Molecular Cell, Volume *28*

Supplemental Data

A Universal Framework

for Regulatory Element Discovery

across All Genomes and Data Types

Olivier Elemento, Noam Slonim, and Saeed Tavazoie

Supplemental Results

In what follows, we present additional information about the motifs obtained when applying FIRE to the yeast and human clustering partitions (stress and tissue expression datasets). Then, we describe the results obtained when applying FIRE to several additional datasets: a large compendium of gene expression profiles in *C. elegans*, and pre-specified groups of genes in *D. melanogaster*, including groups of genes with similar spatio-temporal patterns of gene expression revealed by *in situ* hybridization. The ability to analyze such groups of genes is important, as it is common practice to simply specify a relatively small set of presumably co-expressed genes, using some *ad hoc* criteria, and to attempt to elucidate the regulatory elements responsible for this assumed co-expression.

Additional results for yeast and human gene clustering partitions

Expression variance explained by the yeast predicted motifs

As shown in Figure 3, the motifs predicted by FIRE are in general highly over-represented in a small set of co-expression clusters. To further evaluate the significance of the motifs discovered by FIRE, we consider the percentage of genes that have a given motif both in clusters where the motif is significantly over-represented (denoted “active” clusters – such clusters are shown with a red frame in Figure 3) and in the remaining clusters (“non-active” clusters). These two fractions, shown in Table S2, for the 23 yeast motifs presented in Figure 3, indicate the level of variance in the expression data that can be explained by the discovered motifs. For example, the top motif (PAC) in Figure 3 has active occurrences in 47% of promoters associated with 5 different clusters (marked by red frames in the first row in Figure 3), and additional non-active occurrences in only 7% of the promoters associated with other clusters. Table S2 shows that, in general, the percentage of the motif's non-active occurrences is relatively low, suggesting that indeed much of the variance in the expression data can be explained by the predicted motifs.

Comparison with conservation-based motif discovery approaches

To further validate our predictions, we have undertaken a systematic comparison between our predicted motifs and the motifs predicted using alignment-based phylogenetic footprinting in yeast (Kellis et al., 2003) and mammals (Xie et al., 2005).

Motif comparisons were performed using CompareACE (Hughes et al., 2000). Two motifs were considered equivalent if their CompareACE score was greater than 0.7. At this threshold, in yeast, 3/12 (DNA) and 1/2 (RNA) of the FIRE-predicted motifs that match a known (experimentally verified) motif do not match any motif predicted in Kellis *et al*. In addition, 1/5 (DNA) and 3/4 (RNA) of the novel FIRE-predicted motifs do not match any motif reported in (Kellis et al.*,* 2003).

Moreover, the ability of FIRE to detect entirely novel motifs is fully revealed in mammals. In particular, for the human tissue expression data, although many of the FIRE-predicted motifs that match known motifs in TRANSFAC or JASPAR also match conserved motifs discovered in (Xie et al., 2005), when we examine FIRE-predicted motifs that do not match known motifs in these databases, we observe the opposite trend. Specifically, among these FIRE-predicted motifs, 32/52 (DNA) and 21/36 (RNA) do not match any motif in (Xie et al., 2005). We therefore conclude that FIRE predicts many motifs that were not already predicted by prior comparative analysis. More generally, we believe that FIRE analyses should be seen as complementary to comparative genomics methods, providing additional support for conserved motifs, and further pointing out less conserved, possibly species-specific motifs that are relevant to the expression data under analysis.

On the sensitivity/specificity tradeoff

The tradeoff between specificity and sensitivity is a fundamental issue that must be addressed when designing motif prediction approaches. In FIRE, we have chosen to favor specificity over sensitivity. Thus, the output of FIRE typically consists of a small, experimentally tractable set of high confidence motifs. We believe that this emphasis on specificity represents an important distinction between FIRE and previously described motif discovery methods that will play a significant role in making FIRE the tool of choice for motif discovery.

However, it is possible (and straightforward) to decrease the default stringency level in FIRE, and thus to increase the number of predicted motifs (at the possible cost of reducing specificity). The simplest way is to use a less stringent robustness index threshold. Figure S15 shows the number of predicted motifs for the yeast dataset of Gasch *et al*., at different robustness index thresholds (from 0 to 10; 6 is the default value). Clearly, lower thresholds increase the number of predicted motifs. Importantly, several of the lower scoring motifs are supported by GO enrichments of their target genes and/or match known (experimentally validated) yeast motifs, suggesting that many of the predicted "weaker" motifs are also true positives. We note that the ability of FIRE to detect weak motifs is possibly due to the fact that several clusters (or expression bins for continuous data) may contribute to the mutual information. Thus, a motif does not strictly need to be highly over-represented in a single cluster to be detected by FIRE (as opposed to other cluster-based approaches).

Application to *C. elegans*

We examined a dataset containing 551 *C. elegans* expression microarrays reported in (Kim et al., 2001) and focused on the 11,562 genes with reported expression levels in at least 400 microarrays. Following precisely the same analysis discussed above for yeast data, we clustered the 11,562 genes into 108 non-overlapping clusters using Iclust. With default parameters, FIRE predicts 61 DNA motifs and 33 RNA motifs (Figure S16) that are highly informative about the corresponding clustering partition, versus an average of 0.06 DNA "motifs" and 0.09 RNA "motifs" when applied to 100 randomly shuffled partitions. Figure S17 shows how much information each seed initially conveys and how much additional information is gained through optimization. The predicted motifs are further partitioned into 46 modules, many of which contain both DNA and RNA motifs (Figure S18), suggesting that cooperation between transcriptional and post-transcriptional regulatory mechanisms plays an important role in shaping the worm transcriptome.

Overall, 13 predicted DNA motifs closely match 13 distinct known motifs in TRANSFAC (Matys et al., 2006) or JASPAR (Vlieghe et al., 2006), and 3 predicted RNA motifs match the 5' extremity of known worm miRNAs (Chan et al., 2005; Lewis et al., 2005), suggesting that they are targeted by these miRNAs. Independent conservation analysis with respect to the *C. briggsae* genome further supports our predictions: nearly half (47%) of the predicted motifs are highly conserved, with a conservation index ≥ 0.95.

A subset of the predicted motifs is presented in Figure S19. The first module is relatively large with 10 motifs, two of which are shown in Figure S19. The most informative motif is a highly conserved C-rich RNA motif, [CU]N[ACU]CCCC, that is over-represented in several clusters, two of which are associated with regulation of growth (c100, p<1e-14) and DNA replication (c68, p<1e-07).

The second module consists of two motifs that match two different GATA-factor binding sites in TRANSFAC, and are found to be preferentially located close to the TSS. While it is possible that both motifs match binding sites for the same GATA-factor, we note that their over-representation patterns are slightly different (Figure 8). Our analysis also suggests that one of these motifs, [CGT]CTTATC[AT], may be involved in regulating the expression of several enzymes with oxidoreductase activity (p<1e-04).

The third module consists of one RNA motif and one DNA motif, [ACT][AC]CGTG[AG][AC][AC], with a remarkable position bias towards the TSS (Figure S20). This motif strongly resembles an E-box, bound by bHLH transcription factors, and its target genes are associated in general with protein modification (p<1e-27), and more specifically with kinase (p<1e-19) and phosphatase activity (p<1e-12). In fact, 17% of the worm genes with protein kinase activity and 29% of genes coding for protein phosphatases have this motif in their 1kb 5' upstream, suggesting some common regulatory control for these activities. This is intriguing as kinases and phosphatases are typically thought to be involved in a multitude of independent biological processes. Further analyses will be needed to understand the involvement of this motif in the transcriptional regulation related to these post-translational modification processes.

A fourth module contains five predicted motifs, two of which are presented in Figure S19; a DNA motif that resembles the binding site for Cut-like homeobox protein in mammals (Andres et al., 1992) and the DNA-replication element in *Drosophila* (Matsukage et al., 1995), and a U-rich highly conserved RNA motif, previously detected based on independent conservation analysis (Chan et al., 2005). The target genes of these two motifs are highly enriched for protein biosynthesis (p<1e-08 and p<1e-30, respectively).

Another predicted RNA motif, [CGU]UGUANAU[ACU], matches almost perfectly the yeast Puf3 binding site (Gerber et al., 2004) and thus might be bound by one (or several) of the PUF proteins in worm (Wickens et al., 2002). Two other predicted RNA motifs, [ACU][AGU]CGGGU and [CU][AU]UAUCN[ACU][ACU], match with high specificity the 5' extremity of known worm miRNAs (Figure S19); both motifs were also reported in (Chan et al., 2005) based on an independent network-level conservation analysis. Two additional predicted RNA motifs, AA[AG]UAAA and UUGUUGA[ACU], are both highly conserved and show a position bias towards the stop codon. GO analysis shows that the former, which resembles a poly-adenylation sequence, is enriched in the 3'UTRs of mRNAs that code for components of the cuticle (p<1e-24), while the latter is enriched in 3'UTRs of mRNAs that code for components of the ribosome (p<1e-11).

Our analysis further detects pairs of motifs for which the relative distance is significantly informative about the cluster indices. For example, as depicted in Figure S21, the two DNA motifs, [ACG]AA[ACG]CGAG and [CGT]ACCGTA[CG][ACT], are jointly present in 130 upstream promoters, 9 of which are associated with cluster c63 (over representation of p<1e-8); in 8 out of their 130 joint occurrences, these two motifs are located exactly 3 nucleotides one after another, in the same relative order towards the TSS, where 6 of these 8 cases are again within promoters associated with c63 (p<1e-6), and all 6 are ribosomal genes (p<1e-9). Although the numbers involved here are small, we speculate that co-regulation of these genes requires both motifs in the proper spatial arrangement. It is also possible that the two motifs are part of a single larger motif. We note, that for all 6 genes, almost identical motifs with an identical arrangement are observed in the promoters of their *C. briggsae* orthologs (data not shown).

Finally, we also applied FIRE to each of the 551 expression arrays independently. These complementary results are available at our Web site.

Analyzing a pre-specified set of genes: examples from *Drosophila*

The scenarios described so far involve applying FIRE to whole genome expression data that had been processed through subsequent analyses, *e.g.*, describing periodic expression in terms of phase values or finding gene clustering partitions. It is however common practice to simply specify a relatively small set of presumably co-expressed genes, using some *ad hoc* criteria, and to attempt to elucidate the regulatory elements responsible for this assumed co-expression. In essence, this simply represents a special case of categorical data with only two categories: the pre-specified set of genes versus all remaining genes. Thus, FIRE is directly applicable to these scenarios as well, as we demonstrate in the following two examples.

Drosophila early embryonic enhancers

Transcriptional regulation during early development of *Drosophila* embryos involves regulatory elements concentrated within particular DNA regions, termed enhancer sequences (see (Arnosti and Kulkarni, 2005) for a recent review). In the following, we use 124 early embryonic fly enhancer sequences collected from the literature (D. Papatsenko, https://webfiles.berkeley.edu/~dap5/), of which 20 are known to be bound by the Bicoid morphogen (Driever and Nusslein-Volhard, 1989; Struhl et al., 1989), giving rise to a simple two-category partition: the 20 Bicoid bound enhancers, versus the remaining 104. Given these data, FIRE predicts a single motif (Table S3) that matches well the previously reported Bicoid binding site (Driever and Nusslein-Volhard, 1989) and is present in 95% of the Bicoid-bound enhancers, but only in 25% of the remaining enhancers. Applying AlignACE to the same data yields a somewhat similar motif (Table S3), but apparently less discriminative: it is present in 80% of the Bicoid-bound enhancers and in 28% of the remaining enhancers (when using ScanACE (Hughes et al., 2000) with default parameters). We further applied a simple leave-one-out test, in which each enhancer was withdrawn and the remaining 123 were analyzed by FIRE to extract a maximally informative motif that was then used to predict whether the withdrawn enhancer is bound by Bicoid or not; the high performance reported in Table S3 supports the robustness of our results in this example. Applying the same analysis to the 20 enhancers bound by Dorsal, another fly morphogen, yielded similar results (see our Web site).

Drosophila in situ hybridization database

Spatial patterns of gene expression, for example during different developmental stages, can be determined using RNA *in situ* hybridization. In this section, we present the results of the analysis of a large-scale *in situ* hybridization study during *D. melanogaster* embryogenesis (Tomancak et al., 2002). Fifty-five groups of genes were analyzed, where each group corresponds to a set of genes with a similar spatio-temporal expression pattern during embryogenesis (*e.g.,* embryonic midgut during stages 13-16) (Tomancak et al., 2002). Altogether, these groups cover ~2,000 genes, with some genes associated with more than one group. We applied FIRE to each of these 55 groups independently, where in each case, genes within the group constitute one category while the remaining genes constitute the other category. Given these data, FIRE predicts a total of 28 DNA motifs and 11 RNA motifs, versus an average of 2.4 and 4.5 DNA and RNA "motifs", respectively, when applied 10 times to the same data after random shuffling. A selection of the obtained motifs is presented in Figure S22, and the complete results are available at our Web site.

Several motifs shown in Figure S22 have been previously described in the literature. For example, the ATCGATA motif matches the DNA replication-related element (DRE) (Matsukage et al., 1995). Our results indicate that this motif is highly enriched upstream of maternal genes, *i.e.*, genes whose mRNAs are deposited into the egg by the mother, and therefore are detected in stage 1-3 embryos. In addition, our analysis shows that this motif tends to be located close to the TSS (Figure S23). The DRE has previously been associated with genes involved in cell proliferation and DNA replication (Matsukage et al., 1995). Cell divisions in stages 1-3 embryos are likely dependent on the products of many of these genes. However, the mRNAs for these genes must be supplied maternally as the embryonic genome is transcriptionally silent (for the most part) during these early stages. Thus, the predicted motif may be involved in transcribing cell proliferation genes within the maternal nurse cells. The corresponding mRNAs would then be selectively shuttled into the oocyte. Another motif, [GT]C[AG]GGT[AT]G[ATG], is found mainly upstream of genes with a non-uniform expression pattern during stages 4-6 ('subset' category). This motif matches the consensus CAGGTAG element that was recently shown experimentally to enhance transcription of early zygotic genes (De Renzis et al., 2007). Other predicted motifs may intervene in later stages. For example, the [CG][ACG]CGATA[AG] motif is found in upstream regions of 35% of the 376 genes expressed in the embryonic midgut during stages 13-16, but only in 18% of promoters of the remaining 1620 genes (p<1e-11). Its sequence suggests that it is bound by a GATA-factor transcriptional regulator. Indeed, previous reports argue that GATA factors play a major role in *Drosophila* endodermal midgut specification (see (Murakami et al., 2005) for review).

Supplemental Experimental Procedures

In what follows, we discuss the FIRE methodology, algorithms, and features in more detail. Information regarding the gene expression data and sequence data analyzed in this work are available at the end of this document. The complete results, as well as all the relevant source code and software, can be downloaded at the FIRE web site at http://tavazoielab.princeton.edu/FIRE/.

Motif and expression profiles

Motif definition

In the current FIRE implementation, motifs are defined through *regular expressions* and using the standard degenerate code. Thus, a motif 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]). As an example, the motif A[CG]T has A at the first position, C or G at the second position, and T at the third position. Using regular expressions for defining motifs allows for a highly efficient search through motif-space. In addition, determining whether a motif is present within a given promoter is straightforward and requires no arbitrary thresholds.

Motif profile

In the following, we assume that we examine *N* genes where each one is associated with a single expression measurement (see below). In addition, for each of these genes, our data include its 5' upstream and 3'UTR sequences. For brevity, we focus here on 5' upstream regions, but the same applies to 3'UTRs. Given a motif, represented as a regular expression, the *motif profile* is defined as a binary vector with *N* elements, where for each gene, "1" indicates that the motif is present in the corresponding promoter and "0" indicates that it is absent. A motif is considered present in a promoter if the promoter contains at least one exact match to its regular expression. For DNA sequences, we consider occurrences of the motif on either strand. For RNA sequences, we only consider the transcribed strand.

Expression profiles

The *expression profile* is also defined as a vector with *N* elements. Each element corresponds to a gene, and indicates the associated expression value for that gene. The expression profile can be discrete or continuous. For example, a *discrete* expression profile can be obtained using the following procedure: we cluster the *N* genes based on multiple microarray data, associate an index to each cluster, and assign each gene to the index of the cluster it belongs to. A *continuous* expression profile may consist of the results of a single microarray experiment, where each gene is associated with a single expression value (*e.g.*, raw intensity value or log-ratio of intensities). Note that a continuous expression profile does not necessarily contain gene expression levels. For example, in the analysis of the *Plasmodium falciparum* data, the continuous expression profile indicates the "phase" of each gene, corresponding to a time point during the developmental cycle where the expression of this gene peaks (Llinas and DeRisi, 2004). Likewise, a discrete expression profile does not necessarily contain gene expression clusters defined from microarrays conditions. For example, it may often contain two categories, where one category consists of a set of genes of interest (*e.g.,* genes sharing the same spatio-temporal pattern of activity, as measured using *in situ* hybridization or genes defined using any *ad hoc* criteria), and the other consists of all remaining genes.

Quantizing continuous expression profiles

The concept of mutual information is well defined, both for continuous and for discrete random variables (Cover and Thomas, 2006). Nonetheless, in practice, estimating the information when continuous variables are involved requires quantizing their values. In this study, we quantize continuous expression profiles into equally populated bins, as described in (Slonim et al., 2005). In FIRE (by default), the number of bins, N_e, is determined using $N_e \cdot N_m \approx 50 \cdot N$, where N_m is the number of bins used in the motif profile, *i.e.*, N_m=2, and N is the total number of genes. This implies that the expected count within each entry of the joint-counts table created for the motif and the expression profiles (see below) is approximately 50, allowing for a relatively reliable estimation of the mutual information (Slonim et al., 2005).

Removing promoters and 3'UTRs of recently duplicated genes

Recently duplicated members of gene families or transposons (*e.g.*, Ty transposons in yeast) often share a significant amount of sequence identity in their promoters (and in their 3'UTRs). Their recently duplicated sequences also tend to cross-hybridize on certain high-throughput expression assays (*e.g.*, microarrays), and therefore often appear as co-expressed. When this occurs, multiple conserved sequences within the promoter of these genes will appear as highly correlated with the expression, constituting spurious motif predictions, as we observed in a preliminary yeast analysis. To address this issue (which is in fact relevant for all expression-based motif finding approaches), FIRE applies, by default, a simple duplicate removal procedure, which guarantees that within each expression category/bin, no pair of promoters and no pair of 3'UTRs will have a MegaBlast local alignment with E-value < 1e-10. In addition, prior to duplicate removal, repeats and low-complexity sequences are systematically masked using RepeatMasker (http://www.repeatmasker.org) and the appropriate species-specific repeat library from RepBase (Jurka et al., 2005).

Motif-expression information

Estimating the mutual information

In FIRE, we generally seek to evaluate whether a candidate motif is informative about the expression profile at hand. Given a motif profile (with two possible values, corresponding to presence and absence) and the expression profile (with N_e possible values), we first generate a joint-counts table, denoted as C, with 2 rows and N_e columns. $C(1,j)$ indicates the number of promoters which contain the motif and are associated with the j^{th} category/bin; $C(2,j)$ indicates the number of promoters which do not contain the motif and are associated with the j^{th} category/bin. The empirical mutual information between the presence/absence of the motif in a promoter and the expression of the corresponding gene, when averaging across all genes, is given by

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

where $P(i,j) = C(i,j)/N$, $P(i) = \sum_{j=1}^{Ne} P(i,j)$, and $P(j) = \sum_{i=1}^{2} P(i,j)$ (Cover and Thomas, 2006).

Evaluating the information significance via randomization tests

To estimate the statistical significance of observed empirical information values, non-parametric randomization tests are applied. Specifically, let I denote the obtained empirical mutual information, *e.g.*, between a given motif profile and the expression profile. Next, we randomly shuffle the expression profile and calculate the information value between the unchanged motif profile and the shuffled expression profile. We repeat the same procedure N_r times to obtain N_r random information values, and consider the original empirical information, I, as statistically significant (with $p < (1/N_r)$), if and only if it is greater than *all* N_r random information values. As detailed below, different N_r values are used by default, depending on the context and on the number of hypotheses tested. In addition, a corresponding Z-score is reported, defined as $Z=(I-<I_{random}>)/\sigma_{random}$ where $<I_{random}>$ is the average random information value and σ_{random} is the corresponding standard deviation. This Z-score is often useful in comparing motifs that pass the randomization test, as it reflects how far the empirical information is, in number of standard deviations, from the average random information. However, we do not use Z-scores directly to determine the significance of mutual information values as the underlying distribution of random information values is not a normal distribution.

Evaluating the information robustness

Another important non-parametric statistical significance test incorporated in FIRE is based on jack-knife re-sampling (Efron, 1979). Specifically, for each predicted motif, N_j jack-knife trials are applied where in each trial a substantial fraction (one third by default) of the genes is randomly removed from the data. An information value is recalculated based on the remaining data, and its statistical significance is evaluated using the randomization test described above (with N_r=10,000 repeats, by default). The *robustness* score of the motif indicates in how many of these jack-knife trials the motif information was found to be statistically significant. By default, we use N_j=10, hence the robustness scores range from 0/10 up to 10/10.

Revealing highly informative motifs

Detecting motif seeds

Finding motifs whose profiles are highly informative about a given expression profile can be approached through different search strategies. The two-step procedure currently implemented in FIRE is reminiscent of procedures used by other motif finding techniques (*e.g.,* (Foat et al., 2005)). Nonetheless, it is used here to optimize an entirely different target function, namely the information between the predicted motifs and expression. The first step amounts to scoring an exhaustive list of simple motif definitions in the form of *k*-mers (non-degenerate sequences of *k* nucleotides), where by default *k* is set to 7; thus, all 7-mers are examined, resulting in a coarse-grained, yet exhaustive exploration of motif space.

When searching for DNA motifs (typically in upstream promoters), reverse complements are removed from the list of all 7-mers as FIRE considers both matches to a motif and to its reverse complement; therefore, effectively, 8,192 7-mers are examined. When searching for RNA motifs (typically in 3'UTRs), FIRE examines all 16,384 7-mers. For each 7-mer, the mutual information between its profile and the expression profile is evaluated. All 7-mers are then sorted based on their information values and a simple and efficient algorithm is used to search for the first 10 consecutive 7-mers whose information is *not* significant, within the sorted list. All 7-mers sorted above these 10 are retained for further analysis, and are henceforth termed motif *seeds*. Recall, that the information associated with a particular 7-mer is considered significant if and only if it passes the randomization test, *i.e.*, if it is greater than all N_r random information values obtained for this 7-mer profile over N_r randomly shuffled expression profiles. To correct for multiple hypothesis testing, N_r is set by default to the number of *k*-mers initially examined, *i.e.,* N_r = 8,192 when searching for DNA motifs and N_r = 16,384 when searching for RNA motifs.

Optimizing seeds into more informative motifs

It is now fully established that DNA- and RNA-binding proteins typically can bind (with different affinities) to multiple slightly distinct sites. Therefore, motif definitions that can capture multiple sites simultaneously are likely to more accurately represent binding sites for these proteins. As mentioned above, to address this problem, motifs are represented in FIRE using the standard degenerate code. The second search stage in FIRE consists of an optimization process that gradually converts the seeds obtained at the previous stage into longer and potentially degenerate motifs that convey more information about the expression profile.

All seeds obtained in the previous stage are sorted based on their information values, and are examined one after the other, starting with the most informative one. If a seed corresponds to a variant of a motif obtained from optimizing previous – more informative – seeds, it is discarded (see below). Otherwise, it is optimized using the following procedure. First (by default), a single position is added to each side of the seed, initialized to the non-informative N character. Thus, the examined motif is now 9 nucleotides long. A single position among these 9 is chosen at random, and all characters of the degenerate code that are consistent with the seed's initial character at that position are tested. For example, if the seed has an A at that position, a new character is selected from [AC], [AG], [AT], [ACG], [ACT], [AGT], and N. Each of these 7 alternatives induces a possibly different motif profile with corresponding information over the expression profile. Among the permitted alternatives (see below) the one that results in a maximally informative motif is selected. This procedure is repeated until convergence, namely, until no further improvements are possible, at all 9 positions. Due to its greedy nature, this process may converge to a local maximum of the information. Thus, the entire optimization is repeated 10 times per motif, ending up with possibly 10 (slightly) different motifs, of which the most informative one is retained.

Avoiding redundant/degenerate output

As mentioned above, multiple seeds may often represent different variations over the binding site of a single protein. To avoid redundant output, FIRE requires that each reported motif provides some novel information over the expression profile while being relatively independent of motifs obtained from optimizing previous seeds. Formally, this is implemented by 1) avoiding the optimization of some seeds and discarding these seeds from the list of predicted motifs, and 2) when optimizing a seed, distinguishing between permitted versus non-permitted characters at each optimization step. 1) and 2) are implemented using the same mechanism. Each time a new motif is considered (either a seed is considered for optimization, or a new character at a given position is considered during optimization), it is accepted for further consideration if and only if it satisfies:

$$I\,(\,motif\,;\,expression \mid prev\_motif\,)\,/\,I\,(\,motif\,;\,prev\_motif\,) > r,$$

where *I(motif ; expression | prev_motif)* is the *conditional information* (Cover and Thomas, 2006) conveyed by this motif profile over the expression profile, given the profile of a previous motif denoted *prev_motif*; *I(motif ; prev_motif)* is the information between the motif profile and the profile of the previous motif; and *r* is a tradeoff parameter. Importantly, this inequality must be satisfied with respect to *all* motifs obtained from optimizing previous seeds, suggesting that the considered motif provides some novel information over expression (*i.e., I(motif ; expression | prev_motif)* is not too low), while, at the same time, showing relatively little dependency with previous motifs (*i.e.*, *I(motif ; prev_motif)* is not too high). The tradeoff parameter, *r*, can be seen as a "knob" that controls the level of output redundancy. Since seeds not satisfying this inequality (with respect to all previous motifs) are discarded, high *r* values will typically result in very few motifs that are clearly different from one another, while low *r* values will typically result in a larger number of predicted motifs and greater redundancy. In practice, it is recommended to explore different *r* values. However, in order to avoid over-fitting our data, we used a single value (*r*=5) in all the FIRE runs reported in this manuscript, chosen based on preliminary tests in yeast.

Reporting only significant and robust motifs

After optimization, each motif is subjected to a randomization test with N_r=10,000 and motifs that do not pass this test are discarded. In addition, each motif is assigned a robustness score, calculated as explained above using N_j=10 and N_r=10,000. By default, FIRE only reports motifs with a robustness score of at least 6/10, but exploring other values is also recommended. For example, in the analysis of the *Drosophila in situ* hybridization data, only motifs with a robustness score of at least 8/10 are reported.

Post-processing: characterization of predicted motifs

Predicting orientation bias

Given a predicted motif, FIRE further examines two binary profiles, each with N elements, termed here the *transcribed strand motif profile* and the *non-transcribed strand motif profile*. In the former, "1" indicates that the motif is present in the promoter on the transcribed strand (*i.e.*, the same strand as the transcribed gene) and "0" indicates otherwise; in the latter, "1" indicates that the motif is present in the promoter on the non-transcribed strand and "0" indicates otherwise. FIRE evaluates the information conveyed by each of these two profiles over the expression profile, assesses their significance using N_r=10,000 randomization tests, and reports an *orientation bias* if and only if only one of these two information values is found to be statistically significant. As RNA motifs can only be on the transcribed strand, the default FIRE setting is to report an RNA motif only if it has a significant orientation bias in the right direction, *i.e.*, only if the transcribed strand motif profile is significantly informative over the expression profile, while the non-transcribed profile is not. This requirement may prevent contaminations from downstream DNA motifs which are less likely to have an orientation bias. Nonetheless, in many cases (*e.g.* for the yeast gene clustering partition), we have observed that all or most RNA motifs predicted by FIRE have a highly significant orientation bias.

Patterns of motif over- and under-representation

Highly informative motifs are generally over- or under-represented in the promoters associated with certain categories/bins. We quantify this using the binomial distribution. Specifically, let N be the total number of promoters (or genes), n the total number of promoters in which the given motif is present, K the number of promoters within a particular category/bin, and x the overlap, namely the number of promoters in this category/bin in which the given motif is present. Then, the probability of observing x genes (or more) with the motif in that category/bin, under the null hypothesis that the motif is distributed across promoters independently of the expression profile, is given by

$$P(X \geq x) = \sum_{t=x}^{K} \binom{K}{t} f^{t} (1-f)^{K-t}$$

where $f=n/N$. We consider that the motif is *over-represented* in this category/bin if and only if $P(X \geq x) < 0.05/N_e$, where N_e is the number of categories/bins in the expression profile, used as a Bonferroni correction for multiple hypothesis testing. We consider that the motif is *under-represented* in that category/bin if and only if $P(0 \leq X \leq x) < 0.05/N_e$ where $P(0 \leq X \leq x)$ is calculated using the same formula as above.

In many of the tests described below, it is useful to distinguish between motif occurrences that are more likely to represent functional binding sites versus occurrences that are more likely to represent non-functional sites (*e.g.*, due to being located in nonaccessible DNA regions or due to any other reason). To address that, we henceforth refer to motif occurrences in categories/bins in which the motif is over-represented as *active*, and to all other occurrences as *non-active*.

Predicting position bias

For each predicted motif, FIRE examines the subset of promoters in which it is present. Assuming there are n_m such promoters in our data, FIRE generates a *position profile* for this motif with n_m elements, each indicating (for each of these promoters) the average distance between all motif occurrences and the TSS. This position profile is quantized into N_b equally populated bins and FIRE evaluates the empirical information between the quantized profile and a binary version of the original expression profile with the same n_m genes (termed *binary expression profile*), defined as "1" for categories/bins in which the motif is over-represented, and "0" otherwise. Thus, the obtained information quantifies how informative the motif's relative position in the promoter is, regarding the distinction between its "active" versus its "non-active" occurrences. A randomization test (as described above) with N_r=10,000 randomizations of the binary expression profile is applied to determine the significance of this information. This is repeated for N_b=2,3,4,5, and a *position bias* is reported if and only if, in at least one of these four trials, the obtained information is ranked in the top percentile of the 10,000 random information values. This significance test is less stringent than the one used to predict motifs, as it merely serves to point out possibly interesting trends in the motif relative position. Also, notice that Data Processing Inequality (Cover and Thomas, 2006) implies that, if the relative position of the motif is informative about the binary expression profile, it is also informative about the original expression profile.

Predicting functional interactions

Putative functional interactions between pairs of motifs are predicted by FIRE by asking whether the presence of one motif in a promoter is informative about the presence of another motif. Given two predicted motifs, FIRE generates a filtered version of each of their profiles where promoters with "non active" motif occurrences (*i.e.*, not in categories/bins where the motif is over-represented) are assigned a "0" instead of a "1". The mutual information between the resulting filtered motif profiles, termed here the *interaction information*, is evaluated, and a randomization test with N_r=10,000 repeats is used to determine its significance. An interaction is predicted if and only if the interaction information is greater than all 10,000 random information values. Note that this information may be significant also due to a negative correlation between the two motifs, namely, when the presence of one motif implies the absence of the other. These cases are reported as well, but are distinguished from the positive correlation cases in the FIRE interaction heat-map (see below).

Modules of predicted motifs are generated as follows. All motifs are sorted based on the significance of their information values over the expression profile, *i.e.*, based on their Z-scores, and are partitioned into *assigned* versus *non-assigned* motifs, where originally the set of assigned motifs is empty while the set of non-assigned motifs consist of all predicted motifs. Next, the most significant motif in the non-assigned set is picked as the core of a new module and is moved to the assigned set. All motifs in the non-assigned set are examined in the order of their Z-scores, and a motif is added to the new module (and correspondingly moved to the assigned set) if and only if it has significant interaction information due to a positive correlation with each and every motif already present in that module. This is repeated until the non-assigned set is empty. By construction, in each of the resulting modules, all motifs are significantly informative about one another due to a positive correlation.

Predicting motif co-localization

If the interaction information between two predicted motifs is found to be significant due to a positive correlation, FIRE further examines whether these two motifs tend to co-localize when both are present within the same promoter. Specifically, given all promoters in which both motifs are present, the *intersection expression profile* is defined as “1” if both motifs have “active” occurrences within the promoter, and “0” otherwise. The *co-localization profile* is defined as a continuous profile that indicates the minimal distance between both motif occurrences within each promoter in which both are present. This profile is then quantized into N_b equally populated bins and the information between the quantized profile and the intersection expression profile is evaluated. Intuitively, this information quantifies how informative the minimal distance between the two motifs is over the distinction between promoters were both motif occurrences are “active” versus the remaining promoters. A randomization test with N_r=10,000 repeats is applied to determine the significance of this information. This is repeated for N_b=2,3,4,5, and a *co-localization* is reported if and only if the obtained information is ranked in the top percentile of the 10,000 random information values in at least one of these four trials.

Gene Ontology analysis

We define the *target genes* of a predicted motif as all genes whose promoter contain the motif and are associated with a category/bin where the motif is over-represented. In other words, these are the genes whose promoters contain the “active” motif occurrences. For the species discussed in this article, for each predicted motif, FIRE automatically determines whether its target genes significantly overlap with any Gene Ontology (GO) category, as significant overlaps may hint at the biological role of this motif. The overlap significance is determined using the hyper-geometric distribution (Tavazoie et al., 1999), and a motif is defined as enriched with a particular GO category if and only if the associated *p*-value is smaller than 0.05, after correcting for multiple hypothesis testing (using the number of GO categories tested as a Bonferroni correction). The results are reported through an automatically generated table (*e.g.*, Table S1). A similar analysis is done for each category/bin of genes within the expression profile, and is reported on top of each column in the FIRE p-value heat-map (*e.g.*, Figure 3).

Assessing motif conservation

For the species discussed in this article, FIRE automatically determines a *conservation index* for each predicted motif. First, the network-level conservation score of a predicted motif is calculated with respect to a different related genome, as described in (Elemento and Tavazoie, 2005; Pritsker et al., 2004). Briefly, only genes with non-ambiguous orthologs in both species are considered; the sets of genes, in each species, bearing the motif in their promoter region (or 3'UTR) are determined, and the conservation score is defined as the negative logarithm of the hyper-geometric *p*-value for the overlap between these two sets (see (Elemento and Tavazoie, 2005) for more details). This score is compared to the scores obtained for all 7-mers (8,192 for promoters, 16,384 for 3'UTRs), and the motif conservation index is defined as the fraction of these 7-mers with a smaller conservation score. Thus, a conservation index of 1.0 implies that the motif is more conserved than all 7-mers, according to this evaluation procedure.

Comparing predicted motifs to known regulatory elements

Motifs predicted by FIRE are systematically compared to known regulatory elements, when relevant data are available. While analyzing yeast data, predicted motifs are compared to a database compiled from diverse sources (literature, high-quality ChIP-chip assays, etc.), originally used in (Pritsker et al., 2004). While analyzing data for worm, fly, mouse, and human, predicted motifs are compared to TRANSFAC (Matys et al., 2006) (release 4) and JASPAR (Vlieghe et al., 2006). To assess whether a predicted motif matches a known regulatory element, we use CompareACE (Hughes et al., 2000) to find the Pearson correlation between weights of the position weight matrix associated with each regulatory element in these databases, and the position weight matrix representing the predicted motif (where A is converted to [1,0,0,0], [AG] to [0.5,0,0.5,0], etc.). The name of the known regulatory element that is most similar to the predicted motif is reported, assuming that their similarity, *i.e.*, their Pearson correlation, is greater than the stringent cutoff score of 0.8.

Comparing predicted RNA motifs to miRNA targets

It is now widely assumed that metazoan miRNAs typically bind their target 3'UTRs through exact or near-exact complementarity between their 5' extremity and the 3'UTRs. Therefore, in FIRE, predicted metazoan RNA motifs are systematically compared to the 5' extremity of known miRNA sequences within the corresponding species. Similar to what CompareACE does, we try to find a 7-nucleotide (or more) exact match between the motif and the reverse complement of the first 8-nucloetides within the miRNAs. For each motif, we count the number of miRNAs that the motif matches. However, as motifs predicted by FIRE may be highly degenerate and match a large number of miRNAs we further apply a simple control where we shuffle the miRNA sequences 100 times and repeat the analysis. A motif is reported to match miRNAs only when it matches significantly more real miRNAs than randomized miRNAs, *i.e.,* if the number of matched real miRNAs is greater than 3 standard deviations above the average number of matched randomized miRNAs.

Automatically generated FIRE figures

FIRE p-value heat-map and FIRE density heat-map

The FIRE *p-value heat-map* is automatically generated and summarizes the most important results in a graphical concise manner. An example for yeast is given in Figure 3. The rows in this heat-map correspond to all predicted motifs while the columns correspond to expression categories/bins (gene expression clusters in Figure 3). By default, only categories/bins in which at least one motif was found to be over- or under-represented are shown. Motifs are partitioned into modules as described above, and modules are separated by white horizontal bars. For each category/bin of genes, the most highly enriched GO annotation is reported. The yellow color-map indicates (in a log10 scale) the over-representation *p*-value (after Bonferroni correction) of a motif in a category/bin where significant events ($p<0.05$) are marked by red frames. For presentation purposes, *p*-values smaller than 1e-20 are set to 1e-20. The blue color-map indicates (in a log10 scale) under-representation *p*-values (after Bonferroni correction) and significant events ($p<0.05$) are marked by blue frames, where again p<1e-20 values are set to 1e-20. Motif logos are used to represent the predicted motifs (after regular expressions are turned into weight matrices, as described above). For each motif, its mutual information, Z-score, robustness, conservation index, the most similar known motif, and the seed that gave rise to it, are also indicated. In addition, predicted position biases are indicated by "Y" and orientation biases in favor of the transcribed or the non-transcribed strand are indicated by "→" or "←", respectively. The FIRE *density heat-map* (*e.g.*, Figure S4) has the same format, but instead of reporting *p*-values of over/under-representation, indicates the actual fraction of promoters within each category/bin that contain each motif.

FIRE interaction heat-map

The FIRE *interaction heat-map* (*e.g.*, Figure 4) is automatically generated to highlight putative functional relations between predicted motifs. The light (yellow) color map indicates the interaction information between each pair of motifs when this information is due to a positive correlation. The dark (red) color-map indicates the interaction information between each pair of motifs when this information is due to a negative correlation. Putative functional modules are separated by black lines. Significant interaction information values (p<1e-4) that involve two DNA motifs and two RNA motifs are marked by blue and pink frames, respectively. Significant interactions that involve a DNA motif and an RNA motif are marked by green frames. Significant co-localization events are marked by "+".

FIRE motif maps

When a position bias or an orientation bias is observed for a predicted motif, FIRE automatically generates a corresponding *motif map* figure that highlights the nature of the observed bias (*e.g.*, Figure S6). This figure depicts all promoters (or 3'UTRs) in which the motif is present and all occurrences of the motif within each of these promoters.

Genes in the expression bin/category where the motif is most over-represented are shown at the top of the figure. Other bins/categories are subsequently shown in descending order of over-representation significance. Motif maps are also generated for pairs of motifs which avoid being in the same promoters (*e.g.*, Figure S10), or which co-occur. The latter includes cases where motifs co-localize on the DNA or the RNA (*e.g.*, Figure S21).

FIRE text report files

In addition to the above figures, FIRE also generates by default text files that are aimed at facilitating experimental follow-ups. In particular, all occurrences of each predicted motif are reported, along with the corresponding gene, sequence context, position within the promoter/3'UTR, strand (for DNA motifs), *etc*. Promoters in which the predicted motif is present are sorted according to a combined 'putative functionality' score that reflects whether the motif is over-represented within the expression category/bin the gene belongs to, and whether it is located on the preferred strand and/or at the preferred distance from the TSS (or stop codon for RNA motifs).

Modular implementation and command lines

A modular implementation

The relevant FIRE software is implemented via several modules than can be used independently. For example, given expression data and a set of predicted motifs *not* obtained by FIRE analysis, but rather from any other source (*e.g.*, experimentally validated motifs), it is straightforward to generate figures like the FIRE *p*-value heat-map or the FIRE interaction heat-map, in order to highlight various aspects related to these motifs in the context of the available expression data. See the FIRE Web site for more details.

Executing FIRE

For all the species discussed in this article, executing FIRE with default parameters involves a simple command line:

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

where <sp> indicates the species, <inp> indicates the input expression profile (see below), 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). For example, the following command line will reproduce our results for the yeast clustering partition (yeast_gasch_IclustPos.txt is available on our Web site):

```
perl fire.pl --species=yeast --expfile=yeast_gasch_IclustPos.txt --exptyspe=discrete
```

The FIRE program, documentation and all results presented in this article can be downloaded from http://tavazoielab.princeton.edu/FIRE/. A preliminary Web interface to FIRE is also available at http://quantbio-tools.princeton.edu/cgi-bin/FIRE/form.pl.

Captions for Supplemental Figures

Figure S1: Informative co-occurrence and co-avoidance.

Mutual information is also used to characterize the level of interaction between predicted motifs. A functional interaction between motifs is predicted when the presence of one motif in the promoter implies the presence of the other motif (left panel), leading to a significant information signal between the two motif profiles. However, mutual information can also capture other dependencies, *e.g.*, scenarios when the presence of one motif implies the absence of another motif (right panel). In addition, the same concept can be used to capture dependencies between DNA motifs and RNA motifs, pointing to possible interactions between transcriptional and post transcriptional processes.

Figure S2: Informative co-localization.

The concept of mutual information is further used to capture significant co-localization of pairs of motifs. Specifically, if the combination of two motifs is functional only when both are located in the same vicinity, we expect that the minimal distance between the two motifs will convey significant information over the expression data being analyzed. In the depicted example, the two motifs tend to co-localize among promoters of genes within cluster 2.

Figure S3: Schematic overview of the optimization procedure.

Starting from a candidate motif (seed), our optimization procedure uses a greedy algorithm to explore the surrounding motif space, in search for a more informative and possibly more general motif representation. This optimization is constrained in order to avoid being attracted to highly informative motifs obtained by optimizing previous seeds. Specifically, at each optimization step, we require that the examined motif provides a significant amount of novel information over the expression, given all already optimized motifs at this point. See the Supplementary Methods Section for more details.

Figure S4: Density heat map for motifs predicted for yeast gene clustering partition.

The format of this figure is identical to the format of Figure 3. However, this heat map represents motif densities within each cluster. The density is defined as the fraction of promoters in each cluster in which the motif is present at least once. Significant over- and under-representation events detected through p-value estimation (Figure 3) are highlighted using red and blue frames, respectively.

Figure S5: Gaining information through optimization – results for yeast.

This figure depicts the information initially conveyed by each seed (darker colors), and the additional information gained through optimizing this seed (lighter colors). DNA motifs are depicted in red color, while RNA motifs are depicted in green.

Figure S6: Motif map for the predicted yeast motif matching PAC.

Position and orientation of the predicted motif that matches the PAC binding site are presented for all *S. cerevisiae* 600bp upstream regions in clusters c66 and c8, in which this motif is most over-represented, and in other clusters (below the thick black line), in which the motif is not over-represented. The position bias detected by FIRE reflects a strong tendency of this motif to be located towards the beginning of the gene (the ATG codon). The full motif map, with all occurrences of this motif, is available at our Web site and illustrates how this tendency is much more dominant in clusters in which the motif is over-represented, leading to a significant information signal between motif positions and cluster indices.

Figure S7: Motif map for the predicted yeast motif matching RRPE.

Position and orientation of the predicted motif that matches the RRPE binding site are presented for all *S. cerevisiae* 600bp upstream regions in clusters c66 and c8, in which this motif is most over-represented, and in other clusters (below the thick black line), in which the motif is not over-represented. The position bias detected by FIRE reflects a strong tendency of this motif to be located towards the beginning of the gene (the ATG codon), although not as close as the motif matching PAC (Figure S6). The full motif map, with all occurrences of this motif, is available at our Web site and illustrates how this tendency is much more dominant in clusters in which the motif is over-represented.

Figure S8: Motif map for the predicted yeast motif matching Rap1 binding site.

Position and orientation of the predicted motif that matches the Rap1 binding site are presented for all *S. cerevisiae* 600bp upstream regions in cluster c9 and in several other clusters. The motif is over-represented in c9 but not in the other clusters. The position bias detected by FIRE is different from the position biases of the motifs matching PAC (Figure S6) and RRPE (Figure S7). Indeed, this motif seem to be preferentially located in an interval ranging from -200bp to -500bp. This preferential location, and in particular its dominance in cluster c9, is reflected by a statistically significant information between motif position and cluster indices.

Figure S9: Motif map for the predicted yeast motif matching Puf3 binding site.

Position and orientation of the predicted RNA motif that matches the Puf3 binding site are presented for all *S. cerevisiae* 300nt 3'UTRs in clusters c53 and c70, in which this motif is most over-represented, and in other clusters (below the thick black line), in which the motif is not over-represented. The position bias detected by FIRE reflects a strong tendency of this motif to be located close to the stop codon. Notice, however that a position bias observed for an RNA motif may also reflect a correlation between the actual 3'UTR length and expression, as opposed to a biophysical constraint over motif location.

Figure S10: Combined motif map for the predicted yeast motifs matching PAC and Msn2/4 binding sites.

Position and orientation of the predicted motifs that match PAC and Msn2/4 binding sites are presented for all *S. cerevisiae* 600bp upstream regions in clusters c8, c43, and c66. PAC is most over-represented in c8 and c66, while Msn2/4 is most over-represented in c43. As detected by FIRE, both motifs are typically not found within the same promoter, leading to statistically significant information, where the presence of one motif implies the absence of the other.

Figure S11: All predicted DNA and RNA motifs for the human genes clustering partition.

17,390 human genes were clustered based on the human tissue expression data in (Su et al., 2004) and the obtained clustering partition was analyzed by FIRE. All predicted motifs are presented in the figure. The format of this figure is identical to the format of Figure 3. Motif names are based on the closest motif in JASPAR or TRANSFAC (with compareACE score > 0.8). MiRNAs whose 5' extremity matches 3'UTR elements with high specificity are also reported (see Supplementary Methods).

Figure S12: Predicted interactions among predicted human motifs.

17,390 human genes were clustered based on the human tissue expression data in (Su et al., 2004) and the obtained clustering partition was analyzed by FIRE. The interactions between all predicted motifs are presented in the figure. The format of this figure is identical to the format of Figure 4.

Figure S13: Motif map for the predicted human motif matching the NF-Y binding site.

Position and orientation of a predicted motif matching the NF-Y binding site are presented for all human 1000bp upstream regions in cluster c114, in which this motif is most over-represented, and in other clusters (below the thick black line), in which the motif is not over-represented. The position bias detected by FIRE reflects a strong tendency of this motif to be located close to the TSS.

Figure S14: Average expression patterns of human genes in tissue specific clusters.

Genes within cluster c0 are highly expressed almost exclusively in adult and fetal liver. Genes within cluster c112 are highly expressed in heart, skeletal muscle, and tongue tissues. Motifs associated with these two clusters are described in the main text.

Figure S15: number of predicted motifs by FIRE (DNA+RNA) when analyzing the Gasch *et al*. yeast stress dataset, at different robustness index thresholds.

The figure also shows the number of motifs supported by GO functional categories (red), the number of motifs matching known motifs (green), and the expected number of false positives calculated from 100 shuffled clustering partitions (blue).

Figure S16: All predicted DNA and RNA motifs for worm genes clustering partition.

11,562 worm genes were clustered based on the expression data in (Kim et al., 2001) and the obtained clustering partition was analyzed by FIRE. All predicted motifs are presented in the figure. The format of this figure is identical to the format of Figure 3. Motif names are based on the closest motif in JASPAR or TRANSFAC (with compareACE score > 0.8). MiRNAs whose 5' extremity matches 3'UTR elements with high specificity are also reported (see Supplementary Methods).

Figure S17: Gaining information through optimization – results for worm.

This figure depicts the information initially conveyed by each seed (darker colors), and the additional information gained through optimizing this seed (lighter colors). DNA motifs are depicted in red color, while RNA motifs are depicted in green.

Figure S18: Predicted interactions among all predicted worm motifs.

11,562 worm genes were clustered based on the expression data in (Kim et al., 2001) and the obtained clustering partition was analyzed by FIRE. The interactions between all predicted motifs are presented in the figure. The format of this figure is identical to the format of Figure 4.

Figure S19: Predicted DNA and RNA motifs for worm genes clustering partition.

11,562 worm genes were clustered based on the expression data in (Kim et al., 2001). The format of this figure is identical to the format of Figure 3. Due to space limitations, only a selection of the predicted motifs is presented. The complete figure is given as Figure S15. DNA motif names are reported based on the closest known motif in JASPAR or TRANSFAC, with compareACE score > 0.8. MiRNAs whose 5' extremity matches 3'UTR elements with high specificity are also reported (see Supplementary Methods).

Figure S20: Motif map for a predicted E-box-like worm motif.

Position and orientation of a predicted E-box like motif are presented for all *C. elegans* 1000bp upstream regions in clusters c30 and c103, in which this motif is most over-represented, and in other clusters (below the thick black line), in which the motif is not over-represented. The position bias detected by FIRE reflects a strong tendency of this motif towards the TSS.

Figure S21: Combined motif map for two predicted co-localizing worm motifs.

Position and orientation of two predicted motifs within *C. elegans* 1000bp upstream regions in which both motifs are present. In c63, where both motifs are over-represented, their minimal distance is almost always exactly 3 nucleotides, leading to statistically significant information between this minimal distance and the cluster indices, detected by FIRE.

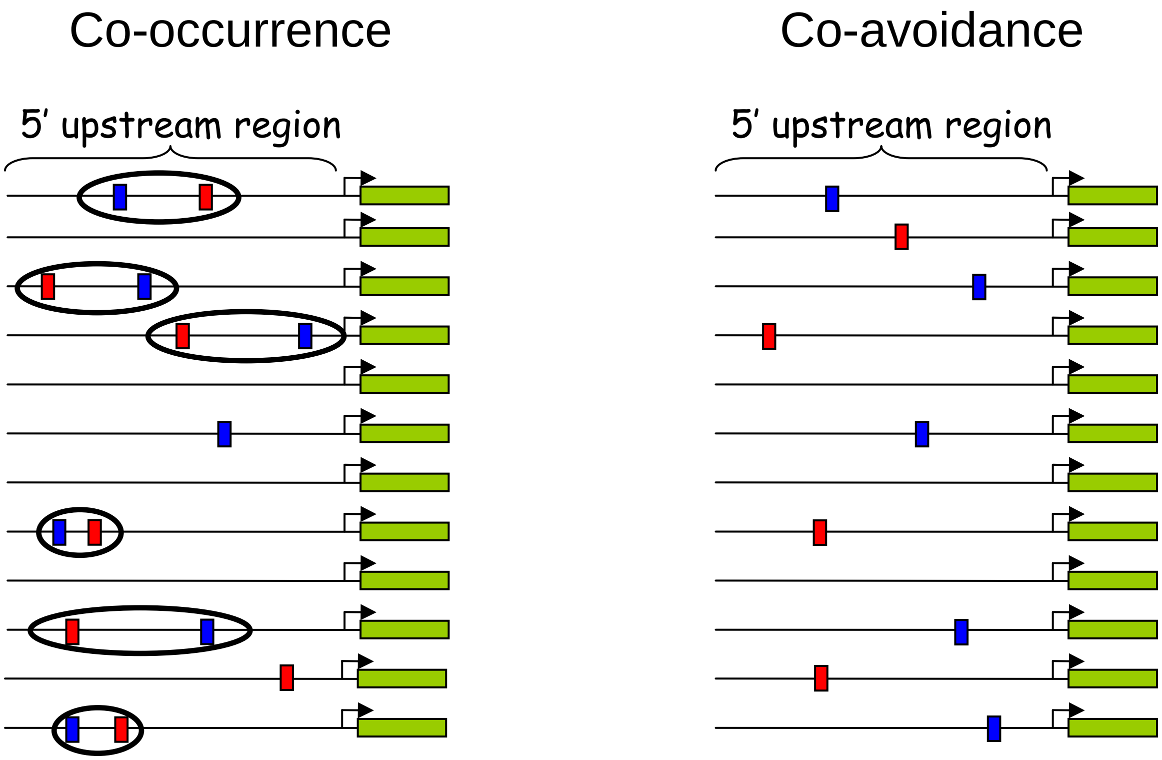
Figure S22: Predicted motifs for *Drosophila in situ* hybridizations data.

FIRE was applied independently to 55 binary expression profiles, which describe whether genes are expressed (or not) in a specific tissue and at a specific stage during *Drosophila* embryogenesis. All motifs predicted within these 55 runs were clustered based on their pairwise CompareACE scores, and the motif with the maximal Z-score within each cluster was chosen as the cluster representative. A selection of these representatives is shown in the figure that indicates the information robustness scores obtained for these motifs for 8 different spatio-temporal expression profiles.

Figure S23: Motif map for the predicted fly motif matching the DRE element.

Position and orientation of the predicted motif that matches the DRE element are presented for all *D. melanogaster* 1000bp upstream regions in clusters c1 and c0. The expression profile analyzed here only contains two clusters, with genes in c1 showing a relatively uniform expression during stages 1-3 in the embryo (implying these are maternal genes). Expression of genes in c0 is not detected at stages 1-3. The position bias detected by FIRE reflects a strong tendency of this motif to be located towards the TSS, especially for genes in c1.

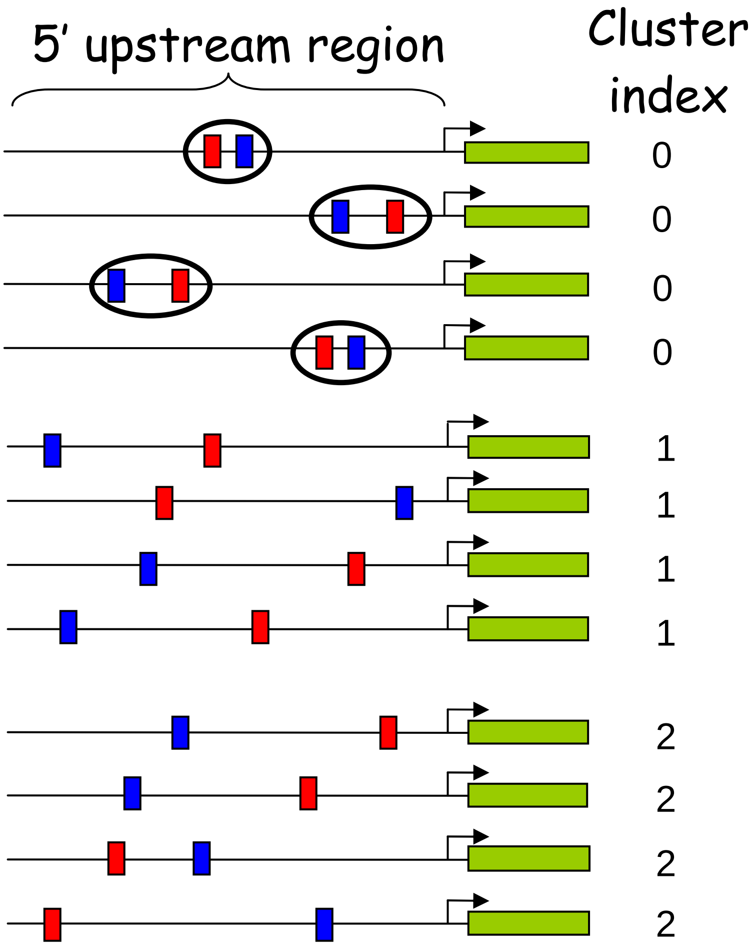
Figure S1



Co-occurrence

Co-avoidance

Figure S2



Co-localization

Figure S3

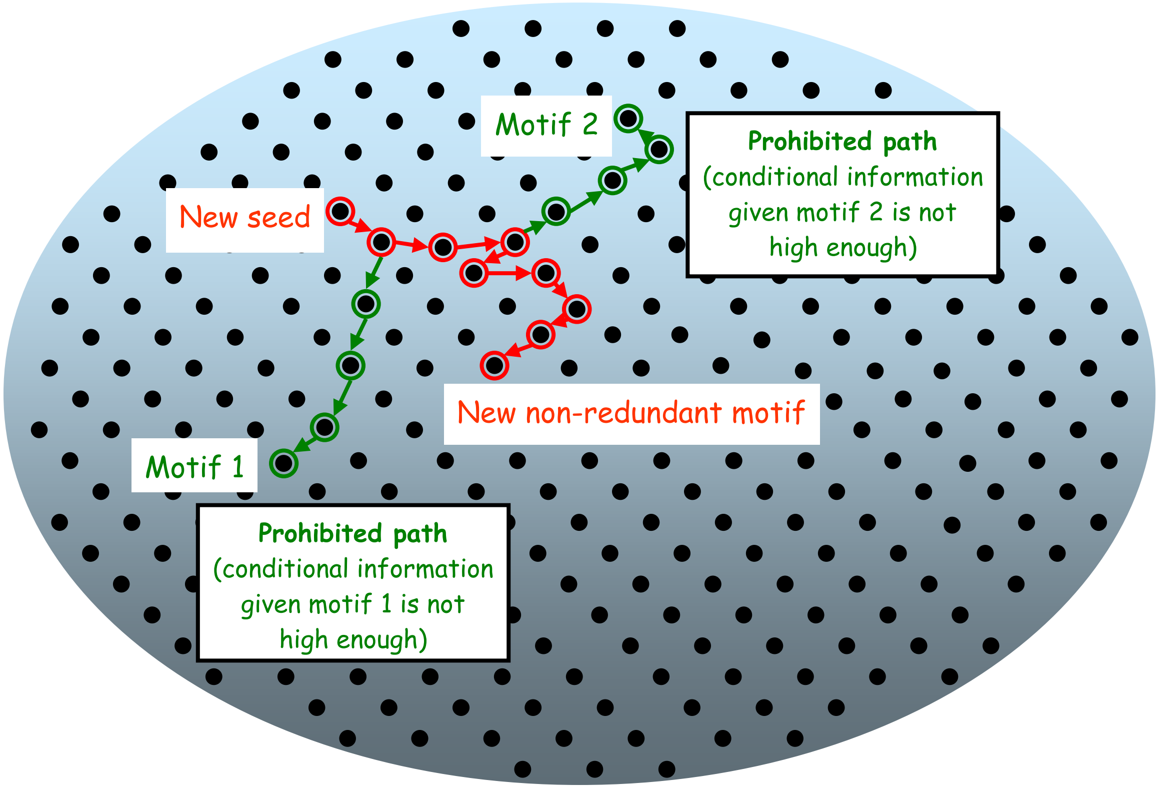
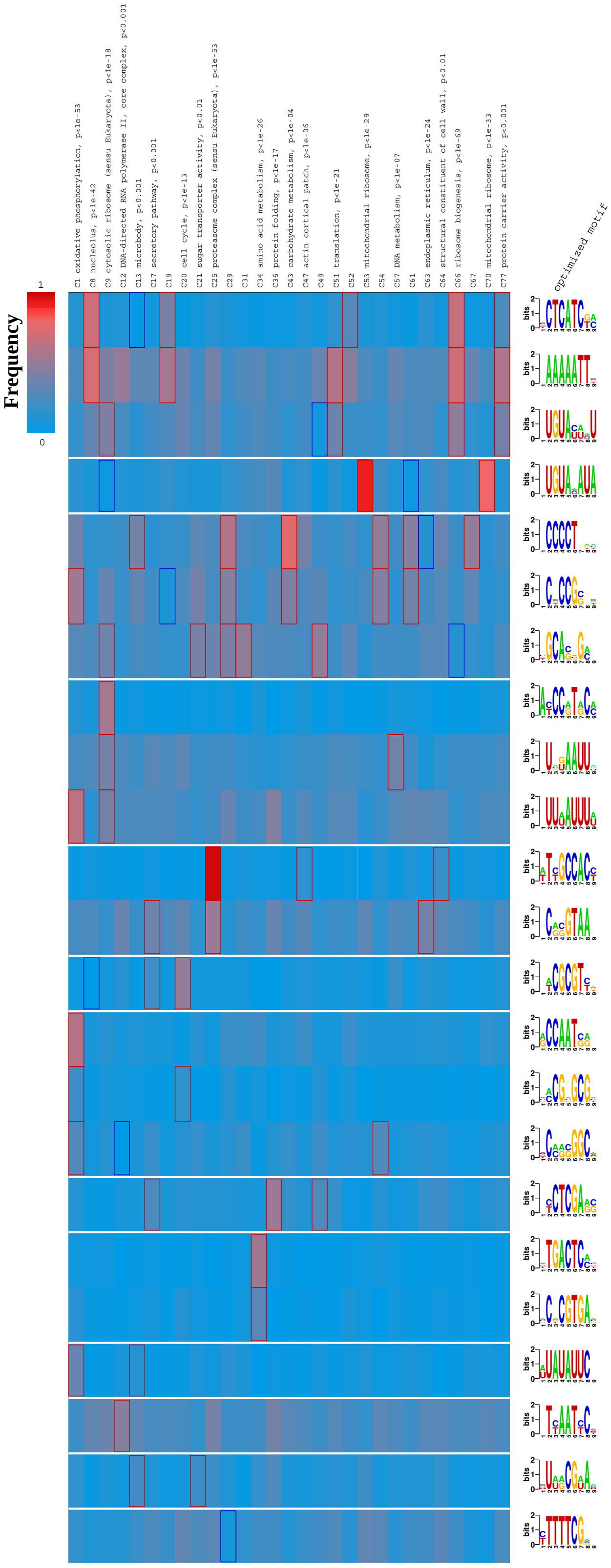

Motif 2
Prohibited path
(conditional information
given motif 2 is not
high enough)
New seed
New non-redundant motif
Motif 1
Prohibited path
(conditional information
given motif 1 is not
high enough)

Figure S4



location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5’	0.081	45.0	10/10	Y	-	1.00	CTCATCG	PAC
5’	0.042	20.9	10/10	Y	-	1.00	AAAAATT	RRPE
3’UTR	0.032	14.7	10/10	Y	→	1.00	UGUAUAA	PUF4
3’UTR	0.055	29.7	10/10	Y	→	1.00	UGUAAAU	PUF3
5’	0.047	24.0	10/10	Y	-	1.00	CCCCTTA	MSN24
5’	0.029	12.6	10/10	Y	-	1.00	CTCCGCA	-
5’	0.025	9.8	9/10	-	→	0.98	GCACAGC	-
5’	0.043	20.4	10/10	Y	→	0.99	TCCGTAC	RAP1
3’UTR	0.025	9.8	9/10	Y	→	0.98	UAUAAUU	-
3’UTR	0.024	9.7	9/10	-	→	1.00	UUUAUUU	-
5’	0.040	18.7	10/10	Y	-	1.00	TTGCCAC	RPN4
5’	0.020	6.7	7/10	-	-	1.00	CGGGTAA	REB1
5’	0.028	11.8	10/10	-	-	1.00	ACGCGTC	MBP1
5’	0.028	11.4	10/10	Y	→	0.99	CCAATCA	HAP4
5’	0.024	8.6	8/10	-	←	0.96	ACGCGCG	-
5’	0.022	7.6	7/10	-	-	0.97	CCACGGC	-
5’	0.024	8.8	8/10	Y	→	1.00	TCTCGAA	XBP1
5’	0.023	7.9	6/10	Y	→	0.98	TGACTCA	BAS1
5’	0.022	7.8	9/10	-	→	1.00	CACGTGA	CBF1
3’UTR	0.021	7.1	8/10	-	→	1.00	UAUAUUC	-
5’	0.021	7.2	7/10	-	-	0.95	TCAATTC	-
3’UTR	0.021	7.2	6/10	-	→	0.98	UUACGAA	-
5’	0.020	6.8	8/10	-	←	0.96	TTTTCGC	SWI4

Figure S5

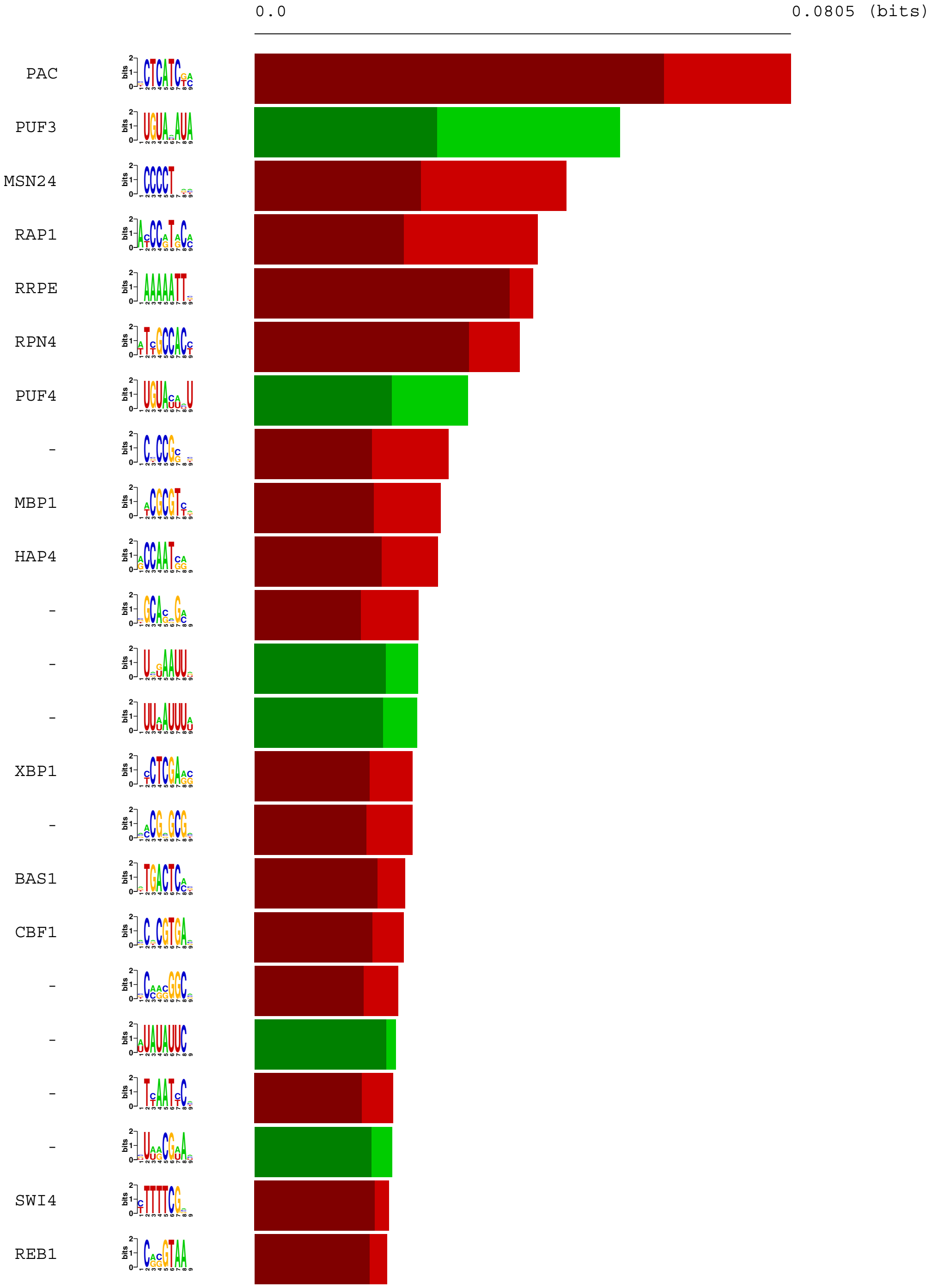
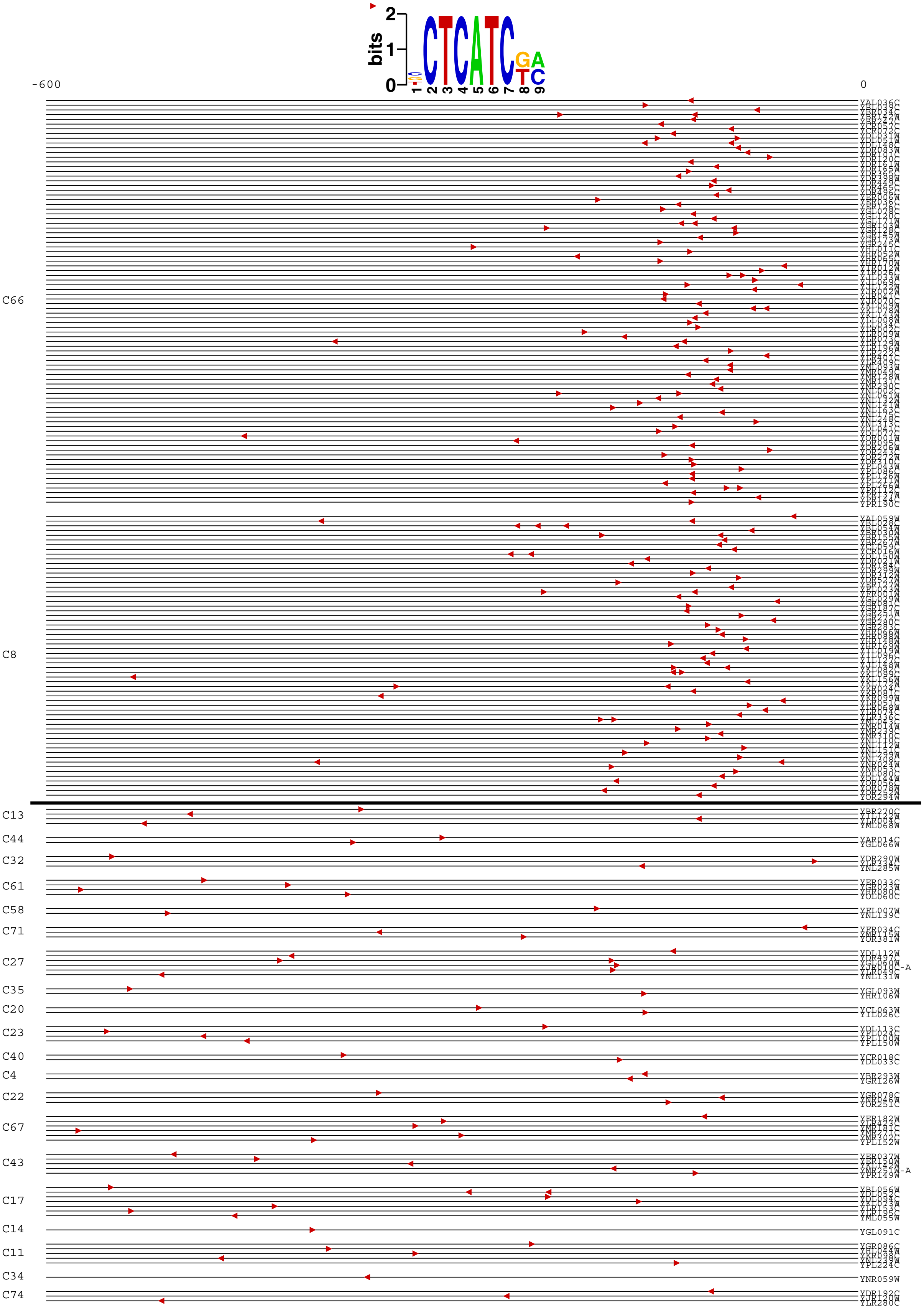

0.0
0.0805 (bits)
PAC
PUF3
MSN24
RAP1
RRPE
RPN4
PUF4
-
MBP1
HAP4
-
-
-
XBP1
-
BAS1
CBF1
-
-
-
-
SWI4
REB1

Figure S6



bits
2
1
0
1
2
3
4
5
6
7
8
9
-600
0
C66
YAL036C
YBL039C
YBR034C
YBR142W
YBR247C
YCR057C
YCR072C
YDL031W
YDL051W
YDL148C
YDR083W
YDR101C
YDR120C
YDR161W
YDR165W
YDR365C
YDR398W
YDR449C
YDR465C
YDR496C
YER006W
YER036C
YER126C
YGL078C
YGL120C
YGL171W
YGR103W
YGR128C
YGR145W
YGR173W
YGR245C
YHL011C
YHR052W
YHR065C
YHR170W
YIR012W
YIR026C
YJL033W
YJL069C
YJL122W
YJR002W
YJR041C
YJR070C
YKL009W
YKL078W
YKL143W
YLL008W
YLL034C
YLR002C
YLR009W
YLR073C
YLR129W
YLR196W
YLR222C
YLR401C
YLR409C
YML093W
YMR049C
YMR128W
YMR131C
YMR290C
YNL002C
YNL061W
YNL132W
YNL141W
YNL163C
YNL175C
YNL248C
YNL313C
YOL041C
YOL077C
YOR001W
YOR095C
YOR206W
YOR243C
YOR272W
YOR310C
YPL043W
YPL086C
YPL126W
YPL211W
YPL266W
YPR112C
YPR137W
YPR144C
YPR190C
C8
YAL059W
YBL028C
YBL054W
YBR030W
YBR155W
YBR267W
YCL059C
YCR016W
YDL150W
YDR021W
YDR184C
YDR299W
YDR312W
YDR527W
YER127W
YFL023W
YFR001W
YGL029W
YGR081C
YGR187C
YGR251W
YGR272C
YGR280C
YGR283C
YHR066W
YHR088W
YHR148W
YHR169W
YIL019W
YIL096C
YIL127C
YJL148W
YKL082C
YKL099C
YKL156W
YKL172W
YKR024C
YKR081C
YKR099W
YLR051C
YLR068W
YLR074C
YLR336C
YML043C
YMR014W
YMR239C
YMR310C
YNL110C
YNL112W
YNL151C
YNL299W
YNL308C
YNR024W
YNR053C
YOL080C
YOL144W
YOR056C
YOR078W
YOR252W
YOR294W
C13
YBR270C
YIL122W
YLR004C
YML068W
C44
YAR014C
YGL066W
C32
YDR290W
YLR334C
YNL285W
C61
YER033C
YGR023W
YHR080C
YOL060C
C58
YFL007W
YNL139C
C71
YER034C
YMR115W
YOR381W
C27
YDL112W
YDR497C
YGL060W
YJR010C-A
YLR049C
YNL131W
C35
YGL093W
YHR106W
C20
YCL063W
YIL026C
C23
YDL113C
YFL024C
YPL100W
YPL150W
C40
YCR018C
YDL033C
C4
YBR293W
YGR126W
C22
YGR078C
YNR046W
YOR251C
C67
YER182W
YLR423C
YMR181C
YMR271C
YMR302C
YPL152W
C43
YER037W
YER150W
YKL142W
YMR251W-A
YPR149W
C17
YBL056W
YDL052C
YDL094C
YKL073W
YLR153C
YLR195C
YML055W
C14
YGL091C
C11
YGR086C
YHL044W
YKR098C
YNL239W
YPL224C
C34
YNR059W
C74
YDR192C
YJR120W
YLR280C

Figure S7

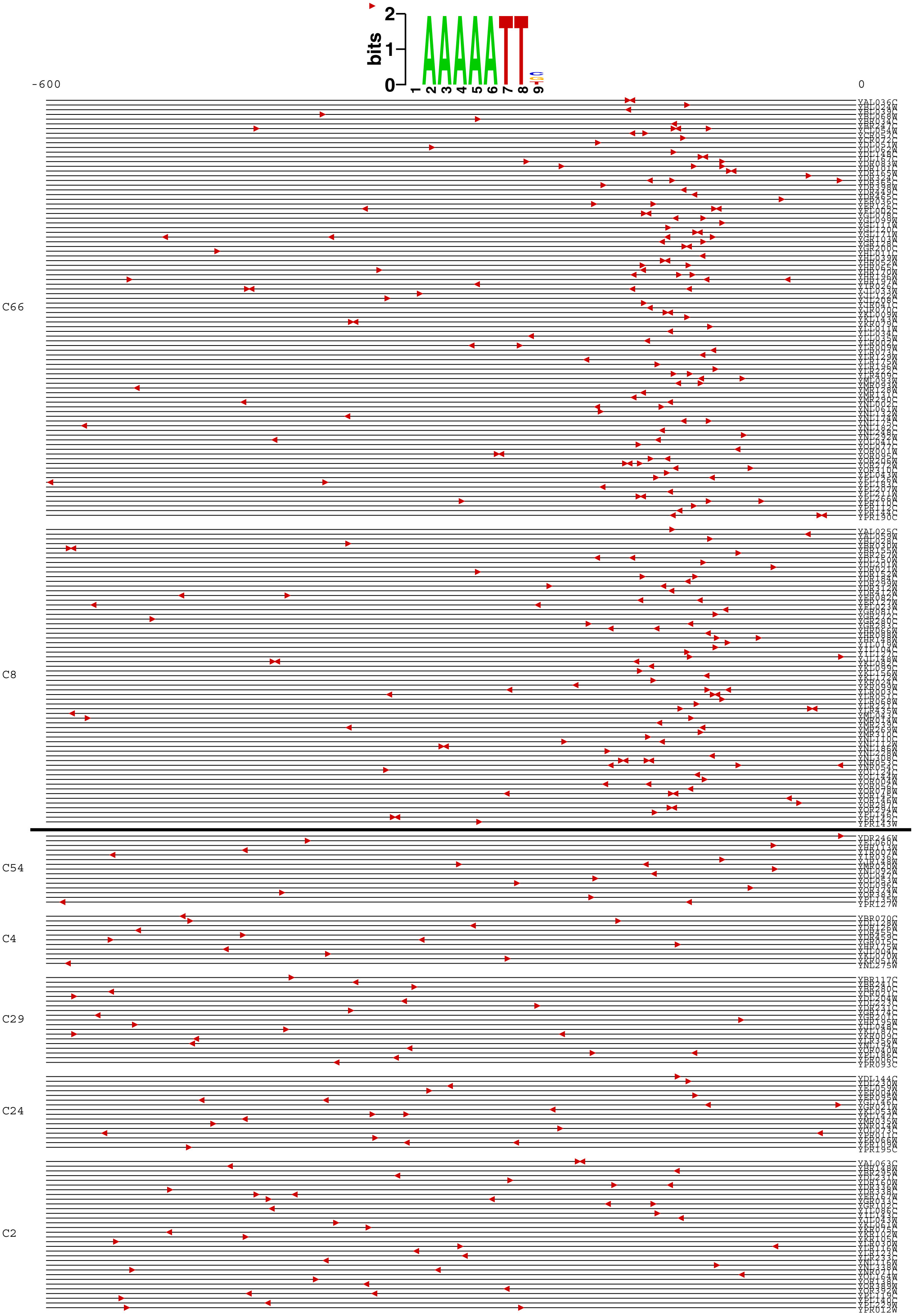

2
1
0
bits
1
2
3
4
5
6
7
8
9
-600
0
C66
C8
C54
C4
C29
C24
C2
YAL036C
YPR190C
YAL025C
YPR143W
YDR246W
YPR127W
YBR070C
YNL275W
YBR117C
YPR093C
YDL144C
YPR195C
YAL063C
YPR012W

Figure S8

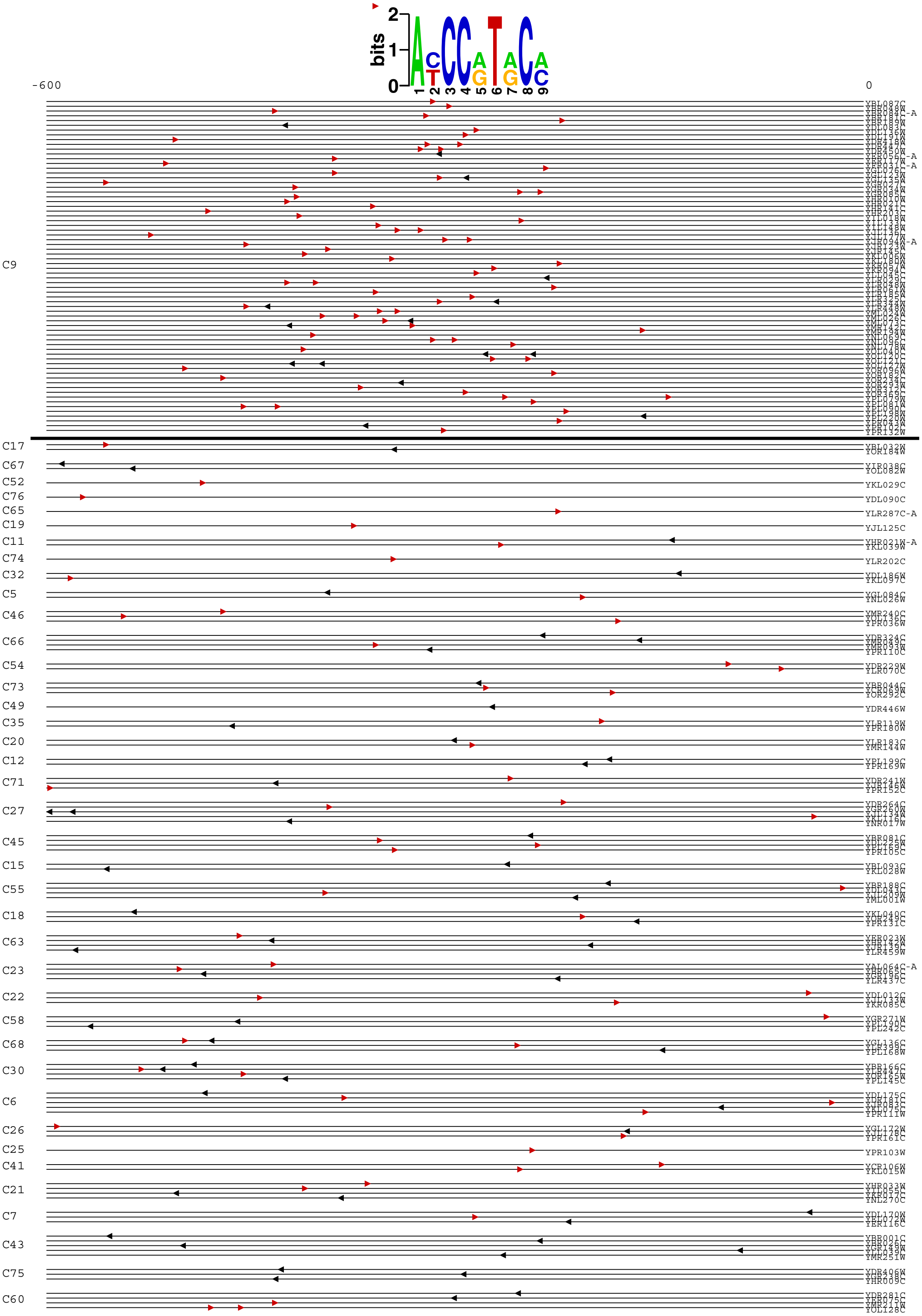

bits
2
1
0
1 2 3 4 5 6 7 8 9
-600
0
C9
C17
C67
C52
C76
C65
C19
C11
C74
C32
C5
C46
C66
C54
C73
C49
C35
C20
C12
C71
C27
C45
C15
C55
C18
C63
C23
C22
C58
C68
C30
C6
C26
C25
C41
C21
C7
C43
C75
C60
YBL087C
YBR048W
YBR084C-A
YBR181C
YBR189W
YDL083C
YDL136W
YDL191W
YDR418W
YDR447C
YDR450W
YER056C-A
YER117W
YFR031C-A
YGL076C
YGL123W
YGL135W
YGR027C
YGR034W
YGR085C
YHR010W
YHR021C
YHR141C
YHR203C
YIL018W
YIL133C
YIL148W
YJL136C
YJL177W
YJR094W-A
YJR123W
YJR145C
YKL006W
YKL180W
YKR057W
YKR094C
YLL045C
YLR029C
YLR048W
YLR061W
YLR185W
YLR325C
YLR344W
YLR448W
YML024W
YML026C
YML073C
YMR142C
YMR194W
YNL069C
YNL096C
YNL178W
YOL040C
YOL120C
YOL121C
YOL127W
YOR096W
YOR182C
YOR234C
YOR293W
YOR312C
YOR369C
YPL079W
YPL081W
YPL090C
YPL198W
YPL220W
YPR043W
YPR102C
YPR132W
YBL032W
YOR184W
YIR038C
YOL082W
YKL029C
YDL090C
YLR287C-A
YJL125C
YHR021W-A
YKL039W
YLR202C
YDL186W
YKL097C
YGL084C
YNL026W
YMR240C
YOL136C
YPR036W
YDR324C
YMR049C
YMR093W
YPR110C
YDR229W
YLR070C
YBR044C
YCR069W
YOR292C
YDR446W
YLR119W
YPR180W
YLR183C
YMR144W
YPL199C
YPR169W
YDR241W
YJR146W
YPR152C
YDR264C
YGR260W
YJL134W
YKL116C
YNR017W
YBR081C
YDL225W
YPL169C
YPR105C
YBL093C
YKL028W
YBR188C
YDL043C
YJL209W
YML001W
YKL040C
YOR249C
YPR131C
YER023W
YHR142W
YJR139C
YLR459W
YAL064C-A
YBR065C
YGR196C
YLR437C
YDL012C
YJL133W
YKR085C
YGR271W
YPL190C
YPL242C
YGL136C
YLR399C
YPL168W
YBR166C
YLR447C
YOR165W
YPL145C
YDL175C
YDR181C
YJR083C
YKL075C
YPR111W
YGL172W
YJL178C
YPR161C
YPR103W
YCR106W
YKL015W
YHR033W
YIL055C
YKR017C
YNL270C
YDL170W
YEL072W
YER116C
YBR001C
YBR026C
YGR149W
YLL039C
YMR251W
YDR406W
YGR238C
YHR009C
YDR281C
YER075C
YMR211W
YOL128C

Figure S9

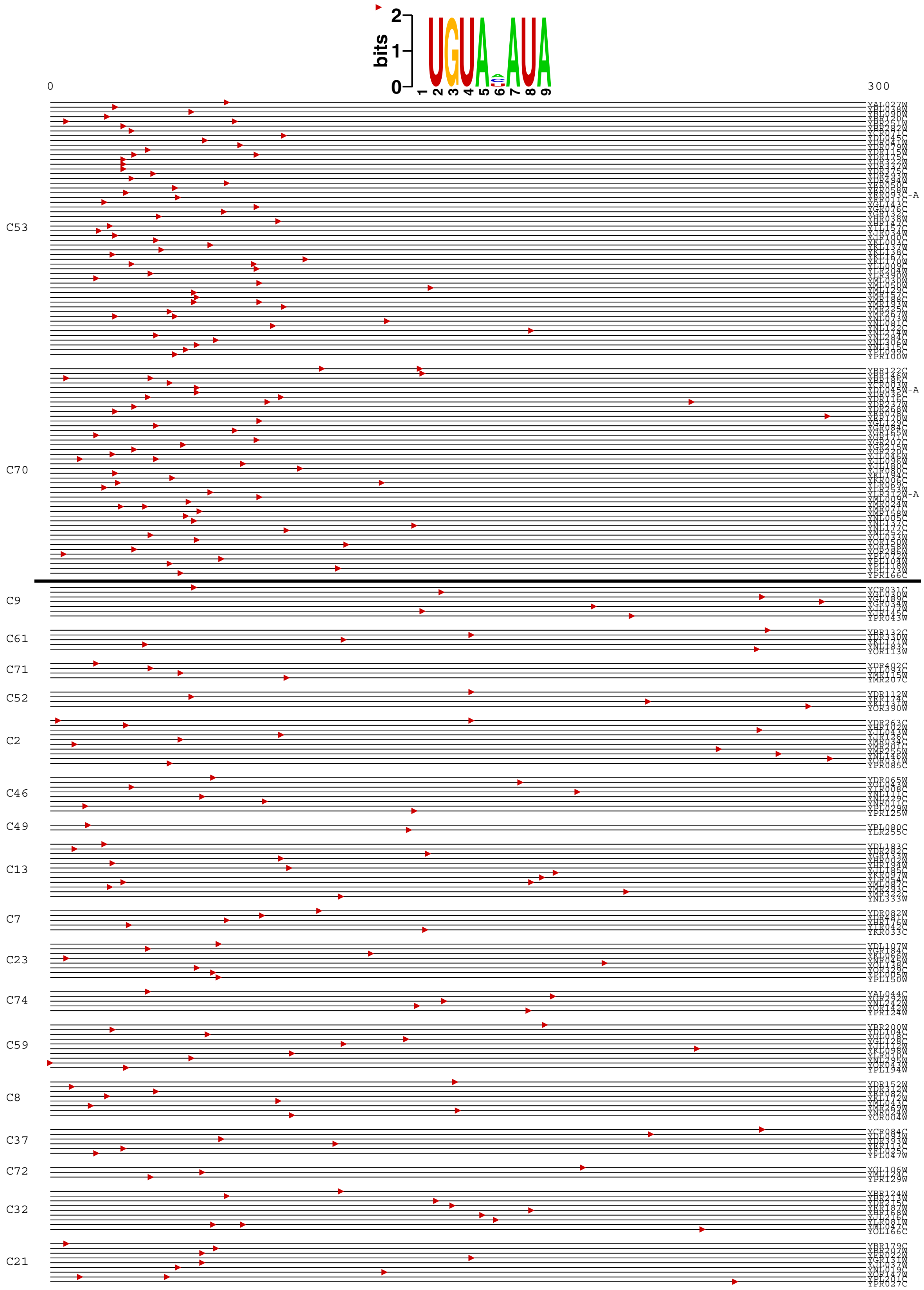

bits
2
1
0
UGUA
AUA
1
2
3
4
5
6
7
8
9
0
300
C53
YAL027W
YBL038W
YBL090W
YBR120C
YBR251W
YBR282W
YCR071C
YDL045C
YDR041W
YDR079W
YDR115W
YDR175C
YDR322W
YDR337W
YDR375C
YDR493W
YDR494W
YER050C
YER058W
YER093C-A
YFR011C
YGL143C
YGR076C
YGR132C
YHR038W
YHR147C
YIL157C
YJR034W
YJR100C
YKL003C
YKL137W
YKL138C
YKL167C
YKL170W
YLL009C
YLR204W
YLR390W
YML030W
YML050W
YML129C
YMR157C
YMR188C
YMR193W
YMR225C
YMR267W
YNL073W
YNL081C
YNL122C
YNL214W
YNL284C
YNL306W
YNL315C
YPL099C
YPR100W
C70
YBR122C
YBR146W
YBR185C
YCR003W
YDL045W-A
YDR036C
YDR116C
YDR237W
YDR268W
YER078C
YER170W
YGL129C
YGR084C
YGR165W
YGR171C
YGR207C
YGR215W
YGR220C
YJL046W
YJL096W
YJL180C
YJR080C
YKL194C
YKR006C
YLR069C
YLR253W
YLR312W-A
YML009C
YMR024W
YMR071C
YMR158W
YNL005C
YNL137C
YNL177C
YNL252C
YOL033W
YOR150W
YOR158W
YOR286W
YPL072W
YPL104W
YPL118W
YPL173W
YPR166C
C9
YCR031C
YGL030W
YGL189C
YGR034W
YJL177W
YJR145C
YPR043W
C61
YBR132C
YDR330W
YKL171W
YNL183C
YOR113W
C71
YDR402C
YIL093C
YMR115W
YMR207C
C52
YDR112W
YER174C
YKL131W
YOR390W
C2
YDR263C
YHR102W
YJL043W
YJR126C
YMR034C
YMR201C
YMR255W
YNL146W
YOR031W
YPR085C
C46
YDR065W
YGL043W
YIR008C
YNL111C
YNL229C
YNR011C
YPL029W
YPR125W
C49
YBL080C
YLR255C
C13
YDL183C
YDR282C
YGR133W
YHR002W
YHR194W
YJL185C
YKR097W
YLR054C
YML087C
YMR293C
YMR322C
YNL333W
C7
YDR082W
YDR481C
YHR176W
YIR042C
YKR033C
C23
YDL107W
YGR184C
YKL066W
YNR045W
YOL138C
YOR329C
YPL005W
YPL150W
C74
YAL044C
YGR292W
YNL242W
YOR142W
YPR124W
C59
YBR200W
YDL104C
YGL018C
YGL128C
YJL112W
YKL098W
YLR010C
YNL295W
YOR043W
YPL194W
C8
YDR152W
YDR312W
YER082C
YKL172W
YML043C
YMR269W
YNR024W
YOR004W
C37
YCR084C
YDL093W
YDR393W
YER113C
YFL025C
YFL047W
C72
YGL106W
YML124C
YPR129W
C32
YBR124W
YBR213W
YDR215C
YER187W
YHR168W
YJL216C
YLR081W
YML047C
YOL166C
C21
YBR179C
YBR207W
YER022W
YGR131W
YJL037W
YNL019C
YOR147W
YPL201C
YPR027C

Figure S10

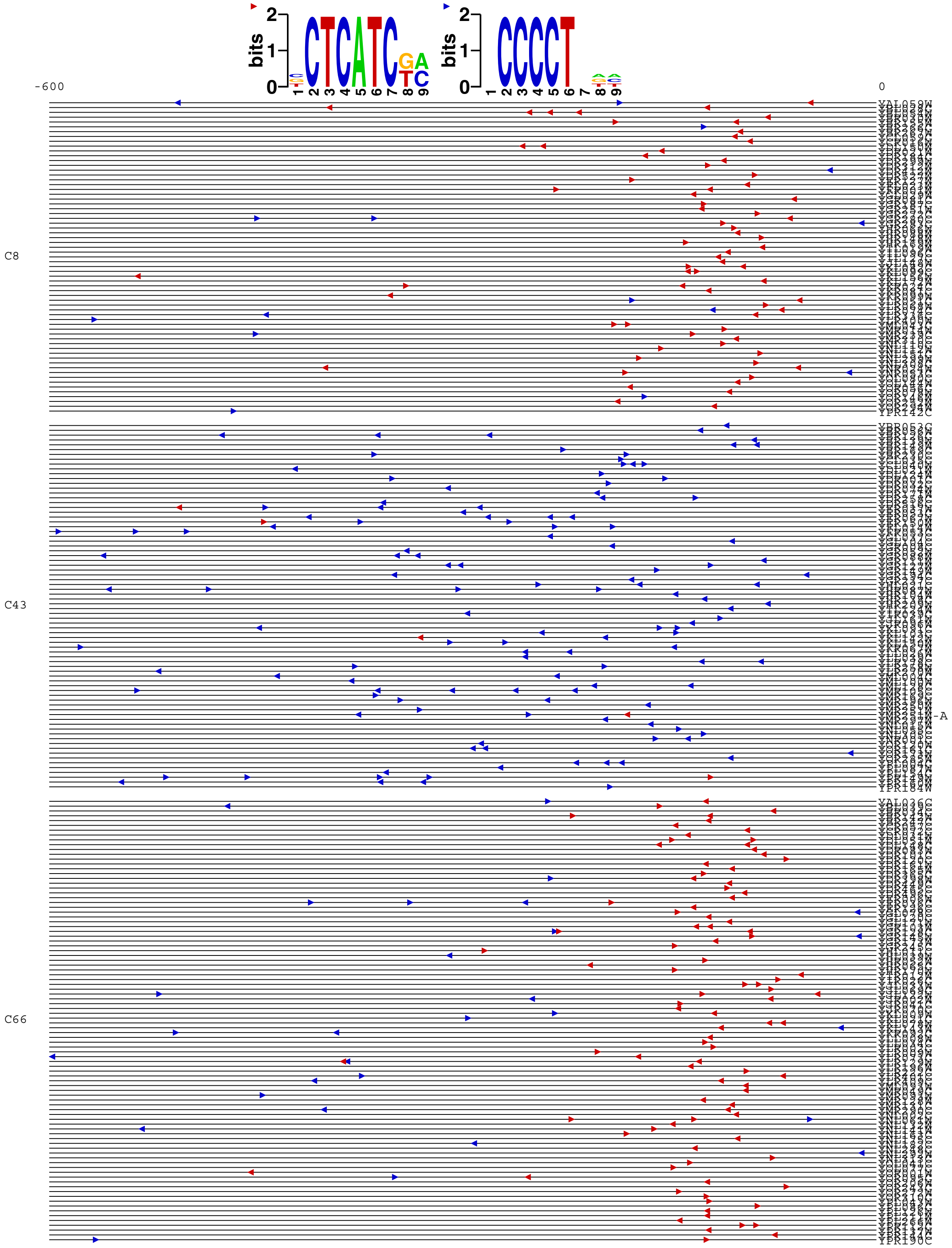
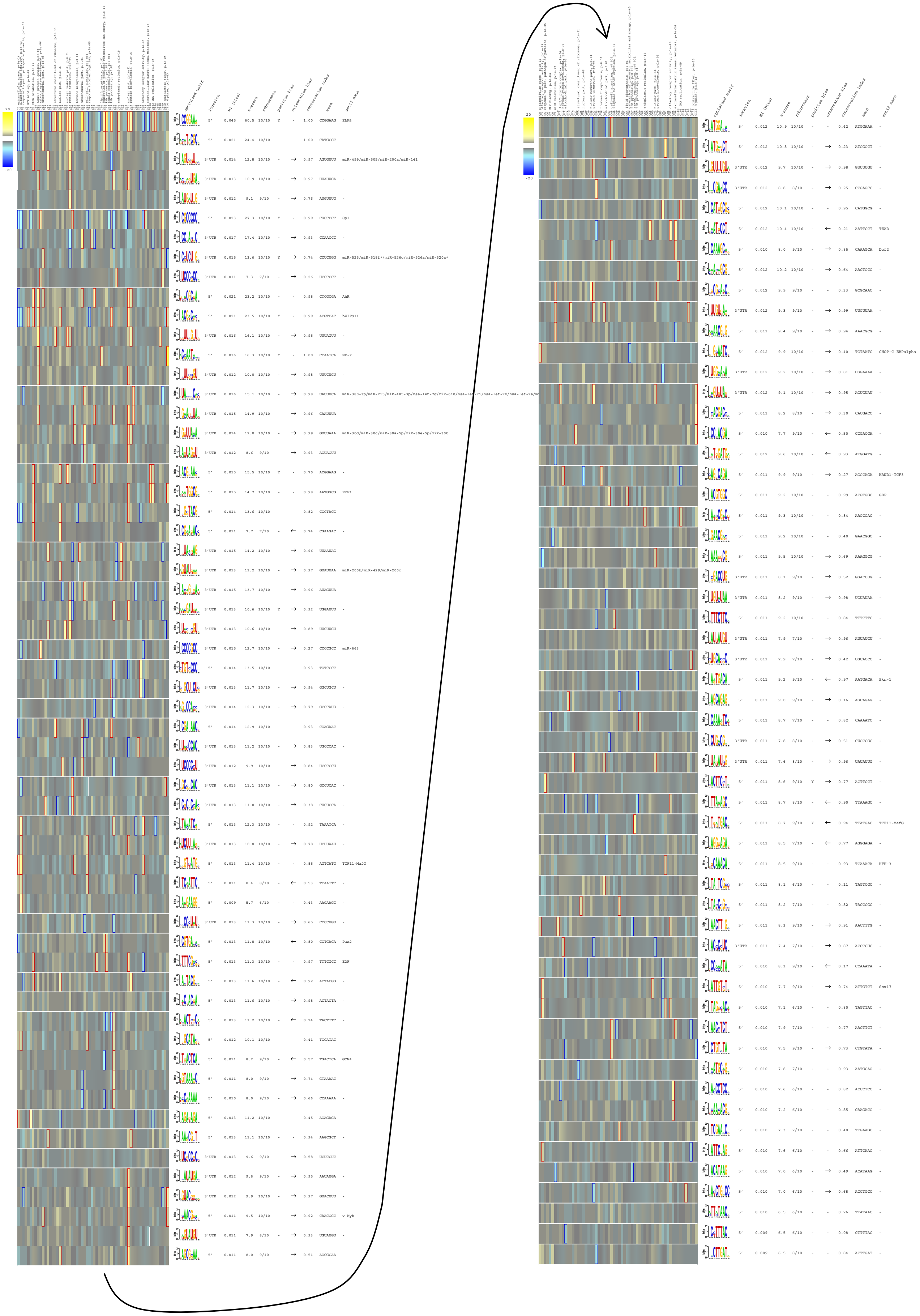

bits
2
1
0
1 2 3 4 5 6 7 8 9
bits
2
1
0
1 2 3 4 5 6 7 8 9
-600
0
C8
C43
C66
YAL059W
YPR142C
YBR052C
YPR184W
YAL036C
YPR190C
A

Figure S11



location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5′	0.045	60.5	10/10	Y	-	1.00	CCGGAAG	ELK4
5′	0.021	24.4	10/10	-	-	1.00	CATGCGC	-
3′UTR	0.014	12.8	10/10	-	→	0.97	AGUGUUU	miR-499/miR-505/miR-200a/miR-141
3′UTR	0.013	10.9	10/10	-	→	0.97	UUAUUGA	-
3′UTR	0.012	9.1	9/10	-	→	0.76	AUGUUUG	-
5′	0.023	27.3	10/10	Y	-	0.99	CGCCCCC	Sp1
3′UTR	0.017	17.4	10/10	-	→	0.93	CCAACCC	-
3′UTR	0.015	13.6	10/10	Y	→	0.74	CCUCUGG	miR-525/miR-518f*/miR-526c/miR-526a/miR-520a*
3′UTR	0.011	7.3	7/10	-	→	0.26	UCCCCCC	-
5′	0.021	23.2	10/10	-	-	0.98	CTCGCGA	AhR
5′	0.021	23.5	10/10	Y	-	0.99	ACGTCAC	bZIP911
3′UTR	0.016	16.1	10/10	-	→	0.95	UUUAGUU	-
5′	0.016	16.3	10/10	Y	-	1.00	CCAATCA	NF-Y
3′UTR	0.012	10.0	10/10	-	→	0.98	UUUCUGU	-
3′UTR	0.016	15.1	10/10	-	→	0.98	UAUUUCA	miR-380-3p/miR-215/miR-485-3p/hsa-let-7g/miR-610/hsa-let-7i/hsa-let-7b/hsa-let-7a/m
3′UTR	0.015	14.9	10/10	-	→	0.96	GAAUUUA	-
3′UTR	0.014	12.0	10/10	-	→	0.99	GUUUAAA	miR-30d/miR-30c/miR-30a-5p/miR-30e-5p/miR-30b
3′UTR	0.012	8.6	9/10	-	→	0.93	AGUAGUU	-
5′	0.015	15.5	10/10	Y	-	0.70	ACGGAAG	-
5′	0.015	14.7	10/10	-	-	0.98	AATGGCG	E2F1
5′	0.014	13.6	10/10	-	-	0.82	CGCTACG	-
5′	0.011	7.7	7/10	-	←	0.74	CGAAGAC	-
3′UTR	0.015	14.2	10/10	-	→	0.96	UUAAUAG	-
3′UTR	0.013	11.2	10/10	-	→	0.97	GUAUUAA	miR-200b/miR-429/miR-200c
3′UTR	0.015	13.7	10/10	-	→	0.96	AUAGUUA	-
3′UTR	0.013	10.6	10/10	Y	→	0.92	UGGAUUU	-
3′UTR	0.013	10.6	10/10	-	→	0.89	UUCUUGU	-
3′UTR	0.015	12.7	10/10	-	→	0.27	CCCCGCC	miR-663
5′	0.014	13.5	10/10	-	-	0.93	TGTCCCC	-
3′UTR	0.013	11.7	10/10	-	→	0.94	GGCUGCU	-
3′UTR	0.014	12.3	10/10	-	→	0.79	GCCCAUG	-
5′	0.014	12.9	10/10	-	-	0.93	CGAGAAC	-
3′UTR	0.013	11.2	10/10	-	→	0.83	UGCCCAC	-
3′UTR	0.012	9.9	10/10	-	→	0.84	UCCCCCU	-
3′UTR	0.013	11.1	10/10	-	→	0.80	GCCUCAC	-
3′UTR	0.013	11.0	10/10	-	→	0.38	CUCUCCA	-
5′	0.013	12.3	10/10	-	-	0.92	TAAATCA	-
3′UTR	0.013	10.8	10/10	-	→	0.78	UCUUAAU	-
5′	0.013	11.4	10/10	-	-	0.85	AGTCATG	TCF11-MafG
5′	0.011	8.4	8/10	-	←	0.53	TCAATTC	-
5′	0.009	5.7	6/10	-	-	0.43	AAGAAGG	-
3′UTR	0.013	11.3	10/10	-	→	0.65	CCCCUGU	-
5′	0.013	11.8	10/10	-	←	0.80	CGTGACA	Pax2
5′	0.013	11.3	10/10	-	-	0.97	TTTCGCC	E2F
5′	0.013	11.6	10/10	-	←	0.92	ACTACGG	-
5′	0.013	11.6	10/10	-	→	0.98	ACTACTA	-
5′	0.013	11.2	10/10	-	←	0.24	TACTTTC	-
5′	0.012	10.1	10/10	-	-	0.41	TGCATAC	-
5′	0.011	8.2	9/10	-	←	0.57	TGACTCA	GCN4
5′	0.011	8.0	9/10	-	→	0.74	GTAAAAC	-
5′	0.010	8.0	9/10	-	→	0.66	CCAAAAA	-
5′	0.013	11.2	10/10	-	-	0.45	AGAGAGA	-
5′	0.013	11.1	10/10	-	-	0.94	AAGCGCT	-
3′UTR	0.013	9.6	9/10	-	→	0.58	UCUCCUC	-
3′UTR	0.012	9.6	9/10	-	→	0.95	AAUAUGA	-
3′UTR	0.012	9.9	10/10	-	→	0.97	GUACUUU	-
5′	0.011	9.5	10/10	-	→	0.92	CAACGGC	v-Myb
3′UTR	0.011	7.9	8/10	-	→	0.93	UGUAUGU	-
5′	0.011	8.0	9/10	-	→	0.51	AGCGCAA	-

location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
5′	0.012	10.9	10/10	-	-	0.42	ATGGAAA	-
5′	0.012	10.8	10/10	-	→	0.23	ATGGGCT	-
3′UTR	0.012	9.7	10/10	-	→	0.98	GUUUUGU	-
3′UTR	0.012	8.8	8/10	-	→	0.25	CCGAGCC	-
5′	0.012	10.1	10/10	-	-	0.95	CATGGCG	-
5′	0.012	10.4	10/10	-	←	0.21	AATTCCT	TEAD
5′	0.010	8.0	9/10	-	→	0.85	CAAAGCA	Dof2
5′	0.012	10.2	10/10	-	→	0.64	AACTGCG	-
5′	0.012	9.9	9/10	-	-	0.33	GCGCAAC	-
3′UTR	0.012	9.3	9/10	-	→	0.99	UUGUUAA	-
5′	0.011	9.4	9/10	-	→	0.94	AAACGCG	-
5′	0.012	9.9	10/10	-	→	0.40	TGTAATC	CHOP-C_EBPalpha
3′UTR	0.012	9.2	10/10	-	→	0.81	UGGAAAA	-
3′UTR	0.012	9.1	10/10	-	→	0.95	AGUGUAU	-
5′	0.011	8.2	8/10	-	→	0.30	CACGACC	-
5′	0.010	7.7	9/10	-	←	0.50	CCGACGA	-
5′	0.012	9.6	10/10	-	←	0.93	ATGGATG	-
5′	0.011	9.9	9/10	-	→	0.27	AGGCAGA	HAND1-TCF3
5′	0.011	9.2	10/10	-	-	0.99	ACGTGGC	GBP
5′	0.011	9.3	10/10	-	-	0.84	AAGCGAC	-
5′	0.011	9.2	10/10	-	-	0.40	GAACGGC	-
5′	0.011	9.5	10/10	-	→	0.69	AAAGGCG	-
3′UTR	0.011	8.1	9/10	-	→	0.52	GGACCUG	-
3′UTR	0.011	8.2	9/10	-	→	0.98	UGUAUAA	-
5′	0.011	9.2	10/10	-	-	0.84	TTTCTTC	-
3′UTR	0.011	7.9	7/10	-	→	0.96	AUUAUGU	-
3′UTR	0.011	7.9	7/10	-	→	0.42	UGCACCC	-
5′	0.011	9.2	9/10	-	←	0.97	AATGACA	Skn-1
5′	0.011	9.0	9/10	-	→	0.16	AGCAGAG	-
5′	0.011	8.7	7/10	-	-	0.82	CAAAATC	-
3′UTR	0.011	7.8	8/10	-	→	0.51	CUGCCGC	-
3′UTR	0.011	7.6	8/10	-	→	0.96	UAUAUUG	-
5′	0.011	8.6	9/10	Y	→	0.77	ACTTCCT	-
5′	0.011	8.7	8/10	-	←	0.90	TTAAAGC	-
5′	0.011	8.7	9/10	Y	←	0.94	TTATGAC	TCF11-MafG
5′	0.011	8.5	7/10	-	←	0.77	AGGGAGA	-
5′	0.011	8.5	9/10	-	-	0.93	TCAAACA	HFH-3
5′	0.011	8.1	6/10	-	-	0.11	TAGTCGC	-
5′	0.011	8.2	7/10	-	-	0.82	TACCCGC	-
5′	0.011	8.3	9/10	-	→	0.91	AACTTTG	-
3′UTR	0.011	7.4	7/10	-	→	0.87	ACCCCUC	-
5′	0.010	8.1	9/10	-	←	0.17	CCAAATA	-
5′	0.010	7.7	9/10	-	→	0.74	ATTGTCT	Sox17
5′	0.010	7.1	6/10	-	-	0.80	TAGTTAC	-
5′	0.010	7.9	7/10	-	-	0.77	AACTTCT	-
5′	0.010	7.5	9/10	-	→	0.73	CTGTATA	-
5′	0.010	7.8	7/10	-	-	0.93	AATGCAG	-
5′	0.010	7.6	6/10	-	-	0.82	ACCCTCC	-
5′	0.010	7.2	6/10	-	-	0.85	CAAGACG	-
5′	0.010	7.3	7/10	-	-	0.48	TCGAAGC	-
5′	0.010	7.6	6/10	-	-	0.66	ATTCAAG	-
5′	0.010	7.0	6/10	-	→	0.49	ACATAAG	-
5′	0.010	7.0	6/10	-	→	0.68	ACCTGCC	-
5′	0.010	6.5	6/10	-	-	0.26	TTATAAC	-
5′	0.009	6.5	6/10	-	-	0.08	CTTTTAC	-
5′	0.009	6.5	8/10	-	-	0.84	ACTTGAT	-

Figure S12

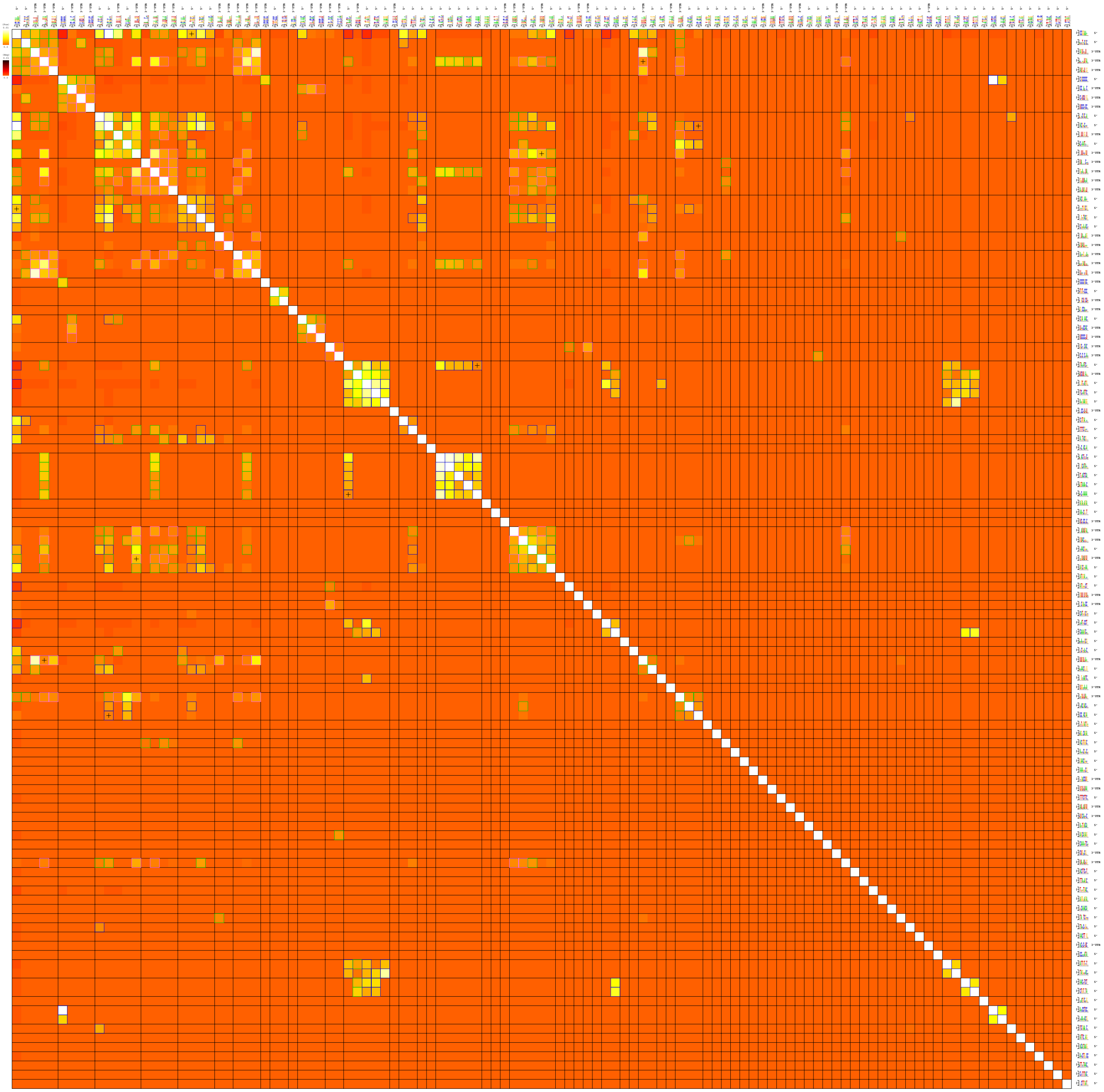


Figure S13

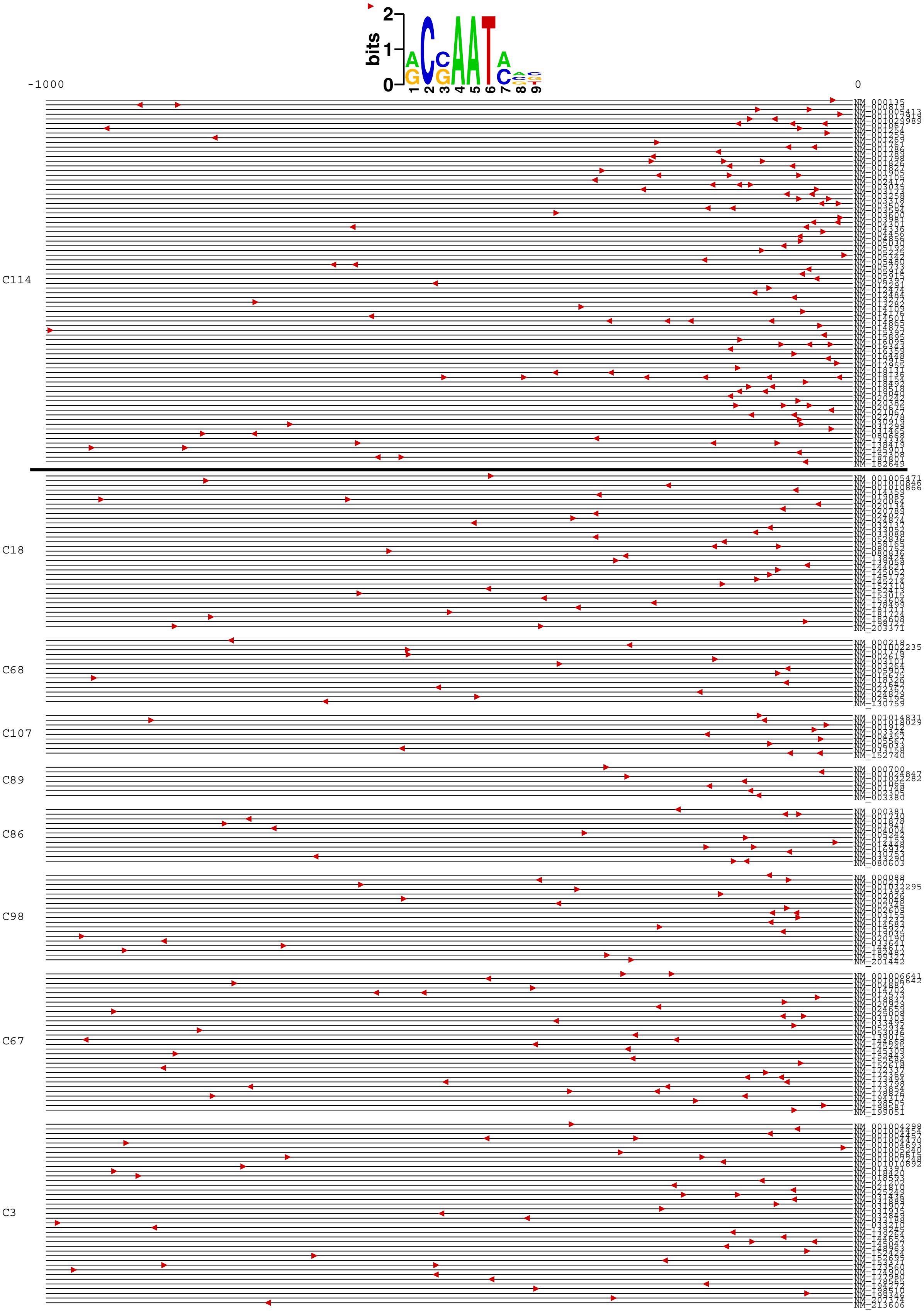
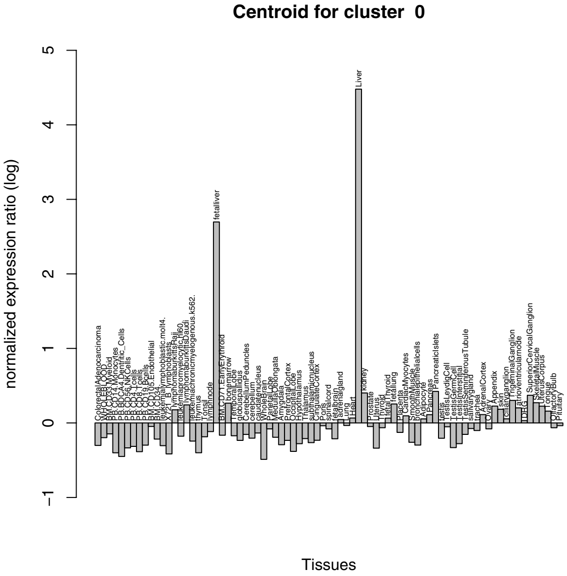
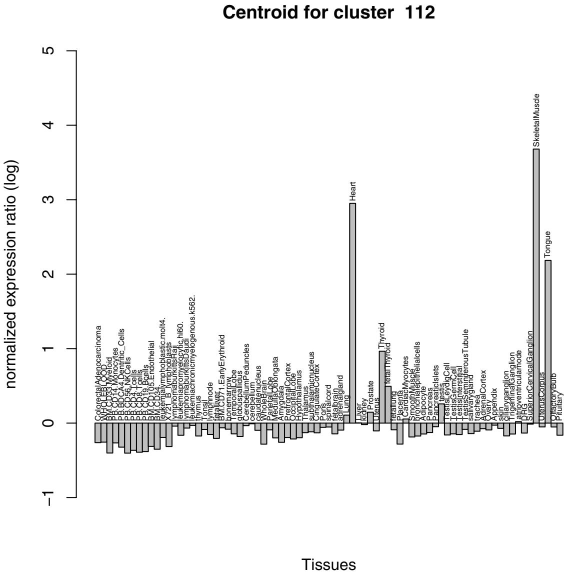

bits
2
1
0
1
2
3
4
5
6
7
8
9
-1000
0
C114
C18
C68
C107
C89
C86
C98
C67
C3
NM_000135
NM_000819
NM_001005413
NM_001017919
NM_001029989
NM_182849
NM_001005471
NM_001010846
NM_001010866
NM_203371
NM_000218
NM_001002235
NM_130759
NM_001014831
NM_001018029
NM_152740
NM_000700
NM_001024847
NM_001033082
NM_003380
NM_000381
NM_080603
NM_000098
NM_001012295
NM_201442
NM_001006641
NM_001006642
NM_198081
NM_001004298
NM_001004354
NM_001004470
NM_001010832
NM_213606

Figure S14

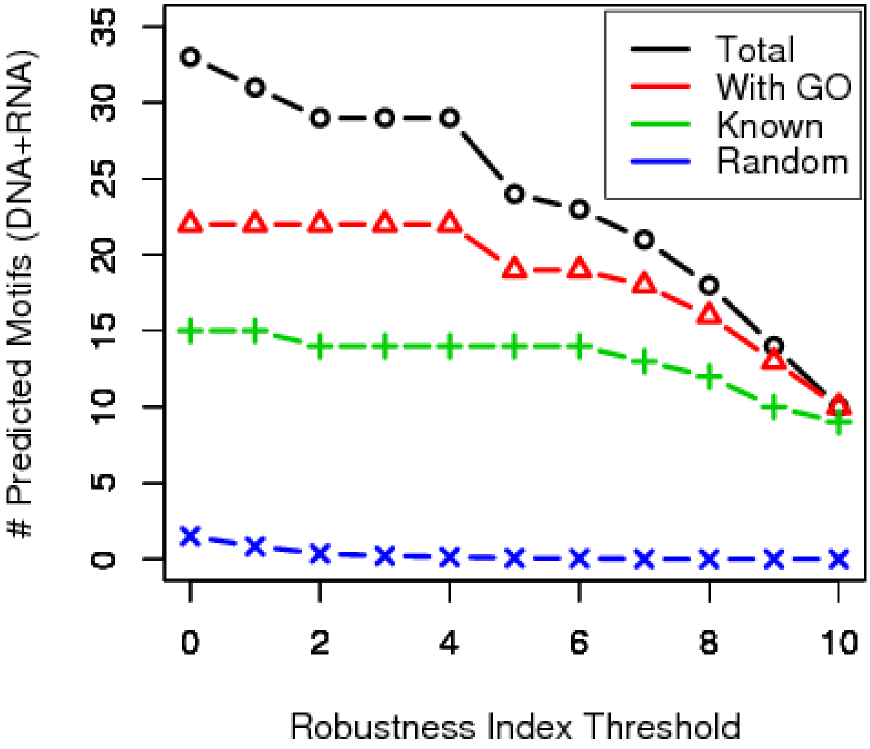


Centroid for cluster 0
normalized expression ratio (log)
Tissues
5
4
3
2
1
0
−1
ColorectalAdenocarcinoma
WHOLEBLOOD
BM.CD33.Myeloid
PB.CD14.Monocytes
PB.BDCA4.Dentritic_Cells
PB.CD56.NKCells
PB.CD4.Tcells
PB.CD8.Tcells
PB.CD19.Bcells
BM.CD105.Endothelial
BM.CD34.
leukemialymphoblastic.molt4.
X721_B_lymphoblasts
lymphomaburkittsRaji
leukemiapromyelocytic.hl60.
lymphomaburkittsDaudi
leukemiachronicmyelogenous.k562.
thymus
Tonsil
lymphnode
fetalliver
BM.CD71.EarlyErythroid
bonemarrow
TemporalLobe
globuspallidus
CerebellumPeduncles
cerebellum
caudatenucleus
WholeBrain
ParietalLobe
MedullaOblongata
Amygdala
PrefrontalCortex
OccipitalLobe
Hypothalamus
Thalamus
subthalamicnucleus
CingulateCortex
Pons
spinalcord
fetalbrain
adrenalgland
Lung
Heart
Liver
kidney
Prostate
Uterus
Thyroid
fetalThyroid
fetallung
Placenta
CardiacMyocytes
SmoothMuscle
bronchialepithelialcells
Adipocyte
Pancreas
PancreaticIslets
testis
TestisLeydigCell
TestisGermCell
TestisInterstitial
TestisSeminiferousTubule
salivarygland
trachea
AdrenalCortex
Ovary
Appendix
skin
ciliaryganglion
TrigeminalGanglion
atrioventricularnode
DRG
SuperiorCervicalGanglion
SkeletalMuscle
UterusCorpus
Tongue
OlfactoryBulb
Pituitary



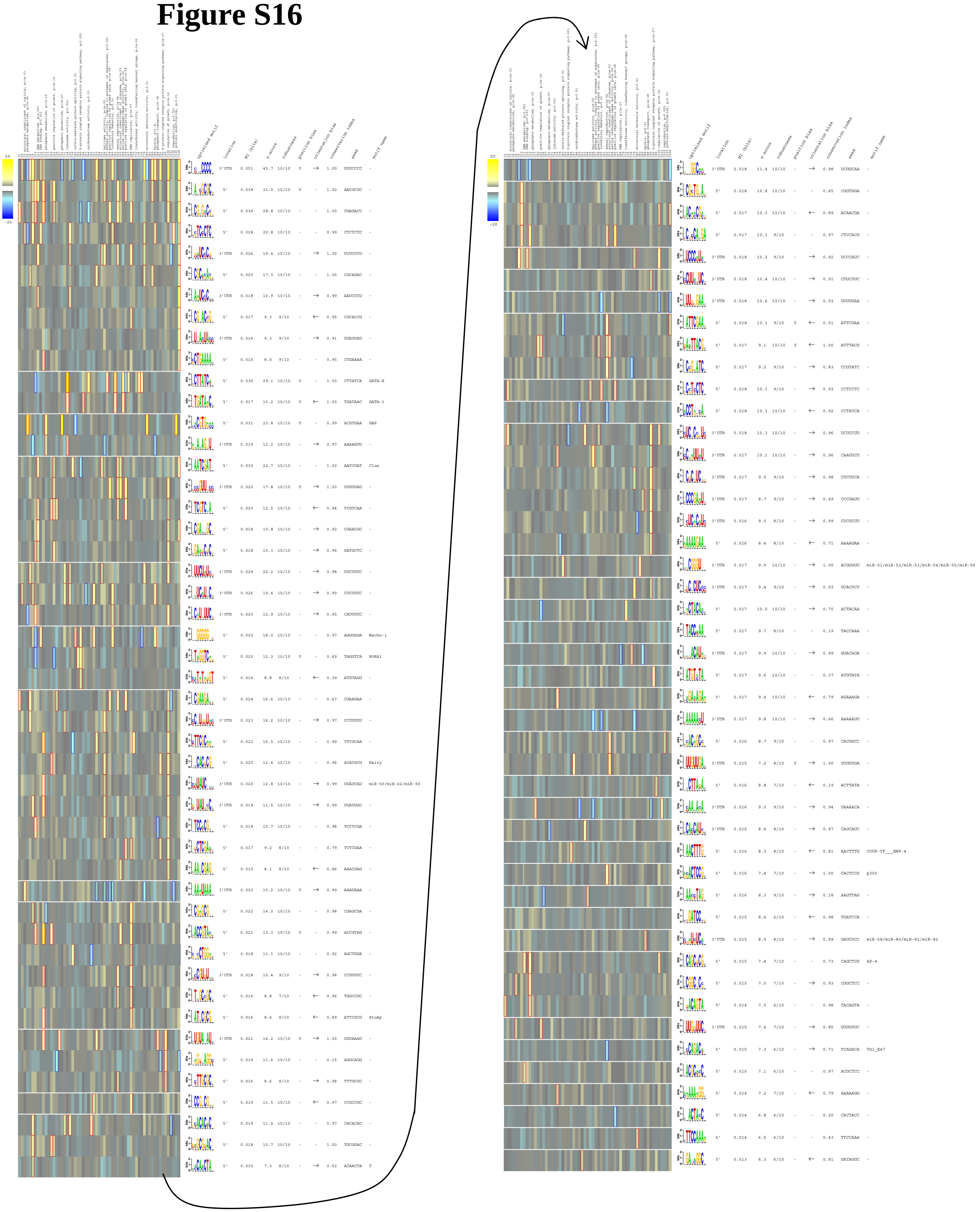
Centroid for cluster 112
normalized expression ratio (log)
Tissues
5
4
3
2
1
0
−1
ColorectalAdenocarcinoma
WHOLEBLOOD
BM.CD33.Myeloid
PB.CD14.Monocytes
PB.BDCA4.Dentritic_Cells
PB.CD56.NKCells
PB.CD4.Tcells
PB.CD8.Tcells
PB.CD19.Bcells
BM.CD105.Endothelial
BM.CD34.
leukemialymphoblastic.molt4.
X721_B_lymphoblasts
lymphomaburkittsRaji
leukemiapromyelocytic.hl60.
lymphomaburkittsDaudi
leukemiachronicmyelogenous.k562.
thymus
Tonsil
lymphnode
fetalliver
BM.CD71.EarlyErythroid
bonemarrow
TemporalLobe
globuspallidus
CerebellumPeduncles
cerebellum
caudatenucleus
WholeBrain
ParietalLobe
MedullaOblongata
Amygdala
PrefrontalCortex
OccipitalLobe
Hypothalamus
Thalamus
subthalamicnucleus
CingulateCortex
Pons
spinalcord
fetalbrain
adrenalgland
Lung
Heart
Liver
kidney
Prostate
Uterus
Thyroid
fetalThyroid
fetallung
Placenta
CardiacMyocytes
SmoothMuscle
bronchialepithelialcells
Adipocyte
Pancreas
PancreaticIslets
testis
TestisLeydigCell
TestisGermCell
TestisInterstitial
TestisSeminiferousTubule
salivarygland
trachea
AdrenalCortex
Ovary
Appendix
skin
ciliaryganglion
TrigeminalGanglion
atrioventricularnode
DRG
SuperiorCervicalGanglion
SkeletalMuscle
UterusCorpus
Tongue
OlfactoryBulb
Pituitary

Figure S15



Total
With GO
Known
Random
# Predicted Motifs (DNA+RNA)
0
5
10
15
20
25
30
35
0
2
4
6
8
10
Robustness Index Threshold

Figure S16



location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
3'UTR	0.051	45.7	10/10	Y	→	1.00	UUUCCCC	-
5'	0.038	31.0	10/10	Y	-	1.00	AACGCGC	-
5'	0.036	28.8	10/10	-	-	1.00	CGAGACC	-
5'	0.028	20.8	10/10	-	-	0.99	CTCTCTC	-
3'UTR	0.026	19.4	10/10	-	→	1.00	UCUCUCU	-
5'	0.025	17.3	10/10	-	-	1.00	CGCAGAC	-
3'UTR	0.018	10.9	10/10	-	→	0.99	AAUCUCU	-
5'	0.017	9.3	9/10	-	←	0.95	CGCACCG	-
3'UTR	0.016	9.3	9/10	-	→	0.91	UUAUUAU	-
5'	0.015	8.0	9/10	-	-	0.95	CTGAAAA	-
5'	0.036	29.1	10/10	Y	-	1.00	CTTATCA	GATA-X
5'	0.017	10.2	10/10	Y	←	1.00	TGATAAC	GATA-1
5'	0.031	23.8	10/10	Y	-	0.99	ACGTGAA	GBP
3'UTR	0.019	12.2	10/10	-	→	0.97	AAAAGUU	-
5'	0.030	22.7	10/10	-	-	1.00	AATCGAT	Clox
3'UTR	0.025	17.8	10/10	Y	→	1.00	UUGUUAU	-
5'	0.020	12.5	10/10	-	←	0.94	TCGTCAA	-
5'	0.018	10.8	10/10	-	→	0.92	CGAACGC	-
5'	0.018	10.3	10/10	-	→	0.96	GATGCTC	-
3'UTR	0.029	22.2	10/10	-	→	0.98	UUCUUUC	-
3'UTR	0.026	19.4	10/10	-	→	0.99	CUCUUUC	-
3'UTR	0.020	12.9	10/10	-	→	0.95	CAUUUUC	-
5'	0.025	18.0	10/10	-	-	0.97	AGGGGGA	Macho-1
5'	0.020	12.3	10/10	Y	-	0.69	TAGGTCA	RORA1
5'	0.016	8.8	8/10	-	←	0.39	ATGTAGG	-
5'	0.024	16.6	10/10	-	-	0.67	CGAAGAA	-
3'UTR	0.023	16.2	10/10	-	→	0.97	CCUUUUU	-
5'	0.023	15.5	10/10	-	-	0.99	TTCGCAA	-
5'	0.020	12.4	10/10	-	-	0.96	ACACGCG	Hairy
3'UTR	0.020	12.8	10/10	-	→	0.99	UUAUCAU	miR-50/miR-62/miR-90
3'UTR	0.019	11.5	10/10	-	→	0.99	UUAUUGC	-
5'	0.018	10.7	10/10	-	-	0.98	TCCTCGA	-
5'	0.017	9.2	8/10	-	-	0.79	TCTCGAA	-
5'	0.015	8.1	8/10	-	←	0.86	AAACGAG	-
3'UTR	0.022	15.2	10/10	Y	→	0.99	AAAUAAA	-
5'	0.022	14.3	10/10	-	-	0.98	CGAGCGA	-
5'	0.021	13.3	10/10	Y	-	0.99	ACCGTAG	-
5'	0.018	11.1	10/10	-	-	0.92	AGCTGGA	-
3'UTR	0.018	10.4	9/10	-	→	0.98	CCGUUUC	-
5'	0.016	8.8	7/10	-	←	0.96	TGGCCGC	-
5'	0.016	8.4	8/10	-	←	0.89	ATTCGCG	StuAp
3'UTR	0.021	14.2	10/10	Y	→	1.00	UGUAAAU	-
5'	0.019	11.6	10/10	-	-	0.15	AGGCAGG	-
5'	0.016	8.6	8/10	-	→	0.98	TTTGCGC	-
5'	0.019	11.5	10/10	-	←	0.97	CCGCCGC	-
5'	0.019	11.4	10/10	-	-	0.97	CACACAC	-
5'	0.018	10.7	10/10	-	-	1.00	TGCGGAC	-
5'	0.015	7.3	8/10	-	→	0.52	ACAACTA	T

location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
3'UTR	0.018	11.4	10/10	-	→	0.98	UCGGCAA	-
5'	0.018	10.8	10/10	-	-	0.65	CGGTGGA	-
5'	0.017	10.3	10/10	-	←	0.89	ACAACGA	-
5'	0.017	10.1	9/10	-	-	0.97	CTCCACG	-
3'UTR	0.018	10.2	9/10	-	→	0.92	UCCCAUC	-
3'UTR	0.018	10.4	10/10	-	→	0.91	CUUCUUC	-
3'UTR	0.018	10.6	10/10	-	→	0.93	UUUUGAA	-
5'	0.018	10.1	9/10	Y	←	0.51	ATTCGAA	-
5'	0.017	9.1	10/10	Y	←	1.00	ATTTACG	-
5'	0.017	9.2	9/10	-	→	0.83	CCGTATC	-
5'	0.018	10.1	9/10	-	→	0.93	CCTCCTC	-
5'	0.018	10.3	10/10	-	←	0.92	CCTATCA	-
3'UTR	0.018	10.3	10/10	-	→	0.96	UCUCCUU	-
3'UTR	0.017	10.1	10/10	-	→	0.96	CAAUUCU	-
3'UTR	0.017	9.5	9/10	-	→	0.98	CUCUUCA	-
3'UTR	0.017	8.7	9/10	-	→	0.49	CCCGAUU	-
3'UTR	0.016	9.0	8/10	-	→	0.99	CUCGCUU	-
5'	0.016	8.4	8/10	-	←	0.71	AAAAGAA	-
3'UTR	0.017	9.9	10/10	-	→	1.00	ACGGGUU	miR-51/miR-52/miR-53/miR-54/miR-55/miR-56
3'UTR	0.017	9.4	9/10	-	→	0.93	UCACUCU	-
5'	0.017	10.0	10/10	-	→	0.70	ACTACAA	-
5'	0.017	9.7	8/10	-	-	0.19	TACCAAA	-
3'UTR	0.017	9.9	10/10	-	→	0.99	GUACAUA	-
5'	0.017	9.6	10/10	-	-	0.37	ATGTATA	-
5'	0.017	9.4	10/10	-	←	0.79	AGAAAGA	-
3'UTR	0.017	9.8	10/10	-	→	0.66	AAAAAUU	-
5'	0.016	8.7	9/10	-	-	0.97	CACGGCC	-
3'UTR	0.015	7.2	8/10	Y	→	1.00	UUGUUGA	-
5'	0.016	8.8	7/10	-	←	0.19	ACTTATA	-
3'UTR	0.016	9.0	9/10	-	→	0.94	UAAAACA	-
3'UTR	0.016	8.4	8/10	-	→	0.97	CAUCAUC	-
5'	0.016	8.3	8/10	-	←	0.81	AACTTTG	COUP-TF___HNF-4
5'	0.016	7.4	7/10	-	→	1.00	CACTCCG	p300
5'	0.016	8.3	9/10	-	→	0.18	AAGTTAG	-
5'	0.015	8.4	6/10	-	←	0.98	TGATCCA	-
3'UTR	0.015	8.5	8/10	-	→	0.99	GAUCUCC	miR-58/miR-80/miR-81/miR-82
5'	0.015	7.4	7/10	-	-	0.73	CAGCTCG	AP-4
5'	0.015	7.0	7/10	-	→	0.93	CGGCTCC	-
5'	0.014	7.0	6/10	-	-	0.98	TACAGTA	-
3'UTR	0.015	7.4	7/10	-	→	0.85	UUGUUUC	-
5'	0.015	7.3	6/10	-	→	0.71	TCAGACA	Th1_E47
5'	0.015	7.1	6/10	-	-	0.97	ACGCTCC	-
5'	0.014	7.2	7/10	-	←	0.79	AAAAAGG	-
5'	0.014	6.8	6/10	-	-	0.20	CACTACC	-
5'	0.014	6.5	6/10	-	-	0.43	TTCCAAA	-
5'	0.013	6.3	6/10	-	←	0.81	GATAGGC	-

Figure S17

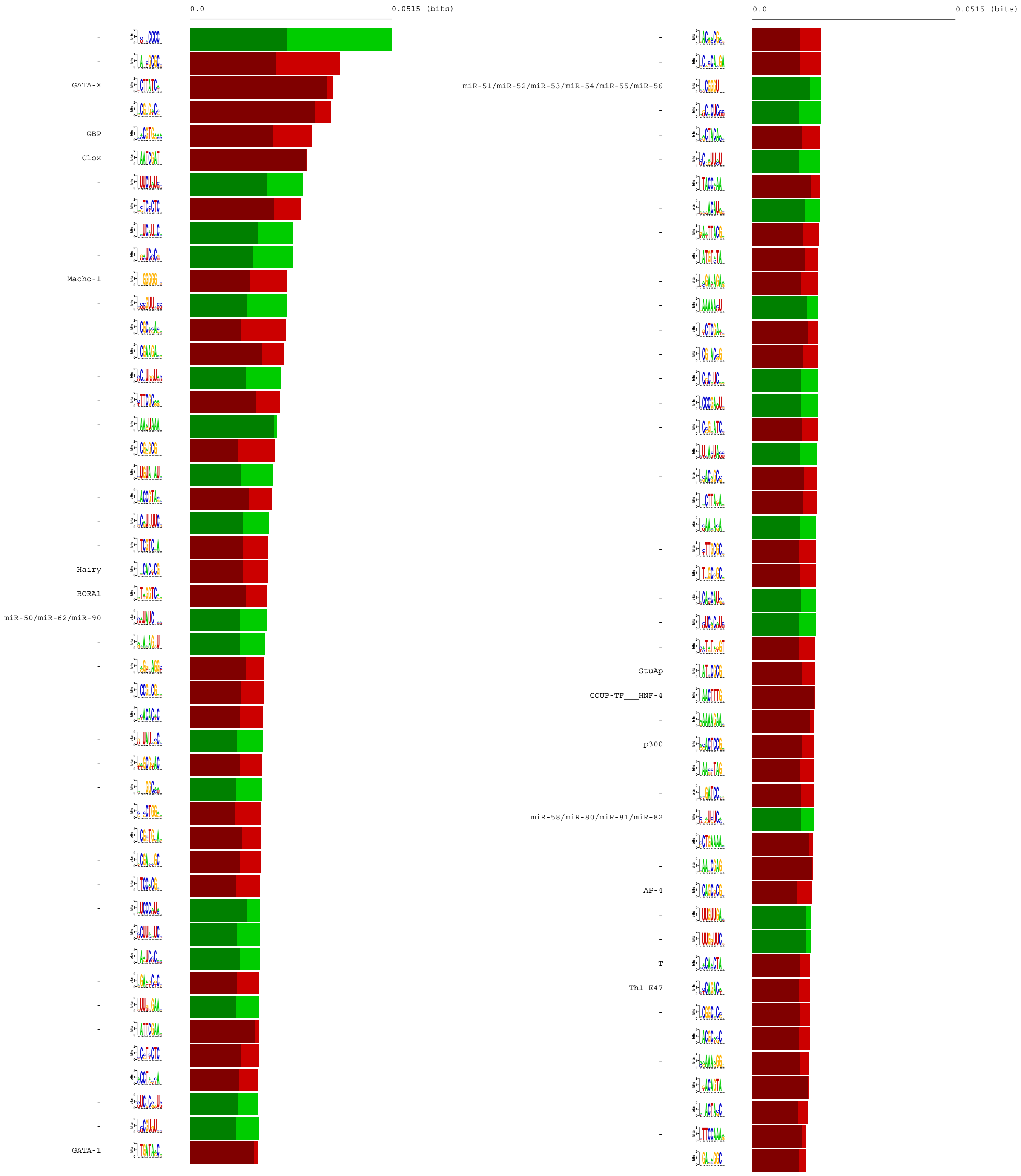

0.0
0.0515 (bits)
GATA-X
GBP
Clox
Macho-1
Hairy
RORA1
miR-50/miR-62/miR-90
GATA-1
0.0
0.0515 (bits)
miR-51/miR-52/miR-53/miR-54/miR-55/miR-56
StuAp
COUP-TF___HNF-4
p300
miR-58/miR-80/miR-81/miR-82
AP-4
T
Th1_E47

Figure S18

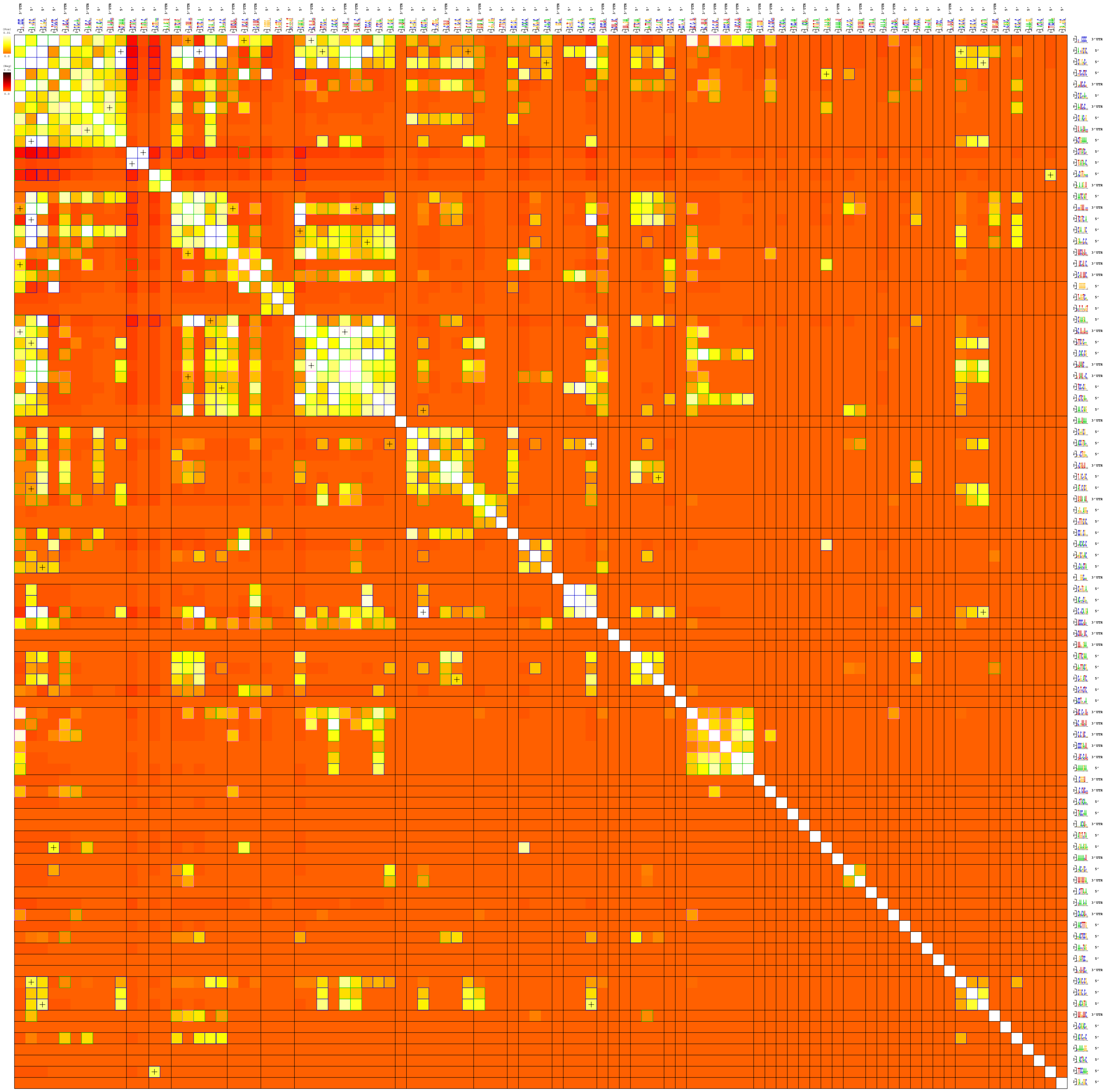
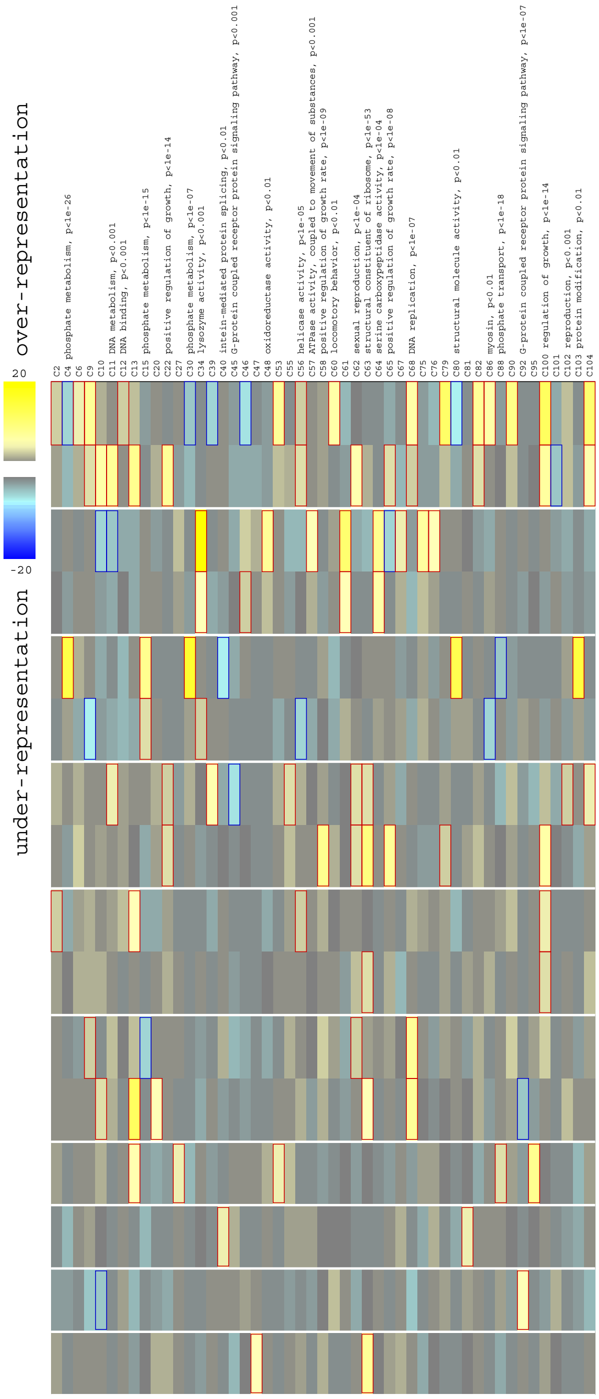


Figure S19



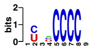
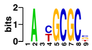
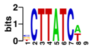
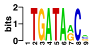
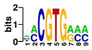
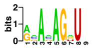
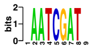
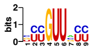
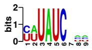
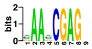
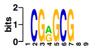
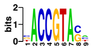
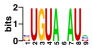
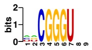
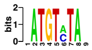
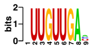
optimized motif	location	MI (bits)	z-score	robustness	position bias	orientation bias	conservation index	seed	motif name
	3'UTR	0.051	45.7	10/10	Y	→	1.00	UUUCCCC	-
	5'	0.038	31.0	10/10	Y	-	1.00	AACGCGC	-
	5'	0.036	29.1	10/10	Y	-	1.00	CTTATCA	GATA-X
	5'	0.017	10.2	10/10	Y	←	1.00	TGATAAC	GATA-1
	5'	0.031	23.8	10/10	Y	-	0.99	ACGTGAA	GBP
	3'UTR	0.019	12.2	10/10	-	→	0.97	AAAAGUU	-
	5'	0.030	22.7	10/10	-	-	1.00	AATCGAT	Clox
	3'UTR	0.025	17.8	10/10	Y	→	1.00	UUGUUAU	-
	3'UTR	0.020	12.8	10/10	-	→	0.99	UUAUCAU	miR-50/miR-62/miR-90
	5'	0.015	8.1	8/10	-	←	0.86	AAACGAG	-
	5'	0.022	14.3	10/10	-	-	0.98	CGAGCGA	-
	5'	0.021	13.3	10/10	Y	-	0.99	ACCGTAG	-
	3'UTR	0.021	14.2	10/10	Y	→	1.00	UGUAAAU	-
	3'UTR	0.017	9.9	10/10	-	→	1.00	ACGGGUU	miR-51 to miR-56
	5'	0.017	9.6	10/10	-	-	0.37	ATGTATA	-
	3'UTR	0.015	7.2	8/10	Y	→	1.00	UUGUUGA	-

Figure S20

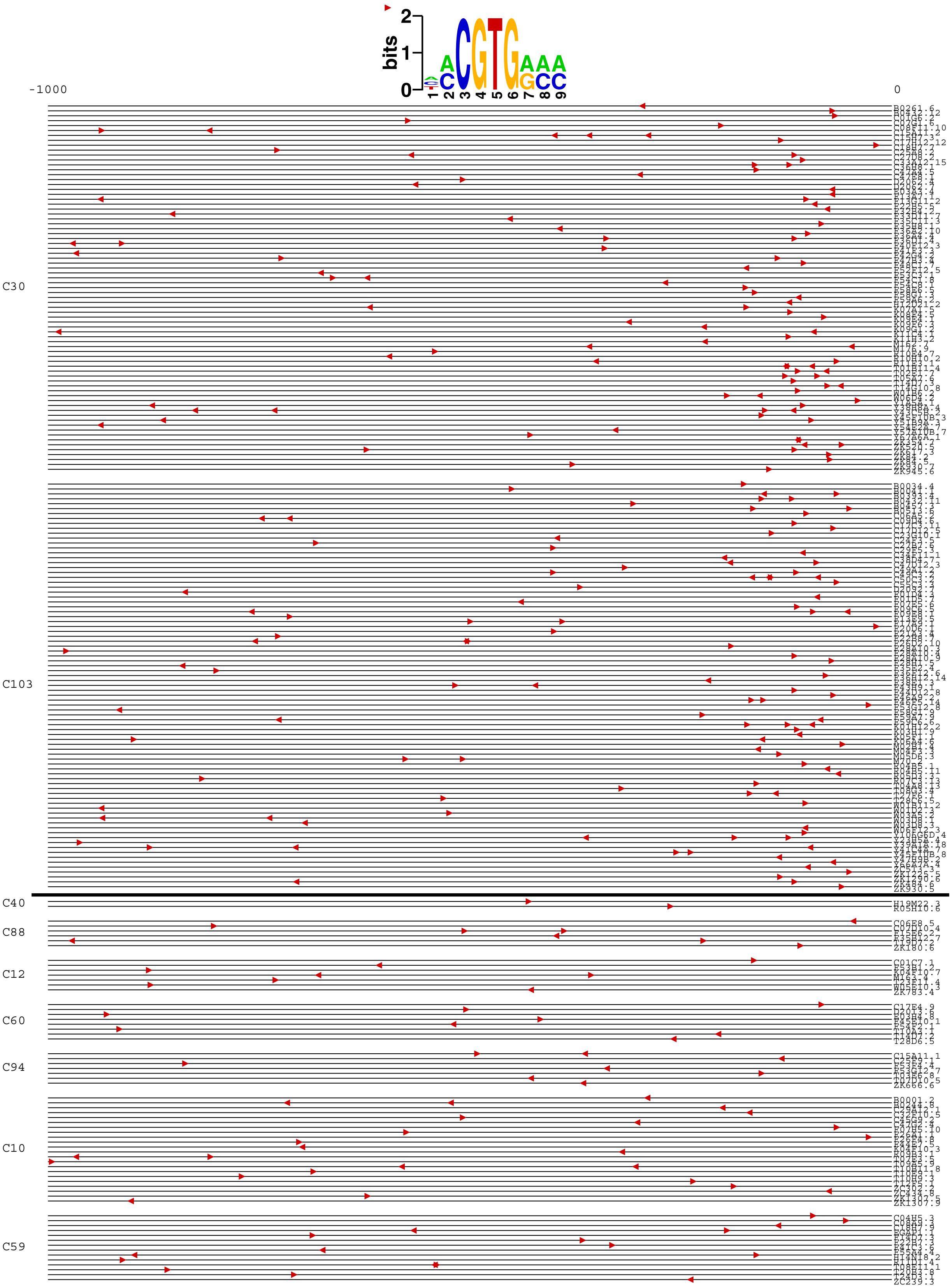

bits
2
1
0
1 2 3 4 5 6 7 8 9
-1000
0
C30
B0261.6
B0432.12
C01G6.2
C07G1.6
C08F11.10
C15A11.2
C15H7.3
C17H12.12
C18H7.7
C25A8.2
C27D8.2
C33A12.15
C36H8.1
C47A4.5
C47E8.1
D2062.4
D2062.7
E03A3.4
F13A7.1
F13G11.2
F22B5.5
F32B4.2
F33D11.7
F35C11.3
F35H8.1
F36A2.10
F36A4.4
F36D1.4
F40F12.3
F41E3.3
F42G4.2
F47B3.4
F48C1.7
F52E12.5
F53C3.1
F54C1.8
F54C8.1
F58E6.5
F58G1.3
F59A6.2
H12D21.2
K07A1.5
K08E4.5
K09E4.1
K09E6.3
K09G1.2
K11C4.1
K11H3.2
M162.7
M176.9
R10E4.7
R10H10.2
R11E3.1
T01B11.4
T02E1.7
T05A7.6
T14D7.3
T14G10.8
W01B6.2
W06D4.2
Y1A5A.1
Y38H8A.4
Y43C5B.2
Y45F10B.3
Y51B9A.3
Y54E2A.7
Y57A10B.7
Y67A6A.1
ZK354.7
ZK520.5
ZK617.3
ZK84.2
ZK84.5
ZK930.7
ZK945.6
C103
B0034.4
B0041.1
B0393.4
B0432.11
B0457.3
B0513.6
C06A5.2
C09D4.6
C17C3.11
C17D12.5
C23G10.1
C24F3.5
C27B7.6
C29E5.3
C34F11.1
C38D4.7
C47D12.3
C49A1.2
C49C3.2
C50C3.2
C55C3.3
D2092.7
F01D4.3
F01D5.7
F07E5.6
F09C8.5
F09E8.1
F13E9.5
F17A9.1
F20D6.1
F21A3.4
F22B8.7
F26D2.10
F28A10.3
F28A10.4
F28A10.9
F28H1.5
F35E2.4
F36F2.6
F36H12.14
F38E1.3
F43H9.1
F44D12.8
F46A9.2
F46F5.14
F53G12.8
F58G1.9
F59A7.9
F59C6.6
K01H12.2
K03H1.9
K05F1.1
K06A4.6
M02B1.4
M04F3.3
M05D6.3
M70.2
R04B5.1
R04B5.11
R05D3.3
R07C3.13
T04A8.13
T08G3.4
T27F6.1
T28C6.5
W01B11.2
W01D2.3
W03A5.2
W03D8.1
W03D8.3
W06F12.3
Y106G6D.4
Y23H5A.4
Y39A1A.18
Y41C4A.7
Y45F10B.8
Y47H9B.2
Y66A7A.4
ZC513.3
ZK1225.5
ZK1290.6
ZK484.6
ZK930.5
C40
H19M22.3
R05H10.6
C88
C06E8.5
C07D10.4
F15E6.2
F35B12.7
T19D7.2
ZK180.6
C12
C01C7.1
F53B1.2
K04F10.7
M163.4
T23E11.4
W05E10.3
ZK783.4
C60
C17E4.9
D2013.6
E03H4.8
F45E10.1
F54F2.1
T10A3.1
T14D7.2
T28D6.5
C94
C15A11.1
C25F9.1
F53F4.4
F53G12.7
T03E6.8
T07D10.5
ZK666.6
C10
B0001.2
B0244.8
C29A12.1
C32F10.5
C45G9.2
C47G2.4
F07H5.10
F26A1.1
F26E4.8
F44E7.5
K04F10.3
R09B3.1
T07E3.5
T09A5.9
T10B11.8
T10E9.1
T10H9.3
T12F5.1
ZC302.2
ZC434.8
ZK1307.5
ZK1307.9
C59
C04H5.3
C08A9.3
C18H7.9
EGAP1.1
F14D7.3
F22B7.3
F41C3.6
F55A4.4
H14N18.2
R11D1.4
T08E11.1
T20B3.8
T24D3.1
ZC239.3

Figure S21

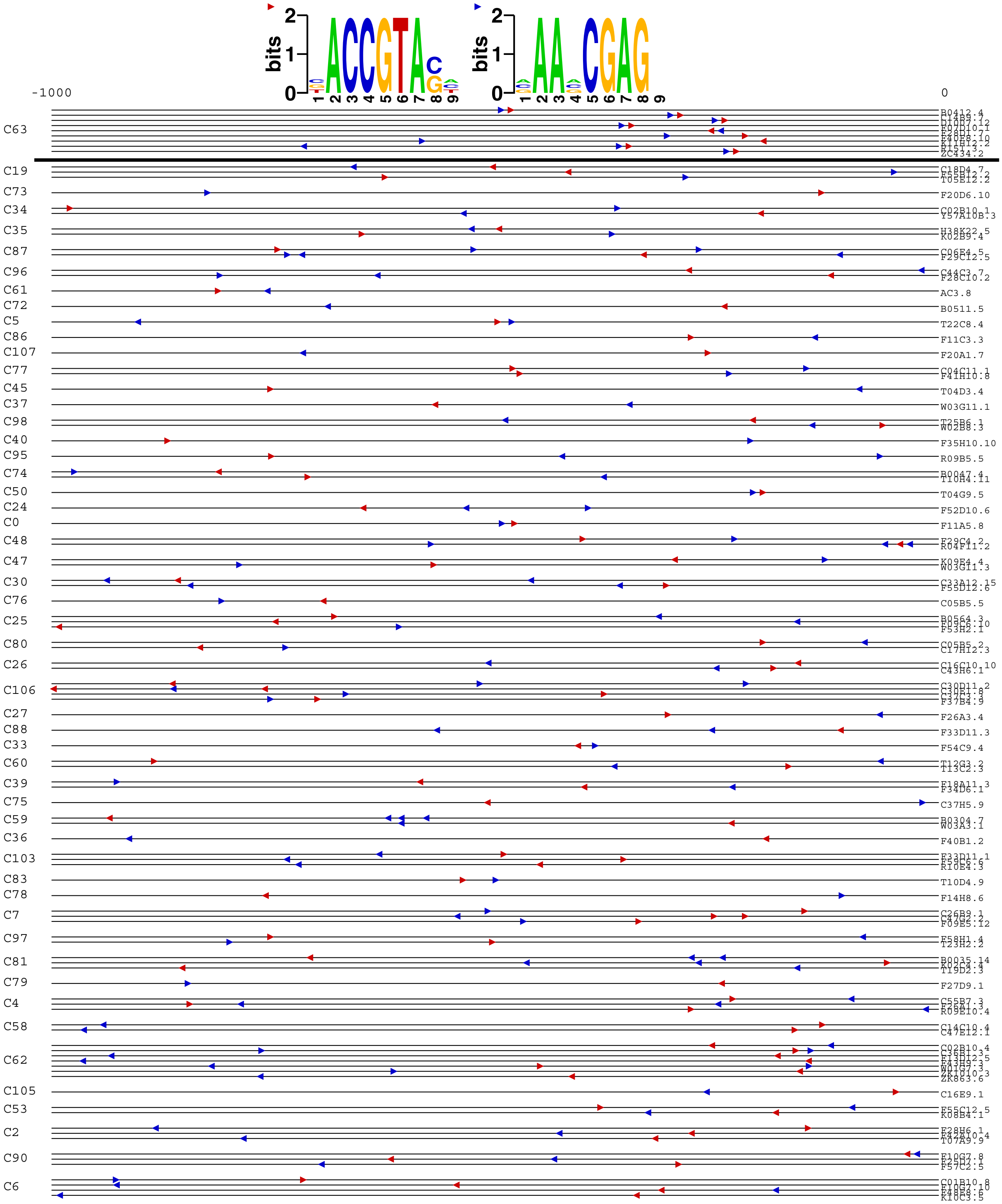

bits
2
1
0
1 2 3 4 5 6 7 8 9
ACCGTACG
bits
2
1
0
1 2 3 4 5 6 7 8 9
AACGAG
-1000
0
C63
B0412.4
C14B9.7
D1007.12
F07D10.1
F28D1.7
F40F8.10
K11H12.2
R151.3
ZC434.2
C19
C18D4.7
F55B12.2
T05E12.2
C73
F20D6.10
C34
C02B10.1
Y57A10B.3
C35
H38K22.5
K02B9.4
C87
C06E4.5
F29C12.5
C96
C44C3.7
F28C10.2
C61
AC3.8
C72
B0511.5
C5
T22C8.4
C86
F11C3.3
C107
F20A1.7
C77
C04C11.1
F41H10.8
C45
T04D3.4
C37
W03G11.1
C98
T25B6.1
W02B8.3
C40
F35H10.10
C95
R09B5.5
C74
B0047.4
T10H4.11
C50
T04G9.5
C24
F52D10.6
C0
F11A5.8
C48
F29C4.2
R04F11.2
C47
K09E4.4
W03G11.3
C30
C33A12.15
F55D12.6
C76
C05B5.5
C25
B0564.3
F09C6.10
F53H2.1
C80
C05B5.2
C17H12.3
C26
C16C10.10
C43H6.1
C106
C30D11.2
C30E1.8
C37C3.3
F37B4.9
C27
F26A3.4
C88
F33D11.3
C33
F54C9.4
C60
T12G3.2
T13C2.3
C39
F18A11.3
F34D6.1
C75
C37H5.9
C59
B0304.7
W03A3.1
C36
F40B1.2
C103
F33D11.1
F59C6.6
R10E4.3
C83
T10D4.9
C78
F14H8.6
C7
C26B9.1
C47G2.2
F09E5.12
C97
F58H1.4
T23H2.2
C81
B0035.14
K02C4.4
T19D2.3
C79
F27D9.1
C4
C55B7.3
F26A1.3
R09E10.4
C58
C14C10.4
C47E12.1
C62
C02B10.4
C36B1.3
F13D12.5
F43H9.3
W01G7.3
ZK1010.3
ZK863.6
C105
C16E9.1
C53
F55C12.5
K08B4.1
C2
F28H6.1
F42A10.4
T07A9.9
C90
F10G7.8
F25D7.1
F57C2.5
C6
C01B10.8
F10G7.10
F48E8.6
K10C3.5

Figure S22

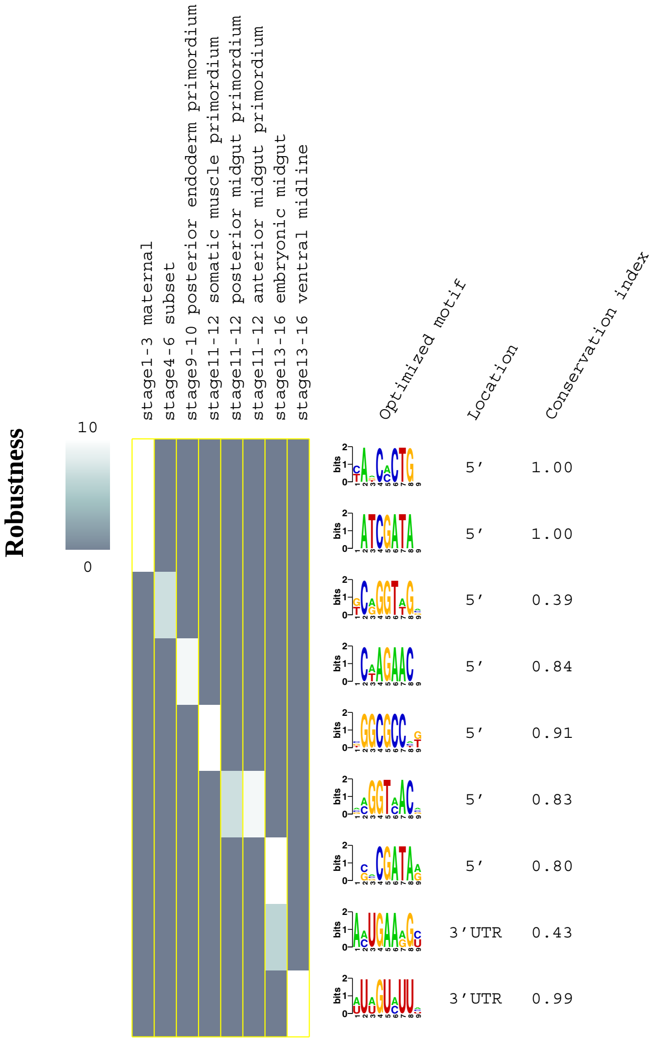
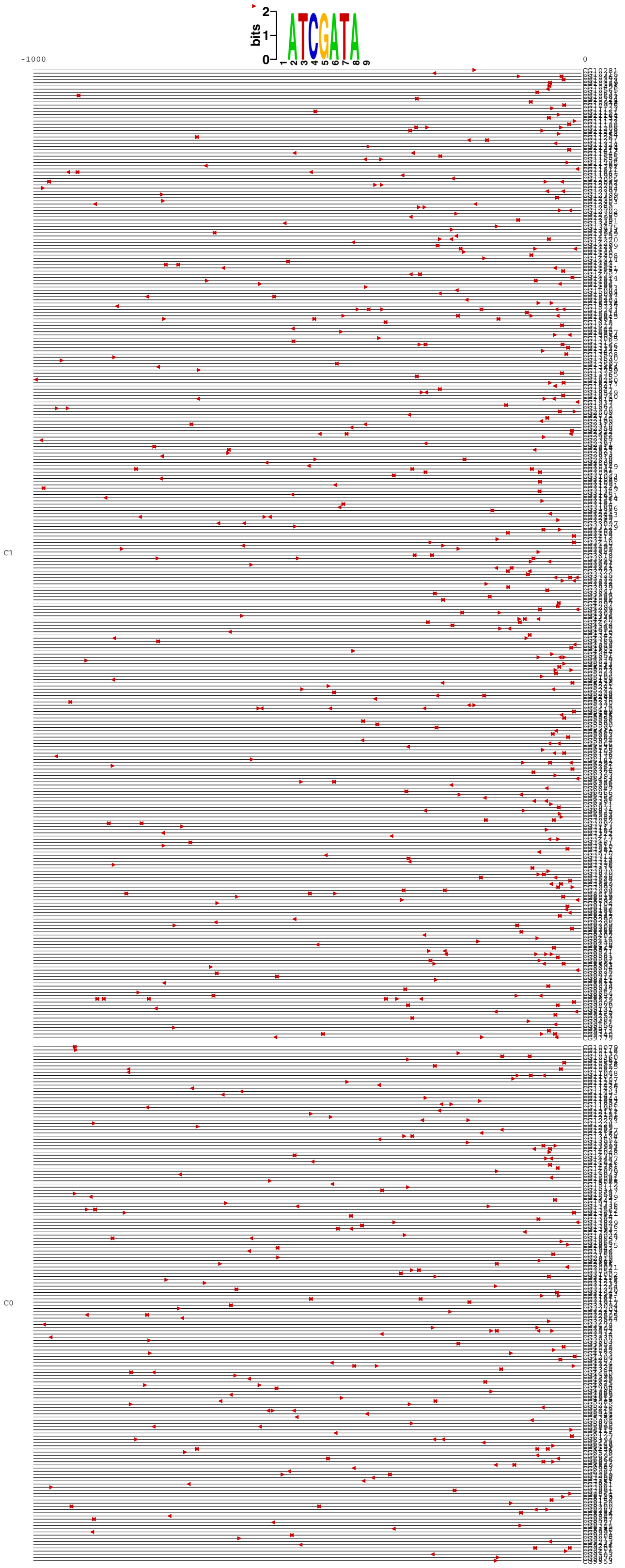

Robustness
10
0
stage1-3 maternal
stage4-6 subset
stage9-10 posterior endoderm primordium
stage11-12 somatic muscle primordium
stage11-12 posterior midgut primordium
stage11-12 anterior midgut primordium
stage13-16 embryonic midgut
stage13-16 ventral midline
Optimized motif
Location
Conservation index
5' 1.00
5' 1.00
5' 0.39
5' 0.84
5' 0.91
5' 0.83
5' 0.80
3'UTR 0.43
3'UTR 0.99

Figure S23



Captions for Supplemental Tables

Table S1: Most significant Gene Ontology enrichments for all predicted yeast motifs.

The target genes of each predicted motif are defined as genes whose promoter contain the motif and are associated with clusters in which the motif is over-represented. Enrichment of the target genes of each motif is determined with respect to Gene Ontology (GO) categories using the hypergeometric distribution, and the most significant GO category is indicated in this table along with its p-value after correcting for multiple hypotheses testing using a Bonferroni correction.

Table S2: Fraction of genes that have a given FIRE motif both in clusters where the motif is significantly over-represented ("active" clusters) and in the remaining clusters (non-active clusters).

These two fractions provide an indication of the amount of variance in the expression data that can be explained by the discovered motifs. This table shows that, in general, the percentage of the motif's non-active occurrences is relatively low, suggesting that much of the variance in the expression data is explained by the predicted motifs.

Table S3: Motif discovery from *Drosophila* enhancer data.

FIRE was applied to 124 early embryonic *Drosophila* enhancer sequences collected from the literature (D. Papatsenko, https://webfiles.berkeley.edu/~dap5/) of which 20 are known to be bound by the Bicoid morphogen. The predicted motif (on the left) matches well the previously reported Bicoid binding site and is more discriminative between Bicoid-bound enhancers and other enhancers compared to the somewhat similar motif obtained by applying AlignACE (on the right) to the same data. In a leave-one-out test, each enhancer was withdrawn and the remaining 123 were analyzed by FIRE to extract a maximally informative motif that was then used to predict whether the withdrawn enhancer is bound by Bicoid or not. The success rate (SR), sensitivity (SE) and specificity (SP) are shown.

Table S1

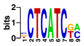
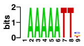
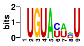
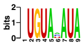
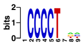
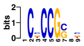
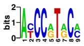
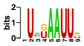
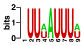
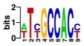
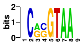
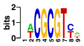
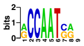
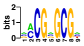
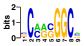
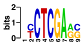
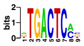
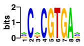
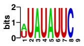
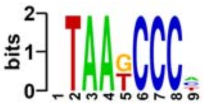
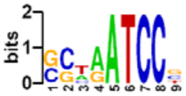
Motif	Location	Most significant functional enrichment
	5′	ribosome biogenesis, p<1e-09
	5′	cytoplasm organization and biogenesis, p<1e-09
	3′UTR	cytoplasm organization and biogenesis, p<1e-42
	3′UTR	organellar ribosome, p<1e-71
	5′	carbohydrate metabolism, p<1e-06
	5′	oxidative phosphorylation, p<1e-17
	5′	cytosolic ribosome (sensu Eukaryota), p<1e-10
	3′UTR	cytosolic ribosome (sensu Eukaryota), p<1e-54
	3′UTR	cytosolic ribosome (sensu Eukaryota), p<1e-48
	5′	proteasome complex (sensu Eukaryota), p<1e-44
	5′	modification-dependent macromolecule catabolism, p<1e-12
	5′	DNA replication, p<1e-07
	5′	oxidative phosphorylation, p<1e-38
	5′	oxidative phosphorylation, p<0.001
	5′	oxidative phosphorylation, p<1e-05
	5′	protein folding, p<0.01
	5′	amino acid metabolism, p<1e-14
	5′	sulfur utilization, p<1e-05
	3′UTR	oxidative phosphorylation, p<1e-22

Table S2

Motif	# active clusters	Fraction in active clusters	Fraction in non-active clusters
[CGT]CTCATC[GT][AC]	5	0.47	0.07
.AAAAATT[CGT]	5	0.63	0.30
.UGUA[CU][AU][ACU]U	4	0.41	0.20
.UGUA[ACU]AUA	2	0.86	0.13
.CCCCT.[AGT][ACT]	6	0.51	0.23
.C[CGT]CCG[CG].[CGT]	4	0.45	0.23
[CGT]GCA[CG][ACG]G[AC].	4	0.44	0.24
A[CT]CC[AG]T[AG]C[AC]	1	0.55	0.05
.U[ACG][GU]AAUU[AGU]	2	0.39	0.20
.UU[AU]AUUU[AU]	2	0.47	0.23
[AT]T[CT]GCCAC[CT]	3	0.32	0.04
.C[AG][CG]GTAA.	3	0.40	0.22
.[AT]CGCGT[CT][AGT]	2	0.31	0.09
[AG]CCAAT[CG][AG].	1	0.59	0.11
[ACG][AC]CG[ACG]GCG[ACT]	2	0.18	0.04
[CGT]C[AC][AG][CG]GGC[ACG]	2	0.27	0.10
.[CT]CTCGA[AG][CG]	2	0.29	0.10
[AGT]TGACTC[AC][CGT]	1	0.53	0.04
[ACG]C[AGT]CGTGA[ACG]	1	0.30	0.04
[AU]UAUAUUC.	1	0.34	0.06
.T[CT]AAT[CT]C[ACT]	1	0.46	0.26
[CGU]U[AU][AG]CG[AU]A[ACU]	1	0.22	0.09

Table S3

	FIRE motif	FIRE cross-val			AlignACE motif
		SR	SE	SP	
Bicoid		69%	50%	73%	

Data Sources

S. cerevisiae **(FIRE species name: yeast)**	
Expression dataset	173 microarrays that measure mRNA levels in response to different environmental stress conditions, originally published in (Gasch et al., 2000); downloaded from SMD (http://genome-www5.stanford.edu/).
Sequence data	Genome sequence (SGD1) downloaded from Ensembl (http://www.ensembl.org/index.html) on Aug 16th 2006.
Gene annotation	Downloaded from Ensembl on Aug 16th 2006.
Gene Identifiers	ORF names (*e.g.*, YAL001C)
Upstream regions	600bp upstream of ATG.
3'UTRs	300nt downstream of Stop codon.
Gene Ontology	Downloaded from http://www.geneontology.org on Sep 19th 2006.
Comparative genomics	Comparisons were made with *S. bayanus* 600bp upstream regions and 300nt 3'UTRs, downloaded from SGD on Aug 16th 2006. Protein sequences downloaded from Ensembl (*S. cerevisiae*) and SGD (*S. bayanus*).
Known motifs	A list of yeast motifs manually compiled by M. Beer, used previously in (Pritsker et al., 2004); two 3'UTR motifs were added to this list based on (Gerber et al., 2004).

P. falciparum **(FIRE species name: malaria)**	
Expression dataset	46 microarrays that measure mRNA levels at different time points during the parasite developmental cycle, along with gene "phase" values, originally published in (Bozdech et al., 2003); downloaded from http://malaria.ucsf.edu/supplementaldata/Datasets/ Overview_Dataset.txt.
Sequence data	Genome sequence downloaded from PlasmoDB (http://www.plasmodb.org/plasmo/home.jsp) on June 15th 2006.
Gene annotation	Downloaded from PlasmoDB on June 15th 2006.
Gene Identifiers	ORF names (*e.g.*, PFB0245c)
Upstream regions	1kb upstream of ATG.
3'UTRs	1kb downstream of Stop codon.
Gene Ontology	Downloaded from http://www.geneontology.org on Sep 22nd 2006.
Comparative genomics	Comparisons were made with *P. yoelli*. Genome sequence, downloaded from PlasmoDB on Dec 2nd 2005. Since no gene annotation was available at that time, we annotated the genome using the *P. falciparum* protein sequences and an

	(unpublished) annotation script based on Blast and Genewise (Birney et al., 2004). 1kb upstream regions and 3'UTRs were then extracted.
Known motifs	NA.

***C. elegans* (FIRE species name: worm)**	
Expression dataset	551 microarrays, originally published in (Kim et al., 2001); downloaded from http://cmgm.stanford.edu/~kimlab/topomap/kimbig
Sequence data	Genome sequence downloaded from Ensembl on Aug 3rd 2006.
Gene annotation	Downloaded from Ensembl on Aug 3rd 2006.
Gene Identifiers	ORF names (*e.g.*, 2L52.1)
Upstream regions	1kb upstream of TSS; if several TSSs are annotated, the farthest upstream one was used.
3'UTRs	300nt downstream of stop codon. If several stop codons are annotated, the farthest downstream one was used.
Gene Ontology	Downloaded from http://www.geneontology.org on Aug 4th 2006
Comparative genomics	Comparisons were made with *C. briggsae*. Genome, gene annotation, and protein sequences were downloaded from ftp://ftp.sanger.ac.uk/pub/wormbase/cbriggsae/cb25.agp8 on Sep 19th 2006. 1kb upstream regions and 300nt 3'UTRs were then extracted from the genomic sequence.
Known motifs	Compiled from TRANSFAC Release 4.0 (http://www.genome.jp/dbget-bin/show_tfmatrix) and JASPAR, Dec 2006 (http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl).
miRNAs	117 miRNA sequences, downloaded from miRBase (Griffiths-Jones, 2004) on Oct 20th 2006.

***D. melanogaster* (FIRE species name: drosophila)**	
Expression dataset	*In situ* hybridization data, originally published in (Tomancak et al., 2002) downloaded from http://www.fruitfly.org/cgi-bin/ex/insitu.pl; 124 early embryonic enhancer sequences downloaded from https://webfiles.berkeley.edu/~dap5/ .
Sequence data	Genome sequence (BDGP 4.3) downloaded from Ensembl on Sep 5th 2006.
Gene annotation	Downloaded from Ensembl on Sep 5th 2006.
Gene Identifiers	CG identifiers (*e.g.*, CG2986).
Upstream regions	1kb upstream of TSS; if several TSSs are annotated, the farthest upstream one was used.

3'UTRs	300nt downstream of stop codon. If several stop codons are annotated, the farthest downstream one was used.
Gene Ontology	Downloaded from http://www.geneontology.org on Oct 9th 2006.
Comparative genomics	Comparisons were made with *D. pseudoobscura*. Genome, gene annotation, and protein sequences were downloaded from Flybase on Jan 12th 2007. 1kb upstream regions and 300nt 3'UTRs were then extracted from the genomic sequence.
Known motifs	Same as for worm.
miRNAs	73 miRNA sequences, downloaded from miRBase (Griffiths-Jones, 2004) on Oct 19th 2006.

M. musculus **(FIRE species name: mouse)**	
Expression dataset	62 microarrays that measure mRNA levels in different mouse tissues (each array represents the average of two repeats), originally published in (Su et al., 2004); downloaded from http://wombat.gnf.org/index.html.
Sequence data	Genome sequence was downloaded from UCSC Genome Browser on Nov 5th 2006 (chromFaMasked.tar.gz).
Gene annotation	Downloaded from UCSC on Nov 5th 2006 (refGene.txt).
Gene Identifiers	RefSeq identifiers (*e.g.*, NM_144958).
Upstream regions	1kb upstream of TSS; if several TSSs are annotated, the farthest upstream one was used.
3'UTRs	300nt downstream of stop codon. If several stop codons are annotated, the farthest downstream one was used.
Gene Ontology	Downloaded from http://www.geneontology.org on Nov 7th 2006.
Comparative genomics	Comparisons were made with *G. gallus*. Genome, gene annotation, and protein sequences were downloaded from Ensembl on Nov 11th 2006. 1kb upstream regions and 300nt 3'UTRs were then extracted from the genomic sequence.
Known motifs	Same as for worm and fly.
miRNAs	375 miRNA sequences, downloaded from miRBase (Griffiths-Jones, 2004) on Jan 8th 2007.

H. sapiens **(FIRE species name: human)**	
Expression dataset	79 microarrays that measure mRNA levels in different human tissues (each array represents the average of two repeats), originally published in (Su et al., 2004); downloaded from http://wombat.gnf.org/index.html.
Sequence data	Genome sequence was downloaded from UCSC Genome Browser on Oct 1st 2006 (chromFaMasked.tar.gz).
Gene annotation	Downloaded from UCSC on Oct 1st 2006 (refGene.txt).

Gene Identifiers	RefSeq identifiers (*e.g.*, NM_018117).
Upstream regions	1kb upstream of TSS; if several TSSs are annotated, the farthest upstream one was used.
3'UTRs	300nt downstream of stop codon. If several stop codons are annotated, the farthest downstream one was used.
Gene Ontology	Downloaded from http://www.geneontology.org on July 3rd 2006.
Comparative genomics	Comparisons were made with *G. gallus*. Genome, gene annotation, and protein sequences were downloaded from Ensembl on Nov 11th 2006. 1kb upstream regions and 300nt 3'UTRs were then extracted from the genomic sequence.
Known motifs	Same as for worm, fly, and mouse.
miRNAs	420 miRNA sequences, downloaded from miRBase (Griffiths-Jones, 2004) on Oct 19th 2006.

Supplemental References

Andres, V., Nadal-Ginard, B., and Mahdavi, V. (1992). Clox, a mammalian homeobox gene related to Drosophila cut, encodes DNA-binding regulatory proteins differentially expressed during development. Development *116*, 321-334.

Arnosti, D. N., and Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem *94*, 890-898.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res *14*, 988-995.

Bozdech, Z., Llinas, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol *1*, E5.

Chan, C. S., Elemento, O., and Tavazoie, S. (2005). Revealing posttranscriptional regulatory elements through network-level conservation. PLoS Comput Biol *1*, e69.

Cover, T., and Thomas, J. (2006). Elements of Information Theory, 2nd edn (Hoboken, NJ, Wiley-Interscience).

De Renzis, S., Elemento, O., Tavazoie, S., and Wieschaus, E. (2007). Unmasking Activation of the Zygotic Genome Using Chromosomal Deletions in the Drosophila Embryo. PLoS Biol (in press).

Driever, W., and Nusslein-Volhard, C. (1989). The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo. Nature *337*, 138-143.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The annals of statistics *7*, 1-26.

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.

Foat, B. C., Houshmandi, S. S., Olivas, W. M., and Bussemaker, H. J. (2005). Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proc Natl Acad Sci U S A *102*, 17675-17680.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell *11*, 4241-4257.

Gerber, A., Herschlag, D., and Brown, P. (2004). Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol *2*, E79.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res *32 Database issue*, D109-111.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol *296*, 1205-1214.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res *110*, 462-467.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *423*, 241-254.

Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., and Davidson, G. S. (2001). A gene expression map for Caenorhabditis elegans. Science *293*, 2087-2092.

Lewis, B., Burge, C., and Bartel, D. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Llinas, M., and DeRisi, J. L. (2004). Pernicious plans revealed: Plasmodium falciparum genome wide expression analysis. Curr Opin Microbiol *7*, 382-387.

Matsukage, A., Hirose, F., Hayashi, Y., Hamada, K., and Yamaguchi, M. (1995). The DRE sequence TATCGATA, a putative promoter-activating element for Drosophila melanogaster cell-proliferation-related genes. Gene *166*, 233-236.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., *et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-110.

Murakami, R., Okumura, T., and Uchiyama, H. (2005). GATA factors as key regulatory molecules in the development of Drosophila endoderm. Dev Growth Differ *47*, 581-589.

Pritsker, M., Liu, Y., Beer, M., and Tavazoie, S. (2004). Whole-genome discovery of transcription factor binding sites by network-level conservation. Genome Res *14*, 99-108.

Slonim, N., Atwal, G., Tkacik, G., and Bialek, W. (2005). Estimating mutual information and multi-information in large networks.

Struhl, G., Struhl, K., and Macdonald, P. M. (1989). The gradient morphogen bicoid is a concentration-dependent transcriptional activator. Cell *57*, 1259-1273.

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.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281-285.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E., and Rubin, G. M. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol *3*, R88.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006). A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res *34*, D95-97.

Wickens, M., Bernstein, D., Kimble, J., and Parker, R. (2002). A PUF family portrait: 3'UTR regulation as a way of life. Trends Genet *18*, 150-157.

Xie, X., Lu, J., Kulbokas, E., Golub, T., Mootha, V., Lindblad-Toh, K., Lander, E., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.