of predictions based on the associations in the Cancer Pathway-Regulatory Interaction Map.

Supplemental Figures

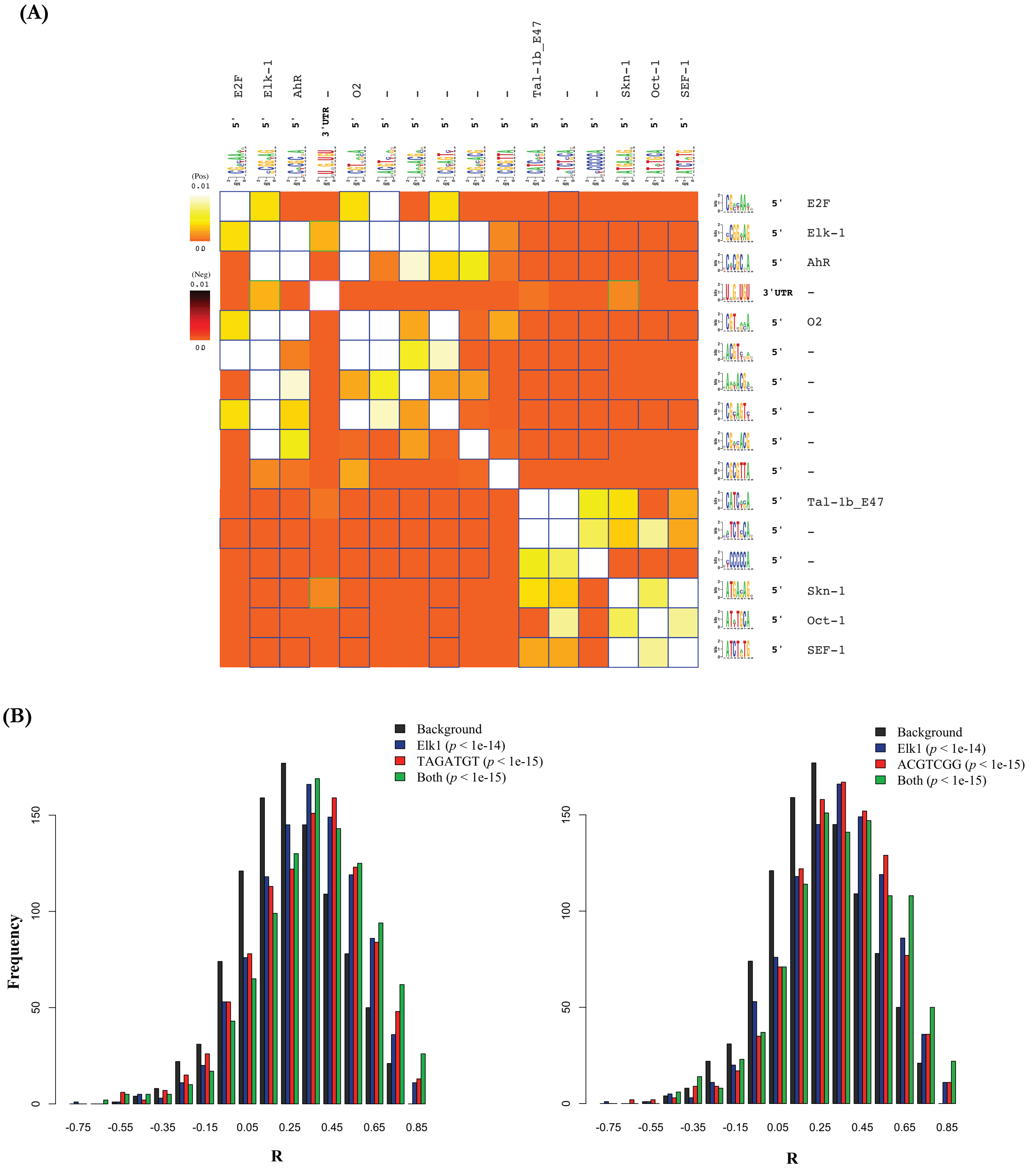


Figure S1

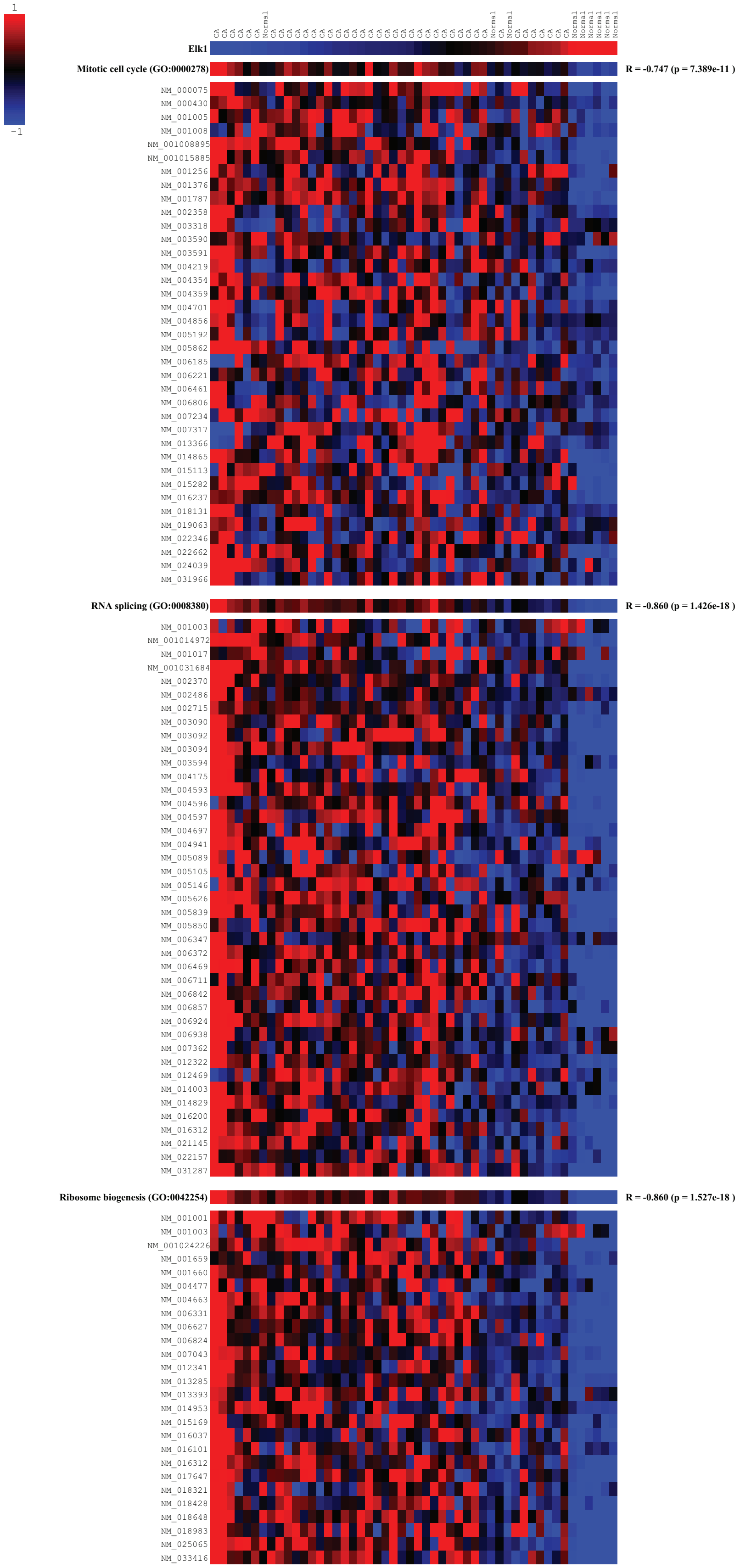

1
-1
CA
CA
CA
CA
CA
CA
Normal
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
Normal
CA
Normal
CA
CA
CA
CA
CA
CA
CA
Normal
Normal
Normal
Normal
Normal
Normal
Elk1
Mitotic cell cycle (GO:0000278)
R = -0.747 (p = 7.389e-11)
NM_000075
NM_000430
NM_001005
NM_001008
NM_001008895
NM_001015885
NM_001256
NM_001376
NM_001787
NM_002358
NM_003318
NM_003590
NM_003591
NM_004219
NM_004354
NM_004359
NM_004701
NM_004856
NM_005192
NM_005862
NM_006185
NM_006221
NM_006461
NM_006806
NM_007234
NM_007317
NM_013366
NM_014865
NM_015113
NM_015282
NM_016237
NM_018131
NM_019063
NM_022346
NM_022662
NM_024039
NM_031966
RNA splicing (GO:0008380)
R = -0.860 (p = 1.426e-18)
NM_001003
NM_001014972
NM_001017
NM_001031684
NM_002370
NM_002486
NM_002715
NM_003090
NM_003092
NM_003094
NM_003594
NM_004175
NM_004593
NM_004596
NM_004597
NM_004697
NM_004941
NM_005089
NM_005105
NM_005146
NM_005626
NM_005839
NM_005850
NM_006347
NM_006372
NM_006469
NM_006711
NM_006842
NM_006857
NM_006924
NM_006938
NM_007362
NM_012322
NM_012469
NM_014003
NM_014829
NM_016200
NM_016312
NM_021145
NM_022157
NM_031287
Ribosome biogenesis (GO:0042254)
R = -0.860 (p = 1.527e-18)
NM_001001
NM_001003
NM_001024226
NM_001659
NM_001660
NM_004477
NM_004663
NM_006331
NM_006627
NM_006824
NM_007043
NM_012341
NM_013285
NM_013393
NM_014953
NM_015169
NM_016037
NM_016101
NM_016312
NM_017647
NM_018321
NM_018428
NM_018648
NM_018983
NM_025065
NM_033416

Figure S2

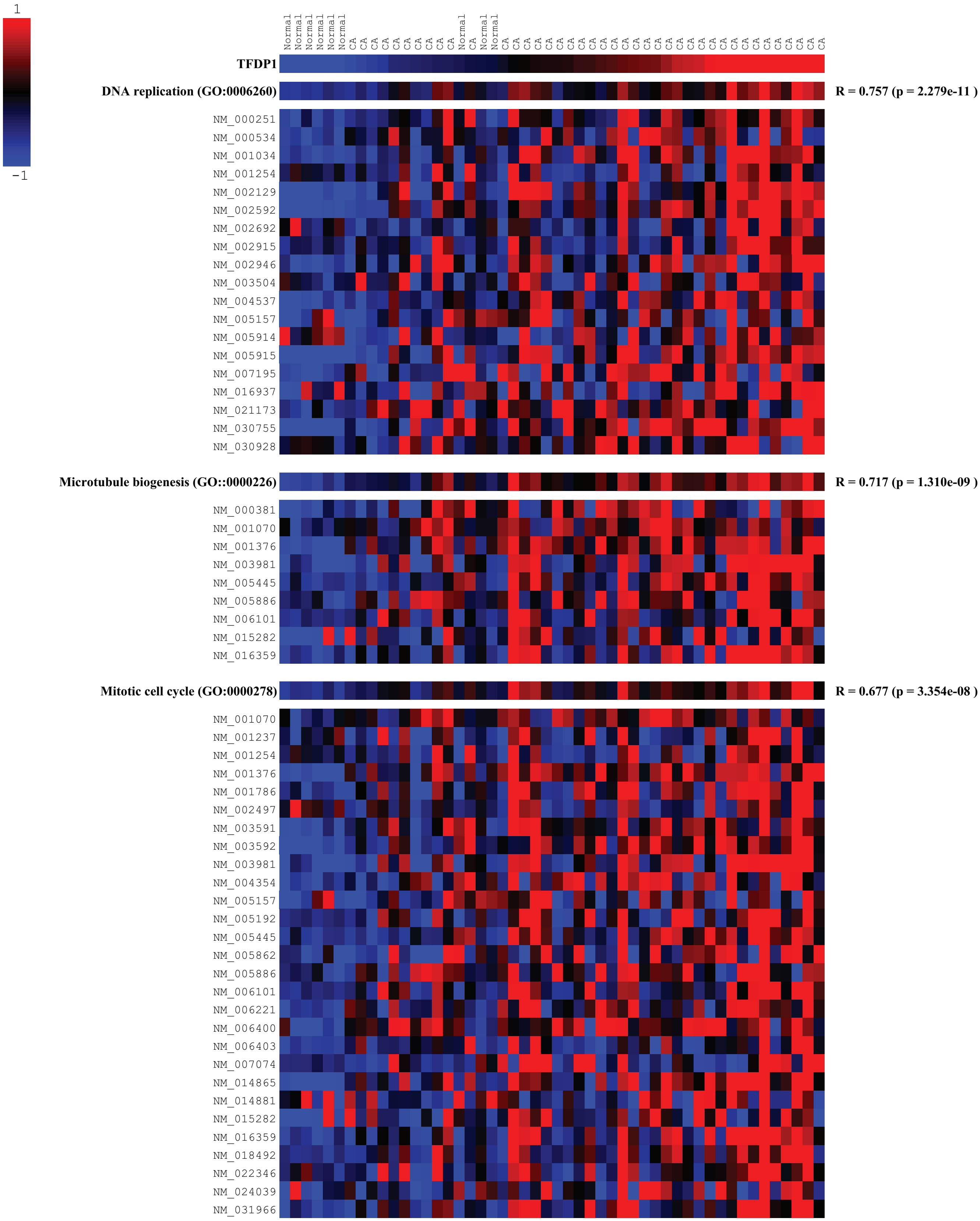

1
-1
Normal
Normal
Normal
Normal
Normal
Normal
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
Normal
CA
Normal
Normal
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
TFDP1
DNA replication (GO:0006260)
R = 0.757 (p = 2.279e-11)
NM_000251
NM_000534
NM_001034
NM_001254
NM_002129
NM_002592
NM_002692
NM_002915
NM_002946
NM_003504
NM_004537
NM_005157
NM_005914
NM_005915
NM_007195
NM_016937
NM_021173
NM_030755
NM_030928
Microtubule biogenesis (GO::0000226)
R = 0.717 (p = 1.310e-09)
NM_000381
NM_001070
NM_001376
NM_003981
NM_005445
NM_005886
NM_006101
NM_015282
NM_016359
Mitotic cell cycle (GO:0000278)
R = 0.677 (p = 3.354e-08)
NM_001070
NM_001237
NM_001254
NM_001376
NM_001786
NM_002497
NM_003591
NM_003592
NM_003981
NM_004354
NM_005157
NM_005192
NM_005445
NM_005862
NM_005886
NM_006101
NM_006221
NM_006400
NM_006403
NM_007074
NM_014865
NM_014881
NM_015282
NM_016359
NM_018492
NM_022346
NM_024039
NM_031966

Figure S3

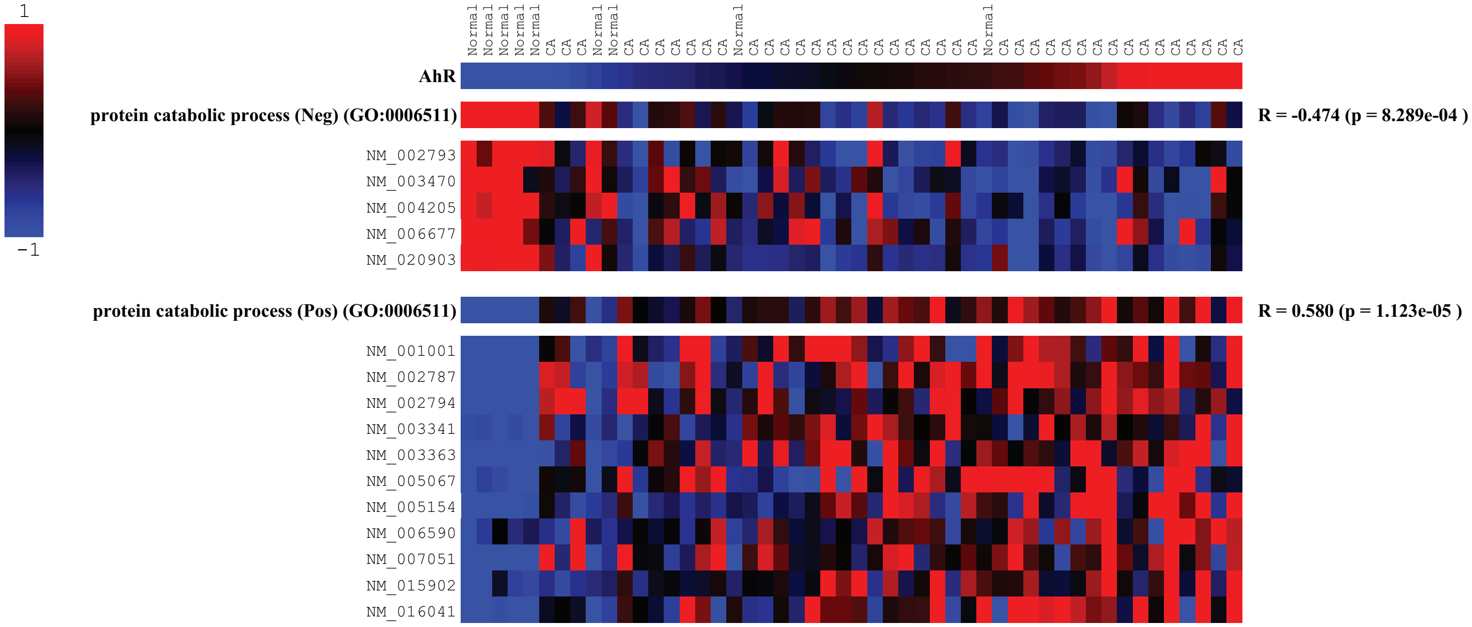

1
-1
Normal
Normal
Normal
Normal
Normal
CA
CA
CA
Normal
Normal
CA
CA
CA
CA
CA
CA
CA
Normal
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
Normal
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
CA
AhR
protein catabolic process (Neg) (GO:0006511)
R = -0.474 (p = 8.289e-04)
NM_002793
NM_003470
NM_004205
NM_006677
NM_020903
protein catabolic process (Pos) (GO:0006511)
R = 0.580 (p = 1.123e-05)
NM_001001
NM_002787
NM_002794
NM_003341
NM_003363
NM_005067
NM_005154
NM_006590
NM_007051
NM_015902
NM_016041

Figure S4

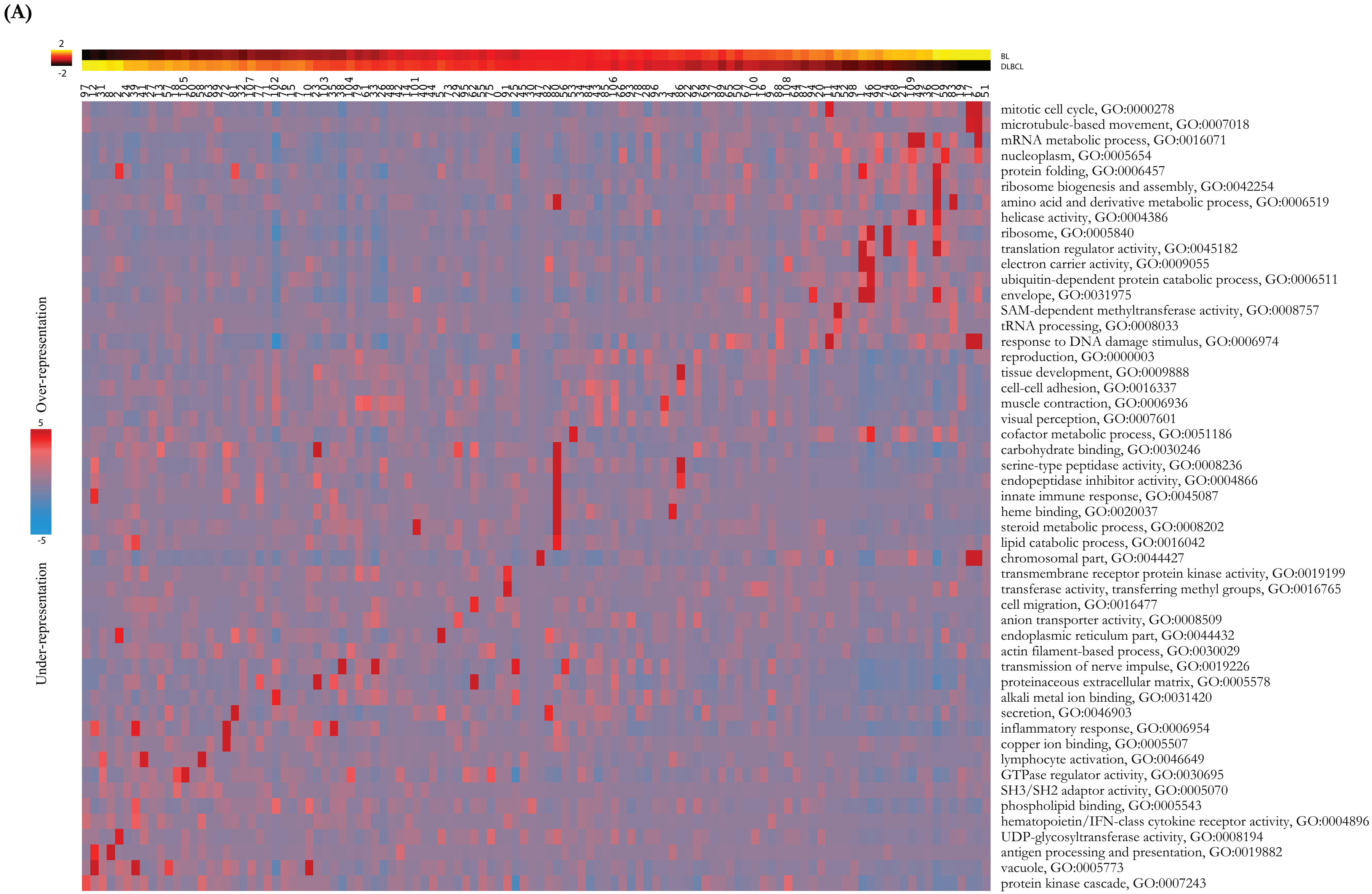

(A)
2
-2
BL
DLBCL
Over-representation
Under-representation
5
-5
mitotic cell cycle, GO:0000278
microtubule-based movement, GO:0007018
mRNA metabolic process, GO:0016071
nucleoplasm, GO:0005654
protein folding, GO:0006457
ribosome biogenesis and assembly, GO:0042254
amino acid and derivative metabolic process, GO:0006519
helicase activity, GO:0004386
ribosome, GO:0005840
translation regulator activity, GO:0045182
electron carrier activity, GO:0009055
ubiquitin-dependent protein catabolic process, GO:0006511
envelope, GO:0031975
SAM-dependent methyltransferase activity, GO:0008757
tRNA processing, GO:0008033
response to DNA damage stimulus, GO:0006974
reproduction, GO:0000003
tissue development, GO:0009888
cell-cell adhesion, GO:0016337
muscle contraction, GO:0006936
visual perception, GO:0007601
cofactor metabolic process, GO:0051186
carbohydrate binding, GO:0030246
serine-type peptidase activity, GO:0008236
endopeptidase inhibitor activity, GO:0004866
innate immune response, GO:0045087
heme binding, GO:0020037
steroid metabolic process, GO:0008202
lipid catabolic process, GO:0016042
chromosomal part, GO:0044427
transmembrane receptor protein kinase activity, GO:0019199
transferase activity, transferring methyl groups, GO:0016765
cell migration, GO:0016477
anion transporter activity, GO:0008509
endoplasmic reticulum part, GO:0044432
actin filament-based process, GO:0030029
transmission of nerve impulse, GO:0019226
proteinaceous extracellular matrix, GO:0005578
alkali metal ion binding, GO:0031420
secretion, GO:0046903
inflammatory response, GO:0006954
copper ion binding, GO:0005507
lymphocyte activation, GO:0046649
GTPase regulator activity, GO:0030695
SH3/SH2 adaptor activity, GO:0005070
phospholipid binding, GO:0005543
hematopoietin/IFN-class cytokine receptor activity, GO:0004896
UDP-glycosyltransferase activity, GO:0008194
antigen processing and presentation, GO:0019882
vacuole, GO:0005773
protein kinase cascade, GO:0007243

Figure S5A

(B)

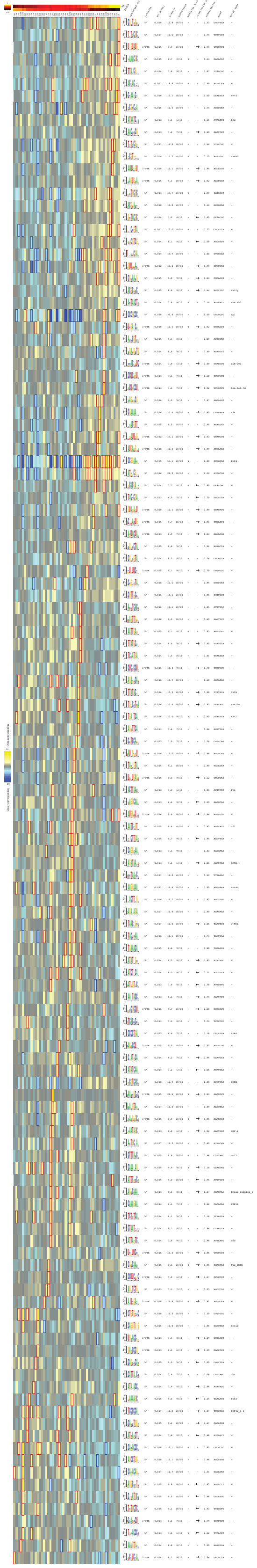


Figure S5B

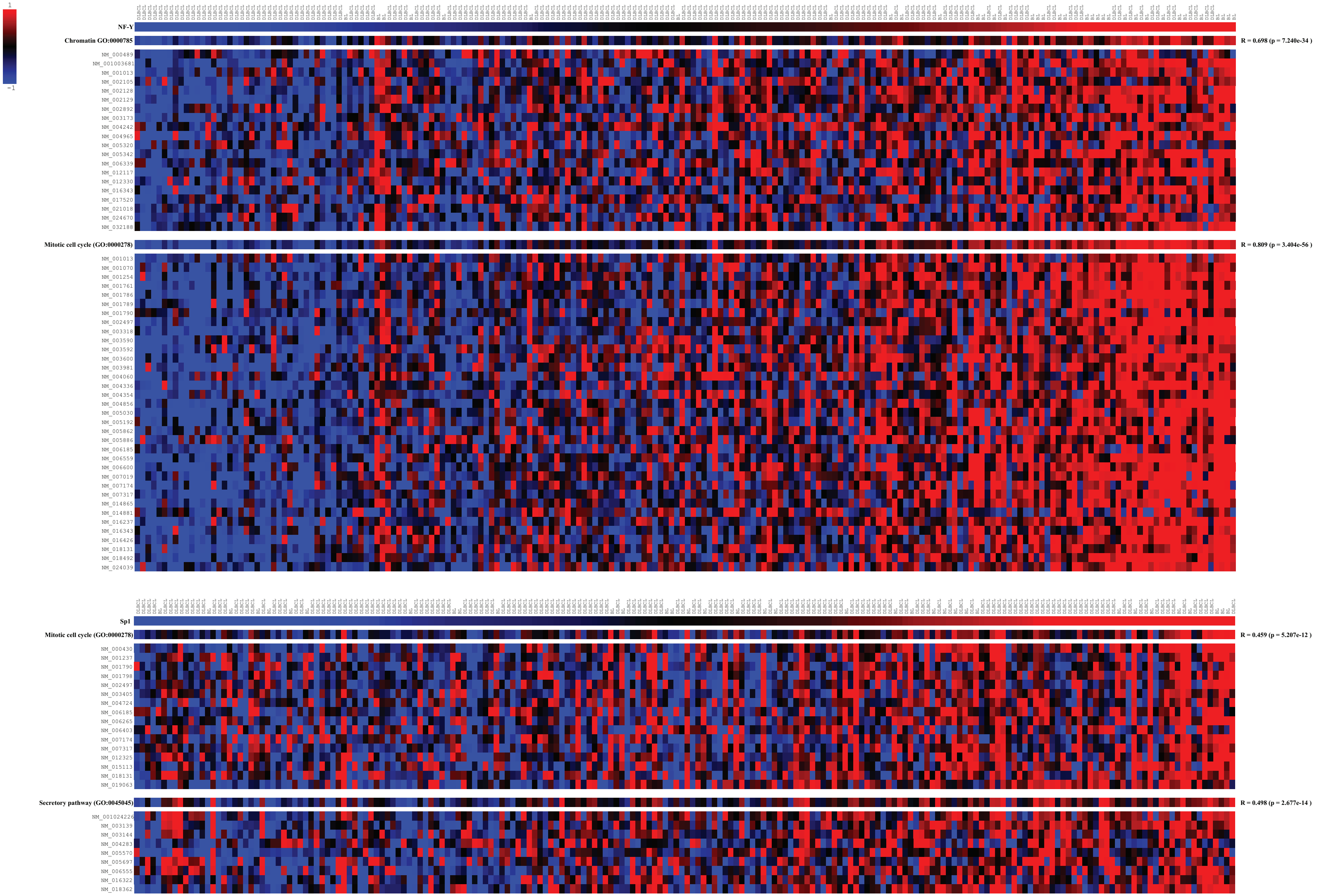

1
−1
NF-Y
Chromatin GO:0000785
R = 0.698 (p = 7.240e-34)
NM_000489
NM_001003681
NM_001013
NM_002105
NM_002128
NM_002129
NM_002892
NM_003173
NM_004242
NM_004965
NM_005320
NM_005342
NM_006339
NM_012117
NM_012330
NM_016343
NM_017520
NM_021018
NM_024670
NM_032188
Mitotic cell cycle (GO:0000278)
R = 0.809 (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 S6

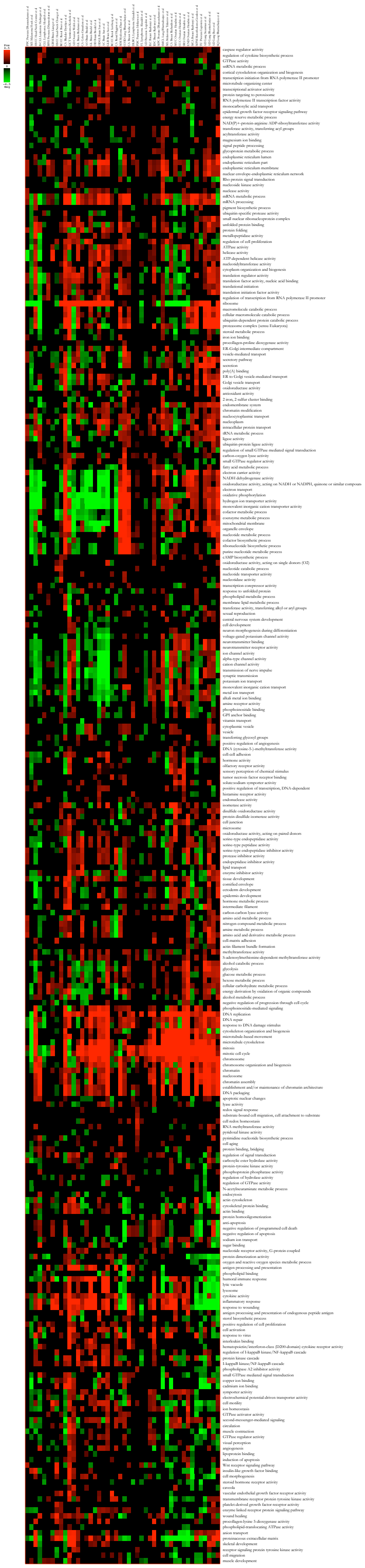

Pos
4.5
0
-4.5
Neg
PPC Prostate Dhanasekaran et al
ME Melanoma Hoek et al
HSCC Head-Neck Cromer et al
B-CLL Leukemia Haslinger et al
CLL Lymphoma Alizadeh et al
BPH Prostate Dhanasekaran et al
GBM Brain Liang et al
HSCC Head-Neck Chung et al
RCCC Renal Boer et al
CA Bladder Dyrskjot et al
GCT Seminoma Korkola et al
AD Ovarian Welsh et al
GL Brain Rickman et al
CA Colon Graudens et al
AO Brain Bredel et al
GL Brain Bredel et al
OD Brain Bredel et al
ODGL Brain Sun et al
AC Brain Sun et al
GLB Brain Sun et al
RCCC Renal Lenburg et al
CA Renal Higgins et al
MM Myeloma Zhan et al
COID Lung Bhattacharjee et al
CA Breast Sotiriou et al
DLBCL Lymphoma Alizadeh et al
PDC Pancreas Ishikawa et al
FL Lymphoma Alizadeh et al
AD Pancreas Logsdon et al
ILC Breast Radvanyi et al
IDC Breast Radvanyi et al
MPC Prostate Dhanasekaran et al
SMCL Lung Bhattacharjee et al
ML Melanoma Talantov et al
CA Breast Richardson et al
MUC Ovarian Hendrix et al
CCC Ovarian Hendrix et al
SRS Ovarian Hendrix et al
END Ovarian Hendrix et al
MCA Breast Radvanyi et al
MPM Mesothelioma Gordon et al
TU Prostate Lapointe et al
AD Lung Stearman et al
AD Lung Bhattacharjee et al
AD Lung Beer et al
SQ Lung Bhattacharjee et al
caspase regulator activity
regulation of cytokine biosynthetic process
GTPase activity
snRNA metabolic process
cortical cytoskeleton organization and biogenesis
transcription initiation from RNA polymerase II promoter
microtubule organizing center
transcriptional activator activity
protein targeting to peroxisome
RNA polymerase II transcription factor activity
monocarboxylic acid transport
epidermal growth factor receptor signaling pathway
energy reserve metabolic process
NAD(P)+-protein-arginine ADP-ribosyltransferase activity
transferase activity, transferring acyl groups
acyltransferase activity
magnesium ion binding
signal peptide processing
glycoprotein metabolic process
endoplasmic reticulum lumen
endoplasmic reticulum part
endoplasmic reticulum membrane
nuclear envelope-endoplasmic reticulum network
Rho protein signal transduction
nucleoside kinase activity
nuclease activity
mRNA metabolic process
mRNA processing
pigment biosynthetic process
ubiquitin-specific protease activity
small nuclear ribonucleoprotein complex
unfolded protein binding
protein folding
metallopeptidase activity
regulation of cell proliferation
ATPase activity
helicase activity
ATP-dependent helicase activity
nucleotidyltransferase activity
cytoplasm organization and biogenesis
translation regulator activity
translation factor activity, nucleic acid binding
translational initiation
translation initiation factor activity
regulation of transcription from RNA polymerase II promoter
ribosome
macromolecule catabolic process
cellular macromolecule catabolic process
ubiquitin-dependent protein catabolic process
proteasome complex (sensu Eukaryota)
steroid metabolic process
iron ion binding
procollagen-proline dioxygenase activity
ER-Golgi intermediate compartment
vesicle-mediated transport
secretory pathway
secretion
poly(A) binding
ER to Golgi vesicle-mediated transport
Golgi vesicle transport
oxidoreductase activity
antioxidant activity
2 iron, 2 sulfur cluster binding
endomembrane system
chromatin modification
nucleocytoplasmic transport
nucleoplasm
intracellular protein transport
tRNA metabolic process
ligase activity
ubiquitin-protein ligase activity
regulation of small GTPase mediated signal transduction
carbon-oxygen lyase activity
small GTPase regulator activity
fatty acid metabolic process
electron carrier activity
NADH dehydrogenase activity
oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound
electron transport
oxidative phosphorylation
hydrogen ion transporter activity
monovalent inorganic cation transporter activity
cofactor metabolic process
coenzyme metabolic process
mitochondrial membrane
organelle envelope
nucleotide metabolic process
cofactor biosynthetic process
ribonucleotide biosynthetic process
purine nucleotide metabolic process
cAMP biosynthetic process
oxidoreductase activity, acting on single donors (O2)
nucleotide catabolic process
nucleotide transporter activity
nucleotidase activity
transcription corepressor activity
response to unfolded protein
phospholipid metabolic process
membrane lipid metabolic process
transferase activity, transferring alkyl or aryl groups
sexual reproduction
central nervous system development
cell development
neuron morphogenesis during differentiation
voltage-gated potassium channel activity
neurotransmitter binding
neurotransmitter receptor activity
ion channel activity
alpha-type channel activity
cation channel activity
transmission of nerve impulse
synaptic transmission
potassium ion transport
monovalent inorganic cation transport
metal ion transport
alkali metal ion binding
amine receptor activity
phosphoinositide binding
GPI anchor binding
vitamin transport
cytoplasmic vesicle
vesicle
transferring glycosyl groups
positive regulation of angiogenesis
DNA (cytosine-5-)-methyltransferase activity
cell-cell adhesion
hormone activity
olfactory receptor activity
sensory perception of chemical stimulus
tumor necrosis factor receptor binding
solute:sodium symporter activity
positive regulation of transcription, DNA-dependent
histamine receptor activity
endonuclease activity
isomerase activity
disulfide oxidoreductase activity
protein disulfide isomerase activity
cell junction
microsome
oxidoreductase activity, acting on paired donors
serine-type endopeptidase activity
serine-type peptidase activity
serine-type endopeptidase inhibitor activity
protease inhibitor activity
endopeptidase inhibitor activity
lipid transport
enzyme inhibitor activity
tissue development
cornified envelope
ectoderm development
epidermis development
hormone metabolic process
intermediate filament
carbon-carbon lyase activity
amino acid metabolic process
nitrogen compound metabolic process
amine metabolic process
amino acid and derivative metabolic process
cell-matrix adhesion
actin filament bundle formation
methyltransferase activity
S-adenosylmethionine-dependent methyltransferase activity
alcohol catabolic process
glycolysis
glucose metabolic process
hexose metabolic process
cellular carbohydrate metabolic process
energy derivation by oxidation of organic compounds
alcohol metabolic process
negative regulation of progression through cell cycle
phosphoinositide-mediated signaling
DNA replication
DNA repair
response to DNA damage stimulus
cytoskeleton organization and biogenesis
microtubule-based movement
microtubule cytoskeleton
mitosis
mitotic cell cycle
chromosome
chromosome organization and biogenesis
chromatin
nucleosome
chromatin assembly
establishment and/or maintenance of chromatin architecture
DNA packaging
apoptotic nuclear changes
lyase activity
redox signal response
substrate-bound cell migration, cell attachment to substrate
cell redox homeostasis
RNA methyltransferase activity
pyridoxal kinase activity
pyrimidine nucleotide biosynthetic process
cell aging
protein binding, bridging
regulation of signal transduction
carboxylic ester hydrolase activity
protein-tyrosine kinase activity
phosphoprotein phosphatase activity
regulation of hydrolase activity
regulation of GTPase activity
N-acetylneuraminate metabolic process
endocytosis
actin cytoskeleton
cytoskeletal protein binding
actin binding
protein homooligomerization
anti-apoptosis
negative regulation of programmed cell death
negative regulation of apoptosis
sodium ion transport
sugar binding
nucleotide receptor activity, G-protein coupled
protein dimerization activity
oxygen and reactive oxygen species metabolic process
antigen processing and presentation
phospholipid binding
humoral immune response
lytic vacuole
lysosome
cytokine activity
inflammatory response
response to wounding
antigen processing and presentation of endogenous peptide antigen
sterol biosynthetic process
positive regulation of cell proliferation
cell activation
response to virus
interleukin binding
hematopoietin/interferon-class (D200-domain) cytokine receptor activity
regulation of I-kappaB kinase/NF-kappaB cascade
protein kinase cascade
I-kappaB kinase/NF-kappaB cascade
phospholipase A2 inhibitor activity
small GTPase mediated signal transduction
copper ion binding
cadmium ion binding
symporter activity
electrochemical potential-driven transporter activity
cell motility
ion homeostasis
GTPase activator activity
second-messenger-mediated signaling
circulation
muscle contraction
GTPase regulator activity
visual perception
angiogenesis
lipoprotein binding
induction of apoptosis
Wnt receptor signaling pathway
insulin-like growth factor binding
cell morphogenesis
steroid hormone receptor activity
caveola
vascular endothelial growth factor receptor activity
transmembrane receptor protein tyrosine kinase activity
platelet-derived growth factor receptor activity
enzyme linked receptor protein signaling pathway
wound healing
procollagen-lysine 5-dioxygenase activity
phospholipid-translocating ATPase activity
anion transport
proteinaceous extracellular matrix
skeletal development
receptor signaling protein tyrosine kinase activity
cell migration
muscle development

Figure S7

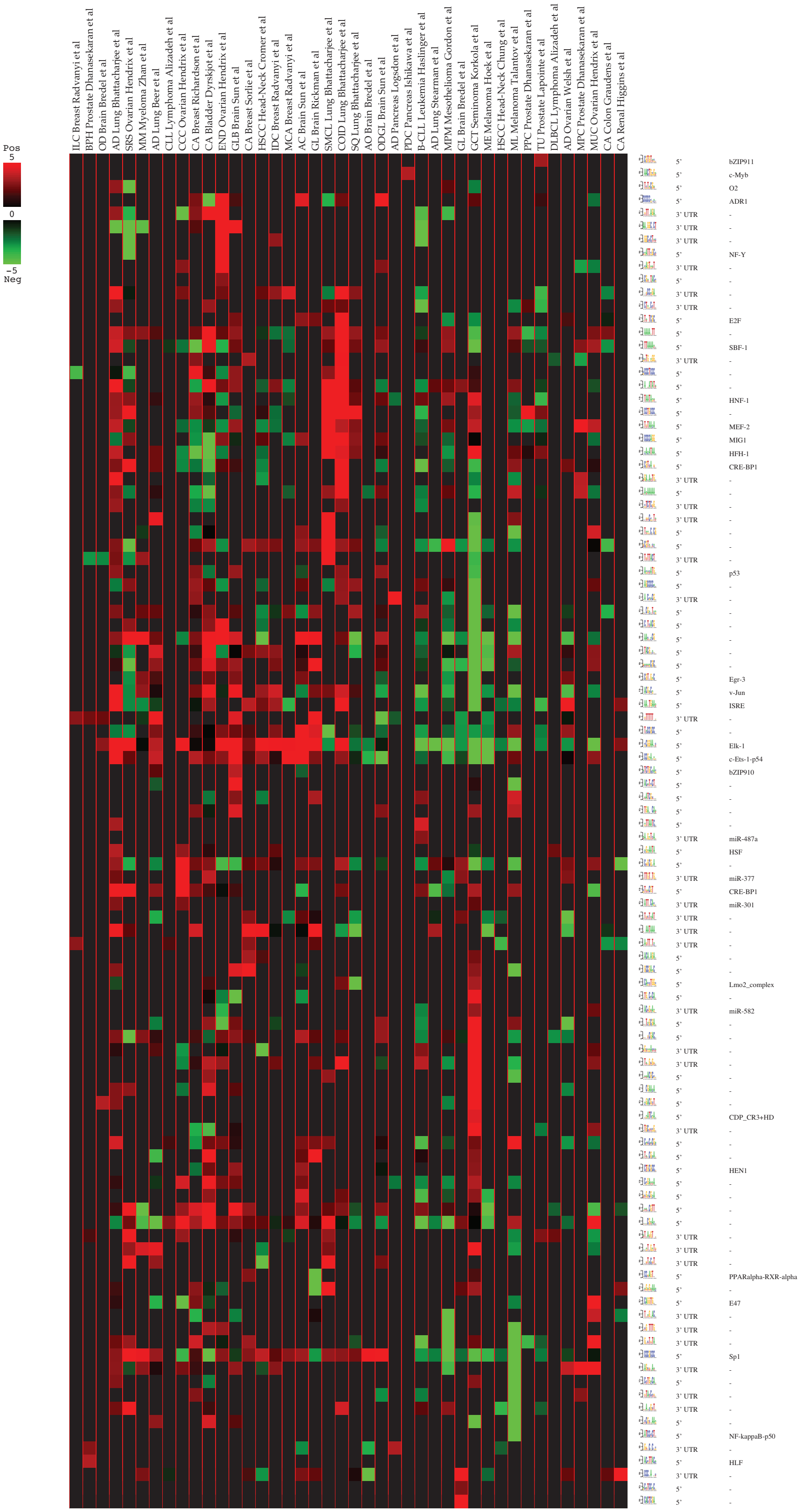
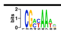
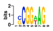
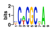
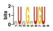
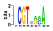
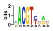
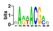
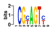
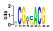
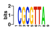
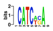
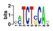
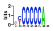
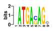
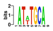
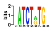
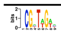
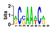
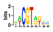
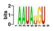
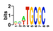
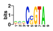
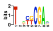
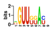
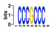
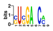
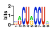
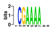
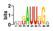
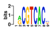
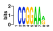
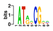
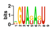
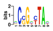
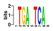
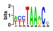
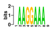
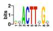
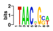
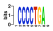
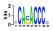
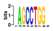

Pos
5
0
-5
Neg
ILC Breast Radvanyi et al
BPH Prostate Dhanasekaran et al
OD Brain Bredel et al
AD Lung Bhattacharjee et al
SRS Ovarian Hendrix et al
MM Myeloma Zhan et al
AD Lung Beer et al
CLL Lymphoma Alizadeh et al
CCC Ovarian Hendrix et al
CA Breast Richardson et al
CA Bladder Dyrskjot et al
END Ovarian Hendrix et al
GLB Brain Sun et al
CA Breast Sorlie et al
HSCC Head-Neck Cromer et al
IDC Breast Radvanyi et al
MCA Breast Radvanyi et al
AC Brain Sun et al
GL Brain Rickman et al
SMCL Lung Bhattacharjee et al
COID Lung Bhattacharjee et al
SQ Lung Bhattacharjee et al
AO Brain Bredel et al
ODGL Brain Sun et al
AD Pancreas Logsdon et al
PDC Pancreas Ishikawa et al
B-CLL Leukemia Haslinger et al
AD Lung Stearman et al
MPM Mesothelioma Gordon et al
GL Brain Bredel et al
GCT Seminoma Korkola et al
ME Melanoma Hoek et al
HSCC Head-Neck Chung et al
ML Melanoma Talantov et al
PPC Prostate Dhanasekaran et al
TU Prostate Lapointe et al
DLBCL Lymphoma Alizadeh et al
AD Ovarian Welsh et al
MPC Prostate Dhanasekaran et al
MUC Ovarian Hendrix et al
CA Colon Graudens et al
CA Renal Higgins et al
5’ bZIP911
5’ c-Myb
5’ O2
5’ ADR1
3’ UTR -
3’ UTR -
3’ UTR -
5’ NF-Y
3’ UTR -
5’ -
3’ UTR -
3’ UTR -
5’ E2F
5’ -
5’ SBF-1
3’ UTR -
5’ -
5’ -
5’ HNF-1
5’ -
5’ MEF-2
5’ MIG1
5’ HFH-1
5’ CRE-BP1
3’ UTR -
5’ -
3’ UTR -
3’ UTR -
5’ -
5’ -
3’ UTR -
5’ p53
5’ -
3’ UTR -
5’ -
5’ -
5’ -
5’ -
5’ -
5’ Egr-3
5’ v-Jun
5’ ISRE
3’ UTR -
5’ -
5’ Elk-1
5’ c-Ets-1-p54
5’ bZIP910
5’ -
5’ -
5’ -
5’ -
3’ UTR miR-487a
5’ HSF
5’ -
3’ UTR miR-377
5’ CRE-BP1
3’ UTR miR-301
3’ UTR -
3’ UTR -
3’ UTR -
5’ -
5’ -
5’ Lmo2_complex
5’ -
3’ UTR miR-582
5’ -
5’ -
3’ UTR -
3’ UTR -
5’ -
5’ -
5’ -
5’ CDP_CR3+HD
3’ UTR -
5’ -
5’ -
5’ HEN1
5’ -
5’ -
5’ -
5’ -
3’ UTR -
3’ UTR -
3’ UTR -
5’ PPARalpha-RXR-alpha
5’ -
5’ E47
3’ UTR -
3’ UTR -
3’ UTR -
5’ Sp1
3’ UTR -
5’ -
3’ UTR -
3’ UTR -
5’ -
5’ NF-kappaB-p50
3’ UTR -
5’ HLF
3’ UTR -
5’ -
5’ -

Figure S8

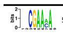
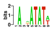
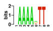
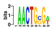
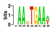
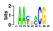
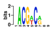
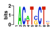
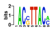
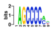
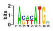
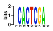
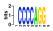
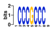
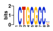
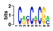
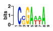
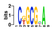
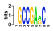
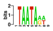
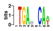
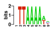
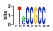
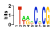
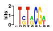
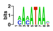
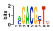
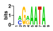
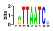
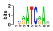
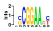
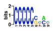
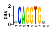
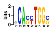
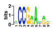
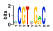
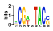
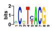
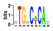
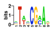
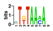
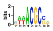
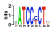
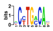
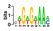
Bladder Carcinoma

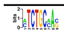
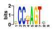
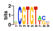
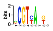
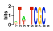
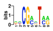
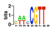
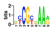
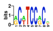
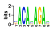
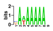
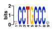
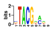
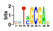
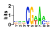
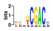
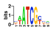
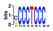
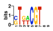
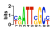
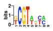
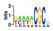
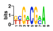
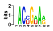
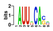
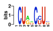
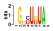
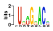
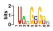
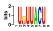
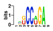
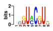
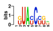
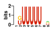
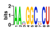
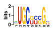
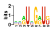
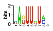
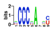
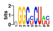
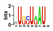
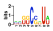
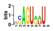
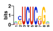
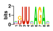
Motif		Conservation Score
	5′	0.86
	5′	1
	5′	0.93
	3′	0.96
	5′	0.98
	5′	0.52
	5′	0.94
	5′	0.55
	5′	0.96
	5′	0.23
	5′	0.96
	5′	0.56
	5′	0.86
	5′	0.57
	5′	0.97
	5′	0.68

BL vs. DLBCL

Motif		Conservation Score
	5′	0.33
	5′	1
	5′	0.7
	3′	0.94
	5′	0.99
	5′	0.72
	5′	0.84
	3′	0.99
	5′	1
	3′	0.92
	3′	0.92
	5′	0.65
	3′	0.93
	5′	1
	5′	1
	5′	1
	3′	0.99
	5′	0.95
	5′	0.4
	5′	0.99
	5′	0.55
	5′	0.97
	5′	0.46
	5′	0.16
	5′	0.92
	5′	0.96

Cancer Regulatory Motifs

Motif		Conservation Score
	5′	0.96
	5′	1.00
	5′	0.98
	5′	0.14
	5′	0.96
	5′	0.96
	5′	0.99
	5′	0.97
	5′	0.45
	5′	0.93
	5′	0.88
	5′	0.15
	5′	0.95
	5′	1.00
	5′	0.94
	5′	0.80
	5′	0.80
	5′	0.99
	5′	0.92
	5′	1.00
	5′	0.97
	5′	1.00
	5′	1.00
	5′	0.38
	5′	0.38
	5′	1.00
	5′	0.58
	5′	0.72
	5′	0.95
	5′	0.98
	5′	1.00
	5′	0.98
	5′	0.82
	5′	0.86
	5′	1.00
	5′	0.87
	5′	0.11
	5′	0.94
	5′	0.77
	5′	0.83
	5′	1.00
	5′	0.92
	5′	0.73
	5′	0.42
	5′	0.98

Motif		Conservation Score
	5′	0.88
	5′	0.44
	5′	0.22
	5′	0.99
	5′	0.94
	5′	0.99
	5′	0.74
	5′	0.75
	5′	0.88
	5′	0.72
	5′	0.97
	5′	0.92
	5′	0.94
	5′	1.00
	5′	0.98
	5′	0.96
	5′	0.87
	5′	0.95
	5′	0.93
	5′	1.00
	5′	0.43
	5′	0.99
	5′	0.84
	3′	0.98
	3′	0.97
	3′	0.90
	3′	0.99
	3′	0.99
	3′	0.98
	3′	0.98
	3′	0.80
	3′	0.96
	3′	0.99
	3′	1.00
	3′	0.85
	3′	0.35
	3′	0.94
	3′	1.00
	3′	0.99
	3′	0.72
	3′	1.00
	3′	0.96
	3′	0.98
	3′	0.44
	3′	0.94

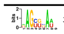
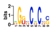
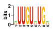
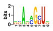
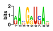
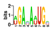
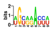
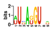
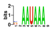
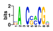
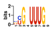
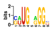
Motif		Conservation Score
	3′	0.95
	3′	0.99
	3′	0.92
	3′	0.97
	3′	0.82
	3′	0.74
	3′	0.99
	3′	0.99
	3′	1.00
	3′	0.46
	3′	0.96
	3′	0.79

Figure S9

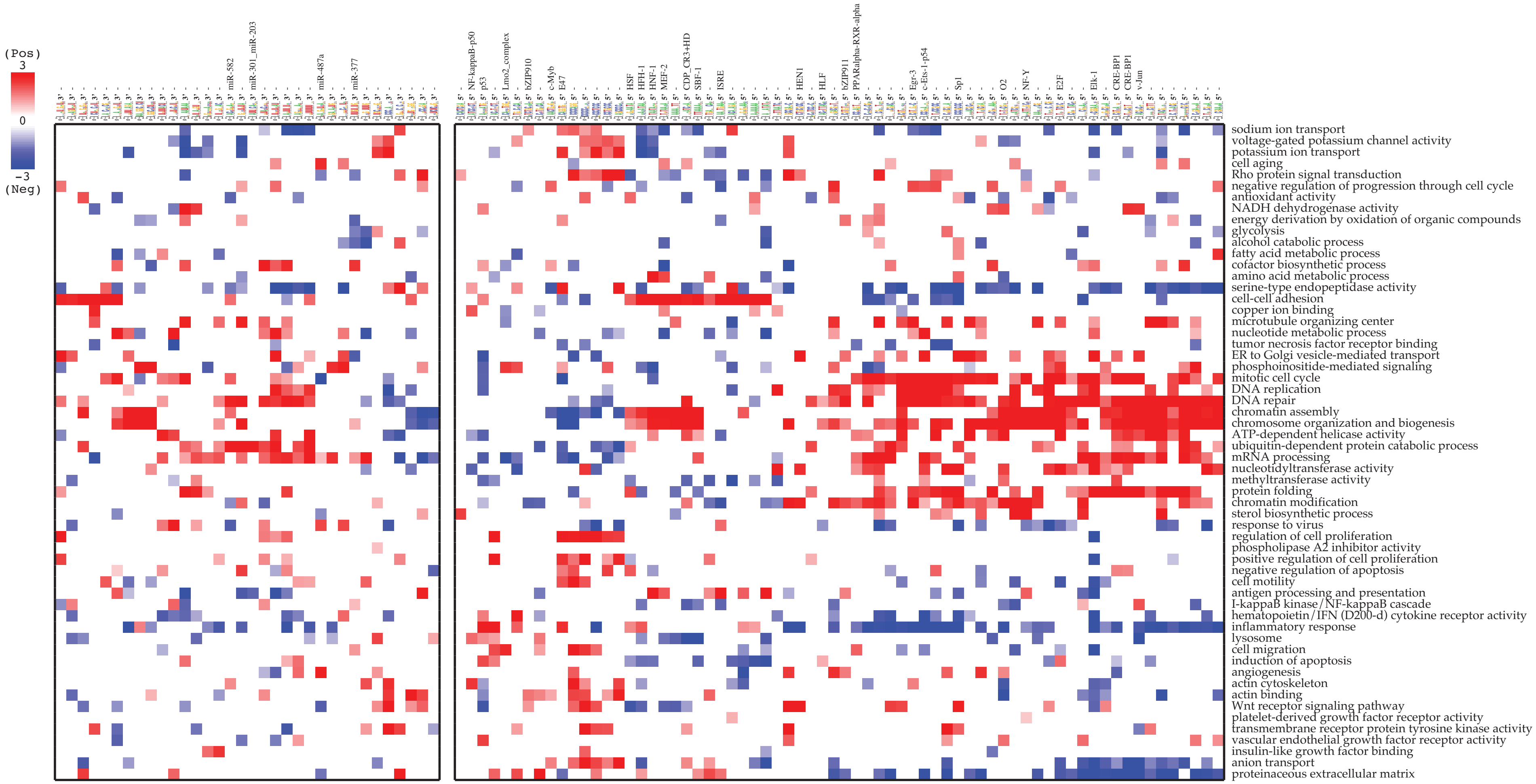

(Pos)
3
0
-3
(Neg)
miR-582
miR-301_miR-203
miR-487a
miR-377
NF-kappaB-p50
p53
Lmo2_complex
bZIP910
c-Myb
E47
HSF
HFH-1
HNF-1
MEF-2
CDP_CR3+HD
SBF-1
ISRE
HEN1
HLF
bZIP911
PPARalpha-RXR-alpha
Egr-3
c-Ets-1-p54
Sp1
O2
NF-Y
E2F
Elk-1
CRE-BP1
CRE-BP1
v-Jun
sodium ion transport
voltage-gated potassium channel 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 S10

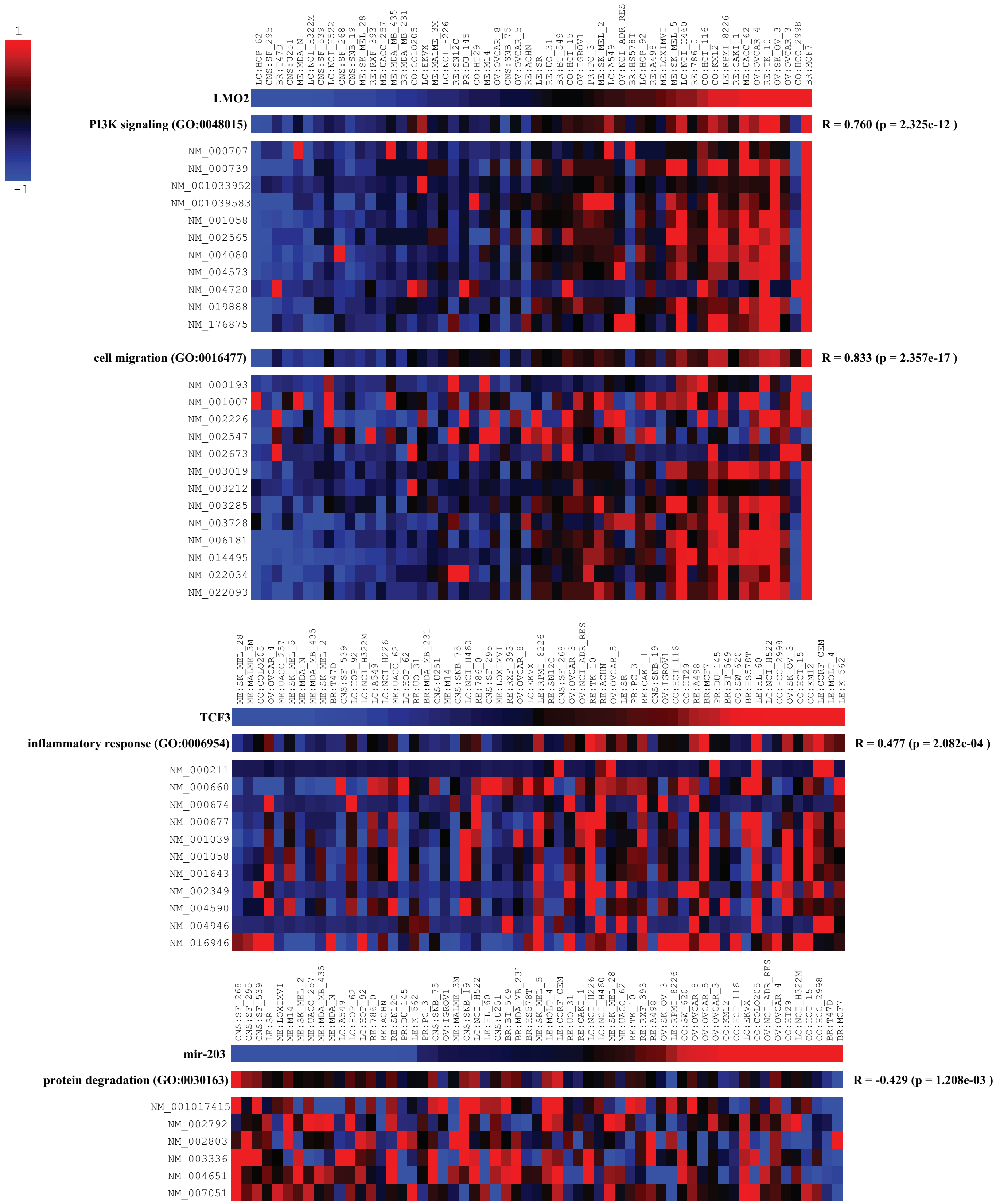

1
-1
LC:HOP_62
CNS:SF_295
BR:T47D
CNS:U251
ME:MDA_N
LC:NCI_H322M
CNS:SF_539
LC:NCI_H522
CNS:SF_268
CNS:SNB_19
ME:SK_MEL_28
RE:RXF_393
ME:UACC_257
ME:MDA_MB_435
BR:MDA_MB_231
CO:COLO205
LC:EKVX
ME:MALME_3M
LC:NCI_H226
RE:SN12C
PR:DU_145
CO:HT29
ME:M14
OV:OVCAR_8
CNS:SNB_75
OV:OVCAR_5
RE:ACHN
LE:SR
RE:UO_31
BR:BT_549
CO:HCT_15
OV:IGROV1
PR:PC_3
ME:SK_MEL_2
LC:A549
OV:NCI_ADR_RES
BR:HS578T
LC:HOP_92
RE:A498
ME:LOXIMVI
ME:SK_MEL_5
LC:NCI_H460
RE:786_0
CO:HCT_116
CO:KM12
LE:RPMI_8226
RE:CAKI_1
ME:UACC_62
OV:OVCAR_4
RE:TK_10
OV:SK_OV_3
OV:OVCAR_3
CO:HCC_2998
BR:MCF7
LMO2
PI3K signaling (GO:0048015)
R = 0.760 (p = 2.325e-12)
NM_000707
NM_000739
NM_001033952
NM_001039583
NM_001058
NM_002565
NM_004080
NM_004573
NM_004720
NM_019888
NM_176875
cell migration (GO:0016477)
R = 0.833 (p = 2.357e-17)
NM_000193
NM_001007
NM_002226
NM_002547
NM_002673
NM_003019
NM_003212
NM_003285
NM_003728
NM_006181
NM_014495
NM_022034
NM_022093
ME:SK_MEL_28
ME:MALME_3M
CO:COLO205
OV:OVCAR_4
ME:UACC_257
ME:SK_MEL_5
ME:MDA_N
ME:MDA_MB_435
ME:SK_MEL_2
BR:T47D
CNS:SF_539
LC:HOP_92
LC:NCI_H322M
LC:A549
LC:NCI_H226
ME:UACC_62
LC:HOP_62
RE:UO_31
BR:MDA_MB_231
CNS:U251
ME:M14
CNS:SNB_75
LC:NCI_H460
RE:786_0
CNS:SF_295
ME:LOXIMVI
RE:RXF_393
OV:OVCAR_8
LC:EKVX
LE:RPMI_8226
RE:SN12C
CNS:SF_268
OV:OVCAR_3
OV:NCI_ADR_RES
RE:TK_10
RE:ACHN
OV:OVCAR_5
LE:SR
PR:PC_3
RE:CAKI_1
CNS:SNB_19
OV:IGROV1
CO:HCT_116
CO:HT29
RE:A498
BR:MCF7
PR:DU_145
BR:BT_549
CO:SW_620
BR:HS578T
LE:HL_60
LC:NCI_H522
CO:HCC_2998
OV:SK_OV_3
CO:HCT_15
CO:KM12
LE:CCRF_CEM
LE:MOLT_4
LE:K_562
TCF3
inflammatory response (GO:0006954)
R = 0.477 (p = 2.082e-04)
NM_000211
NM_000660
NM_000674
NM_000677
NM_001039
NM_001058
NM_001643
NM_002349
NM_004590
NM_004946
NM_016946
CNS:SF_268
CNS:SF_295
CNS:SF_539
LE:SR
ME:LOXIMVI
ME:M14
ME:SK_MEL_2
ME:UACC_257
ME:MDA_MB_435
ME:MDA_N
LC:A549
LC:HOP_62
LC:HOP_92
RE:786_0
RE:ACHN
RE:SN12C
PR:DU_145
LE:K_562
PR:PC_3
CNS:SNB_75
OV:IGROV1
ME:MALME_3M
CNS:SNB_19
LC:NCI_H522
LE:HL_60
CNS:U251
BR:BT_549
BR:MDA_MB_231
BR:HS578T
ME:SK_MEL_5
LE:MOLT_4
LE:CCRF_CEM
RE:UO_31
RE:CAKI_1
LC:NCI_H226
LC:NCI_H460
ME:SK_MEL_28
ME:UACC_62
RE:TK_10
RE:RXF_393
RE:A498
OV:SK_OV_3
LE:RPMI_8226
CO:SW_620
OV:OVCAR_8
OV:OVCAR_5
OV:OVCAR_3
CO:KM12
CO:HCT_116
LC:EKVX
CO:COLO205
OV:NCI_ADR_RES
OV:OVCAR_4
CO:HT29
LC:NCI_H322M
CO:HCT_15
CO:HCC_2998
BR:T47D
BR:MCF7
mir-203
protein degradation (GO:0030163)
R = -0.429 (p = 1.208e-03)
NM_001017415
NM_002792
NM_002803
NM_003336
NM_004651
NM_007051

Figure S11

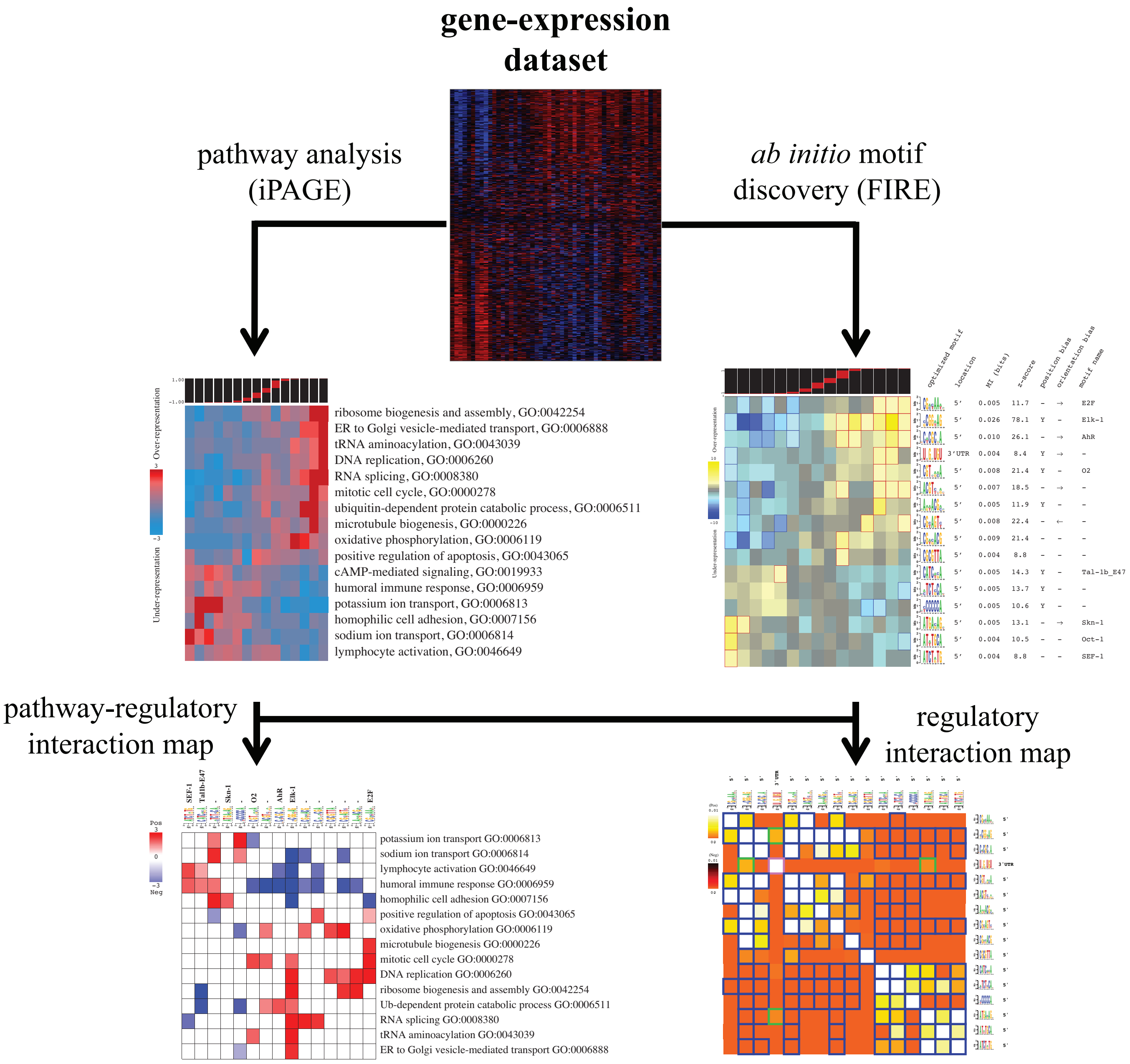
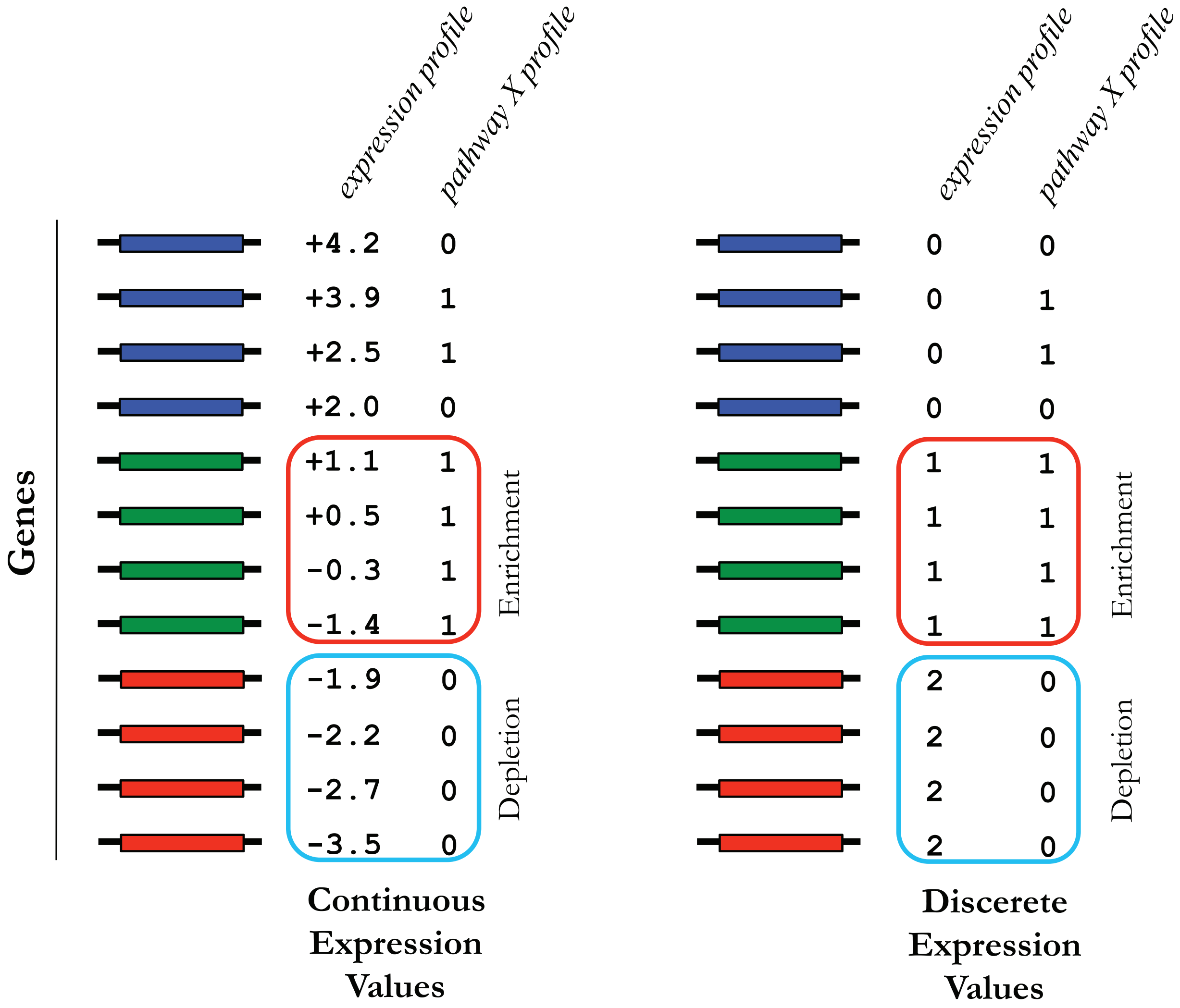

gene-expression dataset
pathway analysis (iPAGE)
ab initio motif discovery (FIRE)
pathway-regulatory interaction map
regulatory interaction map
Over-representation
Under-representation
ribosome biogenesis and assembly, GO:0042254
ER to Golgi vesicle-mediated transport, GO:0006888
tRNA aminoacylation, 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
optimized motif
location
MI (bits)
z-score
position bias
orientation bias
motif name
5' 0.005 11.7 - → E2F
5' 0.026 78.1 Y - Elk-1
5' 0.010 26.1 - → AhR
3'UTR 0.004 8.4 Y → -
5' 0.008 21.4 Y - O2
5' 0.007 18.5 - → -
5' 0.005 11.9 Y - -
5' 0.008 22.4 - ← -
5' 0.009 21.4 - - -
5' 0.004 8.8 - - -
5' 0.005 14.3 Y - Tal-1b_E47
5' 0.005 13.7 Y - -
5' 0.005 10.6 Y - -
5' 0.005 13.1 - → Skn-1
5' 0.004 10.5 - - Oct-1
5' 0.004 8.8 - - SEF-1
SEF-1
Tal1b-E47
Skn-1
O2
AhR
Elk-1
E2F
Pos
Neg
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:0000226
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
3'UTR

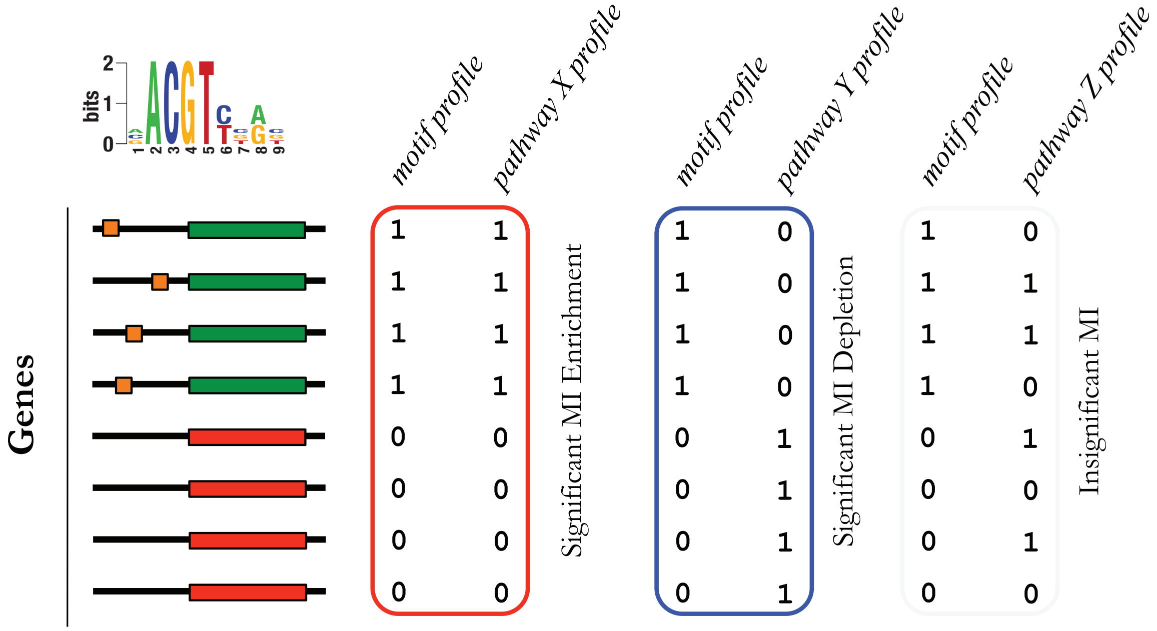
Figure S12



expression profile	pathway X profile
+4.2	0
+3.9	1
+2.5	1
+2.0	0
+1.1	1
+0.5	1
-0.3	1
-1.4	1
-1.9	0
-2.2	0
-2.7	0
-3.5	0

expression profile	pathway X profile
0	0
0	1
0	1
0	0
1	1
1	1
1	1
1	1
2	0
2	0
2	0
2	0

Figure S13



motif profile	pathway X profile	motif profile	pathway Y profile	motif profile	pathway Z profile
1	1	1	0	1	0
1	1	1	0	1	1
1	1	1	0	1	1
1	1	1	0	1	0
0	0	0	1	0	1
0	0	0	1	0	0
0	0	0	1	0	1
0	0	0	1	0	0

Figure S14

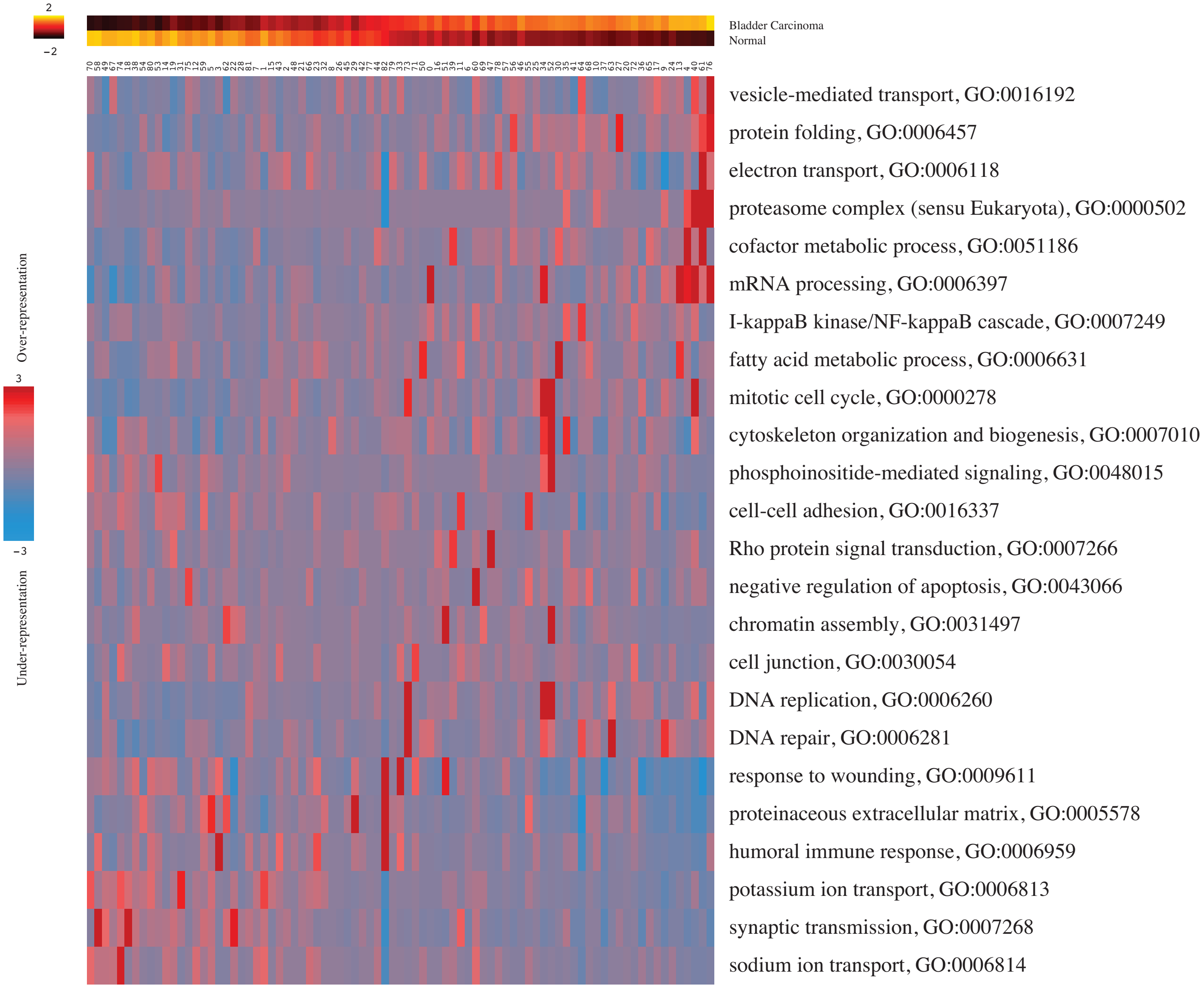
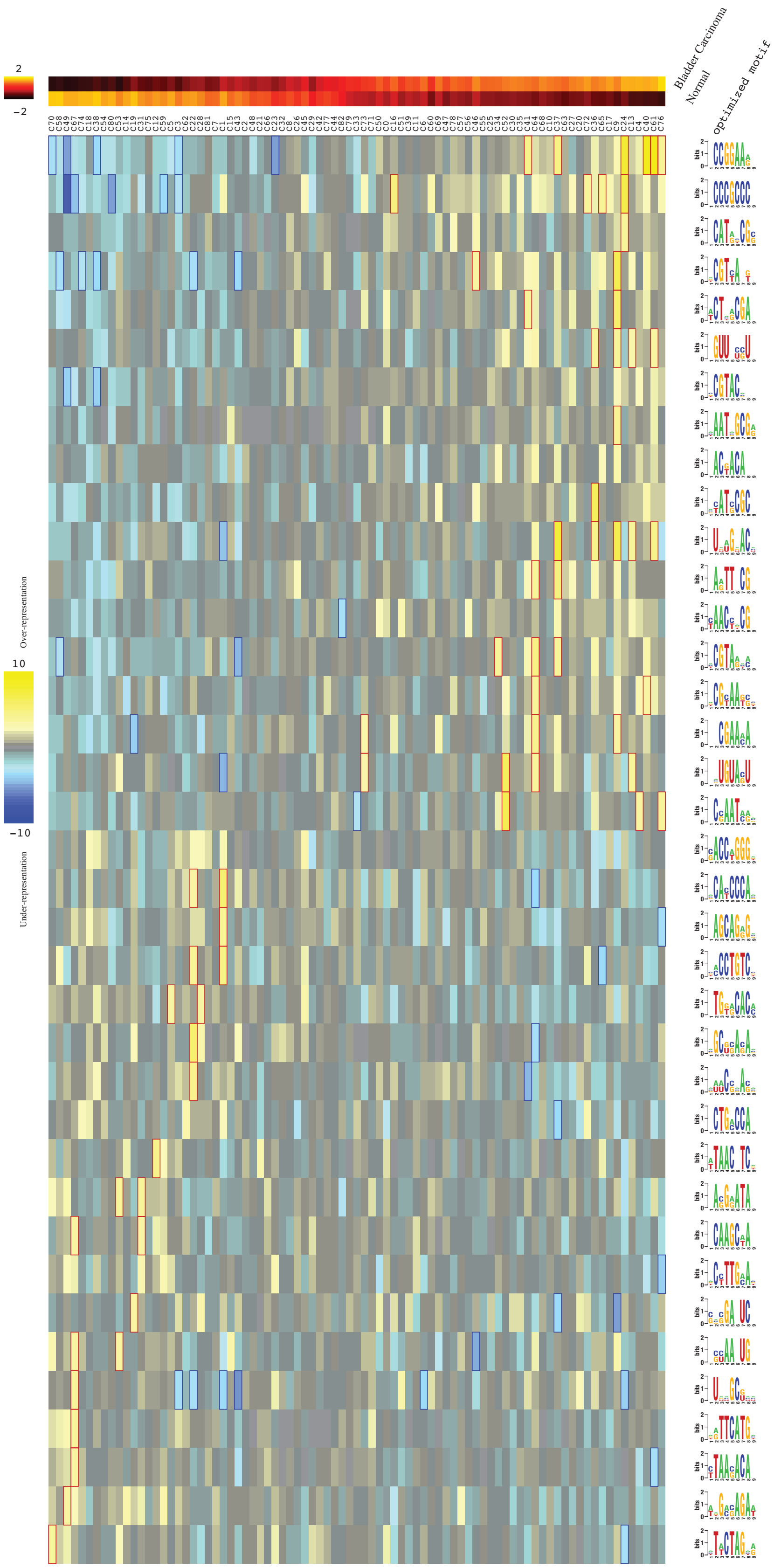

2
−2
Bladder Carcinoma
Normal
Over-representation
3
−3
Under-representation
70 58 49 67 74 18 38 54 80 53 14 19 31 75 12 59 5 3 62 22 28 81 7 1 15 43 2 48 21 66 23 32 8 26 45 29 42 77 44 82 79 33 73 71 50 0 16 51 39 11 6 60 69 47 78 57 56 46 55 25 34 52 30 35 41 64 68 10 37 63 27 20 72 36 65 17 9 24 13 4 40 61 76
vesicle-mediated transport, GO:0016192
protein folding, GO:0006457
electron transport, GO:0006118
proteasome complex (sensu Eukaryota), GO:0000502
cofactor metabolic process, GO:0051186
mRNA processing, GO:0006397
I-kappaB kinase/NF-kappaB cascade, GO:0007249
fatty acid metabolic process, GO:0006631
mitotic cell cycle, GO:0000278
cytoskeleton organization and biogenesis, GO:0007010
phosphoinositide-mediated signaling, GO:0048015
cell-cell adhesion, GO:0016337
Rho protein signal transduction, GO:0007266
negative regulation of apoptosis, GO:0043066
chromatin assembly, GO:0031497
cell junction, GO:0030054
DNA replication, GO:0006260
DNA repair, GO:0006281
response to wounding, GO:0009611
proteinaceous extracellular matrix, GO:0005578
humoral immune response, GO:0006959
potassium ion transport, GO:0006813
synaptic transmission, GO:0007268
sodium ion transport, GO:0006814

Figure S15



location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5′	0.029	28.2	10/10	Y	-	1.00	CCGGAAG	Elk-1
5′	0.028	26.9	10/10	-	-	1.00	CCCGCCC	Sp1
5′	0.013	8.1	6/10	-	←	0.97	CATGGCG	p53
5′	0.018	14.9	10/10	-	-	0.99	CGTCACG	CRE-BP1
5′	0.014	9.9	10/10	-	-	1.00	CTCGCGA	-
3′UTR	0.014	10.2	10/10	-	→	0.91	GUUAUGU	-
5′	0.012	7.6	8/10	-	→	0.75	CGTACCC	-
5′	0.012	7.1	5/10	-	→	0.90	AATGGCG	-
5′	0.011	7.1	6/10	-	→	0.97	ACTACAT	-
5′	0.014	10.5	10/10	-	←	1.00	CATGCGC	-
3′UTR	0.017	13.6	10/10	-	→	0.99	UUUGUAC	-
5′	0.013	9.0	9/10	-	→	0.80	AGTTCCG	STAT3
5′	0.012	7.5	7/10	-	→	0.73	AACCTCG	MEF-2
5′	0.014	10.2	10/10	-	→	0.93	CGTAGTC	Freac-2
5′	0.014	9.9	8/10	-	→	0.94	CGCAAGC	HLF
5′	0.013	9.8	9/10	-	→	0.96	CCGAAAA	E2F
3′UTR	0.015	11.1	10/10	-	→	0.97	AUGUACU	-
5′	0.011	6.9	5/10	Y	-	1.00	CCAATCA	NF-Y
5′	0.014	10.4	10/10	-	→	0.94	ACCTGGG	NRSF
5′	0.015	11.1	10/10	-	←	0.97	CACCCCA	SREBP-1
5′	0.013	8.9	8/10	-	-	0.88	AGCAGGG	HEN1
5′	0.013	8.7	8/10	-	←	0.67	CCCTGTC	Sn
5′	0.013	8.7	8/10	-	→	0.90	TGGACAC	bZIP911
3′UTR	0.013	8.8	10/10	-	→	0.82	GCUGAGA	-
3′UTR	0.013	8.7	9/10	-	→	0.61	UUCCCAC	-
5′	0.011	7.4	8/10	-	-	0.95	CTGCCCA	SRF
5′	0.012	7.4	6/10	-	-	0.71	TAACTTC	CRE-BP1_c-Jun
5′	0.012	8.1	6/10	-	-	0.23	AGGAATA	SEF-1
5′	0.011	6.5	5/10	-	←	0.06	CAAGCAA	MIF-1
5′	0.011	6.4	6/10	-	-	0.32	CCTTGCA	HNF-4
3′UTR	0.015	10.5	10/10	-	→	0.76	GGGACUC	-
3′UTR	0.013	9.7	10/10	-	→	0.94	GUAAAUG	-
3′UTR	0.013	9.5	8/10	-	→	0.97	UUUGCUU	-
5′	0.013	8.2	7/10	-	-	0.99	ATTCATG	HSF
5′	0.012	7.0	7/10	-	-	0.79	TAAGACA	-
5′	0.012	8.0	6/10	-	←	0.68	AGAGAGA	E47
5′	0.012	7.4	5/10	-	←	0.79	TACTAGA	HSF

Figure S16

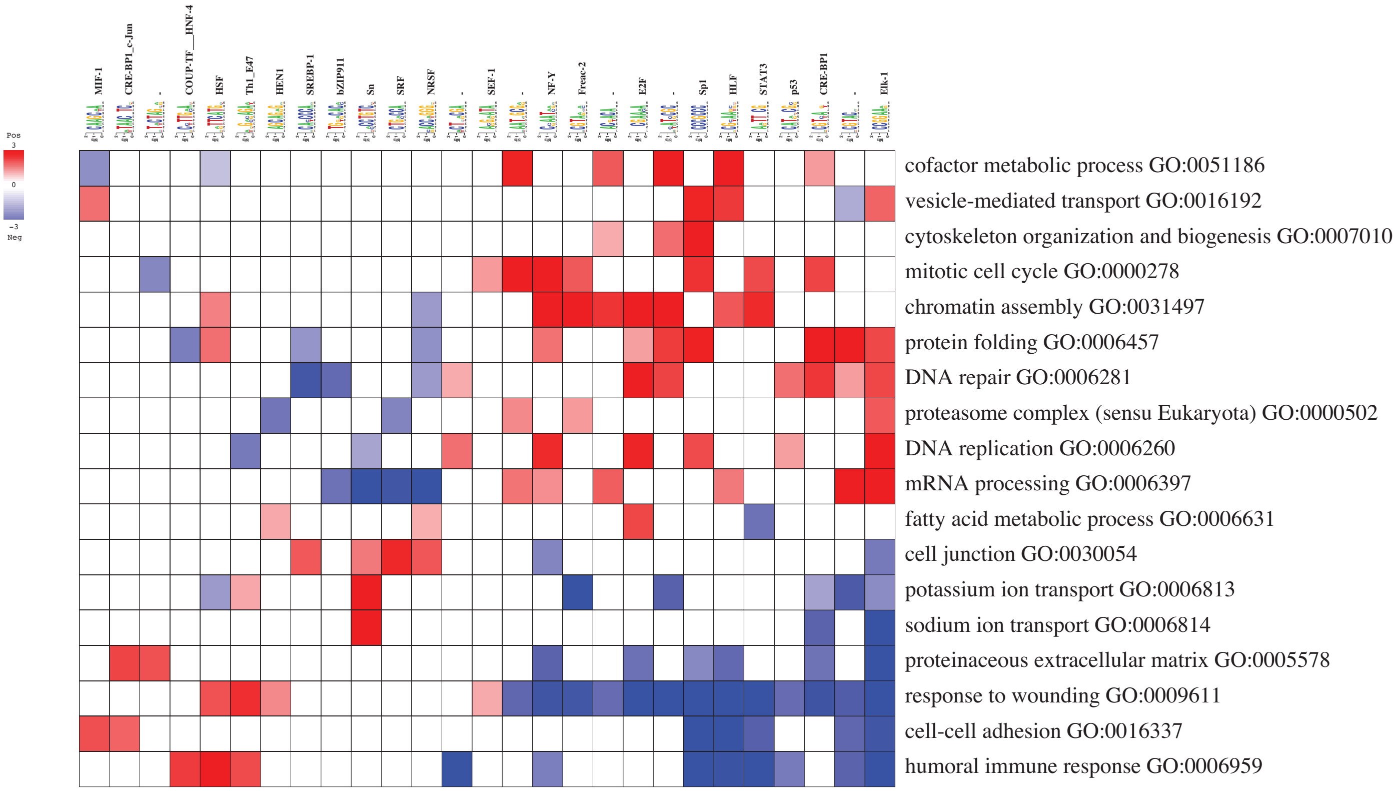

Pos
3
0
-3
Neg
MIF-1
CRE-BP1_c-Jun
COUP-TF___HNF-4
HSF
Th1_E47
HEN1
SREBP-1
bZIP911
Sn
SRF
NRSF
SEF-1
NF-Y
Freac-2
E2F
Sp1
HLF
STAT3
p53
CRE-BP1
Elk-1
cofactor metabolic process GO:0051186
vesicle-mediated transport GO:0016192
cytoskeleton organization and biogenesis GO:0007010
mitotic cell cycle GO:0000278
chromatin assembly GO:0031497
protein folding GO:0006457
DNA repair GO:0006281
proteasome complex (sensu Eukaryota) GO:0000502
DNA replication GO:0006260
mRNA processing GO:0006397
fatty acid metabolic process GO:0006631
cell junction GO:0030054
potassium ion transport GO:0006813
sodium ion transport GO:0006814
proteinaceous extracellular matrix GO:0005578
response to wounding GO:0009611
cell-cell adhesion GO:0016337
humoral immune response GO:0006959

Figure S17

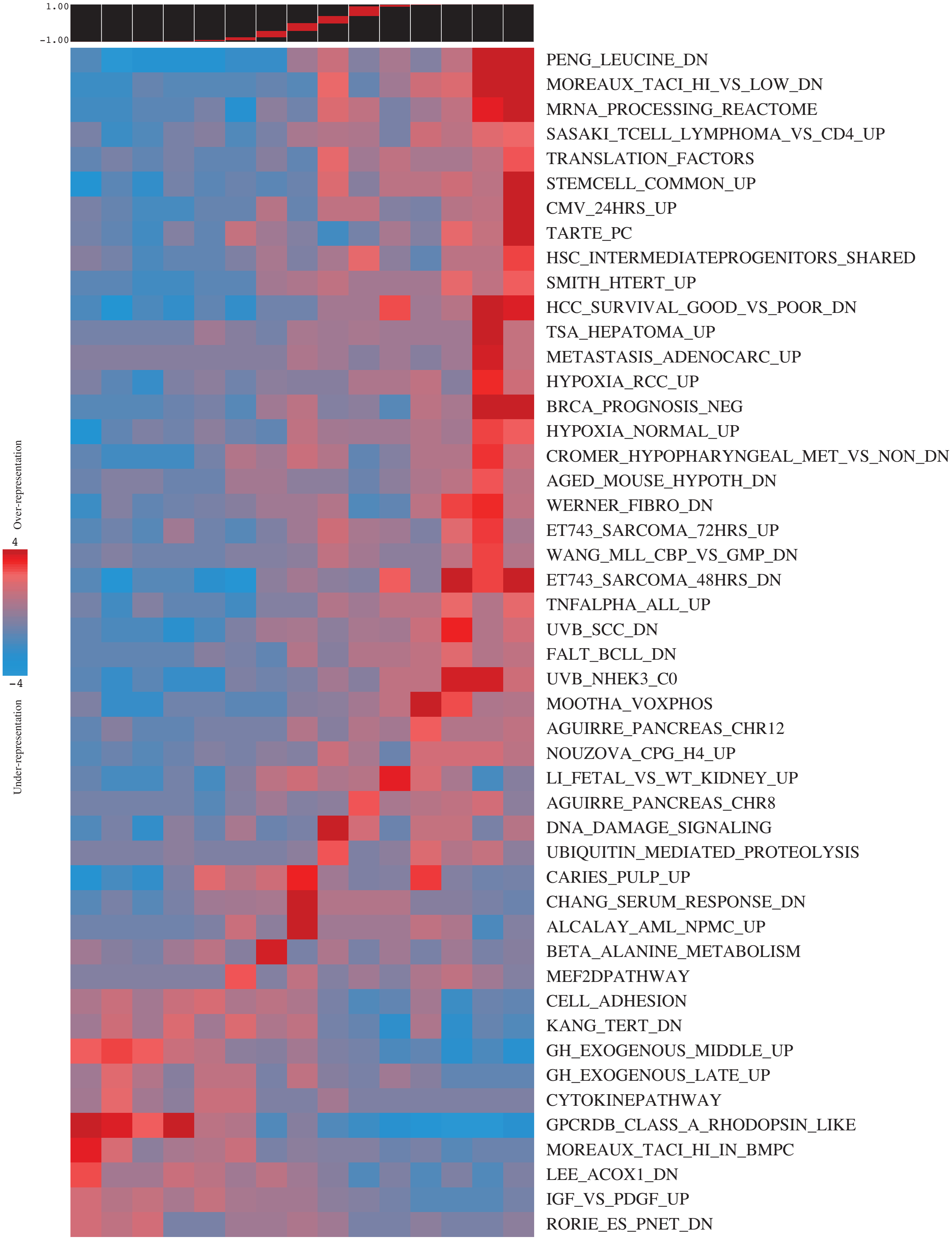

1.00
-1.00
Over-representation
4
-4
Under-representation
PENG_LEUCINE_DN
MOREAUX_TACI_HI_VS_LOW_DN
MRNA_PROCESSING_REACTOME
SASAKI_TCELL_LYMPHOMA_VS_CD4_UP
TRANSLATION_FACTORS
STEMCELL_COMMON_UP
CMV_24HRS_UP
TARTE_PC
HSC_INTERMEDIATEPROGENITORS_SHARED
SMITH_HTERT_UP
HCC_SURVIVAL_GOOD_VS_POOR_DN
TSA_HEPATOMA_UP
METASTASIS_ADENOCARC_UP
HYPOXIA_RCC_UP
BRCA_PROGNOSIS_NEG
HYPOXIA_NORMAL_UP
CROMER_HYPOPHARYNGEAL_MET_VS_NON_DN
AGED_MOUSE_HYPOTH_DN
WERNER_FIBRO_DN
ET743_SARCOMA_72HRS_UP
WANG_MLL_CBP_VS_GMP_DN
ET743_SARCOMA_48HRS_DN
TNFALPHA_ALL_UP
UVB_SCC_DN
FALT_BCLL_DN
UVB_NHEK3_C0
MOOTHA_VOXPHOS
AGUIRRE_PANCREAS_CHR12
NOUZOVA_CPG_H4_UP
LI_FETAL_VS_WT_KIDNEY_UP
AGUIRRE_PANCREAS_CHR8
DNA_DAMAGE_SIGNALING
UBIQUITIN_MEDIATED_PROTEOLYSIS
CARIES_PULP_UP
CHANG_SERUM_RESPONSE_DN
ALCALAY_AML_NPMC_UP
BETA_ALANINE_METABOLISM
MEF2DPATHWAY
CELL_ADHESION
KANG_TERT_DN
GH_EXOGENOUS_MIDDLE_UP
GH_EXOGENOUS_LATE_UP
CYTOKINEPATHWAY
GPCRDB_CLASS_A_RHODOPSIN_LIKE
MOREAUX_TACI_HI_IN_BMPC
LEE_ACOX1_DN
IGF_VS_PDGF_UP
RORIE_ES_PNET_DN

Figure S18

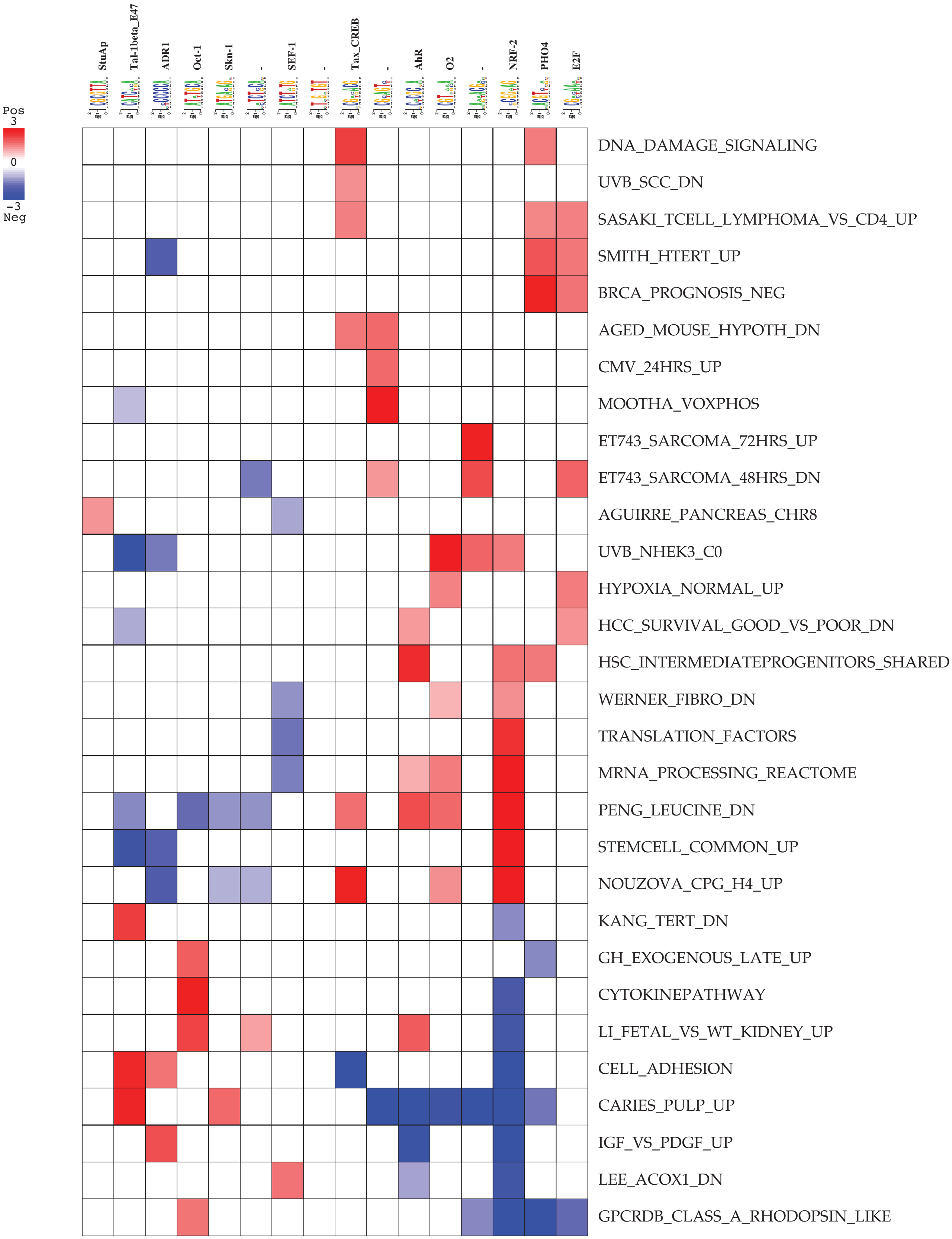

Pos
3
0
-3
Neg
StuAp
Tal-1beta_E47
ADR1
Oct-1
Skn-1
-
SEF-1
-
Tax_CREB
-
AhR
O2
-
NRF-2
PHO4
E2F
DNA_DAMAGE_SIGNALING
UVB_SCC_DN
SASAKI_TCELL_LYMPHOMA_VS_CD4_UP
SMITH_HTERT_UP
BRCA_PROGNOSIS_NEG
AGED_MOUSE_HYPOTH_DN
CMV_24HRS_UP
MOOTHA_VOXPHOS
ET743_SARCOMA_72HRS_UP
ET743_SARCOMA_48HRS_DN
AGUIRRE_PANCREAS_CHR8
UVB_NHEK3_C0
HYPOXIA_NORMAL_UP
HCC_SURVIVAL_GOOD_VS_POOR_DN
HSC_INTERMEDIATEPROGENITORS_SHARED
WERNER_FIBRO_DN
TRANSLATION_FACTORS
MRNA_PROCESSING_REACTOME
PENG_LEUCINE_DN
STEMCELL_COMMON_UP
NOUZOVA_CPG_H4_UP
KANG_TERT_DN
GH_EXOGENOUS_LATE_UP
CYTOKINEPATHWAY
LI_FETAL_VS_WT_KIDNEY_UP
CELL_ADHESION
CARIES_PULP_UP
IGF_VS_PDGF_UP
LEE_ACOX1_DN
GPCRDB_CLASS_A_RHODOPSIN_LIKE

Figure S19

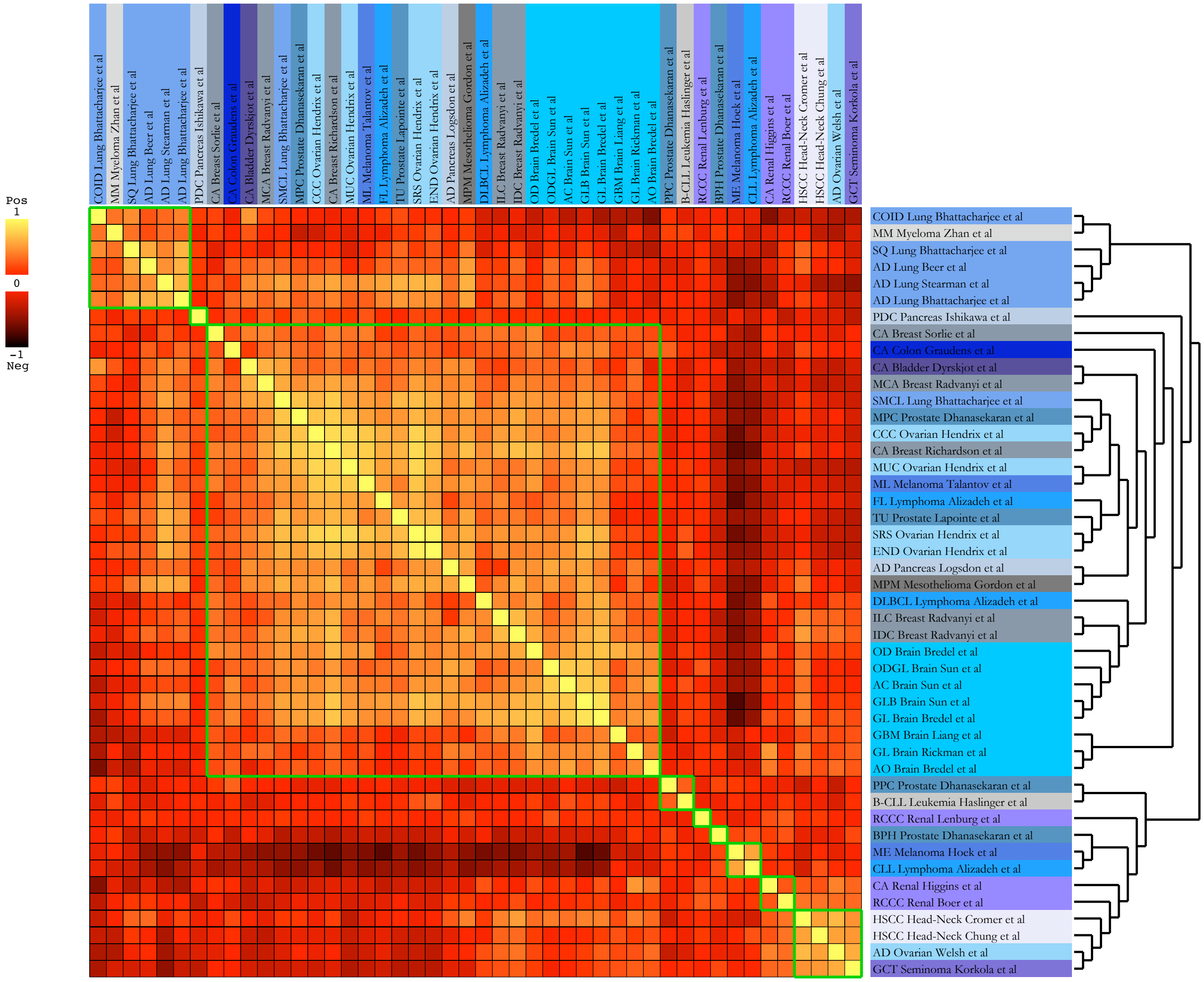

COID Lung Bhattacharjee et al
MM Myeloma Zhan et al
SQ Lung Bhattacharjee et al
AD Lung Beer et al
AD Lung Stearman et al
AD Lung Bhattacharjee et al
PDC Pancreas Ishikawa et al
CA Breast Sorlie et al
CA Colon Graudens et al
CA Bladder Dyrskjot et al
MCA Breast Radvanyi et al
SMCL Lung Bhattacharjee et al
MPC Prostate Dhanasekaran et al
CCC Ovarian Hendrix et al
CA Breast Richardson et al
MUC Ovarian Hendrix et al
ML Melanoma Talantov et al
FL Lymphoma Alizadeh et al
TU Prostate Lapointe et al
SRS Ovarian Hendrix et al
END Ovarian Hendrix et al
AD Pancreas Logsdon et al
MPM Mesothelioma Gordon et al
DLBCL Lymphoma Alizadeh et al
ILC Breast Radvanyi et al
IDC Breast Radvanyi et al
OD Brain Bredel et al
ODGL Brain Sun et al
AC Brain Sun et al
GLB Brain Sun et al
GL Brain Bredel et al
GBM Brain Liang et al
GL Brain Rickman et al
AO Brain Bredel et al
PPC Prostate Dhanasekaran et al
B-CLL Leukemia Haslinger et al
RCCC Renal Lenburg et al
BPH Prostate Dhanasekaran et al
ME Melanoma Hoek et al
CLL Lymphoma Alizadeh et al
CA Renal Higgins et al
RCCC Renal Boer et al
HSCC Head-Neck Cromer et al
HSCC Head-Neck Chung et al
AD Ovarian Welsh et al
GCT Seminoma Korkola et al
Pos
1
0
-1
Neg
COID Lung Bhattacharjee et al
MM Myeloma Zhan et al
SQ Lung Bhattacharjee et al
AD Lung Beer et al
AD Lung Stearman et al
AD Lung Bhattacharjee et al
PDC Pancreas Ishikawa et al
CA Breast Sorlie et al
CA Colon Graudens et al
CA Bladder Dyrskjot et al
MCA Breast Radvanyi et al
SMCL Lung Bhattacharjee et al
MPC Prostate Dhanasekaran et al
CCC Ovarian Hendrix et al
CA Breast Richardson et al
MUC Ovarian Hendrix et al
ML Melanoma Talantov et al
FL Lymphoma Alizadeh et al
TU Prostate Lapointe et al
SRS Ovarian Hendrix et al
END Ovarian Hendrix et al
AD Pancreas Logsdon et al
MPM Mesothelioma Gordon et al
DLBCL Lymphoma Alizadeh et al
ILC Breast Radvanyi et al
IDC Breast Radvanyi et al
OD Brain Bredel et al
ODGL Brain Sun et al
AC Brain Sun et al
GLB Brain Sun et al
GL Brain Bredel et al
GBM Brain Liang et al
GL Brain Rickman et al
AO Brain Bredel et al
PPC Prostate Dhanasekaran et al
B-CLL Leukemia Haslinger et al
RCCC Renal Lenburg et al
BPH Prostate Dhanasekaran et al
ME Melanoma Hoek et al
CLL Lymphoma Alizadeh et al
CA Renal Higgins et al
RCCC Renal Boer et al
HSCC Head-Neck Cromer et al
HSCC Head-Neck Chung et al
AD Ovarian Welsh et al
GCT Seminoma Korkola et al

Figure S20

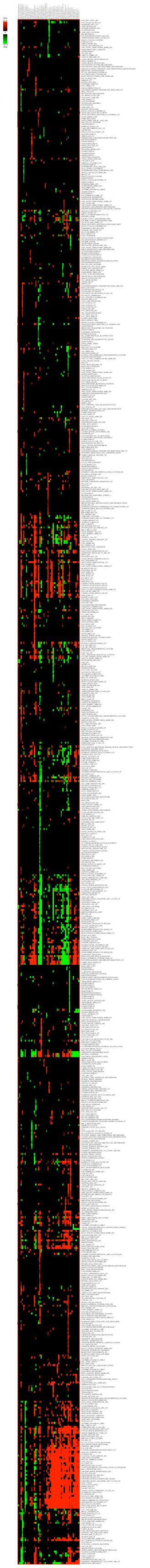
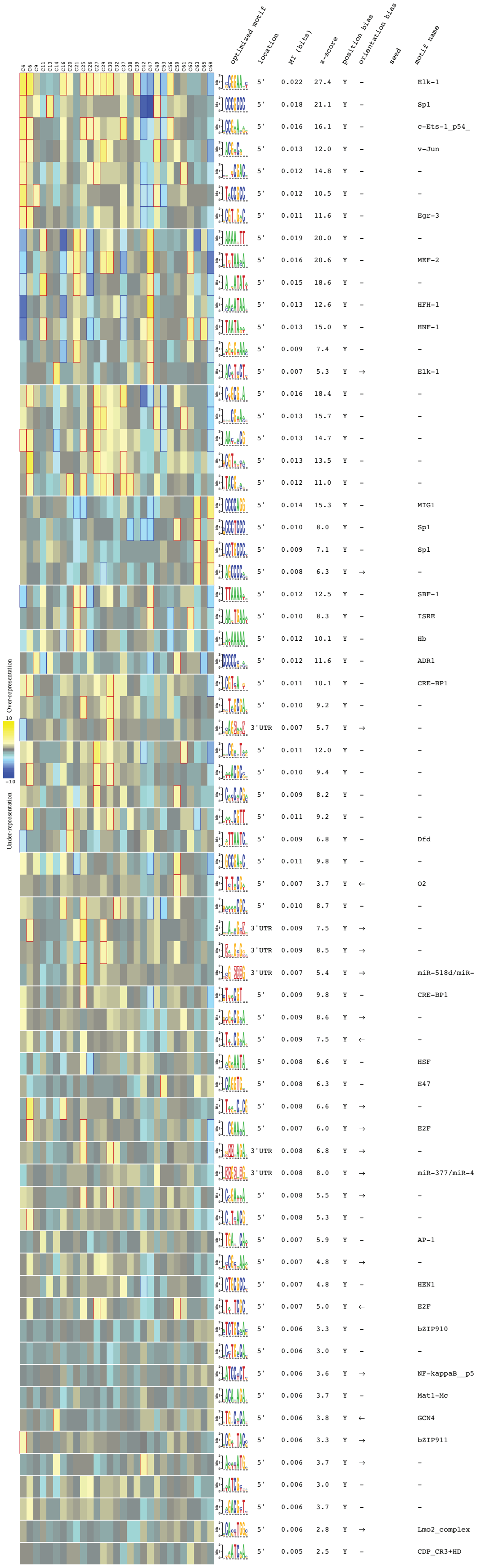
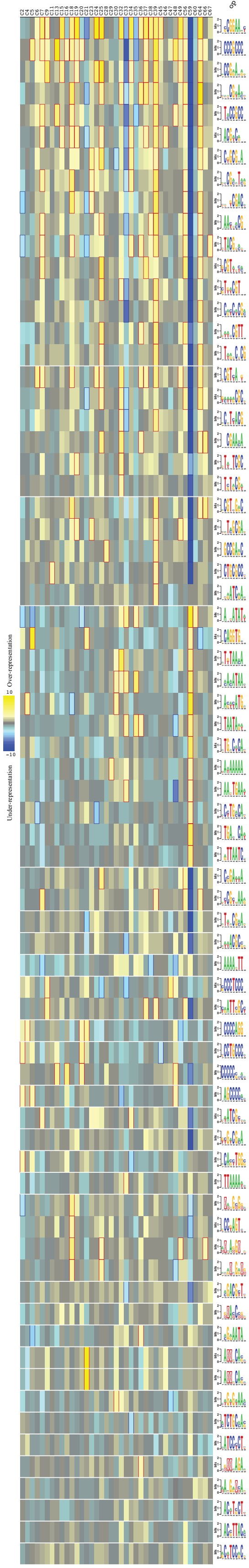


Figure S21



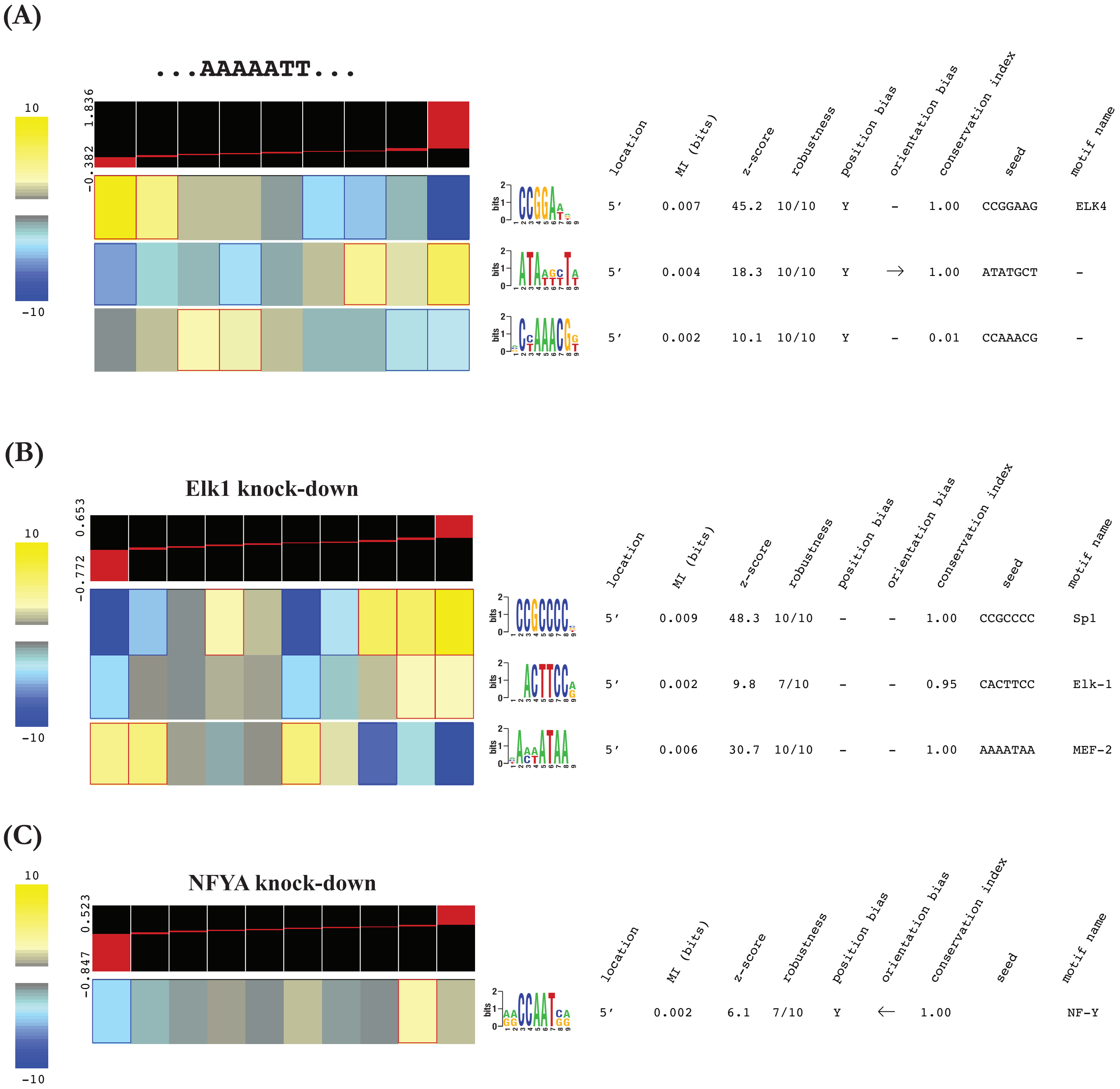
optimized motif	location	MI (bits)	z-score	position bias	orientation bias	seed	motif name
	5'	0.022	27.4	Y	–		Elk-1
	5'	0.018	21.1	Y	–		Sp1
	5'	0.016	16.1	Y	–		c-Ets-1_p54_
	5'	0.013	12.0	Y	–		v-Jun
	5'	0.012	14.8	Y	–		–
	5'	0.012	10.5	Y	–		–
	5'	0.011	11.6	Y	–		Egr-3
	5'	0.019	20.0	Y	–		–
	5'	0.016	20.6	Y	–		MEF-2
	5'	0.015	18.6	Y	–		–
	5'	0.013	12.6	Y	–		HFH-1
	5'	0.013	15.0	Y	–		HNF-1
	5'	0.009	7.4	Y	–		–
	5'	0.007	5.3	Y	→		Elk-1
	5'	0.016	18.4	Y	–		–
	5'	0.013	15.7	Y	–		–
	5'	0.013	14.7	Y	–		–
	5'	0.013	13.5	Y	–		–
	5'	0.012	11.0	Y	–		–
	5'	0.014	15.3	Y	–		MIG1
	5'	0.010	8.0	Y	–		Sp1
	5'	0.009	7.1	Y	–		Sp1
	5'	0.008	6.3	Y	→		–
	5'	0.012	12.5	Y	–		SBF-1
	5'	0.010	8.3	Y	–		ISRE
	5'	0.012	10.1	Y	–		Hb
	5'	0.012	11.6	Y	–		ADR1
	5'	0.011	10.1	Y	–		CRE-BP1
	5'	0.010	9.2	Y	–		–
	3'UTR	0.007	5.7	Y	→		–
	5'	0.011	12.0	Y	–		–
	5'	0.010	9.4	Y	–		–
	5'	0.009	8.2	Y	–		–
	5'	0.011	9.2	Y	–		–
	5'	0.009	6.8	Y	–		Dfd
	5'	0.011	9.8	Y	–		–
	5'	0.007	3.7	Y	←		O2
	5'	0.010	8.7	Y	–		–
	3'UTR	0.009	7.5	Y	→		–
	3'UTR	0.009	8.5	Y	→		–
	3'UTR	0.007	5.4	Y	→		miR-518d/miR-
	5'	0.009	9.8	Y	–		CRE-BP1
	5'	0.009	8.6	Y	→		–
	5'	0.009	7.5	Y	←		–
	5'	0.008	6.6	Y	–		HSF
	5'	0.008	6.3	Y	–		E47
	5'	0.008	6.6	Y	→		–
	5'	0.007	6.0	Y	→		E2F
	3'UTR	0.008	6.8	Y	→		–
	3'UTR	0.008	8.0	Y	→		miR-377/miR-4
	5'	0.008	5.5	Y	→		–
	5'	0.008	5.3	Y	–		–
	5'	0.007	5.9	Y	–		AP-1
	5'	0.007	4.8	Y	→		–
	5'	0.007	4.8	Y	–		HEN1
	5'	0.007	5.0	Y	←		E2F
	5'	0.006	3.3	Y	–		bZIP910
	5'	0.006	3.0	Y	–		–
	5'	0.006	3.6	Y	→		NF-kappaB__p5
	5'	0.006	3.7	Y	–		Mat1-Mc
	5'	0.006	3.8	Y	←		GCN4
	5'	0.006	3.3	Y	→		bZIP911
	5'	0.006	3.7	Y	→		–
	5'	0.006	3.0	Y	–		–
	5'	0.006	3.7	Y	–		–
	5'	0.006	2.8	Y	→		Lmo2_complex
	5'	0.005	2.5	Y	–		CDP_CR3+HD

Figure S22



optimized motif	location	MI (bits)	z-score	position bias	orientation bias	seed	motif nam
	5'	0.035	62.0	Y	-		Elk-1
	5'	0.029	53.4	Y	-		Sp1
	5'	0.023	37.5	Y	-		c-Ets-1_p54_
	5'	0.018	29.2	Y	-		-
	5'	0.018	31.0	Y	-		-
	5'	0.017	30.1	Y	-		v-Jun
	5'	0.017	25.2	Y	-		-
	5'	0.017	28.2	Y	-		-
	5'	0.015	22.1	Y	-		-
	5'	0.013	24.0	Y	-		-
	5'	0.013	18.2	Y	-		-
	5'	0.012	19.9	Y	-		-
	5'	0.012	19.1	Y	-		CRE-BP1
	5'	0.010	11.5	Y	-		-
	5'	0.010	12.6	Y	-		-
	5'	0.010	12.5	Y	-		-
	5'	0.014	27.2	Y	-		CRE-BP1
	5'	0.013	21.5	Y	-		-
	5'	0.011	14.2	Y	-		-
	5'	0.011	14.1	Y	-		E2F
	5'	0.010	12.7	Y	-		E2F
	5'	0.007	7.2	Y	-		O2
	5'	0.012	15.0	Y	-		Egr-3
	5'	0.010	15.1	Y	-		-
	5'	0.009	12.3	Y	-		-
	5'	0.006	6.2	Y	-		HEN1
	5'	0.004	2.8	Y	-		CDP_CR3+HD
	5'	0.009	15.1	Y	-		-
	5'	0.008	10.1	Y	-		E47
	5'	0.008	11.2	Y	-		MEF-2
	5'	0.008	6.9	Y	-		HFH-1
	5'	0.007	9.7	Y	-		-
	5'	0.007	7.3	Y	-		HNF-1
	5'	0.006	7.3	Y	-		GCN4
	5'	0.006	6.5	Y	-		Hb
	5'	0.006	6.5	Y	→		ISRE
	5'	0.006	6.4	Y	-		-
	5'	0.005	4.0	Y	-		AP-1
	5'	0.004	3.1	Y	→		Dfd
	5'	0.009	14.9	Y	-		-
	5'	0.008	8.6	Y	-		-
	5'	0.009	10.5	Y	-		-
	5'	0.008	8.3	Y	-		-
	5'	0.008	8.9	Y	-		-
	5'	0.008	8.9	Y	-		Sp1
	5'	0.007	7.3	Y	-		NF-Y
	5'	0.008	9.0	Y	-		MIG1
	5'	0.007	7.9	Y	-		Sp1
	5'	0.007	9.3	Y	-		ADR1
	5'	0.007	6.3	Y	-		-
	5'	0.007	7.3	Y	-		-
	5'	0.006	6.6	Y	→		-
	5'	0.006	5.9	Y	-		Lmo2_complex
	5'	0.006	6.1	Y	-		SBF-1
	3'UTR	0.006	5.0	Y	→		-
	5'	0.006	4.7	Y	-		PPARalpha_RXR
	3'UTR	0.006	7.0	Y	→		-
	3'UTR	0.005	5.6	Y	→		-
	5'	0.006	6.7	Y	-		-
	3'UTR	0.006	6.1	Y	→		-
	5'	0.005	5.0	Y	→		HSF
	3'UTR	0.005	6.0	Y	→		miR-203
	3'UTR	0.005	6.0	Y	→		miR-203
	5'	0.005	4.2	Y	→		-
	5'	0.005	3.7	Y	←		bZIP910
	5'	0.005	3.8	Y	→		NF-kappaB__p5
	3'UTR	0.005	4.4	Y	→		-
	3'UTR	0.005	3.6	Y	→		miR-487a
	5'	0.004	2.8	Y	-		Elk-1
	5'	0.004	2.4	Y	-		HLF
	5'	0.004	1.9	Y	-		GCR1

Figure S23



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	–

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

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 S24

Supplemental Table

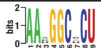
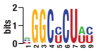
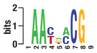
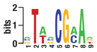
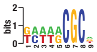
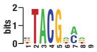
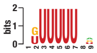
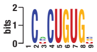
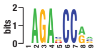
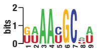
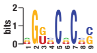
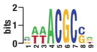
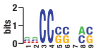
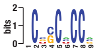
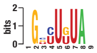
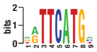
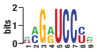
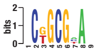
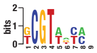
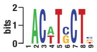
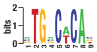
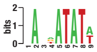
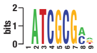
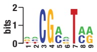
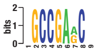
Table S1

	Real Data	Random Data I	Random Data II
BL vs DLBCL	525	1	1
Bladder Carcinoma (cont)	224	4	5
Bladder Carcinoma (disc)	248	1	1

Table S2

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 S3

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

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.

Cover, T., and Thomas, J. (2006). Elements of Information Theory, Second Edition edn (Hoboken, NJ, Wiley-Interscience).

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.

Dyrskjot, L., Kruhoffer, M., Thykjaer, T., Marcussen, N., Jensen, J.L., Moller, K., and Orntoft, T.F. (2004). Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. Cancer Res *64*, 4040-4048.

Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. Mol Cell *28*, 337-350.

Elemento, O., and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol *6*, R18.

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.

Sinha, S., Adler, A.S., Field, Y., Chang, H.Y., and Segal, E. (2008). Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res *18*, 477-488.

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.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

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.