Molecular Systems Biology 5; Article number 268; doi:10.1038/msb.2009.24
Citation: ***Molecular Systems Biology*** **5:268**

www.molecularsystemsbiology.com

molecular systems biology

Discovering structural *cis*-regulatory elements by modeling the behaviors of mRNAs

Barrett C Foat and Gary D Stormo*

Department of Genetics, Center for Genome Sciences, Washington University School of Medicine, St Louis, MO, USA
* Corresponding author. Department of Genetics, Washington University School of Medicine, 4444 Forest Park Ave., Campus Box 8510, St Louis, MO 63108, USA. Tel.: +314 747 5534; Fax: +314 362 2156; E-mail: stormo@genetics.wustl.edu

Received 27.10.08; accepted 17.3.09

Gene expression is regulated at each step from chromatin remodeling through translation and degradation. Several known RNA-binding regulatory proteins interact with specific RNA secondary structures in addition to specific nucleotides. To provide a more comprehensive understanding of the regulation of gene expression, we developed an integrative computational approach that leverages functional genomics data and nucleotide sequences to discover RNA secondary structure-defined *cis*-regulatory elements (SCREs). We applied our structural *cis*-regulatory element detector (StructRED) to microarray and mRNA sequence data from *Saccharomyces cerevisiae, Drosophila melanogaster*, and *Homo sapiens*. We recovered the known specificities of Vts1p in yeast and Smaug in flies. In addition, we discovered six putative SCREs in flies and three in humans. We characterized the SCREs based on their condition-specific regulatory influences, the annotation of the transcripts that contain them, and their locations within transcripts. Overall, we show that modeling functional genomics data in terms of combined RNA structure and sequence motifs is an effective method for discovering the specificities and regulatory roles of RNA-binding proteins.
Molecular Systems Biology 28 April 2009; doi:10.1038/msb.2009.24
Subject Categories: bioinformatics; RNA
Keywords: modeling; mRNA stability; polysome association; post-transcriptional regulation; secondary structure

Introduction

Gene expression is regulated at each step from chromatin remodeling through translation and degradation. Yet, most efforts to understand the regulation of gene expression have been focused on transcription and DNA-binding regulatory proteins. Although regulatory RNAs have received appreciable attention (Bushati and Cohen, 2007; Coppins *et al*, 2007), regulatory elements within mRNAs that are recognized by nucleic acid-binding proteins have been largely ignored until recently (Keene, 2007). This state exists despite observations that suggest changes in mRNA stability may account for half of the changes in mRNA expression in some cells and conditions (Fan *et al*, 2002; Cheadle *et al*, 2005). Moreover, it is a mathematical certainty that mRNAs of average stability can only be rapidly downregulated by altering the mRNA decay rate (see Pérez-Ortín *et al* (2007) for derivation). Thus, one way to execute rapid, large-scale gene expression responses to unpredictable environmental stimuli is through decay-regulating RNA-binding proteins (RBPs), whose activity can be rapidly modulated post-transcriptionally. Early metazoan embryogenesis also requires mRNA stability and translation regulation to orchestrate the activities of maternally deposited transcripts (for review see Vardy and Orr-Weaver, 2007).

Despite the potential importance of RNA secondary structures as binding sites for regulatory RBPs, computational methods for their discovery have failed to keep pace with current functional genomics technology (e.g. microarrays). Over 20 years ago, Sankoff (1985) described an algorithm for the simultaneous alignment and folding of RNA sequences. The Sankoff algorithm was computationally intractable for even modestly sized datasets. Thus, nearly all subsequent computational approaches to finding RNA structural motifs sought shortcuts to making the RNA structural alignment problem computationally feasible. Now, well into the era of functional genomics, RNA structure finding algorithms are still sequence-only methods, having so far failed to use the data-integrative approaches that are becoming increasingly common for the discovery of DNA-binding protein specificities

(Bussemaker *et al*, 2001, 2007; Foat *et al*, 2005, 2006). Small regulatory RNA structures likely play a significant role in post-transcriptional regulation of gene expression. However, due to the lack of computational methods that fully leverage the current wealth of genomics data, we are nearly blind to their existence.

In this work, we present a novel, alignment-free method that discovers secondary structure-defined *cis*-regulatory elements (SCREs) in mRNAs by modeling the effects that their occurrences exert on quantitative measurements of mRNA behavior in the form of microarray data. This process is embodied in a regression-based algorithm called structural *cis*-regulatory element detector (StructRED). We accurately recover the known stem–loop binding specificities of the *Drosophila* RBP Smaug and Vts1p in *Saccharomyces cerevisiae* from mRNA sequences and microarray data. We also provide evidence of the post-translational regulatory activity of Smaug that is consistent with earlier genetic and biochemical characterization. We report other putative structure-sequence specificities that likely play diverse roles in *Drosophila* and humans. Finally, we find that SCREs exist in coding sequences as often as in untranslated regions (UTRs), which presents a caution against the common practice of restricting searches for regulatory elements to non-coding sequences. Overall, we show that structurally defined *cis*-regulatory elements can be discovered through integrative modeling of functional genomics and mRNA sequence data.

Results and discussion

Discovering RNA structural *cis*-regulatory elements by observing their effects on mRNA behavior using StructRED

To understand the later biological inference, we must first provide an overview of our StructRED algorithm (Figure 1) and supporting methods. In this work, we develop a way to infer the activities of RBPs and the small RNA structural *cis*-regulatory elements to which they bind by modeling how the occurrences of SCREs could give rise to the observed genomic mRNA measurements. We start with the simplest model and pose that the observed microarray value for a

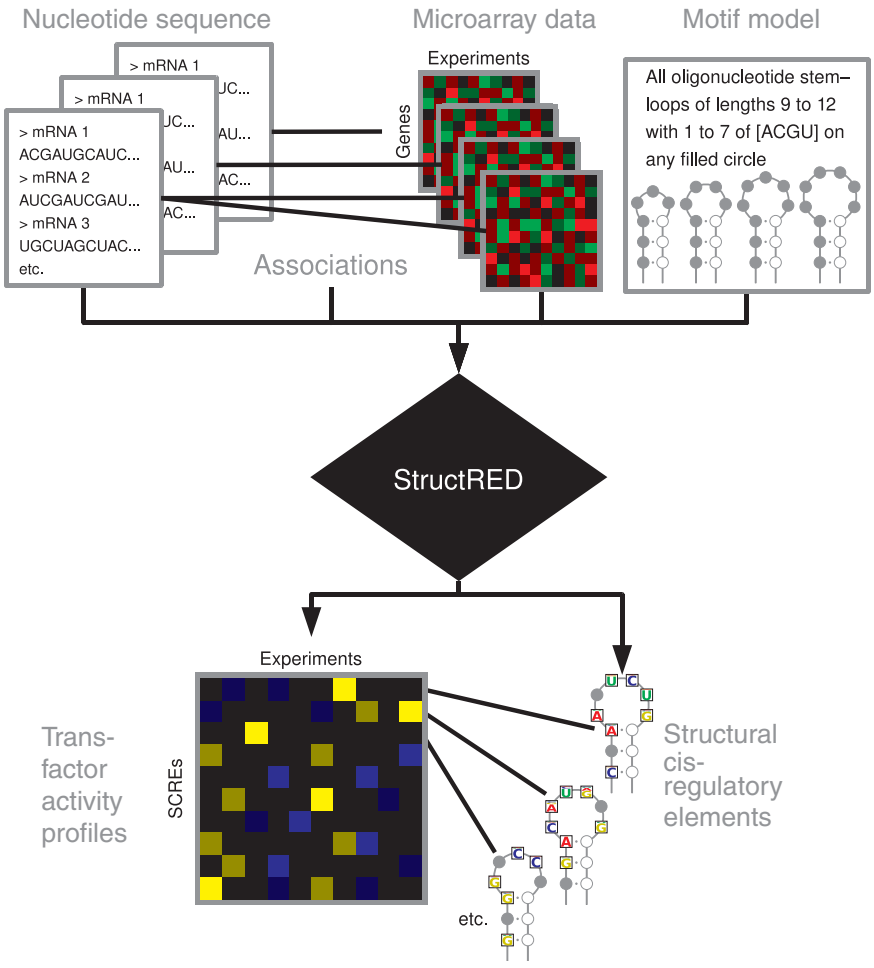

Figure 1 The flow of data for StructRED. As input, StructRED takes one or more multicolumn tables of microarray data and one or more associated FASTA sequence files containing mRNA sequences for the spots on the arrays. The other major input to StructRED is a motif model that defines the search space in which explanatory SCREs can exist. StructRED fits a simple physical model using each motif in the search space to identify the SCREs that best explain the observed microarray measurements. The outputs of StructRED are SCRE matrices and their correlations with each experiment in the input set, called *trans*-factor activity profiles (TFAPs), as they reflect changes in the activities of the regulators that bind to the SCREs.

particular mRNA A_g is the result of a combination of additive effects of the occurrences M_{ng} of multiple SCREs n. We also allow a base level of signal C and some unexplained signal ε_g. Each SCRE has a corresponding effect F_n on the measurement that is related to the regulatory activity of a corresponding *trans*-factor under the measured condition.

$$A_g = \sum_n F_n M_{ng} + C + \varepsilon_g \quad (1)$$

If we had definitions of each of the SCREs n, we could count them in all mRNA sequences and use multiple linear regression to define the global values of all F_n and C. However, as we do not know the SCREs in advance, we define a space of possibilities and test each one to see if its occurrences explain why some mRNAs have different measurements than others. We can then collect each SCRE that is significantly explanatory, in turn, to build an additive model for the measured microarray value of each mRNA. One output of this search process is a list of candidate SCREs n. The sequence information in a discovered SCRE is in the form of a position-specific affinity matrix (PSAM; Foat *et al*, 2005), a kind of weight matrix. This means that 'occurrences' of an SCRE have values that range between zero and one, depending on the weights of each nucleotide at each position. The other output is the inferred activities F_n of the corresponding *trans*-factor in each dataset that we fit with the model. We call the collection of F_n's (divided by their standard errors) for a particular SCRE across multiple microarray experiments the *trans*-factor activity profile (TFAP) for that SCRE, as it reflects changes in activity of the corresponding regulator.

For all analysis done for this study, the SCRE search space included all small stem–loops of 9 to 12 nucleotides in length, closed by three hybridizing nucleotides (allowing G-U pairs), and containing up to seven positions with sequence information. Stem–loops are attractive because, while being true structures, the single-stranded loop allows for additional specific protein–nucleotide interactions with the unpaired Watson-Crick edges of the bases. Despite the simplicity of this stem–loop structural constraint, it provides enough additional specificity information for the discovered SCREs to display significant correlations with one or more microarray datasets. To confirm the utility of the structural constraint using the *Drosophila* SCREs discussed below, we tested only the sequence specificity of each SCRE for its ability to explain the best-correlating dataset for the structurally constrained SCRE. In all cases, the stem–loop-constrained SCRE better explained the data than the sequence-only specificity (Supplementary Table 1).

Modeling gene expression regulation: hold the thresholds

Our equation for the predicted occupancy of an mRNA by a *trans*-factor (see Materials and methods) sums over all possible binding site positions in the mRNA. Why is this more rational than other expressions of sequence scores (e.g. maximum score)? As we illustrate using a recently characterized concrete example of mRNA stability regulation, a sequence score that sums over all possible binding sites is a way of representing the *in vivo* total average occupancy of an mRNA by *trans*-factors, and under reasonable assumptions of regulatory mechanisms, increased occupancy could result in increased regulatory strength. When bound to an mRNA, the Smaug RBP recruits the CCR4–NOT complex, promoting deadenylation and decay of the transcript (Semotok *et al*, 2005). Semotok *et al* (2008) predicted that the Hsp83 transcript contained eight high-affinity binding sites for the Smaug RBP. The authors tested a variety of fragments of the Hsp83 transcript for their sufficiency in conferring instability to the mRNA. They found that a fragment containing six predicted Smaug-binding sites caused the greatest instability, followed by fragments containing three, two, one or no predicted binding sites. Thus, Semotok *et al* (2008) observed a rough correlation between Smaug binding opportunities on the test transcripts and actual destabilization. One can mechanistically motivate this observation in at least two ways: First, if Smaug is at a sufficiently low concentration that none of the sites are saturated, an mRNA with six binding sites will have, on average, a Smaug protein bound to it six times more often than an mRNA with one binding site, resulting in six times more opportunities for deadenylation. Another possibility is that interactions between Smaug and the CCR4–NOT complex are rare (low CCR4–NOT concentration). Thus, even if Smaug-binding sites are always saturated, a transcript with six binding sites, again, has six times more opportunities to bind a deadenylase than a transcript with one site and a single bound Smaug protein.

The above discussion leads us to address a common shortfall in regulatory sequence analysis, which is the application of thresholds, either to the sequence score or to functional genomics data. First, why not treat regulation as an on/off, binary variable? If one were to assume that regulation is an on/off event and that an mRNA requires a threshold number of high-affinity binding sites to be regulated, the appropriate statistical test would be a *t*-test between the sample of mRNAs that were above the threshold for predicted binding versus the sample of mRNAs below the threshold. However, it can be shown that the Pearson correlation, when applied to categorical data, reduces to a *t*-test (Lev, 1949; Tate, 1954). Thus, using standard regression statistics can detect binary or gradual relationships between the predicted occupancy and microarray data equally well. Also if a threshold was imposed on the sequence score when the relationship was indeed gradual, there would be a cost in sensitivity, as informative covariance was discarded. Supplementary Figure 1 shows scatter plots and regression lines for the correlation between occurrences of our *Drosophila* SCREs with mRNA expression data. Moreover, plotted are curves that follow the *t*-value that would be calculated if each predicted occupancy was used as a threshold to perform a *t*-test. The *t*-value calculated by the two-sample *t*-test never becomes much more significant than the *t*-value of the regression (only slightly higher for one SCRE). Moreover, although a reasonable threshold could be chosen in many cases, the multiple hypothesis correction required to choose a good threshold would greatly reduce the statistical power of the analysis. Upon inspection (Supplementary Figure 1), some SCREs (e.g. Smg-4, Smg-5) seem to have a gradual relationship between predicted binding and observed regulation, but others are less clear. Thus, whether regulation of mRNA stability and

translation is executed in a binary or gradual manner remains an open question for future investigation. Nevertheless, our regression-based computational approach based on gradual regulation assumptions is an efficient and powerful method to discover regulatory specificities and activities, which could then serve as starting points for detailed mechanistic studies.

Thresholds on functional genomics data may also discard informative covariance between sequence features and observed mRNA levels. Although the study was performed on transcription factors and not RBPs, Tanay (2006) provided evidence suggesting that transcription factor binding is analog rather than digital, and this variable binding is detectable in corresponding microarray measurements. Such observations combined with our discussion in the last paragraph lead us to question the ultimate sensitivity of common approaches to regulatory sequence analysis where a threshold is placed on the microarray data and then sequence features enriched in the above-threshold set are analyzed. Rabani *et al* (2008) recently performed such an analysis to discover RNA structural elements within mRNAs related to the processes measured in a variety of microarray datasets. Their method discovers a candidate structure that can be seen in as many of the input sequences as possible using a stochastic context-free grammar-based method. Although their approach is powerful and valid, it is essentially placing thresholds on both the sequence score and the microarray values, possibly discarding usable signal from both. Our method does not discover structures as complex as those found by Rabani *et al* (2008), but our biophysically motivated regression approach represents a radical conceptual departure from the sequence-only methods RNA structure-discovery algorithms that have dominated for 25 years (Sankoff, 1985). Moreover, our method is particularly well suited to discovering small structures (<20 nt), which may be too small to provide sufficient signal to sequence-only methods.

Annotating SCREs by scoring the responding mRNAs

Given our model, every sequence window in an mRNA contributes to the predicted occupancy of a transcript by the SCRE-binding regulator. mRNAs that actually contain functional, high-affinity instances of a SCRE, should have both a high occurrence score based on the mRNA sequence and a high predicted occurrence score based on its microarray measurement. Rather than assume an arbitrary sequence score threshold, a ranking based on a combination of these scores should give a sense for how likely it is that a particular mRNA responds to the SCRE sites that it contains (see Materials and methods). Functional gene annotations can be scored for uneven distribution toward highly ranked genes by using the Mann–Whitney–Wilcoxon test. We scored the Gene Ontology annotations (Ashburner *et al*, 2000) in this way for *S. cerevisiae* and *Drosophila*. In addition, *Drosophila* genes have been annotated by allele phenotype (Wilson *et al*, 2008) and *in situ* expression (Tomancak *et al*, 2002) using controlled, hierarchical vocabularies similar to Gene Ontology. Thus, these two annotation systems were similarly scored. The various annotations can give valuable insights into the biological roles of the newly discovered SCREs.

Testing StructRED in *S. cerevisiae*—the specificity of Vts1p

We first turned our attention to analyzing data from the yeast *S. cerevisiae*. One of the few RBPs that is known to recognize a specific SCRE is the yeast protein Vts1p (Aviv *et al*, 2003, 2006a, b; Edwards *et al*, 2006; Johnson and Donaldson, 2006; Oberstrass *et al*, 2006). Vts1p has a RNA-binding specificity that is similar to another, better characterized sterile alpha motif (SAM)-containing protein, Smaug in *Drosophila* (Aviv *et al*, 2003). Vts1p recognizes a small stem–loop structure with a four or five nucleotide loop with a 'G' at the third position of the loop and complementary bases at positions one and four of the loop (Aviv *et al*, 2006a, b; Edwards *et al*, 2006; Johnson and Donaldson, 2006; Oberstrass *et al*, 2006). The biological role of Vts1p is unknown, although there is some evidence that suggests that Vts1p regulates the poly(A) tail length of mRNAs through interaction with the Ccr4–Pop2–Not complex and thus affects mRNA stability (Aviv *et al*, 2003; Oberstrass *et al*, 2006; Rendl *et al*, 2008). Two groups have performed functional genomics experiments to identify possible Vts1p target mRNAs. Motivated by the observation that a reporter transcript containing three Smaug-binding sites was less stable in wild-type yeast than it was in a Δ*vts1* strain, Oberstrass *et al* (2006) used microarrays to look for mRNAs that were differentially expressed in a wild-type versus a Δ*vts1* strain (Gene Expression Omnibus (GEO) accession GSE3859). They confirmed with Northern blots that a few transcripts were present at different levels between the two strains and contained predicted Vts1p-binding sites. Aviv *et al* (2006b) used a pull-down/microarray approach to identify those mRNAs that were most often associated with Vts1p (GEO accession GSE3741). They also confirmed that some of the mRNAs had lower steady state levels in a wild-type strain than in a Δ*vts1* strain.

We applied the StructRED algorithm to search for any stem–loop SCREs in the wild-type versus Δ*vts1* (Oberstrass *et al*, 2006) and the Vts1p pull-down (Aviv *et al*, 2006b) microarray data in addition to approximately 6500 other microarray experiments retrieved from the NCBI GEO (Barrett *et al*, 2007). We confirmed the specificity of Vts1p (Figure 2) using the pull-down microarray data (Aviv *et al*, 2006b; Figure 3A). This Vts1p specificity is in good agreement with the Vts1p specificity shown in earlier work (Aviv *et al*, 2006a, b; Edwards *et al*, 2006; Johnson and Donaldson, 2006; Oberstrass *et al*, 2006). Thus, StructRED successfully performs the task for which it was designed—to detect SCREs based on genome-wide measurements of the effects that their occurrences exert on mRNAs. Those mRNAs that we predict are most likely to contain Vts1p SCREs are enriched for functional categories involving carbohydrate metabolism and transmembrane transport (Supplementary Table 2). However, too little is known about the biological role of Vts1p to draw conclusions from these observations.

Vts1p-binding site occurrences did not significantly explain genome-wide mRNA expression in the wild-type versus Δ*vts1* microarray dataset (Oberstrass *et al*, 2006) or any of the other approximately 6500 other microarray experiments that we analyzed (Supplementary information). This suggests at least three possibilities: first, we discovered the Vts1p SCRE using

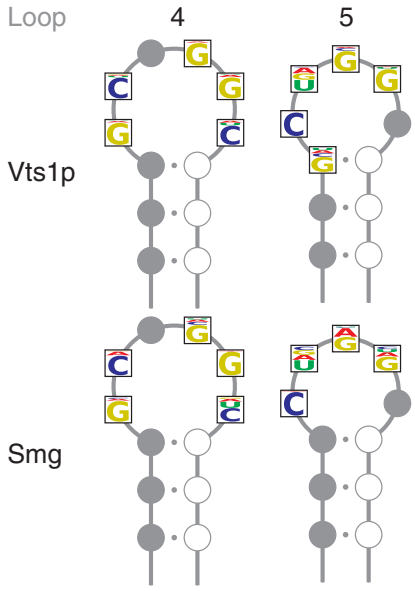

Figure 2 Vts1p and Smaug specificities. Vts1p and Smaug (Smg) are sterile alpha motif (SAM)-containing RNA-binding proteins in *Saccharomyces cerevisiae* and *Drosophila melanogaster*, respectively. These proteins are known to bind stem–loop RNA motifs with loops of four or five nucleotides. The structural logos shown here were discovered using StructRED on relevant microarray data and mRNA sequences without any prior information. The specificities of the SAM-containing proteins are in good agreement with each other and with the known specificities of these proteins. Here, the length four loops are represented by length six loops with a bias toward G-C on the first and last nucleotides of the loop, perhaps indicating a need for a strong G-C pair to close the shorter loop.

pull-down microarray data in which both regulated and unregulated Vts1p-bound mRNAs get pulled down and contribute specificity information. Vts1p may actually regulate only a few genes, and therefore, its specificity may not explain a significant amount of the genome-wide variation in mRNA expression. The second possible explanation is that none of the microarray experiments tested a physiological condition in which mRNA stability regulation by Vts1p causes a significant global expression difference between the measured samples. The final possibility is that the primary role of Vts1p is to regulate translation and that none of the approximately 6500 microarray experiments contained information relevant to this role. The SAM-containing protein Smaug has both mRNA stability (Semotok *et al*, 2005) and translation-regulating effects (Dahanukar *et al*, 1999; Smibert *et al*, 1999), which may suggest that a role in regulating translation is possible for Vts1p.

On to *Drosophila melanogaster*—the specificity and function of Smaug

Smaug is a multifunctional RBP that regulates the stability of at least one transcript (Semotok *et al*, 2005) and regulates the translation of at least one other transcript (Dahanukar *et al*, 1999; Smibert *et al*, 1999) in *Drosophila*. Although the experimentally confirmed target mRNAs of Smaug are few, its regulatory activities have been intensely studied due to its role in translationally repressing maternally deposited *nanos* mRNA everywhere but in the posterior of the early embryo, thus helping to establish the anterior–posterior axis during

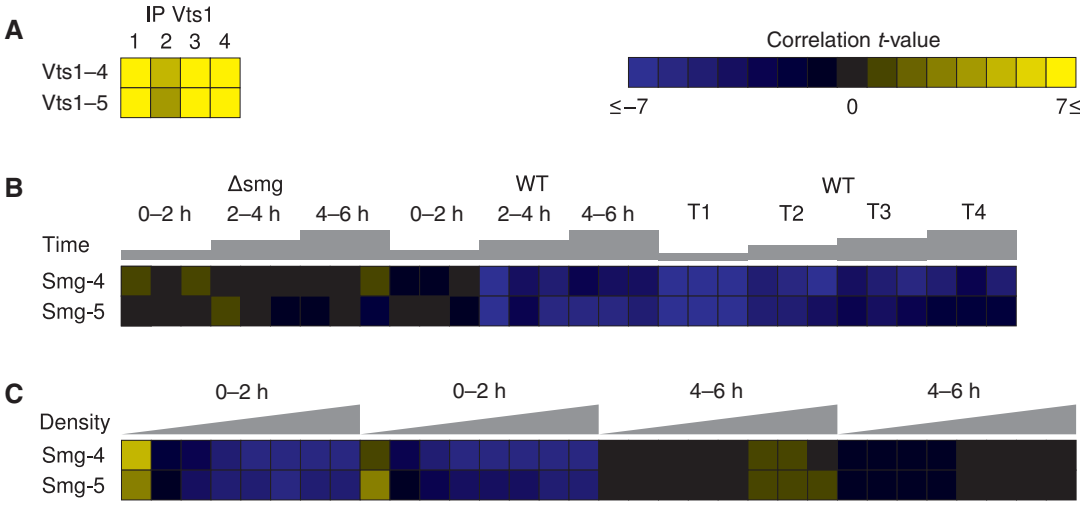

Figure 3 Vts1p and Smaug activities. Each square represents the strength of the correlation between genome-wide occurrences of a SCRE and genome-wide mRNA measurements for a particular microarray experiment. Yellow represents a positive correlation and blue represents a negative correlation. An absolute *t*-value of about 6.7 corresponds to a *P*-value of 0.01, when strictly correcting for the number of motifs tested. (**A**) The Vts1p specificities for the length four loop (Vts1–4) and length five loop (Vts1–5) were discovered using microarray data measured mRNA association with Vts1p in a pull-down experiment in four trials (Aviv *et al*, 2006b). (**B**) The Smaug specificities for the length four (Smg-4) and length five (Smg-5) loops were discovered using mRNA expression microarray data performed over *Drosophila melanogaster* embryonic development. The first two time courses measured the first 6 h of development in Δ*smg* and wild-type (WT) activated eggs (Tadros *et al*, 2007). The third time course (Pilot *et al*, 2006) compares the slow phase (T1), fast phase (T2), cellularization and beginning gastrulation (T3), and end of gastrulation (T4) to embryos before zygotic transcription begins in wild-type (WT) embryos. (**C**) Occurrences of the Smg-4 and Smg-5 specificities also had strong negative correlations (corrected *P*-value <0.001) with ribosome association in the first 2 h of development (Qin *et al*, 2007). Triangles represent increasing density of sucrose gradient fractions, corresponding to increasing numbers of ribosomes.

development (Dahanukar *et al*, 1999; Smibert *et al*, 1999; Semotok *et al*, 2005).

Knowing the importance of post-transcriptional regulation in early embryogenesis (see Vardy and Orr-Weaver, 2007 for review), we applied StructRED to microarray time courses that measured mRNA expression and ribosome association in early *Drosophila* development. We identified the specificity of Smaug (Figure 2), which is strikingly similar to that of Vts1p, considering that Vts1p has little overall homology to Smaug except for 40% sequence identity to the 61 amino-acid SAM domain. The TFAPs for the Smaug SCREs support the established roles of Smaug as both a translation and mRNA stability regulator (Figure 3B and C). The three microarray time courses in Figure 3B examined gene expression in activated *Drosophila* eggs or embryos during early embryogenesis (Pilot *et al*, 2006; Tadros *et al*, 2007; GEO accessions GSE8910, GSE3955). In both wild-type time courses, having high-affinity Smaug-binding sites correlates with reduced mRNA concentration starting at the 2–4 h time point and T1 (slow phase). This observation is consistent with the timing observed earlier for Smaug destabilizing *Hsp83* mRNA (Semotok *et al*, 2005). As this large-scale mRNA destabilization is not observed in the time course data for the Δ*smg* embryos (Figure 3B, left), it is likely that we identified the specificity and activity of Smaug. Tadros *et al* (2007) likewise noted an enrichment in Smaug-binding sites in unstable maternal mRNAs.

We also observe the translation-repressing effects of Smaug. An increasingly common microarray method for gaining insight into translation regulation are polysome association microarrays. In this method, cell lysate is run through a sucrose gradient column. mRNAs associated with different numbers of ribosomes separate into different density fractions in the column, with mRNAs associated with multiple ribosomes, presumably caught while being actively translated, moving into the heavier fractions. Qin *et al* (2007) performed such microarray profiling of mRNA–ribosome association during a time course of the first 10 h of *Drosophila* development (GEO accession GSE5430). By examining the TFAPs of the Smaug specificities over two replicates each of a 0–2 h sample and a 4–6 h sample, we see that mRNAs that are bound by Smaug are being specifically excluded from ribosome association during the 0–2 h time point, as indicated by a strong negative correlation between the Smaug specificity and enrichment in the denser gradient fractions (Figure 3C). This presumed translational repression by Smaug is relieved by the 4–6 h time point. The timing of Smaug-effected translational repression that we observed genome-wide mirrors the characterized ability of Smaug to translationally repress *nanos* mRNAs (Dahanukar *et al*, 1999; Smibert *et al*, 1999). This observation suggests that the temporal regulation of all embryonic mRNA targets by Smaug is similar to that of *nanos* transcripts. Our observed change in Smaug's global regulatory activity is likely due to the earlier observed reduction in Smaug expression after the first 3 h of embryogenesis (Dahanukar *et al*, 1999; Smibert *et al*, 1999).

We scored the ranking of Smaug target genes (see Materials and methods) against Gene Ontology categories, phenotypes, and *in situ* expression annotations. The high-scoring *in situ* categories 'Stages 1–3 maternally deposited' and 'Stages 4–6 rapidly degraded,' reflected the degradation function of Smaug. Other significantly enriched categories (*P*-value $\leqslant 0.001$) for *in situ* expression included categories involving the early stage development of the female reproductive system and categories relating to later development of the nervous system. Significant Gene Ontology categories and allele phenotype annotations similarly support the role of Smaug in the development of the reproductive and nervous systems (Supplementary Table 2).

Tadros *et al* (2007) provided data that suggests that Smaug may regulate the stability of about 712 transcripts. In order for StructRED to detect the specificity and regulatory activity of a *trans*-factor, it must regulate at least tens or hundreds of transcripts to make the correlation statistically significant. Thus, merely detecting the Smaug specificity and activities from genome-wide polysome association data suggests that Smaug has a similarly large role in regulating translation during early embryogenesis. As we have microarray data reflecting both the mRNA stability and translational repression activities of Smaug, we have the opportunity to evaluate whether the mRNA targets that are destabilized by Smaug are the same as those that are translationally repressed. First, we took the Δ*smg* and wild-type time course data from Tadros *et al* (2007) and transformed it so the data compared the wild-type versus Δ*smg* mRNA levels, thus primarily displaying the Smaug-dependent changes in mRNA levels. We also took the 0–2 h time point data (polysome fractions 8–12) from the mRNA–ribosome association time course of Qin *et al* (2007) and subtracted the contribution of all SCREs except those belonging to Smaug, leaving the data depleted for any unrelated translation regulation. Finally, we tested how well having Smaug-dependent mRNA stability regulation was predicted by Smaug-enriched translation repression in early development. We saw a significant correlation between Smaug-dependent destabilization and putatively Smaug-related translation repression (*P*-value $< 10^{-11}$), suggesting that there are at least some transcripts that have both their translation and stability regulated by Smaug.

The predicted specificities and roles of other SCREs in *Drosophila* development

When we applied StructRED to the *Drosophila* development time courses and other microarray data, we did not specifically try to find the Smaug specificity. In fact, in addition to the Smaug SCREs, we discovered six other putative SCREs, which we have labeled Dm1 through Dm6, that have coherent supporting TFAPs and annotation (Figure 4). First, Dm1 and Dm2 were discovered from the same mRNA expression microarray time course for *Drosophila* embryogenesis that we discussed for the Smaug SCREs (Tadros *et al*, 2007). Those transcripts that contain high-affinity instances of Dm1 and Dm2 are expressed at decreasing levels as development proceeds, suggesting that they are involved in destabilizing these transcripts at specific developmental stages. Dm1- and Dm2-containing transcripts have weak enrichments for Gene Ontology categories related to development and protein transport (Supplementary Table 2). Transcripts that contain Dm1 or Dm2 display a bias for expression in the

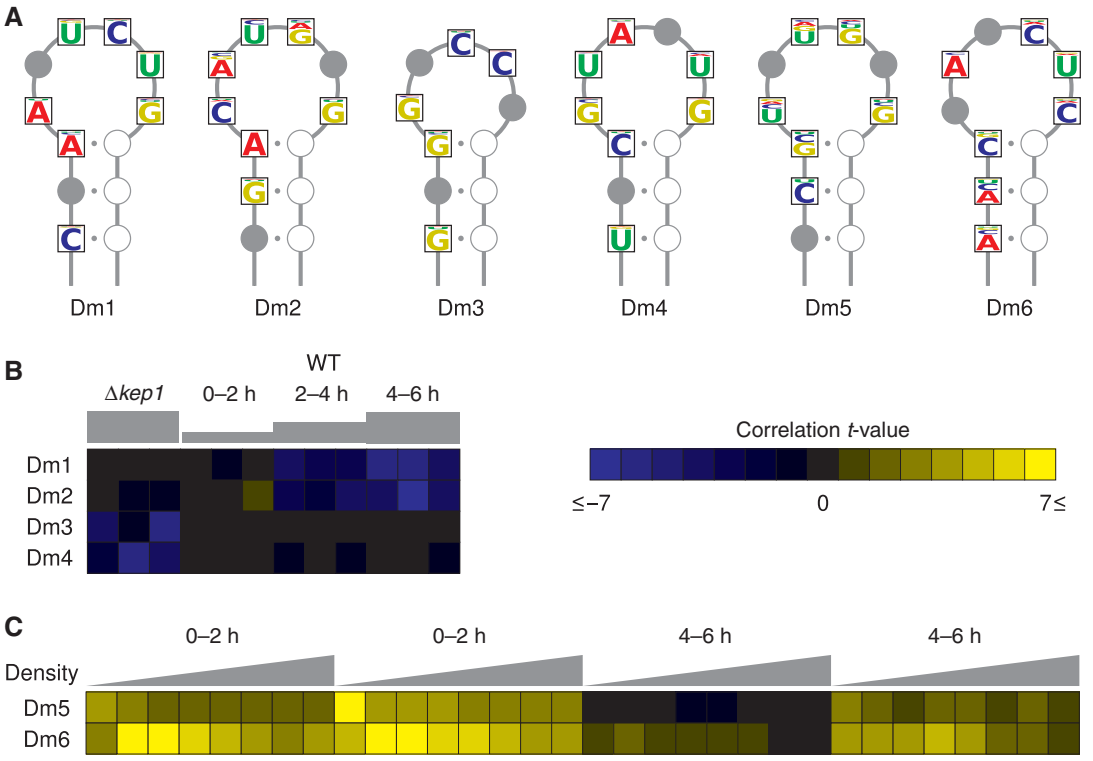

Figure 4 Putative *Drosophila* structural *cis*-regulatory elements. (**A**) The structural logos of the six putative *Drosophila* SCREs. (**B**) Dm1, Dm2, Dm3, and Dm4 were detected using mRNA expression microarray data. Dm1 and Dm2 had strong negative correlations with mRNA levels over early *Drosophila* development. Dm1 and Dm2 did not correlate with mRNA levels in similarly treated Δ*smg* eggs (not shown). Dm3 and Dm4 correlated with mRNA levels changing between wild-type and Δ*kep1* flies (GEO accession GSE6086), suggesting that Dm3 and Dm4 may reflect the specificity of Kep1, an RNA-binding protein. (**C**) Dm5 and Dm6 were detected from microarray data measuring mRNA association with ribosomes in early drosophila development (Qin *et al*, 2007). Triangles represent increasing density of sucrose gradient fractions, corresponding to increasing numbers of ribosomes.

developing female reproductive system. Notably, Dm1 and Dm2 occurrences do not correlate with decreasing expression in Δ*smg* embryos (data not shown), suggesting that the putative destabilizing effect of Dm1 and Dm2 depends on Smaug. However, Dm1 and Dm2 do not show the translational repression activity that we see with the Smaug SCREs.

The Dm3 and Dm4 specificities were detected using microarray data that compared expression in wild-type flies and flies lacking the Kep1 RBP (GEO accession GSE6086), suggesting that they may represent the specificity of Kep1. There were no significant themes among Gene Ontology, phenotype, or *in situ* annotations for Dm3 and Dm4, besides that their targets seem to be membrane associated (Supplementary Table 2). Earlier genetic and biochemical characterization of Kep1 suggests that it may be involved in regulating splicing (Fruscio *et al*, 2003; Robard *et al*, 2006).

Finally, Dm5 and Dm6 both were detected using polysome association data from the early *Drosophila* embryo (Qin *et al*, 2007; GEO accession GSE5430). Both have strong correlations during the 0–2 h time point and have almost no effect by 4–6 h. Dm5 has a strong positive correlation with the lightest fraction, suggesting that transcripts that contain instances of Dm5 are more often without ribosomes or associated with only a partially assembled ribosome, perhaps indicating ribosome pausing in the 5′ UTR. Dm6 increases the likelihood that its transcripts are associated with a moderate number of ribosomes during the 0–2 h time point, suggesting active translation. Annotations for the genes most likely to contain high-affinity Dm5 or Dm6 sites are similar. Gene Ontology annotations for Dm5 and Dm6 center around oogenesis, nervous system development, and appendage development. Concordantly, phenotype annotations for Dm5 and Dm6 responders involve the nervous system, germ band, wing, and leg. *In situ* expression for Dm5 and Dm6 responding genes show them in the ectoderm and neurogenic regions in later embryonic stages (Supplementary Table 2).

SCREs are in coding sequences and UTRs

Cis-regulatory elements in mRNAs are commonly assumed to reside in the 3′ UTR or occasionally the 5′ UTR. Although there are undoubtedly more examples of characterized regulatory elements in UTRs than in coding sequences, this may simply reflect reporting bias. Coding sequences are often assumed only to contain coding information, and as they are under selective pressure to maintain their coding properties, it seems less likely that *cis*-regulatory elements would reside there. However, although actively translating ribosomes may reduce the binding of *trans*-factors, nothing excludes the possibility of *cis*-regulatory elements existing in coding sequences.

Contributing to this problem is the success of cross species comparisons for the identification of putative *cis*-regulatory elements. These methods rely on mutations occurring over evolutionary time to lay bare those parts of the genome that are

under strong selective pressures. The simplest interpretation for conserved sequences in non-coding regions is that they serve regulatory functions for adjacent genes. However, if regulatory elements exist in coding sequences, conservation-based methods need to be more sophisticated to detect them, as they must differentiate between conservation of codons in a coding sequence and conservation of regulatory nucleotide sequence. When looking for *cis*-regulatory elements in coding sequences, regression-based methods like StructRED have a major advantage. As StructRED does not use sequence conservation, all sequences, coding and non-coding, are treated equally. StructRED detects structural *cis*-regulatory elements based on the effect that they have on the mRNAs that contain them, as observed in genome-wide measurements. The functional signal will be equally strong from a SCRE in a coding sequence as in a UTR.

The SCRE discovery in this work was always performed on approximated full-length mRNAs. However, to answer the question of where the discovered SCREs commonly occur in the mRNAs, we scored the occurrences of each SCRE in the 5′ UTRs, 3′ UTRs, and coding sequences separately and then checked which of these mRNA subsequences performed best at explaining the microarray data. If most of the functional SCREs are in the 3′ UTRs, as is commonly assumed, then the TFAPs for the 3′ UTRs alone should be strongly significant and appear similar to the TFAPs when the full-length mRNA sequences are used. For most of the *Drosophila* SCREs, especially the Smaug SCREs, the occurrences that appear in the coding sequences perform best at explaining the microarray measurements of gene expression and polysome association (Figure 5). Thus, most of the functional sites for Dm2, Dm3, Dm4, Dm6, and the Smaug SCREs reside in coding sequences. Recent characterization of Smaug stability regulation of the *Hsp83* transcript showed that all eight predicted binding sites for Smaug do indeed reside in the coding sequence (Semotok *et al*, 2008). Dm1, Dm5, and Dm6 still have appreciable signal in the 3′ UTRs, and Dm5 has signal in the 5′ UTRs. We also calculated the length-normalized scores for the UTRs and the coding sequences for each SCRE. Dm3, Dm4, Dm5, Dm6, and the Smaug SCREs had the highest concentration of binding sites in the same regions that strongly predicted expression (Supplementary Figure 4). Only Dm1 and Dm2 were inconsistent, with Dm1 having a higher density of sites in coding sequences, while the scores in the 3′ UTRs were more predictive, and Dm2 having a higher density of sites in the 3′ UTRs, while the scores in the coding sequences were more predictive. SCREs frequently appearing in coding sequences provides a strong argument for including whole transcripts when searching for *cis*-regulatory elements.

Explaining published observations

Tadros *et al* (2007) investigated the role of Smaug in early *Drosophila* development by measuring genome-wide mRNA expression in both wild-type and Δ*smg* activated eggs. Thus, they identified maternally deposited transcripts that were degraded by both Smaug-dependent and independent mechanisms. The authors do not provide candidate regulators that affect the degradation of the Smaug-independent mRNAs. With the hopes of identifying the specificity of such a regulator, we looked for any of our SCREs that negatively correlated with mRNA levels in both the wild-type and Δ*smg* activated eggs. We found none. However, as part of the preprocessing steps before the SCRE search, we discovered all of the significantly explanatory single-stranded mRNA motifs. The intention of this step was to increase the likelihood that any discovered SCREs reflected real structural elements and not a structure-like reflection of an otherwise single-stranded regulatory element. Although we do not discuss most of those results here, inspecting them for any motifs that correlated with mRNA degradation in the Δ*smg* eggs revealed two such motifs (Figure 6). The Dm7 motif and activity profile may belong to a yet uncharacterized RBP in *Drosophila*. Dm7 has the correct TFAP for a Smaug-independent regulator of mRNA stability during early embryogenesis, having strong negative correlations with mRNA levels as time elapses to 4–6 h in both wild-type and Δ*smg* activated eggs. Dm7 is similar to the UUGUU motifs identified in maternal unstable transcripts by De Renzis *et al* (2007). Another motif that we label 'Pum' is almost certainly the specificity of Pumilio, the founding member of the Puf family of RBPs (Zhang *et al*, 1997). We confirmed that the Pum motif is the Pumilio specificity by testing how well its sequence scores predicted genome-wide association of mRNAs with Pumilio, using the microarray data of (Gerber *et al*, 2006; *P*-value $<10^{-25}$). Strangely, the Pum motif only correlates with the degradation of mRNAs in the Δ*smg* activated eggs. This observation suggests that there is a more complex relationship between the regulatory activities of Pumilio and Smaug that could benefit from future experimental characterization.

A peek into human mRNAs

With hopes of making similar inferences about regulatory RBPs in humans, we applied StructRED to human microarray data that measured genome-wide RBP binding or polysome association. However, we had a few handicaps in performing the analysis. First, rather than well-annotated mRNA sequences as are available for *Drosophila*, we needed to retrieve EST sequences from NCBI Entrez Nucleotide that were relevant to each microarray platform that we used. Moreover, systematic functional annotation of these sequences do not exist. Thus, we could only discover SCREs and follow their regulatory effects through the analyzed microarray conditions.

We discovered three SCREs with functionally coherent TFAPs (Figure 7A and B). Occurrences of the first SCRE, Hs1, correlated with decreased translation in the metastatic colorectal cancer cell line, SW620, versus a non-metastatic cell line from the same patient, SW480, as measured in a polysome association microarray study (Provenzani *et al*, 2006; GEO accession GSE2509). Transcripts containing Hs2 SCREs are expressed at a lower level in U937 cells that have been exposed to 12-myristate 13-acetate (PMA) and caused to differentiate into a macrophage-like state (Kitamura *et al*, 2004; GEO accession GSE1783). Finally, occurrences of Hs3 in mRNAs correlate with increased association with ribosomes in human mammary epithelial cells, regardless of whether translation initiation factor 4F is overexpressed (Larsson *et al*, 2007; GEO accession GSE6043).

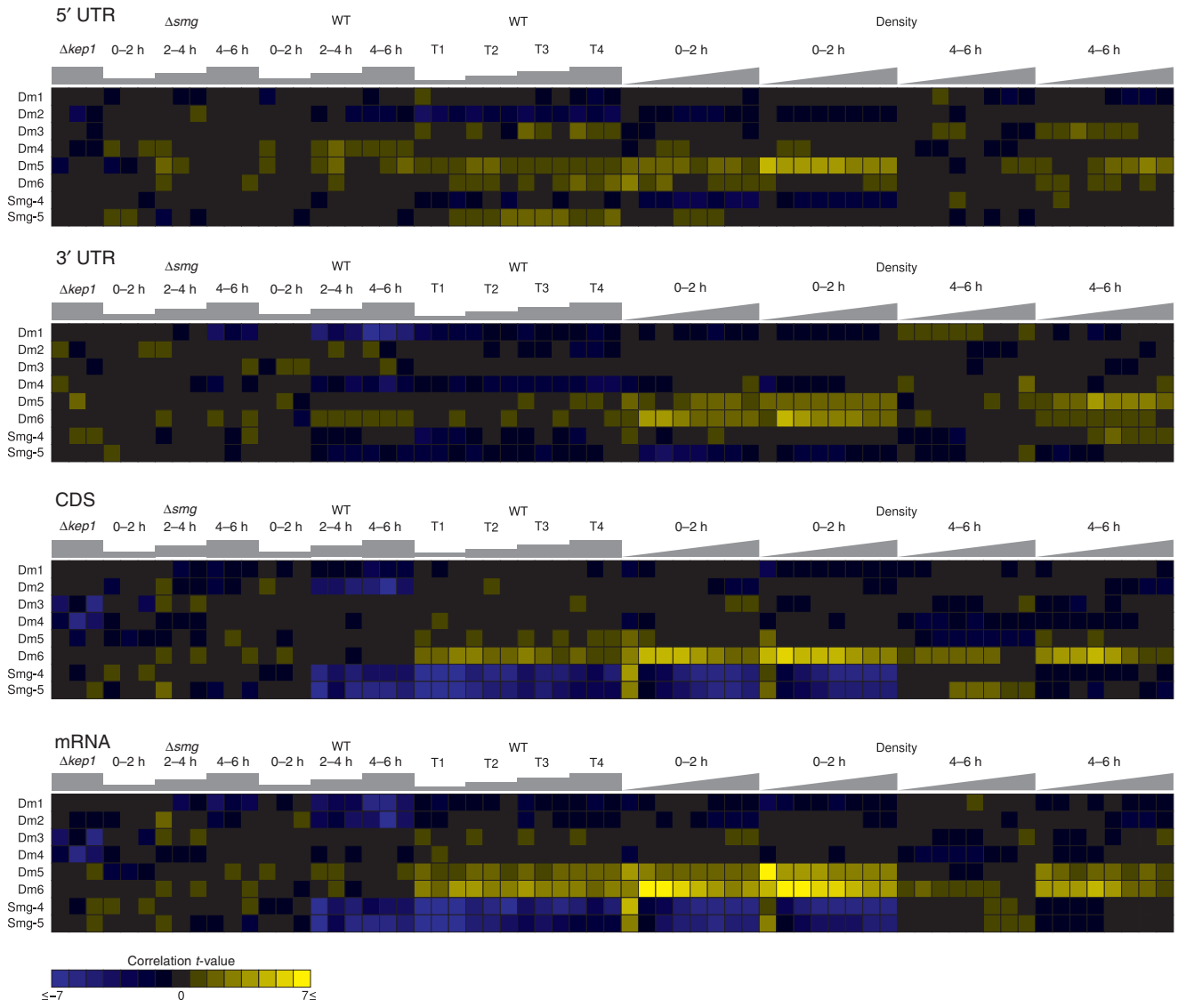

Figure 5 Explanatory structural *cis*-regulatory element content of mRNA regions. These *trans*-factor activity profiles (TFAPs) are for all of the *Drosophila* SCREs over all of the same conditions shown in Figures 3 and 4. However, these TFAPs display how well each SCRE explained the measured RNA levels when occurrences of the SCREs are only scored in the 5′ untranslated regions (UTRs), 3′ UTRs, coding sequences (CDS), or full-length mRNAs. Thus, by comparing each subsequence TFAP to the full-length mRNA TFAP, one can see in which region of mRNAs functional instances of the SCRE tend to exist. Most of the SCREs have their strongest signal in the CDSs, followed by the 3′ UTRs.

Although a Smaug homolog exists in the human genome, we did not detect a Smaug/Vts1p-like specificity in the data that we analyzed. Nevertheless, we could calculate TFAPs for the *Drosophila* Smaug specificities across the human data to look for Smaug activities that were too weak to detect in the original search. Indeed, there were two RBP pull-down microarray studies, one for poly-pyrimidine tract binding protein (PTB; GEO accession GSE6021; Gama-Carvalho *et al*, 2006) and one for Staufen1 and Staufen2 (Furic *et al*, 2008; GEO accessions GSE8437, GSE8438), where pulled-down mRNAs were enriched for Smaug-binding sites (Figure 7C). This suggests that Smaug binds to many of the same targets as PTB, Staufen1, and Staufen2. Our observed correlation between the Smaug specificity and Staufen1 binding is consistent with earlier observations that hSmaug is found in cytoplasmic granules with Staufen1 (Baez and Boccaccio, 2005).

Final thoughts and looking ahead

This work has two major themes: first, we showed that functional genomics data such as microarrays can be used to find *cis*-regulatory elements in mRNAs defined by both sequence and structure by posing a model for their existence and then using the data to identify which modeled SCREs are supported by the data. We implemented this approach in our algorithm StructRED. We successfully applied StructRED to discover known and putative stem–loop *cis*-regulatory elements using only full-length mRNA sequence data and

functional genomics data, without needing to perform any kind of sequence alignment. Second, we applied our StructRED algorithm to data from yeast, flies, and humans both to validate the approach and to contribute significant knowledge to the field of inquiry into post-transcriptional regulation of gene expression. We accurately recover the specificities of the stem–loop RBPs Vts1p from yeast and Smaug from flies. We also discover and computationally characterize putative SCREs in flies and humans involved in diverse roles. Finally, we show that many of our discovered SCREs in flies are found in coding regions as much or more than in UTRs and they are conserved in multiple species of *Drosophila*.

The StructRED algorithm represents a novel method for determining *cis*-regulatory RNA structures. Although the current implementation is limited to finding short stem–loop motifs, it may be extensible to other small structures (dsRNA, internal bulges) and perhaps more complex structures. Given its strengths, we expect that StructRED may become the basis of a class of RNA regulatory element search tools that will expand computational and experimental inquiries into post-transcriptional gene regulation.

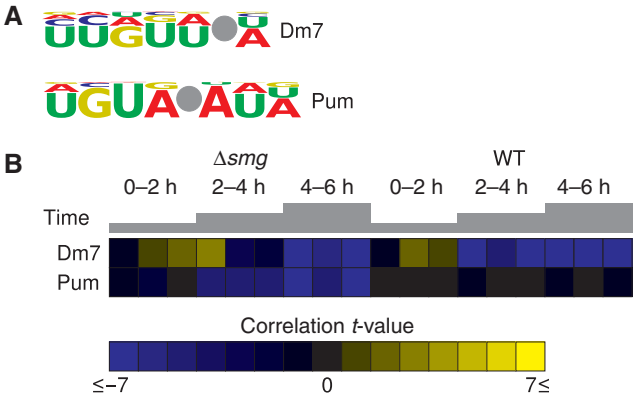

Figure 6 Explaining Smaug-independent mRNA degradation. Not all maternally deposited mRNAs degrade in a Smaug-dependent manner. (**A**) These are the logos of two single-stranded RNA motifs that had strong correlations with decreasing mRNA levels in early embryogenesis in *Drosophila*. The Dm7 motif likely represents the specificity of an unknown *trans*-factor. The Pum motif is likely the specificity of Pumilio. (**B**) Both of these motifs correlated with decreasing mRNA levels in Δ*smg* activated eggs.

Materials and methods

Nucleotide sequence

We performed the search for SCREs on best estimates of full-length mRNAs. Unfortunately, the yeast transcriptome had not been completely sequenced at the time of this analysis, so we did not know

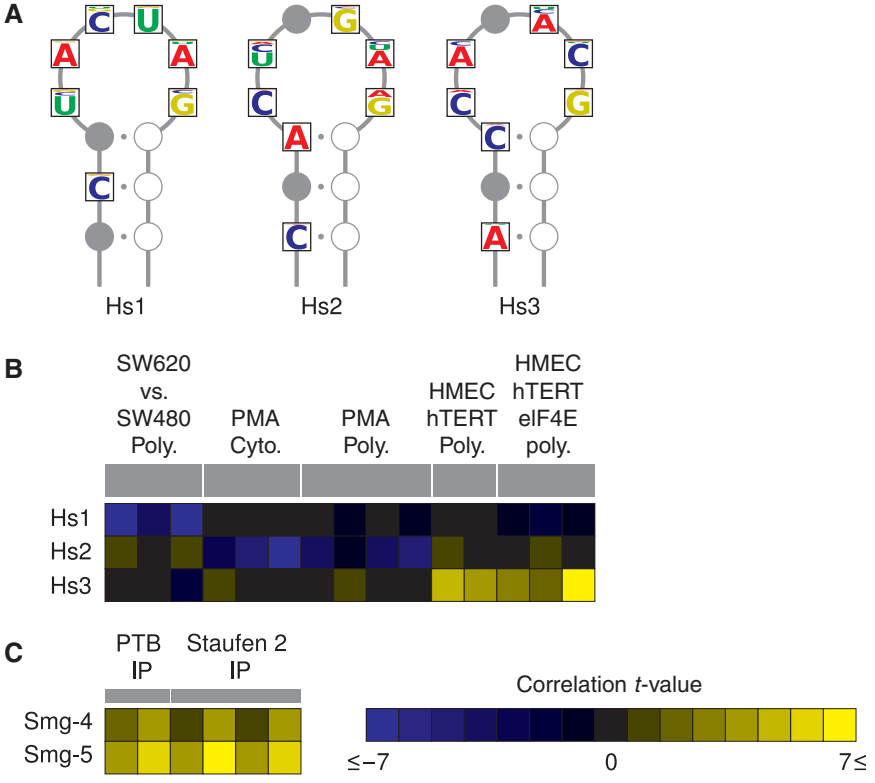

Figure 7 Putative human structural *cis*-regulatory elements. (**A**) We discovered three putative SCREs when applying the StructRED algorithm to human data (Hs1–3). (**B**) Hs1 was discovered using data that measured ribosome association in a metastatic colorectal cancer cell line, SW620, versus a non-metastatic cell line, SW480 (Provenzani *et al*, 2006). Hs2 correlated with decreased mRNA levels in U937 cells that have been exposed to 12-myristate 13-acetate (PMA; Kitamura *et al*, 2004). Hs3 was discovered through a positive correlation with ribosome association in human mammary epithelial cells (with and without overexpressed translation initiation factor 4F, eIF4E; Larsson *et al*, 2007). (**C**) We did not discover a Smaug/Vts1p-like specificity from the human data. However, when the Drosophila Smaug specificities were scored against the human data, we observed significant correlations with the RBP pull-down microarray data for poly-pyrimidine tract binding protein (PTB; GEO accession GSE6021; Gama-Carvalho *et al*, 2006) and Staufen1 and Staufen2 (Furic *et al*, 2008; GEO accessions GSE8437, GSE8438), suggesting that hSmaug shares many target mRNAs with these RBPs.

the transcription start and 3′ processing sites for most genes. However, David *et al* (2006) measured the mRNA levels for every yeast gene using a genome tiling microarray. Thus, they created a nucleotide-resolution map of the transcriptome expressed under log-growth conditions. Also, Miura *et al* (2006) sequenced two yeast cDNA libraries. Although neither dataset contains measurements for every gene, we averaged the annotated UTR lengths from these two sources to provide real full-length mRNAs for about half of the genome. We used the mean 5′ and 3′ UTR lengths from the known half to provide approximate mRNA sequences for the unknown half. *Drosophila* transcript sequences were downloaded from FlyBase (Wilson *et al*, 2008), and the longest transcript representing each gene in *Drosophila melanogaster* was used for further analysis. When estimating full-length mRNAs in other *Drosophila* species, the lengths of the annotated UTRs from *D. melanogaster* were used. Human sequences were downloaded from NCBI Entrez Nucleotide based on associated microarray annotations. To prevent errors in the regression procedure, the sequence sets were purged of redundant entries by using Blast (Altschul *et al*, 1990) to compare each sequence to every other sequence in the same set. One representative sequence was then chosen to represent each group that was hard clustered by having E-values for similarity $<10^{-10}$. An example of problematic sequences that this process removes are transposons. As a particular transposon type can have nearly identical sequences in multiple places in the genome and be represented by multiple microarray measurements, it can cause false correlations.

Genome-wide expression data

Data were downloaded from the NCBI GEO (Barrett *et al*, 2007). All data were purged of extreme outliers by using the Grubbs' test (Grubbs, 1969; *P*-value $\leqslant 10^{-9}$). We used datasets for further analysis if (1) we could automatically resolve the data in terms of systematic ORF identifiers and (2) after removing outliers, the dataset contained data for at least 3000 genes.

StructRED algorithm

RNA StructRED is a data integrative, regression-based algorithm with similarities to the algorithms represented by Stormo *et al* (1986), REDUCE (Bussemaker *et al*, 2001), and MatrixREDUCE (Foat *et al*, 2005, 2006). As input, StructRED takes one or more FastA sequence files with one or more entries per ORF identifier and one or more microarray datasets indexed by the same ORF identifiers. StructRED then builds a model to explain the observed mRNA levels in terms of small stem–loop motifs contained in the associated sequence data. The output of StructRED is a list of candidate SCREs as well as the inferred activity of the corresponding *trans*-factors across the input microarray conditions, a TFAP (Figure 1).

Assume we have spotted cDNA microarray data that provides a $\log_2$-ratio of an experimental condition intensity to a control condition intensity for each gene in the genome. Now let us assume that one or more RBPs are actively regulating under the tested condition and each one binds to a different SCRE. We pose the simplest possible starting model: We assume that binding by each RBP to an mRNA g causes an independent and therefore additive change in the observed $\log_2$-ratio A_g that is proportional by F_n to the number M_{ng} of its specifically bound SCREs n present in that particular mRNA sequence. We also allow for an intercept term C to account for any overall changes in the measured intensities between the two conditions and some experimental noise ε_g. Thus, we have a linear model that expresses the $\log_2$-ratio of an mRNA in terms of the SCREs that it contains:

$$A_g = \sum_n F_n M_{ng} + C + \varepsilon_g \qquad (2)$$

Even if the simplifications of this model are questionable, fitting the microarray data and nucleotide sequence to this model will still detect correspondences between the microarray values and SCRE occurrences. The magnitude of F_n is indicative of the regulatory strength of the *trans*-factor in that particular tested condition. If we normalize these coefficients by their standard error (resulting in a *t*-value) and track them across multiple microarray experiments, we have a TFAP.

If we have a dictionary of possible SCREs, we can discover the real SCREs by iteratively building the model explaining the microarray data through forward parameter selection. First, we score all candidate SCREs in the dictionary by performing ordinary least squares (OLS) regression on the counts of SCREs within mRNA sequences M_{ng} and the respective microarray $\log_2$-ratios A_g. The most explanatory SCRE, the one whose counts best fit the data, is selected from the dictionary, transformed into a matrix (see below), and added to the model. Next, the expression value predicted by the model (which now only contains one SCRE) is subtracted from each expression value, leaving a residual value. Each time using the residuals of the earlier iteration, the process repeats performing OLS on the entire dictionary, adding a SCRE term to the linear model, performing a multivariate fit including all SCRE terms, and calculating the residuals of the current model. When the last SCRE term added to the model does not significantly improve the fit beyond a predefined threshold *P*-value (corrected for multiple hypothesis testing), the process is stopped. The SCREs in the resulting model are the best candidates for real, functional SCREs that play a role in mRNA regulation for the process measured by the microarray experiment.

In practice, the model is not developed for a single experiment at a time. At every iteration, the best scoring SCRE in any experiment is added to the model and then scored for its fit to all experiments, resulting in its TFAP. Thus, a single model is built to explain all input experiment data. If a particular SCRE is not relevant to a particular experiment, its coefficient for that experiment will simply be near zero.

The dictionary of SCREs that this version of StructRED searches includes small stem–loop structures with stem and loop sizes that can be selected at the beginning of a run. The user can specify whether to include sequence information on the stem, loop, or both. The size of the stem–loop SCREs is limited only by available computational power. For all runs in this work, we used stems of length three and loops of lengths three to six nucleotides and allowed sequence information in one to seven positions on the stem and loop.

Concise pseudocode for the StructRED algorithm can be found as Supplementary Figure 2. The full software code will be made available upon request.

Converting oligonucleotide motifs into position-specific affinity matrices

In a PSAM, the definition of the weight w_{jb} for a particular nucleotide b at a particular position j is the ratio of equilibrium association constants K_a for the binding of a *trans*-factor to a reference oligonucleotide S_{ref} and an oligonucleotide that has a single base mutated S_{mut}, changing base position j to base b (see Foat *et al*, 2006 for a derivation):

$$w_{jb} = \frac{K_a(S_{mut})}{K_a(S_{ref})} \qquad (3)$$

To understand the later equations, we need to expand on equation (2). Composing F_n is a protein concentration $[P_n]$, an affinity $K_a(P_n, S_{ref})$ of protein P_n for a reference binding site S_{ref}, and a catch-all term β that includes unmodeled technology parameters.

$$A_g = \sum_n \beta[P_n]K_a(P_n, S_{ref})M_{ng} + C + \varepsilon_g \qquad (4)$$

In the case of scoring exact oligonucleotide structures, rather than weight matrix structures, M_{ng} just represents the count of occurrences of the oligonucleotide structure n in the sequence of mRNA g, and $K_a(P_n, S_{ref})$ becomes the K_a for the interaction of protein P_n with the oligonucleotide structure.

If the occurrences of a reference oligonucleotide S_{ref} and a mutant oligonucleotide S_{mut} were fit for genome-wide correlation with mRNA measurements, they would have two different coefficients F_{ref} and F_{mut}, respectively. The difference in the magnitude of F_{ref} and F_{mut} would be due to the different K_a's for these two sequences, as all else is constant. Thus, for a genome large enough to define F_{ref} and F_{mut}, w_{jb} is the ratio of these coefficients:

$$w_{jb} = \frac{\beta[P_n]K_a(S_{mut})}{\beta[P_n]K_a(S_{ref})} = \frac{F_{mut}}{F_{ref}} \quad (5)$$

StructRED uses this last equation to convert a best scoring oligonucleotide stem–loop motif into a stem–loop matrix SCRE. For each nucleotide in the best scoring stem–loop motif, the w_{jb} value in the matrix equals one. All other values of w_{jb}, are calculated by taking the ratio of the coefficients for each motif differing by a single nucleotide from the best motif versus the coefficient of the best motif. If the ratio is negative, meaning one motif has a positive correlation and one has a negative correlation, it is set to zero, as it suggests that the *trans*-factor cannot bind to the site with that particular single base change. Better estimates of the w_{jb} values are obtained by calculating their geometric mean across all microarray experiments in which the best motif correlated nearly as well with expression as in the best experiment (within one standard error of the best *t*-value). As only significantly correlating data contributes to the SCRE matrix, a matrix will not worsen if more, unrelated data are included as input.

If the assumption that nucleotides contribute independently to the binding of the *trans*-factor is valid, then this calculation loses information from motifs with more than one base change from the best scoring oligonucleotide. However, the above calculation is essentially free when compared with the calculations required in other PSAM-producing methods (Foat *et al*, 2005, 2006). This approximation creates both a faster and a simpler algorithm that maintains most of the information provided by more elaborate PSAM calculations.

Once the PSAM is derived, the sequence score M_{ng} of mRNA g for SCRE n is:

$$M_{ng} = \sum_{i=1}^{N_i} \Omega_n(S_i) \prod_{j=1}^{L_i} w_{jS_i(j)}, \quad (6)$$

where N_i is the number of sequence windows i in mRNA g; L_i is the length of the window i; $w_{jS_i(j)}$ is the weight in the PSAM at position j for the base $S_i(j)$ that appears at position j in the current window sequence S_i; and $\Omega_n(S_i)$ is a function that returns one if the sequence S_i in window i forms the correct structure and zero otherwise. This form of the model, having separate terms for the RNA structure and the nucleotide specificities, assumes that the RNA structure exists independent of the *trans*-factor's binding and that the *trans*-factor then recognizes specific nucleotides in a specific three-dimensional RNA configuration. An 'induced fit' scenario for the binding of the *trans*-factor would require a more complex relationship between the nucleotide specificities and the structure definition. The current StructRED method with a the binary $\Omega_n(S_i)$ function could be considered a simple approximation of a hypothetical algorithm in which $\Omega_n(S_i)$ would return values between zero and one depending on how likely it is that a particular window will form the correct secondary structure. The potential for such an algorithm is explored in Supplementary Figure 3.

Ensuring the discovered elements are structurally defined

To reduce the possibility that discovered SCREs actually represent single-stranded RNA-binding sites that contain inverted repeat sequences, we first fit a model to the microarray data that is composed of all significantly explanatory single-stranded RNA matrices, using a method similar to that described by Foat *et al* (2005). We then performed the search for SCREs using the residual microarray values after subtracting the model composed of single-stranded RNA matrices. This single-stranded matrix search is not strictly necessary. It would be possible to check the SCREs after the analysis to ensure that they could not be represented equally well without the stem–loop constraint. Factoring out single-stranded matrices first simply cuts down on false positives later, possibly at the expense of false negatives. However, any SCRE that could be well represented solely by its sequence composition would be discoverable by many existing sequence analysis tools and would not highlight the strengths of our method. To be certain that, the single-stranded matrix search worked as intended, all *Drosophila* SCREs were tested to make sure that the sequence specificity alone did not explain the data as well as the combined sequence/structure SCRE specificity (*P*-value <0.001; Supplementary Table 1). All discovered *Drosophila* and human SCREs were also checked to confirm that they did not explain mRNA measurements when the SCREs were scored on the strand complementary to the mRNA sequences. This check supported the premise that they are binding sites within mRNA transcripts and not transcriptional regulatory binding sites in UTRs or coding sequence on the DNA.

Structural *cis*-regulatory element responder ranking and annotation

One would expect that for a SCRE-containing mRNA to be a real, regulated target, it would have two major characteristics: It has one or more predicted high-affinity instances of a SCRE, and when the SCRE-binding factor is predicted to have a strong, genome-wide effect, the mRNA is regulated in the predicted direction. Thus, by combining the sequence score of each mRNA for a SCRE with its microarray measurements for important conditions for the SCRE, all genes in the genome can be ranked by how likely they are to contain real instances of the SCRE.

First, for each significantly correlating (*P*-value $\leqslant 0.001$) microarray experiment a for a particular SCRE m, we calculate a residual mRNA expression value e_{ga} that only contains the effect of the SCRE of interest. This value is the original value A_{ga} minus the model for $n-1$ of the SCREs, where the excluded SCRE is the SCRE of interest:

$$e_{mga} = A_{ga} - \sum_{n \neq m} F_{na} M_{ng} - C \quad (7)$$

We then divide e_{mga} by its coefficient F_{ma} to obtain a predicted sequence score $\hat{M}_{mg}$. All genes are then ranked by $\hat{M}_{mg}$ for each experiment a.

The probability of a particular gene g having a rank as good or better than its rank x is simply its rank divided by the number of genes G:

$$P_{top,a} = \frac{x_{ga}}{G} \quad (8)$$

Assuming independence over experiments a, the probability of the gene g having ranks as good or better than its observed ranks in all significantly correlating experiments N_a is:

$$P_{top,all} = \prod_a \frac{x_{ga}}{G} = G^{-N_a} \prod_a x_{ga} \quad (9)$$

Taking the $1/N_a$ power, of this probability gives us a mean probability across all a:

$$P_{top,mean} = G^{-1} \left(\prod_a x_{ga} \right)^{\frac{1}{N_a}} \quad (10)$$

Now if we want to have a mean probability that equally weights the probability of the predicted sequence score with the probability of a ranking y_g or better for the actual sequence score M_{mg}, we take another geometric mean:

$$P_{top,mean,combined} = \sqrt{\frac{y_g}{G} G^{-1} \left(\prod_a x_{ga} \right)^{\frac{1}{N_a}}} = G^{-1} \sqrt{y_g \left(\prod_a x_{ga} \right)^{\frac{1}{N_a}}} \quad (11)$$

In the end all we want is a new ranking z_g by this combined probability, so the G^{-1} scaling factor is irrelevant. Thus, we rank all genes by this two-step, geometric mean of ranks:

$$z_g = \sqrt{y_g \left(\prod_a x_{ga} \right)^{\frac{1}{N_a}}} \quad (12)$$

The only way for an mRNA to have a high responder rank z_g for a particular SCRE m is to have a high actual sequence score M_{mg} and to be predicted to have a high sequence score for the SCRE by $\hat{M}_{mg}$, based on the mRNA's behavior in conditions in which the SCRE is correlated with a large genome-wide effect. The probabilistic motivation of this ranking scheme is similar to that used in the rank product method for microarray analysis (Breitling *et al*, 2004).

Once all mRNAs in the genome are ranked by their responder scores, functional categories of genes are scored for enriched representation among the top-ranked mRNAs by using the Mann–Whitney–Wilcoxon test. Thus, the Gene Ontology (Ashburner *et al*, 2000) annotations for *S. cerevisiae* (Cherry *et al*, 1998) and *D. melanogaster* (Wilson *et al*, 2008) were used to score all category levels of the ontology trees. In addition, the fly community has created hierarchical controlled vocabularies for fly development and anatomy and have created gene annotations for allele phenotypes (Wilson *et al*, 2008) and *in situ* expression (Tomancak *et al*, 2002) using those controlled vocabularies. Thus, these two annotation systems were scored in a similar manner to the Gene Ontology system. No annotations were available for the human sequences.

SCRE location analysis

For the SCRE discovery and TFAP calculation, full-length mRNAs were used. However, it is possible to score subsequences to determine where in transcripts particular SCREs tend to occur genome-wide. Every mRNA was divided up into its 5′ UTR, CDS, and 3′ UTR, and TFAPs were recalculated for each SCRE using only each subsequence individually. Thus, the subsequence TFAP that most resembles the full-length mRNA TFAP is indicative of in which piece of the transcripts the SCRE tends to occur.

We also calculated the genome-wide averaged, length-normalized scores in the CDS and UTRs for each SCRE. Although no inference can be drawn from these calculations, it is interesting to observe if increased predictive power of a sequence region corresponds with increased density of predicted high-affinity binding sites (Supplementary Figure 4).

Visualizing SCRE position-specific affinity matrices with LoopLogo

StructRED produces representations of *cis*-regulatory elements consisting of both a weight matrix and a RNA structure. We developed LoopLogo, a script that generates the structural logos seen in Figures 1, 2, 4, and 7. In this representation, squares mark positions that contain sequence information. The height of each nucleotide character is proportional to its relative affinity. Nucleotides of weaker affinity are smaller and stacked on the larger higher-affinity nucleotides. Dark circles represent nucleotide positions that do not contain nucleotide information. Open circles on the 3′ side of the stem were not allowed to contain independent sequence information, as their sequence is largely specified by the sequence on the 5′ side of the stem (with the exception of G-U pairs). LoopLogo is implemented in Perl and creates Scalable Vector Graphics output.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Craig Smibert, Barak Cohen, Jim Skeath, Yue Zhao, and Ryan Christensen for critical readings of the paper. BCF is a PhRMA Foundation Informatics Fellow and a Washington University School of Medicine, Department of Genetics Fellow. This work was supported by National Institutes of Health grant HG00249 to GDS.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

Aviv T, Amborski AN, Zhao XS, Kwan JJ, Johnson PE, Sicheri F, Donaldson LW (2006a) The NMR and X-ray structures of the Saccharomyces cerevisiae Vts1 SAM domain define a surface for the recognition of RNA hairpins. *J Mol Biol* **356:** 274–279

Aviv T, Lin Z, Ben-Ari G, Smibert CA, Sicheri F (2006b) Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nat Struct Mol Biol* **13:** 168–176

Aviv T, Lin Z, Lau S, Rendl LM, Sicheri F, Smibert CA (2003) The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. *Nat Struct Biol* **10:** 614–621

Baez MV, Boccaccio GL (2005) Mammalian Smaug is a translational repressor that forms cytoplasmic foci similar to stress granules. *J Biol Chem* **280:** 43131–43140

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res* **35:** D760–D765

Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573:** 83–92

Bushati N, Cohen SM (2007) microRNA functions. *Annu Rev Cell Dev Biol* **23:** 175–205

Bussemaker HJ, Foat BC, Ward LD (2007) Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct* **36:** 329–347

Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. *Nat Genet* **27:** 167–171

Cheadle C, Fan J, Cho-Chung YS, Werner T, Ray J, Do L, Gorospe M, Becker KG (2005) Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *BMC Genomics* **6:** 75

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D (1998) SGD: saccharomyces genome database. *Nucleic Acids Res* **10:** 73–79

Coppins RL, Hall KB, Groisman EA (2007) The intricate world of riboswitches. *Curr Opin Microbiol* **10:** 176–181

Dahanukar A, Walker JA, Wharton RP (1999) Smaug, a novel RNA-binding protein that operates a translational switch in Drosophila. *Mol Cell* **4:** 209–218

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **103:** 5320–5325

De Renzis S, Elemento O, Tavazoie S, Wieschaus EF (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biol* **5:** e117

Edwards TA, Butterwick JA, Zeng L, Gupta YK, Wang X, Wharton RP, Palmer AG, Aggarwal AK (2006) Solution structure of the Vts1 SAM domain in the presence of RNA. *J Mol Biol* **356:** 1065–1072

Fan J, Yang X, Wang W, Wood WH, Becker KG, Gorospe M (2002) Global analysis of stress-regulated mRNA turnover by using cDNA arrays. *Proc Natl Acad Sci USA* **99:** 10611–10616

Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci USA* **102:** 17675–17680

Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22:** e141–e149

Fruscio MD, Styhler S, Wikholm E, Boulanger MC, Lasko P, Richard S (2003) Kep1 interacts genetically with dredd/caspase-8, and kep1 mutants alter the balance of dredd isoforms. *Proc Natl Acad Sci USA* **100:** 1814–1819

Furic L, Maher-Laporte M, DesGroseillers L (2008) A genome-wide approach identifies distinct but overlapping subsets of cellular mRNAs associated with Staufen1- and Staufen2-containing ribonucleoprotein complexes. *RNA* **14:** 324–335

Gama-Carvalho M, Barbosa-Morais NL, Brodsky AS, Silver PA, Carmo-Fonseca M (2006) Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *Genome Biol* **7:** R113

Gerber AP, Luschnig S, Krasnow MA, Brown PO, Herschlag D (2006) Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. *Proc Natl Acad Sci USA* **103:** 4487–4492

Grubbs F (1969) Procedures for detecting outlying observations in samples. *Technometrics* **11:** 1–21

Johnson PE, Donaldson LW (2006) RNA recognition by the Vts1p SAM domain. *Nat Struct Mol Biol* **13:** 177–178

Keene JD (2007) RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8:** 533–543

Kitamura H, Nakagawa T, Takayama M, Kimura Y, Hijikata A, Hijika A, Ohara O (2004) Post-transcriptional effects of phorbol 12-myristate 13-acetate on transcriptome of U937 cells. *FEBS Lett* **578:** 180–184

Larsson O, Li S, Issaenko OA, Avdulov S, Peterson M, Smith K, Bitterman PB, Polunovsky VA (2007) Eukaryotic translation initiation factor 4E induced progression of primary human mammary epithelial cells along the cancer pathway is associated with targeted translational deregulation of oncogenic drivers and inhibitors. *Cancer Res* **67:** 6814–6824

Lev J (1949) The point biserial coefficient of correlation. *Ann Math Stat* **20:** 125–126

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci USA* **103:** 17846–17851

Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FHT (2006) Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nat Struct Mol Biol* **13:** 160–167

Pérez-Ortín JE, Alepuz PM, Moreno J (2007) Genomics and gene transcription kinetics in yeast. *Trends Genet* **23:** 250–257

Pilot F, Philippe JM, Lemmers C, Chauvin JP, Lecuit T (2006) Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation. *Development* **133:** 711–723

Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M, Quattrone A (2006) Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis* **27:** 1323–1333

Qin X, Ahn S, Speed TP, Rubin GM (2007) Global analyses of mRNA translational control during early Drosophila embryogenesis. *Genome Biol* **8:** R63

Rabani M, Kertesz M, Segal E (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci USA* **105:** 14885–14890

Rendl LM, Bieman MA, Smibert CA (2008) S. cerevisiae Vts1p induces deadenylation-dependent transcript degradation and interacts with the Ccr4p-Pop2p-Not deadenylase complex. *RNA* **14:** 1328–1336

Robard C, Daviau A, Fruscio MD (2006) Phosphorylation status of the Kep1 protein alters its affinity for its protein binding partner alternative splicing factor ASF/SF2. *Biochem J* **400:** 91–97

Sankoff D (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* **45:** 810–825

Semotok JL, Cooperstock RL, Pinder BD, Vari HK, Lipshitz HD, Smibert CA (2005) Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early Drosophila embryo. *Curr Biol* **15:** 284–294

Semotok JL, Luo H, Cooperstock RL, Karaiskakis A, Vari HK, Smibert CA, Lipshitz HD (2008) Drosophila maternal Hsp83 mRNA destabilization is directed by multiple SMAUG recognition elements in the open reading frame. *Mol Cell Biol* **28:** 6757–6772

Smibert CA, Lie YS, Shillinglaw W, Henzel WJ, Macdonald PM (1999) Smaug, a novel and conserved protein, contributes to repression of nanos mRNA translation *in vitro*. *RNA* **5:** 1535–1547

Stormo GD, Schneider TD, Gold L (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res* **14:** 6661–6679

Tadros W, Goldman AL, Babak T, Menzies F, Vardy L, Orr-Weaver T, Hughes TR, Westwood JT, Smibert CA, Lipshitz HD (2007) SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase. *Dev Cell* **12:** 143–155

Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16:** 962–972

Tate R (1954) Correlation between a discrete and a continuous variable. Point-biserial correlation. *Ann Math Stat* **25:** 603–607

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* **3:** RESEARCH0088

Vardy L, Orr-Weaver TL (2007) Regulating translation of maternal messages: multiple repression mechanisms. *Trends Cell Biol* **17:** 547–554

Wilson RJ, Goodman JL, Strelets VB, Consortium F (2008) FlyBase: integration and improvements to query tools. *Nucleic Acids Res* **36:** D588–D593

Zhang B, Gallegos M, Puoti A, Durkin E, Fields S, Kimble J, Wickens MP (1997) A conserved RNA-binding protein that regulates sexual fates in the C. elegans hermaphrodite germ line. *Nature* **390:** 477–484

