Nucleic Acids Research Advance Access published January 16, 2009

Nucleic Acids Research, 2009, 1–14
doi:10.1093/nar/gkn1064

Allegro: Analyzing expression and sequence in concert to discover regulatory programs

Yonit Halperin, Chaim Linhart, Igor Ulitsky and Ron Shamir*

School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Received November 10, 2008; Revised and Accepted December 16, 2008

ABSTRACT

A major goal of system biology is the characterization of transcription factors and microRNAs (miRNAs) and the transcriptional programs they regulate. We present Allegro, a method for *de-novo* discovery of *cis*-regulatory transcriptional programs through joint analysis of genome-wide expression data and promoter or 3′ UTR sequences. The algorithm uses a novel log-likelihood-based, non-parametric model to describe the expression pattern shared by a group of co-regulated genes. We show that Allegro is more accurate and sensitive than existing techniques, and can simultaneously analyze multiple expression datasets with more than 100 conditions. We apply Allegro on datasets from several species and report on the transcriptional modules it uncovers. Our analysis reveals a novel motif over-represented in the promoters of genes highly expressed in murine oocytes, and several new motifs related to fly development. Finally, using stem-cell expression profiles, we identify three miRNA families with pivotal roles in human embryogenesis.

INTRODUCTION

One of the main challenges in molecular biology is to understand the regulatory program that controls mRNA levels. The key components of this program are transcription factors (TFs), proteins that activate or repress transcription of a gene by binding to short DNA sequences, termed transcription factor binding sites (TFBSs), which usually reside in the gene's promoter. The level of translated mRNA of a gene can also be decreased post-transcriptionally, through annealing of microRNAs (miRNAs) to the 3′ UTR of the mRNA. A key step in reverse engineering regulatory networks is computational analysis of genome-wide measurements of mRNA levels obtained from DNA microarray assays in various environmental conditions, biological samples and time-points (henceforth we use the term *condition* to refer to each microarray assay). The purpose of this analysis is to identify groups of genes that are co-regulated, also termed *transcriptional modules* (TMs), and to characterize their regulators. A two-step approach is most commonly used [see examples in (1,2) and the review in (3)]: In the first step, a clustering procedure is executed to partition the genes into groups believed to be co-regulated, based on expression profile similarity (4). In the second step, a motif discovery tool is applied to search for abundant sequence patterns in the promoters (or 3′ UTRs) of each group that may represent the binding sites (BSs) of TFs (or miRNAs) that regulate the corresponding genes.

Despite extensive research, motif discovery has had limited success due to the short and degenerate nature of BSs, and the high levels of complexity of transcriptional networks, especially in metazoans. Since both the expression profiles and the promoter sequences of the genes carry information regarding their regulation, a methodology that utilizes both sources of information may give better results than the two-step approach. Several studies proposed computational schemes for this parallel analysis. Most of these algorithms use a unified probabilistic model over both gene expression and sequence data, and assume a Gaussian distribution of the expression values (5–7). Additional examples are the algorithms Reduce (8) and Motif Regressor (9), which search for motifs correlated with a *single* condition using linear regression, and assume that the number of BSs and their affinity are linearly correlated with the gene's expression. The algorithm DRIM (10) uses the hypergeometric (HG) score to compute the enrichment of motif occurrences among the top-ranked genes. However, it too is limited to a single condition.

Here we present Allegro (A Log-Likelihood based Engine for Gene expression Regulatory motifs Over-representation discovery), a *de-novo* motif discovery platform for simultaneously detecting gene sets with coherent expression profiles and corresponding over-represented sequence patterns. A graphic overview of the Allegro

*To whom correspondence should be addressed. Tel: +972 3 640 5383; Fax: +972 3 640 5384; Email: rshamir@post.tau.ac.il

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

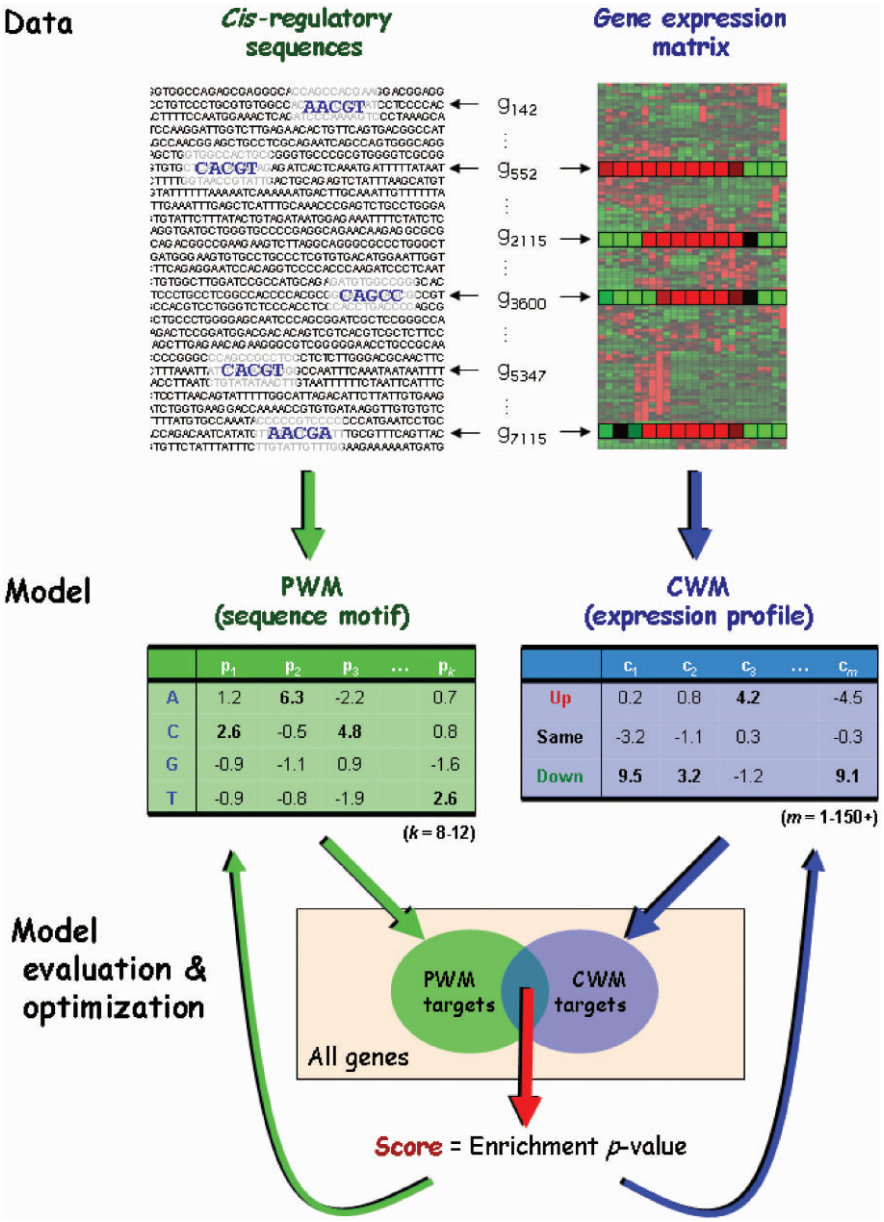

Figure 1. Overview of the Allegro computational approach. Given a genome-wide expression matrix and *cis*-regulatory sequences (promoters or 3′ UTRs), Allegro executes efficient algorithms and statistical analyses to search for transcriptional modules. A transcriptional module is a set of genes sharing a sequence motif, modeled using a PWM, and a common expression profile described using a novel model called CWM. The CWM is analogous to the PWM: it assigns a weight to each discrete expression level in each of the experimental conditions. Allegro uses a multi-phase motif enumeration engine to generate candidate motifs. For each motif, it applies a cross-validation-like procedure to construct a CWM (Supplementary Figure 2), such that there is a significantly large overlap between the targets of the motif (the set of genes whose *cis*-regulatory sequence has an occurrence of the PWM, left arrows at the top) and the targets of the CWM (the genes whose expression levels match the CWM, right arrows). The statistical significance of this overlap is evaluated using one of two enrichment scores: the HG score or the binned enrichment score, which accounts for biases in the length and GC-content of the *cis*-regulatory sequences. The scores obtained by the motifs and their CWMs are iteratively modified to improve the models and eventually converge to high-scoring transcriptional modules.

approach is presented in Figure 1. Unlike existing methods, which rely on statistical assumptions, Allegro uses a novel non-parametric model called *Condition Weight Matrix* (CWM) to describe the expression profile of a group of co-regulated genes. We show that this model represents the expression profiles of sets of co-regulated genes more accurately than do commonly used expression metrics and statistical distributions. Allegro builds upon a motif discovery software platform we recently developed called Amadeus (11). In brief, given a set of co-regulated genes, Amadeus searches for motifs that are over-represented in their *cis*-regulatory sequences with respect to (w.r.t.) the rest of the *cis*-regulatory sequences in the genome or some other background (BG) set (see Supplementary Data for additional information). Allegro utilizes the efficient motif search engine of

Amadeus to enumerate a huge number of candidate motifs and to converge to high-scoring ones. For each candidate motif, Allegro fits a CWM to its putative targets using a cross-validation-like procedure. In order to ascertain whether the motif and the CWM are significantly correlated, Allegro computes one of two enrichment scores: the HG score or the binned enrichment score (11). As we demonstrate, the latter is very useful in cases where the expression profiles are correlated to the length and GC-content of the *cis*-regulatory sequences. Such expression-sequence dependencies are ignored by most existing methods, leading to many false predictions.

To test the performance of our method and highlight its unique features and advantages over existing approaches, we applied Allegro on several large-scale datasets from yeast, fly, mouse and human. In all cases, Allegro successfully recovered binding motifs of TFs and miRNAs that are known to regulate the relevant processes, together with their corresponding expression profiles. In addition, we report on novel transcriptional modules discovered by Allegro in datasets of human and murine tissues, and in *Drosophila* tissues profiled during various stages of development. For example, we discovered a novel motif that is over-represented in the promoters of genes that are highly induced in oocytes and fertilized eggs. Application of Allegro to expression profiles of human stem cell lines highlighted three miRNA families as key players in regulation of cell fate in embryogenesis. The miRNA activities predicted based on these findings are in good agreement with evidence from recent miRNA expression studies. A comparison of our results with those obtained by several current methods for clustering and motif finding indicates that Allegro is more sensitive and accurate. We also demonstrate additional important advantages of our approach, including joint analysis of multiple expression datasets from several organisms, and accounting for correlations between the expression levels of genes and the length and GC-content of their *cis*-regulatory sequences. We believe that Allegro introduces significant novel ideas in computational motif finding and gene expression analysis. On the practical side, our software can serve as an accurate, feature-rich, user-friendly tool for the biological community.

METHODS

Genomic sequences and binding patterns

Promoter sequences (repeat- and coding-sequence-masked) of human, mouse and fly, and 3′ UTR sequences (repeat-masked) of human were extracted from Ensembl (12). Yeast promoters were downloaded from SGD (http://www.yeastgenome.org). Motifs reported by Allegro were compared to known binding patterns of TFs and miRNAs taken from Transfac (13) and miRBase (http://microrna.sanger.ac.uk/sequences), respectively. See Supplementary Data for more details.

Gene expression datasets

The expression dataset for the yeast osmotic-stress response was downloaded from the supplementary material of (14). The values in the data are $\log_2$ of the fold change w.r.t. wild-type (WT) grown on YPD medium at standard osmolarity.

The human cell-cycle dataset was obtained from the supporting web-site of (15) (http://genome-www.stanford.edu/Human-CellCycle/HeLa). Expression values are $\log_2$ of the fold change w.r.t. asynchronously grown HeLa cells.

Human and mouse tissue expression datasets were downloaded from the GNF SymAtlas web-site (http://symatlas.gnf.org/SymAtlas, version 1.2.4, gcRMA-analzyed) (16). We applied quantile normalization (17), as implemented in Expander (18), in order to rescale the expression values in each tissue to a common distribution. We then normalized the values of each gene by computing the $\log_2$ of the fold change w.r.t. the gene's average expression value.

The human stem cells dataset is the stem cell matrix described in (19) (GEO accession number GSE11508). After averaging technical replicates, this dataset contains 124 samples. The full list of cell types used appears in Supplementary Table V. The expression pattern of each gene was normalized to mean 0 and SD 1.

The datasets analyzed in this study are summarized in Supplementary Table I. See Supplementary Data for additional details.

CWM for a motif target set

Denote by B the set of genes in the expression data, and let $e_{g(1)}, \ldots, e_{g(m)}$ denote the *discrete expression levels* (DELs) of gene $g \in B$ ($\forall 1 \leq j \leq m$, $e_{g(j)} \in \{e_1, \ldots, e_l\}$). The background *condition frequency matrix* (CFM), $R = \{r_{i,j}\}$, holds the frequencies of the DELs in each condition across all genes: $r_{i,j} = |\{g \in B \mid e_{g(j)} = e_i\}|/|B|$. For a candidate motif M, denote by T its target set, i.e. the group of genes whose *cis*-regulatory sequences contain an occurrence of M. As described in the Results section, Allegro samples a training set S from T, and constructs a CFM $F = \{f_{i,j}\}$ based on the DELs of the genes in S: $f_{i,j} = |\{g \in S \mid e_{g(j)} = e_i\}|/|S|$. The training-set sampling procedure is described in the Supplementary Data. Allegro then calculates the CWM, which contains the log-ratios between F and R:

$$\forall 1 \leq i \leq l, 1 \leq j \leq m \; CWM(i,j) = \log\left(\frac{f_{i,j}}{r_{i,j}}\right)$$

Allegro uses the CWM to compute the log-likelihood ratio (LLR) score of every gene, as explained below.

LLR score computation

Given the background CFM, $R = \{r_{i,j}\}$, and a CFM, $F = \{f_{i,j}\}$, learnt from the target set of a candidate motif, Allegro computes the LLR score of all the genes, as described in the Results section. The naïve computation takes $O(|B| \cdot |C|)$ time, where B is the set of genes and C is the set of conditions. Different genes may share the same discrete pattern, so the time complexity can be improved to $O(|P| \cdot |C|)$, where P is the set of distinct discrete expression patterns observed in the dataset. For example, in the tissues dataset (16) there are 14 698 human genes but only 2112 distinct expression patterns, so the above

observation gives a 7-fold speedup in this case. Another running time improvement is achieved by reducing the average number of operations per discrete pattern in the LLR computation, as follows. In a preprocessing procedure we build a complete weighted graph, G_P, in which the nodes correspond to the patterns in P, and the weight of an edge is the Hamming distance between the two corresponding patterns. We then find a minimum spanning tree (MST) of G_P, denoted T_P. In order to compute the LLR score of all the patterns in P, we scan T_P in preorder, and use the LLR score of each pattern as a basis for computing the scores of its child nodes. Formally, let $v = (e_{v(1)}, \ldots, e_{v(|C|)})$ be a discrete expression pattern. If v is the root of T_P, the LLR is calculated naïvely, as described in the Results. Otherwise, let $u = (e_{u(1)}, \ldots, e_{u(|C|)})$ be the parent of v in T_P, then:

$$\mathrm{LLP}(v) = \mathrm{LLR}(u) + \sum_{j \in D_{uv}} \left(\log\left(\frac{f_{v(j),j}}{r_{v(j),j}}\right) - \log\left(\frac{f_{u(j),j}}{r_{u(j),j}}\right) \right)$$

where D_{uv} is the set of conditions, in which the DELs in u and v differ ($|D_{uv}|$ is the Hamming distance between u and v). Note that since T_P is scanned in preorder, LLR(u) is calculated before LLR(v), as required. In preprocess, we compute a table that contains the value $\log(f_{i,j}/r_{i,j}) - \log(f_{k,j}/r_{k,j})$ for every pair of DELs, e_i and e_k, and every condition c_j. Using this table, LLR(v) can be calculated given LLR(u) in time $c \cdot |D_{uv}|$, where, c is a very small constant. Thus, the total time complexity of computing the LLR score of all patterns is $O(|P| \cdot d + |C|)$, where d is the average Hamming distance in the MST (the second summand, |C|, is the time for the LLR computation of the root). In the human tissues dataset mentioned above, there are 79 tissues, but the average distance in T_P is only 1.31. Thus, using the MST gives a further 59-fold time improvement.

Enrichment scores

For each candidate motif, we use a subset S of its target genes ($S \subset T$) as a training set for learning a CWM, as described in the Results section. The set of all other genes in the expression data, denoted B_s ($B_s = B \backslash S$), is used for evaluating the fit between the CWM and the motif, as follows. Let W ($W \subseteq B_s$) be the set of genes, whose expression pattern obtained an LLR score higher than the current CWM cutoff (Allegro tries several cutoffs), excluding the training-set genes. Denote by b and w the subset of genes from B_s and from W, respectively, that contain at least one occurrence of the motif in their *cis*-regulatory sequence, i.e. $b = B_s \cap T$ and $w = W \cap T$. Allegro computes one of two supported enrichment scores, as specified by the user, to assess whether the motif is over-represented in W, i.e. whether w is significantly larger than expected, given B_s, W and b. The first score, called the HG enrichment score, uses the HG tail distribution to compute the probability of observing at least $|w|$ sequences in W with a motif hit, under the null hypothesis that the genes in W were drawn randomly, independently and without replacement from B_s:

$$\text{HG score} = \mathrm{HGtail}(|B_s|, |W|, |b|, |w|) = \sum_{i=|w|}^{\min(|W|,|b|)} \frac{\binom{|b|}{i}\binom{|B_s|-|b|}{|W|-i}}{\binom{|B_s|}{|W|}}$$

The second score, called the binned enrichment score, accounts for cases where the expression values are correlated with the length or GC-content of the *cis*-regulatory sequences. In short, the genes are divided into bins according to the length and GC-content of their *cis*-regulatory sequence. The counts of the number of genes in each bin that passed the LLR cutoff and the number of genes with a hit in their sequence are used in order to estimate the overall enrichment. For exact details, see (11).

Clustering and motif-finding tools

K-means (20) and CLICK (21) were executed using the Expander gene expression analysis software (18). K-means was run twice—with $k = 10$, and with $k = 20$. CLICK was run with the 'homogeneity' parameter set to 0.3. Two motif-finding tools, Weeder and Amadeus, were applied on all clusters found by K-means and CLICK, excluding huge clusters with more than 900 genes. Weeder (v1.3) was executed with the 'medium T100 S' parameters and using the BG model files supplied with the software (22). Amadeus (v1.0) was run with its default settings (11).

GO functional analysis

For each motif discovered by Allegro in the tissues datasets, we ran the TANGO algorithm via the Expander software (18) to test whether the CWM targets of the motif are enriched for Gene Ontology biological process terms. TANGO performs a bootstrapping procedure to correct the enrichment p-values for multiple testing and account for the large overlaps between related GO terms. All results reported here obtained a p-value less than 10^{-9} and a corrected p-value less than 10^{-3}.

RESULTS

We developed a novel method, called Allegro, for simultaneous *de novo* discovery of regulatory sequence motifs and the expression profiles they induce in one or more genome-wide gene expression datasets. Given a candidate motif, Allegro learns an expression model that describes the shared expression profile of the genes, whose *cis*-regulatory sequence contains the motif. It then computes a p-value for the over-representation of the motif within the *cis*-regulatory sequences of the genes that best fit the expression profile. We implemented Allegro and integrated it with our Amadeus motif discovery platform. Amadeus executes a series of refinement phases to converge to high-scoring motifs. Each phase receives as input a list of candidate motifs, applies an algorithm for refining the list, and produces a set of improved candidates that constitute the starting point for the next phase. The output of Amadeus is a non-redundant list of top-scoring motifs, modeled using position weight matrices (PWMs).

Additional scoring functions and features available in Amadeus are described in (11). In the current study, motifs in each phase are evaluated using Allegro. Thus, the motifs reported by the algorithm are those that possess the highest correlation to the expression data in terms of the aforementioned *p*-value.

In the following sections we introduce the expression model used by Allegro and demonstrate its advantages over commonly used approaches. We then describe the algorithm Allegro applies to ascribe a *p*-value for a given motif. Finally, we present results of applying Allegro to several large-scale expression datasets representing a diverse set of biological systems and species, and compare them to those obtained by existing tools.

Modeling the expression profile of co-regulated genes

We developed a new method for modeling the expression profile shared by a group of co-regulated genes. Unlike existing approaches, it does not make complex statistical assumptions about the distribution of the expression values in each condition. Furthermore, unlike expression similarity measures employed by clustering techniques, our model is robust against extreme values and can describe profiles that differ across a very small number of conditions. The model is analogous to the PWM model for sequence motifs (23,24), with DNA bases substituted here by discrete expression levels, and the positions along the motif replaced by the experimental conditions.

Given continuous expression values, Allegro first transforms them into *discrete expression levels* (DELs, in short): $e_1, e_2, \ldots, e_l$. The number of expression levels (l) and the range of values that define each one are set by the user. For example, if the expression values are given in $\log_2$ ratios w.r.t. some base condition, then one may use three DELs, as illustrated in Supplementary Figure 1: Expression values above 1.0 are replaced by e_1 (or 'U', for 'Up-regulated'), values between -1.0 and 1.0 are replaced by e_2 (or 'S', for 'Similar to base condition') and values below -1.0 are replaced by e_3 (or 'D', for 'Down-regulated'). The DELs may also be defined using percentiles rather than cutoffs.

Let $c_1, c_2, \ldots, c_m$ be the set of m conditions in the given expression matrix. The expression model assigns to each condition a discrete probability distribution. Define an $l \times m$ matrix, called *condition frequency matrix* (CFM), in which column j holds the distribution of the DELs in condition c_j according to the model. Hence, the value in row i and column j is the probability of generating expression level e_i in condition c_j (Supplementary Figure 1). The background CFM, $R = \{r_{i,j}\}$, is computed from the observed DELs of all given genes; i.e. $r_{i,j}$ is the BG frequency of expression level e_i in condition c_j (see Methods section).

Given another CFM $F = \{f_{i,j}\}$, which models the expression levels of a transcriptional module T, we would like to assign to each gene a score that quantifies its similarity to F. To this end, we use the standard likelihood ratio approach, as follows. Let $e_{g(j)}$ $(1 \le j \le m)$ denote the DEL of gene g in condition c_j. The LLR score of g is the logarithm of the ratio between the probability of observing these expression levels under the assumption that gene g belongs to T, and the probability of observing them under the null hypothesis:

$$\text{LLR(expression of gene } g) = \sum_{j=1}^{m} \log\left(\frac{f_{g(j),j}}{r_{g(j),j}}\right)$$

The $l \times m$ matrix whose entries are $\log(f_{i,j}/r_{i,j})$ is called the CWM. The CWM can be used to classify genes as belonging to the transcriptional module T in the standard way: for a given threshold h, a gene is considered to belong to T if its LLR score is above h. In a sense, the CWM represents an expression motif similarly to the standard sequence motif representation using a PWM. In the next section we explain how the CWM and the threshold h are computed for a putative transcriptional module.

We tested how well the CWM model identifies the expression profile of known transcriptional modules, and compared its performance to that of popular expression metrics: Pearson correlation, Spearman's rank correlation, and Euclidean distance (4). The results show that in most cases (16 out of 18) our model describes the expression profile of TMs more accurately than existing approaches (Supplementary Table II). The experimental procedure and results are detailed in the Supplementary Data.

Learning the expression profile induced by a motif

For each candidate motif, Allegro tries to learn a CWM that describes the expression of (some of) its targets. If the motif represents BSs of a TF that is active in the measured conditions, Allegro will likely find a CWM that is characteristic of the motif's targets; otherwise, the expression values of the target genes are expected to behave like the BG distribution, and no such CWM will be found. Let T denote the set of genes whose *cis*-regulatory sequences contain at least one occurrence, or *hit*, of the motif M. Allegro finds a CWM that models the expression profile of T by executing a cross-validation-like procedure, illustrated in Supplementary Figure 2. First, it samples a training set from T and generates a CFM F based on the frequencies of DELs in that training set. A CWM is computed from F and from the background CFM, as explained earlier (see Methods section). Then, for all genes excluding those in the training set, it computes the LLR score described above. In order to ascertain that the motif M is over-represented in the genes with a high LLR score (i.e. genes whose expression is more similar to the profile represented by F than to the background CFM), Allegro computes one of two enrichment scores developed in Amadeus: the HG score and the binned enrichment score. The latter accounts for biases in the length and nucleotide composition of the regulatory sequences (see Methods section). Note that the training-set genes are ignored when computing the enrichment score in order to avoid over-fitting. The enrichment score is computed for several LLR cutoffs and the best one is chosen and Bonferroni-corrected for multiple testing. Allegro repeats this process for several training sets, which are sampled in a judicious procedure that takes into account both the

expression and sequence data (see Supplementary Data). Finally, Allegro chooses the CWM that yielded the best enrichment score, and this score is set as the *p*-value of the motif. We use the term *CWM targets* to refer to the genes that passed the LLR cutoff of the top-scoring CWM. For an arbitrary motif, only a relatively small fraction of the CWM targets are also targets of the motif (i.e. contain a hit for the PWM in their *cis*-regulatory sequence), whereas, for a biologically relevant motif, the overlap between the set of its PWM targets and the set of its CWM targets is significantly large (in the sense of the enrichment score).

As described earlier, Allegro examines a large number of candidate motifs in a series of refinement phases. The motifs in each phase are ranked according to the above enrichment score. We implemented sophisticated data-structures and algorithms in order to speed-up the CWM learning procedure (see Methods section). The output of the Allegro algorithm is a list of transcriptional modules, each one comprised of a sequence motif (PWM) and an expression profile (CWM) that are highly correlated in terms of the genes they match.

Test case: human cell cycle

Whitfield *et al*. studied cell-cycle regulation using cDNA microarrays that measured gene expression profiles of HeLa cells over five time courses (15). In each time course, the cells were synchronized to the same cell-cycle phase by one of three different methods. In order to identify cell-cycle genes and the phases in which they are active, Whitfield *et al*. quantified the periodicity of the expression levels of each gene using Fourier transform, and compared it to that of known cell-cycle genes. Several studies utilized their findings to analyze the transcriptional programs underlying the cell-cycle phases (25–28).

In order to test the ability of our method to uncover transcriptional modules *ab initio* from a large mammalian dataset, we applied it to the cell-cycle data of Whitfield *et al*. The input to Allegro consisted of expression values across 111 time points and of 1200 bps-long promoter sequences of ~15 000 genes. Consistent with biological knowledge and previous studies, the three top-scoring motifs found by Allegro are the BS patterns of E2F and NF-Y (CCAAT-box), and the motif termed CHR (cell-cycle genes homology region), whose binding protein is yet to be discovered (29) (see Supplementary Data for information on how the motifs are matched to known BS patterns). As shown in Figure 2, the expression of the CWM targets of E2F peaks in the G_1/S phase, whereas genes associated with NF-Y and CHR are active in the G_2 and M phases. Importantly, these results were obtained by analyzing the expression and sequence data alone, without using any prior knowledge on periodicity of human cell-cycle or on known phase-specific genes.

An additional test case on expression data of the innate immune response in mouse is described in the Supplementary Data.

Comparison to the two-step approach: yeast HOG pathway

The *Saccharomyces cerevisiae* high osmolarity glycerol (HOG) pathway is required for osmoadaptation. It contains two branches that activate the protein Hog1 via Pbs2, one containing Ssk1 and the other containing Sho1 and Ste11. O'Rourke *et al*. characterized the roles of Hog1, Pbs2, Ssk1, Sho1 and Ste11 in response to elevated osmolarity using whole-genome expression profiling (14). The expression data contain osmotic shock profiles of the WT strain, and of mutant strains in which components of the HOG pathway were knocked-out. The profiles were monitored at different levels of hyper-osmolarity at several time points. In addition, the transcriptional response of the WT strain to the mating pheromone α-factor was measured at four time points. Overall, the dataset consists of expression values of 5758 genes in 133 conditions.

The seven top-scoring motifs reported by Allegro for this dataset are the RRPE, PAC and STRE (stress response element) motifs, and the BS patterns of Rap1, MBF, Ste12 and Sko1 (Figure 3). Remarkably, all seven motifs are related to osmotic shock (30–33). For example, Msn2 and Msn4 mediate a general stress response through binding to STRE (31,34), and they are also controlled by Hog1 (33). Indeed, the CWM targets of STRE are up-regulated in the time series of exposure to high osmolarity. Another example, which provides further evidence of the sensitivity of our approach, is Sko1, one of the main TFs that control the specific response to hyper-osmolarity (33). Under normal conditions, Sko1 recruits the general repressor complex Tup1–Ssn6 and together they act to repress their target genes. After osmotic shock, Hog1 phosphorylates Sko1, resulting in decreased affinity for Tup1, and Sko1 then activates transcription by an unknown mechanism. Reassuringly, Allegro uncovered the Sko1 binding motif, and its CWM targets are considerably up-regulated in response to high osmolarity only in strains in which Hog1 and Pbs2 were not knocked out. See Supplementary Data for additional analysis of the results.

We applied the standard two-step approach to the HOG dataset to check whether the transcriptional modules discovered by Allegro can also be found using existing techniques. We first performed clustering using three methods—*k*-means with $k = 10$ and $k = 20$ (20), and the CLICK algorithm (21), which resulted in 38 clusters. Four of these clusters were huge (>900 genes, i.e. >20% of the entire gene set) and did not exhibit an interesting expression profile, so we ignored them. We then executed two motif finding tools on each of the 34 remaining clusters: Weeder (22), which out-performed 13 other tools in a large-scale assessment (35), and Amadeus, our recently published software (11). Following (11,35), from each such execution we examined the two top-scoring motifs reported by the motif finder. We thus examined a total of 68 motifs discovered by the clustering and motif-finding pipeline. As listed in Table 1, out of the seven motifs Allegro discovered, only four were found by the two-step approach—RRPE, PAC, MBF and STRE. We also applied the clustering and motif-finding tools developed by Slonim *et al*., Iclust (36) and FIRE (37). Again, only

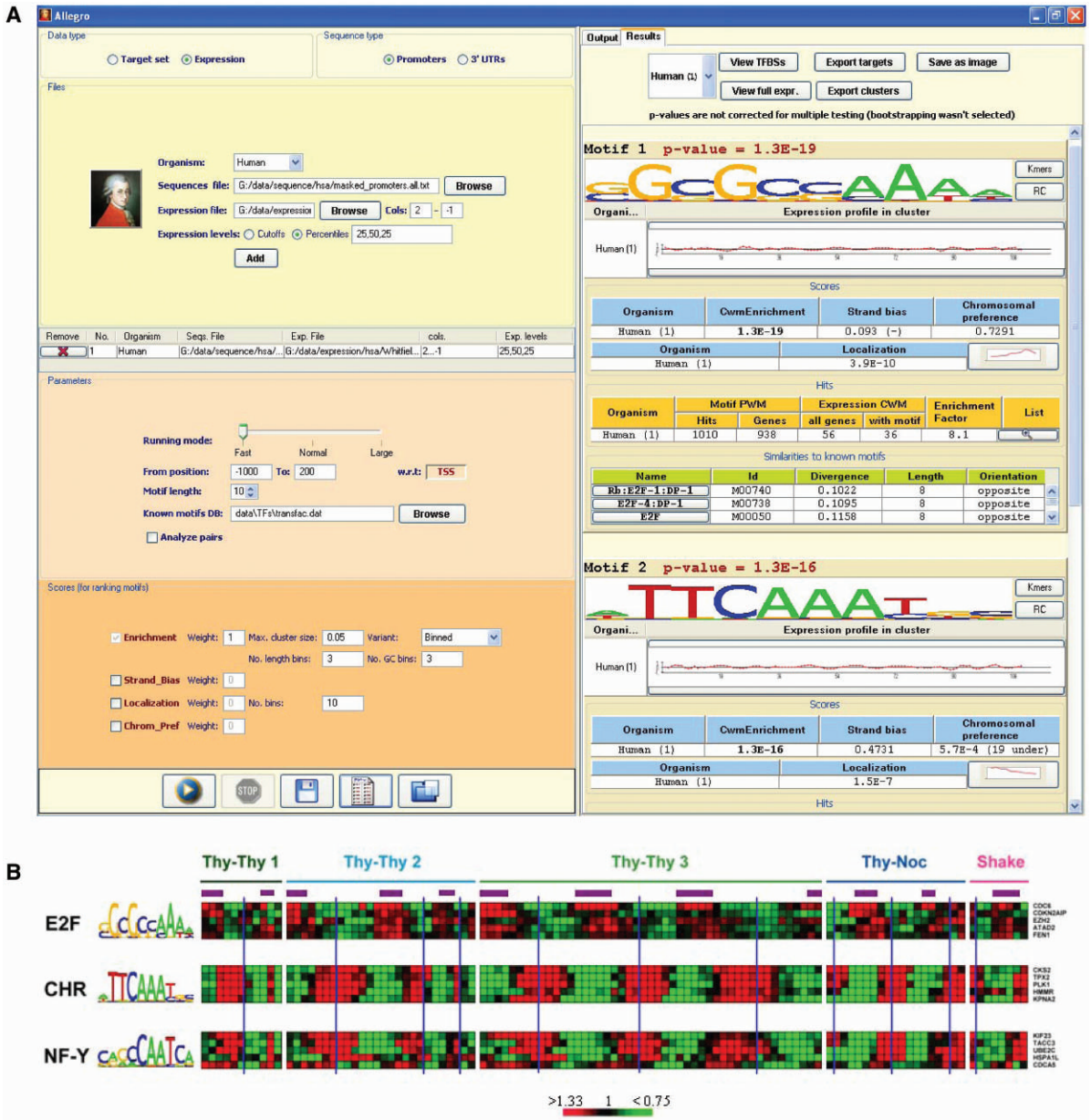

Figure 2. Results of Allegro on the human cell cycle dataset (15). (**A**) Screenshot of Allegro. The left panel presents the input parameters: organism, expression data file, scores, etc. The top-scoring motifs discovered by Allegro are shown in the output panel on the right. Additional information is displayed for each motif, such as the average expression profile of the CWM targets that contain a hit of the motif, statistics on the number of hits and their locations, similar binding patterns from Transfac or miRBase, and more. Here, the three top-scoring motifs reported by Allegro represent the BS patterns of key regulators of the human cell cycle: E2F, CHR (whose binding TF is unknown), and NF-Y (not shown). (**B**) Expression profiles of the five CWM targets with the highest LLR score of the three motifs found by Allegro. High and low expression values w.r.t. time 0 are colored in red and green, respectively. The purple bars represent S phase and the blue vertical lines indicate mitoses, as reported in (15). In agreement with biological knowledge and previous computational analyses (25–28,50), E2F induces genes mainly in the G_1/S phase, whereas CHR and NF-Y are highly specific to the G_2 and G_2/M phases.

four out of the seven motifs Allegro recovered were reported by FIRE. Specifically, the BS motifs of Sko1 and Ste12 were found by Allegro but not by any other method.

We could not compare Allegro to other published methods that infer motifs by simultaneous analysis of sequence and expression data (5–7), either because they are not publicly available or we could not execute them and obtain reasonable results.

Analysis of multiple datasets: tissue-specific regulators

A unique feature of Allegro is simultaneous analysis of multiple datasets from one or more species. Given several expression matrices and corresponding sequence data, Allegro explores the motif search space as described above. For each candidate motif, it computes its enrichment score in each of the datasets separately; i.e. it finds a CWM whose top-scoring genes have a significantly large

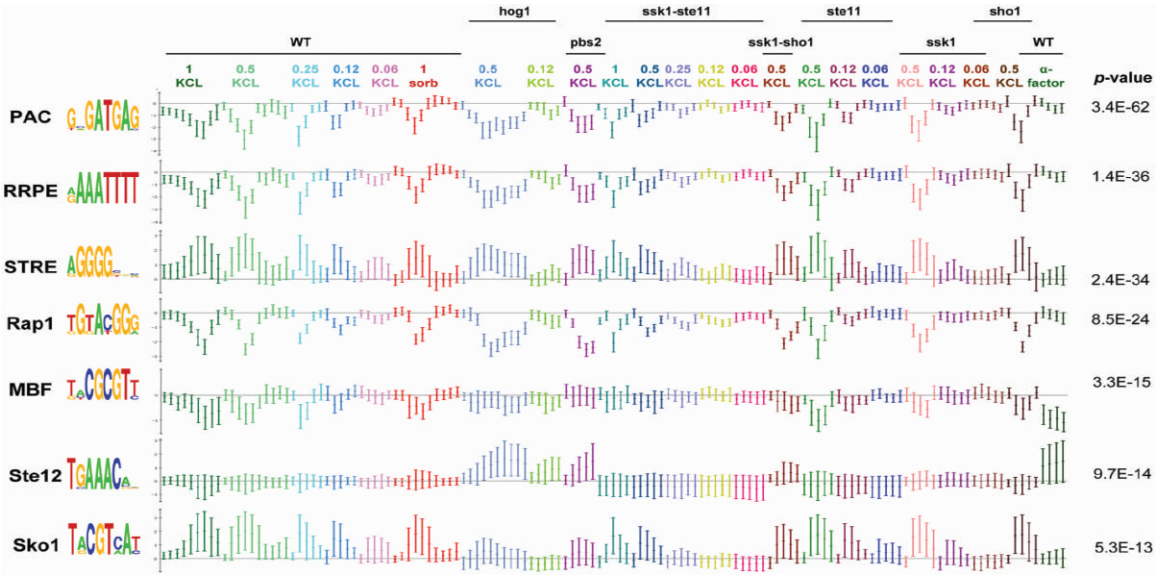

Figure 3. Results of Allegro on the yeast HOG pathway expression dataset (14). Allegro finds the motifs PAC, RRPE, STRE and the binding patterns of Rap1, MBF, Ste12 and Sko1. Each motif is presented together with the average expression profile (±1 SD) of its CWM targets which contain a hit for the motif in their promoter. The titles above the expression series indicate the yeast strain the expression was sampled from: WT, and knockout strains [indicated by the name(s) of the gene(s) that were knocked-out]. The concentrations of KCL and sorbitol are given in molar units.

Table 1. Results of Allegro and existing tools on the yeast HOG MAPK dataset (14)

Biological process	Motif/TF	Reference	K-means/CLICK Amadeus/Weeder	Iclust FIRE	Allegro
General stress response	RRPE	(31,72)	+	+	+
	PAC	(31,72)	+	+	+
	Rap1	(31)	–	+	+
HOG and pheromone response pathways	Sko1	(32,33)	–	–	+
	Ste12	(30,33)	–	–	+
	MBF	(33,73,74)	+	–	+
	Smp1	(32)	–	–	–
	Skn7	(32)	–	–	–
General stress response and HOG pathway	STRE	(31–33)	+	+	+

There are nine TFs and motifs known to be involved in the regulation of genes in the studied conditions. In a single execution, Allegro successfully recovered seven of these binding patterns as the seven top-scoring motifs. In contrast, only four motifs were discovered when the two-step approach was applied using various combinations of existing clustering and motif discovery tools.

overlap with the genes that contain the motif in their *cis*-regulatory sequence. Allegro then combines these scores into a single *p*-value using the *Z*-transform (38).

We tested this feature on the human and mouse gene atlas (16) in search of tissue-specific regulators. Given the expression levels of ~15 000 human and mouse genes across 79 human tissues and 61 mouse tissues, Allegro found known and novel motifs. The main results are summarized in Table 2. The motifs reported by Allegro are non-redundant: for every pair of reported motifs—M_1 and M_2—no more than 5% of their hits overlap, i.e. ≥95% of the occurrences of M_1 do not overlap any occurrence of M_2, and vice versa. Thus, each reported motif is likely to represent a biologically distinct binding pattern.

The top-scoring motif is the binding pattern of CREB/ATF, and its target genes are up-regulated in testis tissues (Supplementary Figure 3). Indeed, CREB is known to activate transcription of genes essential for proper germ cell differentiation (39), and its disruption in mice severely impairs spermatogenesis (40,41). Allegro reported four additional testis-specific motifs: RFX, MYB and two novel motifs (motifs 2–5 in Table 2). Members of the RFX and MYB families are expressed at high levels in the testis (42–46). Interestingly, all three known testis-related TF families—CREB, RFX and MYB—have testis-specific gene products (42,46,47). We performed functional analysis on the sets of CWM targets of the motifs found by Allegro in order to identify GO terms over-represented in these sets (see Methods section). Reassuringly, the CWM targets of all five testis-related motifs in both species are highly enriched for spermatogenesis.

Table 2. Main results of Allegro for the combined analysis of the human and mouse tissue gene atlas datasets (16)

	Logo	TF/motif	*p*-value	Tissues	Gene Ontology (BP; CC)
1		CREB/ATF	10^{-32}	**Testis**: Testis, testis germ cell, testis interstitial, testis Leydig cell, testis seminiferous tubule	Spermato-genesis; Flagellum
2		RFX	10^{-24}		
3		–	10^{-23}		
4		MYB	10^{-21}		
5		–	10^{-15}		
6		MEF2	10^{-29}	**Muscle**: Heart, skeletal muscle, tongue	Muscle contraction; Myofibril
7		ETS/ELF	10^{-27}	**Immune system**: Peripheral blood cells, B/T-cells, lymphnode, BM myeloid, thymus	Immune response; Plasma membrane
8		IRF	10^{-15}		
9		E2F	10^{-23}	**Proliferating cells**: Oocyte, embryo, bone marrow, thymus, lymphoblasts, cancers	Cell cycle, DNA replication; Chromosome
10		NF-Y	10^{-13}		
11		NRF1	10^{-14}		
12		HNF1	10^{-22}	**Digestive tract**: Liver, kidney, pancreas, intestine	Metabolism (carboxylic acid, lipid, amine, ...); Mitochondrion
13		HNF4	10^{-21}		
14		–	10^{-18}	**Keratinocytes**: Epidermis, tongue, digits	Epidermis development, keratinization; Intermediate filament cytoskeleton
15		AP1/FOS	10^{-16}		
16		T-box	10^{-15}		
17		TATA	10^{-14}		
18*		–	10^{-14}	**Oocyte**: Oocyte, fertilized egg	Cell cycle; Nucleus

The table lists all motifs with *p*-value $\leq 10^{-15}$ (combined score for human and mouse datasets), as well as several motifs with high similarity to known binding patterns (TATA, Nrf-1 and NF-Y). Three of the motifs are apparently novel. In addition, a novel motif that obtained a significant *p*-value (10^{-14}) only in the mouse dataset is listed. Similar known binding patterns from the Transfac database are shown in the 'TF/Motif' column. The 'Tissues' column lists the tissues in which the target genes of each motif are up-regulated. Some tissues were sampled in only one of the two organisms. The 'Gene Ontology' column specifies the most enriched biological process (BP) and cellular component (CC) GO terms in the CWM targets of each motif.

**p*-value, tissues and GO terms for motif #18 are based only on the mouse dataset; oocyte and fertilized egg were not sampled in human.

Additional known TF-tissue associations recovered by Allegro include MEF2, whose target genes are induced in heart, skeletal muscle and tongue (48) (Supplementary Figure 4); HNF1 and HNF4, which induce genes in liver, and, to a lesser extent, in kidney, pancreas and intestine (49) (Supplementary Figure 5); and the cell-cycle regulators E2F, NF-Y and NRF1, whose targets are up-regulated in various types of proliferating cells (25,27,50,51). We also found four motifs whose targets are up-regulated in the epidermis and related tissues, such as tongue and digits: the AP1/FOS-binding pattern, T-box, TATA and a novel motif (motifs 14–17 in Table 2). There is evidence of the involvement of FOS and TBP (TATA binding protein) in the regulation of keratinocyte proliferation (52,53).

Allegro discovered a novel motif whose target genes are highly induced in murine oocytes (motif #18 in Table 2, see also Supplementary Figure 6). Oocytes are not among the tested tissues in human, so we do not know whether this enrichment is conserved. A partial list of the putative targets of the motif is given in Supplementary Table III.

To further test the ability of Allegro to simultaneously analyze multiple expression datasets, we applied it on three datasets that recorded expression levels of fly (*Drosophila melanogaster*) genes during various developmental stages (54–56) (see Supplementary Data). Allegro discovered known and novel motifs associated with various developmental profiles. The 20 top-scoring motifs are listed in Supplementary Table IV. Of note, this list includes the top seven core promoter motifs found by Ohler (57), indicating that core promoter *cis*-regulatory elements play an important role in fly development. Another interesting example is the TAGteam motif, which was recently identified and shown experimentally

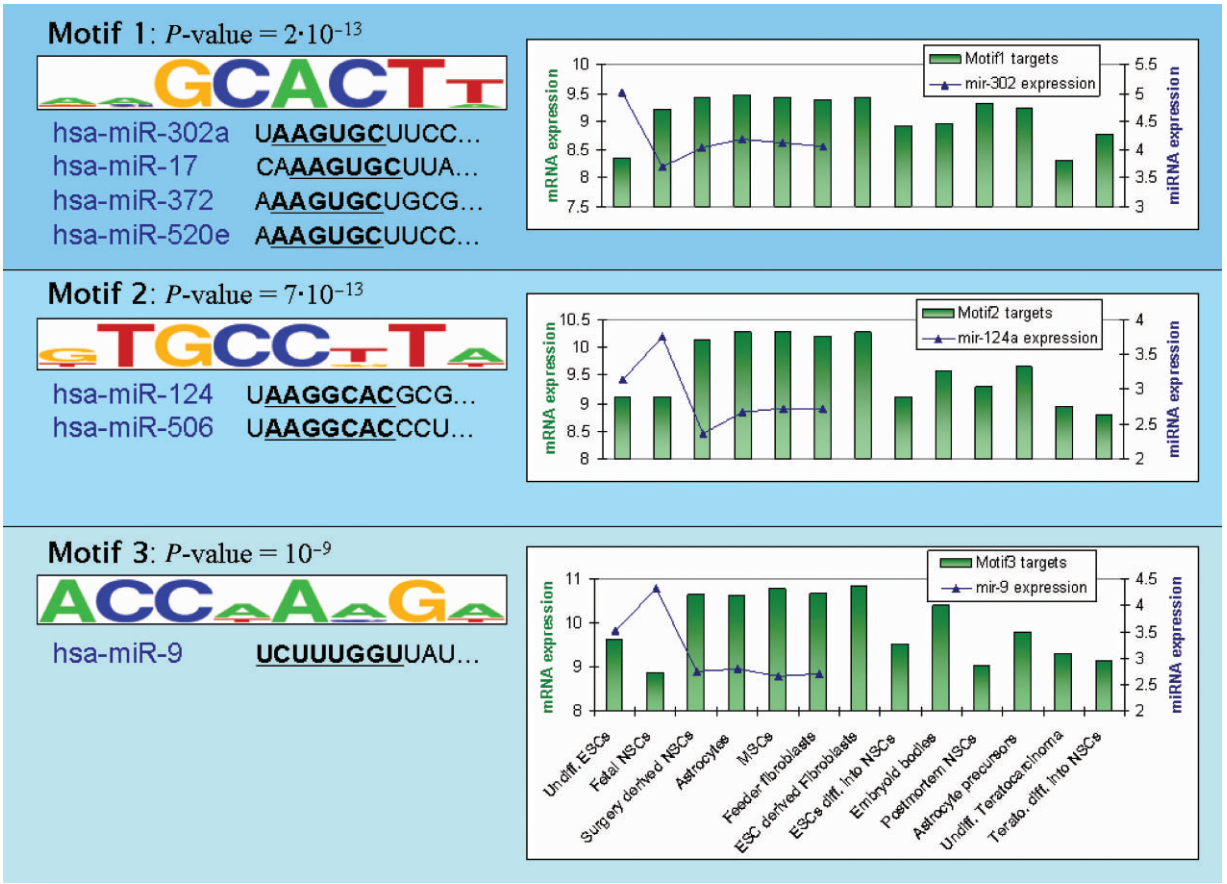

Figure 4. The top three 3′ UTR motifs identified in the stem cells dataset (19). On the left, the motif *p*-value and logo are presented along with the first 11 bases (starting from the 5′ base of the mature microRNA) of miRNAs with a seed that matches the reverse complement of the motif. For the first motif, only one miRNA from each of the four matching miRNA families is presented. For each motif, the graph on the right shows the average expression values (in $\log_2$ scale) of the corresponding CWM targets that contain a hit for the motif. Each bar represents the average expression level in one of the cell types (ESCs/NSCs/MSCs—embryonic/neural/mesenchymal stem cells; 'Undiff.'—Undifferentiated, 'diff.'—differentiated, 'Terato.'—Teratocarcinoma; see also Supplementary Table V; the full expression profile of targets of motif 1 in all 124 samples is shown in Supplementary Figure 9). The graph also shows the expression levels (in $\log_2$ scale) of the matching miRNA(s): mir-302 for motif 1 (average expression over all mir-302 family members), mir-124 for motif 2 and mir-9 for motif 3. miRNA expression levels are presented only for the cell types profiled in (63). Evidently, the expression profiles of the motif targets and those of the matching miRNAs are anti-correlated, increasing our confidence that the discovered motifs represent miRNAs that are active in the relevant cells.

to induce early zygotic genes (58,59). Allegro recovered this motif and the expression profile it induces (Supplementary Figure 7).

3′ UTR analysis: human stem cells

Stem cells, and in particular embryonic stem cells (ESCs), have a unique ability to differentiate into diverse cell types. This multipotency (or pluripotency in case of ESCs) is maintained by a variety of epigenetic mechanisms, including DNA methylation, chromatin modifications and miRNAs (60). Analysis of sequence motifs in 3′ UTRs of genes up- or down-regulated in various types of stem cells carries the promise of identifying key miRNAs maintaining the stem cell differentiation capabilities. Mueller *et al.* (19) profiled gene expression in 124 cell samples, including a variety of stem cells. The analysis of 3′ UTR motifs in this large dataset is hindered by biases in 3′ UTR length and base composition (Supplementary Table V). For example, proliferating cells, such as ESCs, are known to express genes with 3′ UTRs that are much shorter than those of genes expressed in other cell types (61). In contrast, genes specific to the nervous system are known to have particularly long UTRs (62). This leads to an almost 2-fold difference in 3′ UTR length between genes up-regulated in undifferentiated ESCs and genes up-regulated in fetal neural stem cells (NSCs) (Supplementary Figure 8).

We applied Allegro to search for enriched motifs in the 3′ UTRs of the Mueller *et al.* dataset. Due to the biases mentioned above, we used the binned enrichment score to compute the over-representation of each candidate motif in the set of CWM targets fitted to it. The results are presented in Figure 4. The top-scoring motif (GCACTT) is the reverse complement of the hexamer AAGTGC, which appears in the seed sequences of several miRNA families (mir-17, mir-302, mir-290 and mir-515), all of which are among the most highly expressed miRNAs in human and mouse ESCs (63,64). Indeed, genes reported by Allegro as putative targets of these miRNA families are evidently down-regulated in human ESCs compared to

other cell types (Figure 4). Interestingly, as shown in Supplementary Figure 9, these genes are also down-regulated in a subset of NSCs, which were differentiated from ESCs or from teratocarcinoma, indicating that it is possible that the expression of these miRNA families is not down-regulated immediately upon differentiation.

The second most significant motif reported by Allegro is GTGCCTT, which corresponds to the seeds of mir-506 and mir-124a. Inspection of the expression pattern of the CWM targets (Figure 4) shows that genes carrying this motif are generally down-regulated in less differentiated cells (ESCs, NSCs, embryoid bodies and teratocarcinoma) compared to more differentiated ones [mesenchymal stem cells (MSCs), fibroblasts and astrocytes]. Mir-506 did not show any differential expression between ESCs, NSCs and differentiated cells (63), and was not detected in any tissue in a recent comprehensive sequencing effort (62); thus, it is not likely to be the regulator of this gene set. Mir-124a is known to be abundant and functional in the neural cell lineage (65), and is up-regulated in NSCs compared to MSCs and fibroblasts (63). However, it is also up-regulated in NSCs compared to ESCs (63), while the expression levels of the CWM targets do not appear to differ between these two cell types. It is possible, therefore, that the regulation of the CWM targets is carried out by mir-124a alongside other regulatory mechanisms that may or may not involve miRNAs.

The third motif reported by Allegro (ACCAAAG) matches the seed of mir-9. The expression pattern of its targets shows down-regulation in NSCs compared to differentiated cells, with intermediate levels in ESCs and in teratocarcinoma. Mir-9 is expressed specifically in the neural lineage (62,63) and is known to have an active role in neurogenesis (66).

Neither the standard two-step approach (clustering with *k*-means or CLICK, followed by motif finding using Weeder or Amadeus), nor Allegro with the HG enrichment score, recovered the above three motifs. This emphasizes the importance of accounting for sequence biases when conducting *cis*-regulatory motif finding.

DISCUSSION

In this work we present Allegro, a software platform that analyzes genomic sequences and expression datasets to infer transcriptional modules—groups of genes that are co-expressed along all or some of the experimental conditions and share an enriched regulatory motif in their promoters or 3′ UTRs. This single-step methodology, which infers transcriptional modules by simultaneously analyzing the sequence and expression data, utilizes all available information throughout the entire analysis, giving it a clear advantage over the standard two-step approach. Allegro employs a powerful motif enumeration engine and our CWM model to discover sequence motifs and their associated expression profiles without relying on pre-defined types of distribution to model the sequence and expression data. Unlike the vast majority of motif-finding tools, Allegro does not rely on pre-computed *k*-mer counts to construct a sequence model; and, unlike most clustering metrics and existing algorithms for combined sequence-expression analysis, it does not assume a Gaussian distribution of the expression values. Instead, Allegro utilizes the *cis*-regulatory sequences and expression values of all the analyzed genes (typically, the entire genome) as a reference set against which to evaluate the statistical significance of the overlap between each sequence motif and the expression profile fitted to its targets.

Another major contribution of the current study is the CWM, a novel non-parametric model for describing the common expression profile of a group of co-regulated genes. The model gives a likelihood ratio to the group using discrete expression levels. It makes no assumptions about the type of distribution of the expression values, and is robust against extreme values. Unlike similarity metrics, a CWM can describe an expression profile that differs from the background expression levels across a very small number of conditions (even a single condition), and can, in effect, assign a different weight (i.e. contribution) to each condition. Furthermore, a CWM can model more complex transcriptional patterns than existing methods. For example, it can describe the effect of a TF that activates some genes and suppresses others in the same conditions [e.g. Oct4 and Nanog (67)]. As we demonstrated for experimentally derived TF target sets and for functionally related annotated groups of genes, the CWM captures their distinct expression profiles more accurately than commonly used metrics (Supplementary Table II). A detailed discussion on the shortcomings of existing expression similarity measures is given in the Supplementary Data. While in this study we used the CWM in the context of motif finding, it can be applied in other gene expression analysis tasks, such as functional analysis (i.e. identifying GO terms whose genes exhibit a distinct expression profile).

We applied Allegro to several large-scale gene expression datasets in human, mouse, fly and yeast. Our results indicate that in a single run, and without any prior knowledge of known binding patterns or the characteristics of the transcriptional modules (e.g. the number of modules, their size and the overlap between them), Allegro successfully recovers the correct TF/miRNA motifs and reports them as the top-scoring motifs. The transcriptional modules found by Allegro are highly heterogeneous in terms of their expression profiles. For example, the cell-cycle regulators induce very subtle and noisy cyclic patterns in the human cell cycle dataset. The yeast HOG pathway dataset, on the other hand, consists of diverse time-series experiments, and, accordingly, the relevant TFs induce distinct complex expression profiles, some of which differ from the BG distribution in only a small fraction of the conditions.

One of the unique features of Allegro is joint analysis of multiple expression datasets. Unlike some comparative analysis techniques, Allegro does not search for conserved motifs within aligned promoter sequences, since the conservation of TFBSs is, in many cases, very limited across species (68–70). Instead, for each candidate motif it examines, Allegro utilizes the information from all supplied datasets by combining the scores the motif attained on

them into a single *p*-value, thus improving the accuracy of the analysis. We demonstrated this feature on the human and mouse tissues datasets (16), in which Allegro found 18 distinct motifs in various tissue types (Table 2). Notably, some of the tissue-specific motifs obtained borderline *p*-values in one or both species. Many of these motifs were not reported by Allegro when it was applied on each dataset separately (data not shown), underscoring the importance of combined analysis of multiple datasets for increased sensitivity. For example, E2F received a *p*-value of 4×10^{-11} in the human data, which is within the range of random scores given the huge number of candidate motifs considered by the algorithm; the combined human–mouse *p*-value of 10^{-23} is statistically significant. Perhaps this, together with the binned score, is why other methods failed to recover some of the well-known TF-tissue associations. Two cases in point: Elemento *et al*. applied their Iclust and FIRE tools on the human and mouse datasets separately, and did not discover CREB/ATF, RFX, MEF2, IRF, HNF1 and HNF4 (37). When Xie *et al*. searched for conserved promoter elements and tested whether they were tissue-specific (71), they failed to find many of the known TF-tissue associations such as HNF1 and HNF4 in liver, and E2F in proliferating cells. In addition to known TFs, Allegro reported novel motifs that attained statistically significant scores. Experiments are required to verify and study their regulatory roles. Additional novel motifs were discovered by Allegro in fly promoters using three expression datasets of *Drosophila* developmental stages (Supplementary Table IV).

Our analysis of the stem cells dataset demonstrates the ability of Allegro to reverse-engineer transcriptional programs regulated by miRNAs. Using the binned enrichment score, Allegro was able to overcome the two main obstacles in 3′ UTR sequence analysis: length heterogeneity and GC-content bias. The three top-scoring motifs identified by Allegro correspond to three miRNA families, indicating that these families are among the main post-transcriptional regulators in ESCs and NSCs. In particular, the top-scoring motif corresponds to a miRNA seed sequence that was recently shown to be highly dominant in human and mouse ESCs (63,64). The results of Allegro further highlight the importance of the miRNA families carrying this seed sequence in ESC biology. Finally, we show evidence of activity of miRNA carrying this seed sequence in several NSC lines for which miRNA expression profiles are not available. Technologies to accurately measure miRNA expression levels are maturing, but are still inferior in fidelity to mRNA profiling. As we have shown, using sequence analysis and mRNA profiles, we can predict the activity of miRNAs without the direct measurement of miRNA expression.

Due to the flexibility of Allegro's methodology and interface, it is suitable for a broad range of motif discovery tasks. For example, in addition to the HG or binned enrichment score, motifs can be evaluated using other scores we developed previously that measure global features of the distribution of the motif hits: localization along the promoters, strand bias and chromosomal preference (11). Allegro can simultaneously analyze promoter or 3′ UTR sequences and multiple genome-wide expression datasets from several species and combine all available information for optimal results. Running time on a standard PC is between a few minutes and several hours, depending primarily on the size of the expression data. We developed a user-friendly graphical interface, making Allegro accessible to a wide range of users. In order to help the user understand the results of the analysis, Allegro's graphical interface displays additional information and statistics on each reported motif, such as the scores it attained, its putative targets and their expression profile, similar known motifs from Transfac/miRBase, and more. The Allegro software (a standalone Java application) is available at http://acgt.cs.tau.ac.il/allegro.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported in part by the Israel Science Foundation (grant 802/08 and Converging Technologies Program grant 1767.07). Igor Ulitsky was supported in part by a fellowship from the Edmond J. Safra Bioinformatics program at Tel Aviv University. Funding for open access charge: ISF 802/08 and 1767.07.

Conflict of interest statement. None declared.

REFERENCES

1. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
2. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet*., **22**, 281–285.
3. Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev*., **12**, 130–136.
4. Jiang,D., Tang,C. and Zhang,A. (2004) Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng*., **16**, 1370–1386.
5. Holmes,I. and Bruno,W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 202–210.
6. Segal,E., Yelensky,R. and Koller,D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19(Suppl. 1)**, i273–i282.
7. Reiss,D.J., Baliga,N.S. and Bonneau,R. (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **7**, 280.
8. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet*., **27**, 167–171.
9. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
10. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol*., **3**, 39.
11. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Res*., **18**, 1180–1189.

12. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
13. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
14. O'Rourke,S.M. and Herskowitz,I. (2004) Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol. Biol. Cell.*, **15**, 532–542.
15. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
16. 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. USA*, **101**, 6062–6067.
17. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
18. Shamir,R., Maron-Katz,A., Tanay,A., Linhart,C., Steinfeld,I., Sharan,R., Shiloh,Y. and Elkon,R. (2005) EXPANDER–an integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**, 232.
19. Muller,F.J., Laurent,L.C., Kostka,D., Ulitsky,I., Williams,R., Lu,C., Park,I.H., Rao,M.S., Shamir,R., Schwartz,P.H. *et al.* (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**, 401–405.
20. MacQueen,J. (1965) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, CA: University of California Press, pp. 281–297.
21. Sharan,R. and Shamir,R. (2000) CLICK: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 307–316.
22. Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17(Suppl. 1)**, S207–S214.
23. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
24. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
25. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.
26. Tabach,Y., Milyavsky,M., Shats,I., Brosh,R., Zuk,O., Yitzhaky,A., Mantovani,R., Domany,E., Rotter,V. and Pilpel,Y. (2005) The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation. *Mol. Syst. Biol.*, **1**, 0022.
27. Linhart,C., Elkon,R., Shiloh,Y. and Shamir,R. (2005) Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis. *Cell Cycle*, **4**, 1788–1797.
28. Zhu,Z., Shendure,J. and Church,G.M. (2005) Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.
29. Zhu,W., Giangrande,P.H. and Nevins,J.R. (2004) E2Fs link the control of G1/S and G2/M transcription. *Embo J.*, **23**, 4615–4626.
30. Bardwell,L. (2004) A walk-through of the yeast mating pheromone response pathway. *Peptides*, **25**, 1465–1476.
31. 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.
32. Hohmann,S. (2002) Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol. Mol. Biol. Rev.*, **66**, 300–372.
33. O'Rourke,S.M., Herskowitz,I. and O'Shea,E.K. (2002) Yeast go the whole HOG for the hyperosmotic response. *Trends Genet.*, **18**, 405–412.
34. Martinez-Pastor,M.T., Marchler,G., Schuller,C., Marchler-Bauer,A., Ruis,H. and Estruch,F. (1996) The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J.*, **15**, 2227–2235.
35. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
36. Slonim,N., Atwal,G.S., Tkacik,G. and Bialek,W. (2005) Information-based clustering. *Proc. Natl Acad. Sci. USA*, **102**, 18297–18302.
37. Elemento,O., Slonim,N. and Tavazoie,S. (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.
38. Whitlock,M.C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.
39. Don,J. and Stelzer,G. (2002) The expanding family of CREB/CREM transcription factors that are involved with spermatogenesis. *Mol. Cell Endocrinol.*, **187**, 115–124.
40. Hummler,E., Cole,T.J., Blendy,J.A., Ganss,R., Aguzzi,A., Schmid,W., Beermann,F. and Schutz,G. (1994) Targeted mutation of the CREB gene: compensation within the CREB/ATF family of transcription factors. *Proc. Natl Acad. Sci. USA*, **91**, 5647–5651.
41. Blendy,J.A., Kaestner,K.H., Schmid,W., Gass,P. and Schutz,G. (1996) Targeting of the CREB gene leads to up-regulation of a novel CREB mRNA isoform. *EMBO J.*, **15**, 1098–1106.
42. Morotomi-Yano,K., Yano,K., Saito,H., Sun,Z., Iwama,A. and Miki,Y. (2002) Human regulatory factor X 4 (RFX4) is a testis-specific dimeric DNA-binding protein that cooperates with other human RFX members. *J. Biol. Chem.*, **277**, 836–842.
43. Grimes,S.R. (2004) Testis-specific transcriptional control. *Gene*, **343**, 11–22.
44. Mettus,R.V., Litvin,J., Wali,A., Toscani,A., Latham,K., Hatton,K. and Reddy,E.P. (1994) Murine A-myb: evidence for differential splicing and tissue-specific expression. *Oncogene*, **9**, 3077–3086.
45. Oh,I.H. and Reddy,E.P. (1999) The myb gene family in cell growth, differentiation and apoptosis. *Oncogene*, **18**, 3017–3033.
46. Sitzmann,J., Noben-Trauth,K., Kamano,H. and Klempnauer,K.H. (1996) Expression of B-Myb during mouse embryogenesis. *Oncogene*, **12**, 1889–1894.
47. Huang,X., Zhang,J., Lu,L., Yin,L., Xu,M., Wang,Y., Zhou,Z. and Sha,J. (2004) Cloning and expression of a novel CREB mRNA splice variant in human testis. *Reproduction*, **128**, 775–782.
48. Black,B.L. and Olson,E.N. (1998) Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.*, **14**, 167–196.
49. Kuo,C.J., Conley,P.B., Chen,L., Sladek,F.M., Darnell,J.E., , and Crabtree,G.R. (1992) A transcriptional hierarchy involved in mammalian cell-type specification. *Nature*, **355**, 457–461.
50. Dimova,D.K. and Dyson,N.J. (2005) The E2F transcriptional network: old acquaintances with new faces. *Oncogene*, **24**, 2810–2826.
51. Cam,H., Balciunaite,E., Blais,A., Spektor,A., Scarpulla,R.C., Young,R., Kluger,Y. and Dynlacht,B.D. (2004) A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell*, **16**, 399–411.
52. Mehic,D., Bakiri,L., Ghannadan,M., Wagner,E.F. and Tschachler,E. (2005) Fos and jun proteins are specifically expressed during differentiation of human keratinocytes. *J. Invest. Dermatol.*, **124**, 212–220.
53. Fadloun,A., Kobi,D., Pointud,J.C., Indra,A.K., Teletin,M., Bole-Feysot,C., Testoni,B., Mantovani,R., Metzger,D., Mengus,G. *et al.* (2007) The TFIID subunit TAF4 regulates keratinocyte proliferation and has cell-autonomous and non-cell-autonomous tumour suppressor activity in mouse epidermis. *Development*, **134**, 2947–2958.
54. Hooper,S.D., Boue,S., Krause,R., Jensen,L.J., Mason,C.E., Ghanim,M., White,K.P., Furlong,E.E. and Bork,P. (2007) Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. *Mol. Syst. Biol.*, **3**, 72.
55. Arbeitman,M.N., Furlong,E.E., Imam,F., Johnson,E., Null,B.H., Baker,B.S., Krasnow,M.A., Scott,M.P., Davis,R.W. and White,K.P.

(2002) Gene expression during the life cycle of Drosophila melanogaster. *Science*, **297**, 2270–2275.
56. Spellman,P.T. and Rubin,G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.*, **1**, 5.
57. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, 0087.
58. Ten Bosch,J.R., Benavides,J.A. and Cline,T.W. (2006) The TAGt eam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development*, **133**, 1967–1977.
59. De Renzis,S., Elemento,O., Tavazoie,S. and Wieschaus,E.F. (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biol.*, **5**, 117.
60. Bibikova,M., Laurent,L.C., Ren,B., Loring,J.F. and Fan,J.B. (2008) Unraveling epigenetic regulation in embryonic stem cells. *Cell Stem Cell*, **2**, 123–134.
61. Sandberg,R., Neilson,J.R., Sarma,A., Sharp,P.A. and Burge,C.B. (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science*, **320**, 1643–1647.
62. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
63. Laurent,L.C., Chen,J., Ulitsky,I., Mueller,F.J., Lu,C., Shamir,R., Fan,J.B. and Loring,J.F. (2008) Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence. *Stem Cells*, **26**, 1506–1516.
64. Calabrese,J.M., Seila,A.C., Yeo,G.W. and Sharp,P.A. (2007) RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **104**, 18097–18102.
65. Cao,X., Pfaff,S.L. and Gage,F.H. (2007) A functional study of miR-124 in the developing neural tube. *Genes Dev.*, **21**, 531–536.
66. Krichevsky,A.M., Sonntag,K.C., Isacson,O. and Kosik,K.S. (2006) Specific microRNAs modulate embryonic stem cell-derived neurogenesis. *Stem Cells*, **24**, 857–864.
67. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
68. Borneman,A.R., Gianoulis,T.A., Zhang,Z.D., Yu,H., Rozowsky,J., Seringhaus,M.R., Wang,L.Y., Gerstein,M. and Snyder,M. (2007) Divergence of transcription factor binding sites across related yeast species. *Science*, **317**, 815–819.
69. Lin,C.Y., Vega,V.B., Thomsen,J.S., Zhang,T., Kong,S.L., Xie,M., Chiu,K.P., Lipovich,L., Barnett,D.H., Stossi,F. *et al.* (2007) Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet.*, **3**, e87.
70. Odom,D.T., Dowell,R.D., Jacobsen,E.S., Gordon,W., Danford,T.W., MacIsaac,K.D., Rolfe,P.A., Conboy,C.M., Gifford,D.K. and Fraenkel,E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.*, **39**, 730–732.
71. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
72. 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.
73. Escote,X., Zapater,M., Clotet,J. and Posas,F. (2004) Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nat. Cell Biol.*, **6**, 997–1002.
74. Gartner,A., Jovanovic,A., Jeoung,D.I., Bourlat,S., Cross,F.R. and Ammerer,G. (1998) Pheromone-dependent G1 cell cycle arrest requires Far1 phosphorylation, but may not involve inhibition of Cdc28-Cln2 kinase, *in vivo*. *Mol. Cell. Biol.*, **18**, 3681–3691.