Published online 7 March 2017

Nucleic Acids Research, 2017, Vol. 45, No. 10 ***6109–6118***
doi: 10.1093/nar/gkx161

Melting temperature highlights functionally important RNA structure and sequence elements in yeast mRNA coding regions

Fei Qi[1] and Dmitrij Frishman[1,2,*]

[1]Department of Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, Maximus-von-Imhof-Forum 3, D-85354 Freising, Germany and [2]St Petersburg State Polytechnic University, St Petersburg 195251, Russia

Received December 20, 2016; Revised February 04, 2017; Editorial Decision February 24, 2017; Accepted February 24, 2017

ABSTRACT

Secondary structure elements in the coding regions of mRNAs play an important role in gene expression and regulation, but distinguishing functional from non-functional structures remains challenging. Here we investigate the dependence of sequence–structure relationships in the coding regions on temperature based on the recent PARTE data by Wan *et al*. Our main finding is that the regions with high and low thermostability (high T_m and low T_m regions) are under evolutionary pressure to preserve RNA secondary structure and primary sequence, respectively. Sequences of low T_m regions display a higher degree of evolutionary conservation compared to high T_m regions. Low T_m regions are under strong synonymous constraint, while high T_m regions are not. These findings imply that high T_m regions contain thermo-stable functionally important RNA structures, which impose relaxed evolutionary constraint on sequence as long as the base-pairing patterns remain intact. By contrast, low thermostability regions contain single-stranded functionally important conserved RNA sequence elements accessible for binding by other molecules. We also find that theoretically predicted structures of paralogous mRNA pairs become more similar with growing temperature, while experimentally measured structures tend to diverge, which implies that the melting pathways of RNA structures cannot be fully captured by current computational approaches.

INTRODUCTION

Secondary structure elements and global folding patterns play a fundamental role in the function and regulation of RNAs (1–6). For a number of years most of the attention was given to functional secondary structures of non-coding RNAs (ncRNAs), but more recently mRNA structure also moved to the spotlight of genomics and bioinformatics research. mRNA structures have been found to influence multiple stages of gene expression and protein synthesis, including transcription, splicing, RNA transport, translation initiation, elongation and termination, as well as RNA degradation (7–13). Secondary structures of the three functional mRNA domains—5′ UTR, the coding region and 3′ UTR—are largely independent, since base pairs across domain borders are rare (14). A broad variety of functional structural elements were described in UTRs (15–18). While the primary function of the coding regions is to encode amino acid sequences of proteins, they are presumed to contain even more RNA secondary structures than UTRs (1), with some of them already proven to be functional (19–24). The redundancy of the genetic code makes it possible for the coding regions to carry overlapping functions, which manifest themselves at the level of protein and RNA sequences and structures (14,25,26).

In recent years, several experimental approaches have been developed for genome-wide measurement of RNA structures. These methods, including Frag-seq (fragmentation sequencing) (27), PARS (parallel analysis of RNA structure) (1) and SHAPE-seq (selective 2′-hydroxyl acylation analyzed by primer extension sequencing) (28) combine RNA structure probing by structure-specific enzymes and chemical modifications at double- or single-stranded bases with high-throughput next generation sequencing. These approaches can probe millions of molecules at single nucleotide resolution within one experiment and therefore enable comprehensive studies of RNA structures (29). An important question, which arises in this context, is to which extent the results of high-throughput structure probing experiments are reproducible and compatible with each other.

Owing to the availability of RNA structure probing data many of the classical problems in molecular evolution, which have been extensively addressed for protein molecules, can now be examined for mRNAs as well. In

*To whom correspondence should be addressed. Tel: +49 816 171 2134; Fax: +49 816 171 2186; Email: d.frishman@wzw.tum.de

particular, it is of great interest to investigate to which extent secondary and/or tertiary structure of mRNAs constrains sequence variation and how strongly mRNA structures are conserved in evolution, as was done for protein 3D structures long ago (30). Recently, Wan *et al.* published PARTE (Parallel Analysis of RNA structures with Temperature Elevation) experiment, in which secondary structures of yeast RNAs were probed and melting temperatures (T_m) were derived at single nucleotide resolution at five temperatures (13). Using these data, we demonstrate that high and low thermostability regions in the mRNA coding regions highlight functionally important RNA structures and sequence segments, respectively. We report a surprising pattern of structural divergence between sequence-similar mRNAs along the temperature ladder, which cannot be captured by the currently available computational approaches. There is a considerable reproducibility between the high-throughput RNA structure probing experiments, PARS and PARTE.

MATERIALS AND METHODS

Experimental data on secondary structures of yeast mRNAs

Secondary structure profiles of 3002 yeast mRNAs determined at room temperature by PARS (Parallel Analysis of RNA Structures (1)) were downloaded from http://genie.weizmann.ac.il/pubs/PARS10. For each individual nucleotide position of mRNAs a PARS score reflects its likelihood to be in a double-stranded conformation based on the number of sequencing reads upon treatment by two structure-specific enzymes, RNase V1 and nuclease S1, which cleave at double-stranded and single-stranded regions, respectively. A total of 4 405 020 bases in the 3002 mRNAs are covered by PARS scores. For a given mRNA sequence, the vector of its PARS scores is referred to as its PARS structure.

Another mRNA structure dataset used in this work was obtained by a PARTE experiment, in which 4562 yeast mRNAs were structure-probed by RNase V1 at five temperatures (23, 30, 37, 55 and 75°C; two biological replicates were performed for each temperature) (13). PARTE reveals T_m of each base, with double-stranded regions being progressively eliminated as temperature increases. V1 reads resulting from this experiment were downloaded from the GEO database (31) (GSE39680) and their counts were normalized exactly as described by Wan *et al.* (13): (i) for each library peaks were defined as those bases that are covered by more reads than the bases on their left and their right and whose read coverage is greater than the average coverage of bases on the same gene and the average coverage of all bases; (ii) using the PoissonSeq algorithm (32), the library size of each sequencing lane was estimated based on the high confidence peaks observed in both duplicate libraries at at least one of the five temperatures; and (iii) V1 read numbers were normalized by dividing the counts in each sample by the corresponding library size. For each RNA nucleotide position the $\log_2$ value of the mean of the two normalized V1 read numbers from two duplicate samples was treated as its PARTE score (each mean V1 read number was augmented by 0.001 to avoid the undefined logarithm values for those bases where the read count is zero). A total of 7 497 468 bases in the 4562 mRNAs are covered by PARTE scores. For a given mRNA sequence, vectors of its PARTE scores are referred to as its PARTE structure.

Predicted secondary structures of yeast mRNAs

Sequences of 6686 yeast mRNAs were downloaded from SGD (release 57-1-1) (33). Base pairing probabilities for each yeast mRNA were calculated using the RNAfold algorithm from the ViennaRNA package (34). In order to simulate the PARTE experiment, which was carried out at five different temperatures, RNAfold was run five times for each yeast mRNA using five different values of the -T parameter (23,30,35,55,75). For a given mRNA, vectors of its theoretically predicted base pairing probabilities are referred to as its predicted structures.

Yeast paralogous mRNAs

We considered 246 pairs of aligned mRNA sequences of yeast paralogs, as well as the percent identity between aligned coding regions of each pair, as described previously (35). Each of these pairs of paralogous proteins shares over 50% amino acid sequence identity and <10% difference in sequence length.

Distances between secondary structures of yeast paralogs

For a given pair of aligned mRNA sequences, we employed root mean square deviation (RMSD) as the measure of the distance between their experimental structures. RMSD values were calculated between the vectors of PARS or PARTE scores for all aligned positions without gaps. Sequence positions not probed in the experiments (i.e. those with read number 0) were also taken into account in this calculation—their PARS score equals 0 and their PARTE score is $\log_2(0 + 0.001)$. Similarly, distances between predicted structures for a given pair of aligned mRNAs were calculated as RMSD between vectors of predicted base pairing probabilities.

Paired and unpaired bases of yeast mRNAs

We subdivided the bases of yeast mRNAs into two classes according to their PARS scores: (i) paired bases (PARS score ≥ 0) and (ii) unpaired bases (PARS score <0). In all the mRNAs covered by both PARS and PARTE experiments we identified 3 514 124 paired bases and 884 030 unpaired bases.

Melting temperatures of RNA structures in yeast mRNAs

Data on T_m of RNA structures covering over 320 000 bases in yeast mRNAs were kindly provided by Yue Wan and Howard Y. Chang from the Howard Hughes Medical Institute and the Program in Epithelial Biology at Stanford University School of Medicine. For each RNA sequence position the T_m value was calculated as the mean of the two temperatures between which the position transitioned from the double-stranded to the single-stranded state in the PARTE experiment, i.e. 26.5, 33.5, 46 and 65°C for the transitions

between 23 and 30°C, 30 and 37°C, 37 and 55°C, and 55 and 75°C, respectively. Positions that remained double-stranded at 75°C were assigned the T_m value of 80°C (13). We only considered 1262 mRNAs that have at least 5% of bases with probed melting temperatures.

Regions with high or low T_m in pairs of paralogous yeast mRNAs

We applied a sliding window of 100 nt with a step size of 10 bases to the alignments of paralogous yeast mRNAs and calculated average T_m values for each sequence in a window. Only those windows in which the alignment had <10% of gaps and the two aligned sequences both had at least 10% of bases with probed T_m were included in the analysis. We defined two classes of windows—high T_m windows and low T_m windows—dependent on whether the two aligned sequences in a window both had T_m values among the top or bottom 25% of all T_m values. Windows with the middle range of T_m values were excluded from further analysis. For each window, the T_m values of the two aligned sequences were calculated as the arithmetic mean of all the T_m values of their individual bases. Overlapping windows belonging to the same class (high or low T_m) were merged, yielding the total of 167 regions with high T_m and 96 regions with low T_m among the 246 pairs of paralogous mRNAs.

Thermo-stable and meltable positions in pairs of yeast paralogous mRNAs

For each pair of yeast paralogous mRNAs, a position in the alignment was defined as thermo-stable or meltable dependent on whether both aligned bases in this position had $T_m \geq 65$°C or $T_m \leq 46$°C, respectively. This procedure yielded 1323 thermo-stable and 256 413 meltable positions, out of which 734 thermo-stable and 20 790 meltable positions were located in high T_m regions while low T_m regions contained 22 765 meltable positions and no thermo-stable positions.

Synonymous base substitutions between yeast paralogs

The proportion of synonymous (pS) differences between yeast paralogs was estimated by the equation $pS = Sd/S$, where Sd is the number of observed synonymous substitutions and S is the number of potential synonymous substitutions. In this work, we employed the SNAP software (36) to calculate the pS. We compared the pS value of each high/low T_m region (pS_{region}) with the pS value of the entire alignment of yeast paralogous mRNAs ($pS_{alignment}$) by calculating $\Delta pS = pS_{region} - pS_{alignment}$.

Zipcodes in yeast mRNAs

Positions of 12 functional motifs in yeast mRNAs responsible for binding with mRNA transport proteins—the so called zipcodes—were obtained from the study of Jambhekar *et al.* (37). Five of these zipcodes (ASH1-E1min, TPO1N, ERG2N, WSC2C and SRL1C), which are located in the coding regions and covered by the PARTE melting temperature data, were included in further analysis.

RESULTS

Correlation between PARS and PARTE scores

Reproducibility of results is a crucial aspect in the evaluation of experiment strategies. To assess the correlation between the PARTE and PARS data, we computed Spearman's rank correlation coefficients between PARS and PARTE scores over all 4 398 154 bases in those 2995 mRNA sequences that are contained both in the PARS and in the PARTE datasets. The correlation coefficients were relatively low (0.325, 0.323, 0.321, 0.312 and 0.250 at the PARTE temperatures of 23, 30, 37, 55 and 75°C, respectively) but highly significant (all P-values < 2.2e-16). As expected, the highest correlation was detected at 23°C as it is closest to the room temperature at which the PARS experiment was carried out. The lower correlation at high temperatures can be explained by progressive unfolding of RNA structures.

Correlation between PARS and PARTE scores is not surprising given the similarity of these two experimental strategies (1,13). PARS and PARTE scores both reflect the likelihood of individual bases to be in a double-stranded conformation (1,13). The relatively low correlation coefficients are due to a key difference between the PARS and the PARTE experiments—the enzymes used to detect RNA structures. While PARS relies both on RNase V1 and nuclease S1 to probe the bases in a double- and single-stranded conformation, respectively, PARTE only probes bases in a double-stranded conformation by RNase V1 (1,13). Therefore, while PARS can capture the likelihood of bases to be in a single-stranded conformation based on reads stemming from the nuclease S1, the PARTE experiment does not deliver this information. Indeed, the correlation between PARTE and PARS scores was much stronger (correlation coefficient 0.587, P-value < 2.2e-16) when only bases in double-stranded conformation were considered (data not shown).

Dependence of structure divergence on sequence identity

In our previous work, we explored sequence–structure relationships in yeast mRNAs based on PARS data (35). Upon comparing secondary structures between sequence-similar paralogous yeast mRNAs we found that coding regions of mRNAs are not under strong evolutionary pressure to preserve a particular global shape, which implies that global secondary structure of the coding regions does not play a major role in gene regulation. The recent availability of PARTE data for yeast mRNAs (13) has made it possible to investigate sequence–structure divergence at different temperature levels. As seen in Supplementary Figure S1, at all five temperatures the similarity of PARTE structures shows no correlation with the sequence similarity in the range of sequence identity between 50% (the lowest level considered) and roughly 85–90%. In this range the distance between experimental structures of paralogous mRNAs does not differ from the median distance between randomly selected mRNA pairs (dashed horizontal lines in Supplementary Figure S1). By contrast, at sequence identity levels over 85–90% the distance between experimental structures of paralogous mRNAs displays a near linear dependence on sequence identity (Supplementary Figure S2 and Table S1).

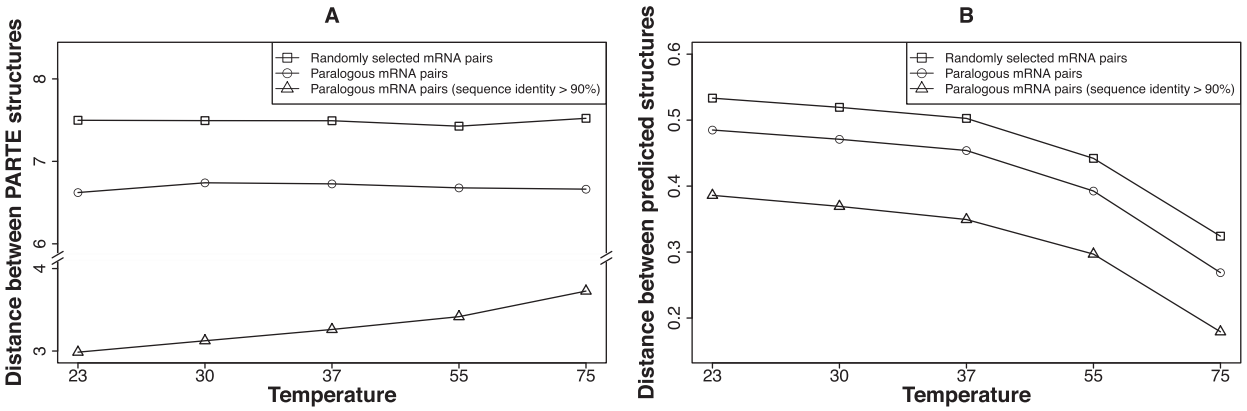

Figure 1. Variation of the distance between secondary structures of paralogous mRNA pairs along the temperature ladder. Points are the median levels of the distance at each temperature. (**A**) Distance between PARTE structures. (**B**) Distance between predicted structures.

This finding is in line with our previous analysis of structure probing data obtained by the PARS method (1), in which we found that the global structural conformation of the coding regions is not crucial for gene expression and regulation. This result is compatible with the notion that mRNA conformation depends on interactions with the solvent as well as with proteins and other ligands and that mRNAs adopt a highly dynamic ensemble of conformations instead of a single global structure (7). An important insight provided by our analysis is that interrogation of mRNA structures by PARS and PARTE leads to qualitatively similar evolutionary conclusions, indicating the reproducibility of the high-throughput RNA structure probing experiments.

Variation of the distance between paralogous mRNA structures along the temperature ladder

The availability of the PARTE structure-probing data opens up the possibility to investigate how the distance between secondary structures of similar RNA molecules varies with temperature and to obtain clues about the RNA structure unfolding pathways during the melting process. We therefore calculated structural distances between paralogous yeast mRNAs along the temperature ladder. Intuitively, one would expect the structural distance to be inversely proportional to temperature: as the temperature grows, more and more base pairs melt and an ever increasing portion of both molecules becomes single-stranded and thus more similar to each other. However, the experimentally determined PARTE structures show a strikingly different behavior. The distances between randomly selected and all paralogous mRNA pairs do not appear to vary with temperature at all, while the distances between highly similar paralogous mRNA pairs actually become larger at higher temperatures (Figure 1A). We speculate that this surprising pattern may, to some extent, be due to the limitations of the experimental approach. First, the PARTE experiment probes the *in vitro* re-folded RNA structures rather than *in vivo* structures (13). Second, as noted by Wan *et al.*, in the PARTE data 20% of the bases show a transition for increased V1 reads at higher temperatures, which may indicate that a considerable proportion of thermo-stable RNA secondary structures became accessible to RNase V1 only upon dissolution of tertiary structures (13). This implies that the differences between these structures were only detected at higher temperatures. We also cannot rule out the possibility that the RNA unfolding pathways during the melting process are actually quite different even between similar molecules, presumably due to complex tertiary interactions and dynamic effects.

This unexpected trend could not be captured by RNAfold predictions. As seen in Figure 1B, the distances between the predicted structures behave exactly as intuitively expected. The distances between the predicted structures of randomly selected, all paralogous and highly similar paralogous mRNA pairs all become smaller as the temperature grows from 23 to 75°C. The same pattern was also obtained with the RNAplfold program, which computes local base pair probabilities (Supplementary Figure S3). This may be due to a number of inherent limitations of the current computational structure prediction approaches, especially when applied to long RNA sequences and at large deviations in temperature from standard conditions. Exponential growth of the number of possible secondary structures with the sequence length necessitates the introduction of approximations into the folding algorithms (38). Modeling pseudoknots and prediction of long-range interactions continue to be an unsolved problem (39). As well, energy calculations are parametrized at 37°C and become less reliable at other temperatures (40,41).

Melting temperature highlights functionally important structure and sequence elements in the coding regions of mRNAs

As discussed above, the global secondary structure of the mRNA coding regions is poorly conserved in evolution and probably does not play a role in gene regulation. Instead, RNA structure is more likely to be functional at the level of local structural elements situated in the coding regions. However, it is very hard to distinguish functionally important structural elements from non-functional ones since every RNA chain tends to fold back on itself for thermodynamic reasons (13). One important and experimentally measurable feature that may be indicative of functionality is the thermostability of RNA structures. It has been demon-

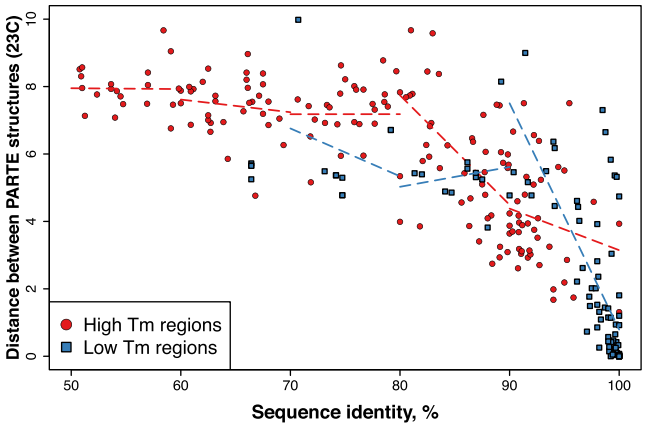

Figure 2. Sequence–structure relationships in the high/low T_m regions of paralogous mRNA pairs. For high T_m regions, the distance between structures shows a linear dependence from sequence identity for sequence identity values over 80% (correlation coefficient −0.54, *P*-value = 2.0e-7). For low T_m regions, the distance between structures shows a linear dependence from sequence identity for the sequence identity values over 90% (correlation coefficient −0.69, *P*-value = 2.1e-11). Linear regression for each 10% range of sequence identity is shown by a dashed line with the corresponding color. PARTE structures at 23°C were used.

strated that in ncRNAs functionally important structures have more stable structures than random RNAs of the same length and dinucleotide frequency (42,43). Many known functional structured RNA regulatory elements were identified in yeast mRNA 3′ UTRs by locating thermostable base pairs (13). It is therefore conceivable that functionally important structural elements in the coding regions of mRNAs could also be discriminated by their thermostability.

Locally stable structures can be gleaned from the genome-wide PARTE experiment, in which secondary structures of yeast RNAs were probed and T_m were derived at single nucleotide resolution at five temperatures (13). However, proving that such local structures actually fulfill a biological function is a challenging task. One approach to this problem could be based on assessing RNA-level selective constraints acting on protein-coding regions, including synonymous constraint and compensatory mutations. These unique patterns of sequence–structure relationships are a hallmark of the functionally important RNA elements in the coding regions.

We identified 167 high T_m regions and 96 low T_m regions in 246 pairs of paralogous mRNAs. As seen in Figure 2, high T_m regions show a much stronger sequence–structure relationship than the low T_m regions in paralogous mRNA pairs. The distance between the structures of high T_m regions depends linearly on their sequence similarity for the sequence identity levels over 80%, while in low T_m regions this dependence only becomes apparent for sequences that share more than 90% identity. Low T_m regions show a higher sequence identity than high T_m regions (Figure 3A), while high T_m regions display a smaller structural distance upon controlling for sequence identity level (Figure 3B). Thus, high Tm and low T_m regions are under evolutionary pressure to preserve secondary RNA structure and primary sequence, respectively, and would therefore be expected to contain functionally important RNA structure elements and sequence segments, respectively. Indeed, high thermostability is a pre-requisite for functionally important RNA structure elements (13,42–44) while low thermostability ensures sufficient accessibility of functionally important RNA sequence elements (45,46). Melting temperature is thus a crucial parameter, which correlates with the distribution of functionally important structure and sequence elements along the coding regions of mRNAs.

Our finding that low T_m regions are more conserved in the nucleotide sequence than high T_m regions does not contradict to the conclusion of Wan *et al.* that thermo-stable bases in yeast mRNAs are significantly more conserved than meltable bases (13). In contrast to the analysis of Wan *et al.*, which was performed at single-nucleotide resolution in full-length mRNA sequences, our results are solely based on low and high T_m regions in the coding portions of mRNAs. We were able to reproduce the results of Wan *et al.* and confirm that at single-base resolution, thermo-stable positions are more conserved than meltable positions when all individual positions of the entire coding regions in yeast paralogous mRNA alignments were examined together (Supplementary Figure S4). However, when only positions located in high T_m and low T_m regions were examined separately, in high T_m regions the thermo-stable positions exhibited higher sequence conservation than meltable positions, while in low T_m regions meltable positions displayed a very high conservation level and thermo-stable position were completely absent (Figures 4 and 5). When considering the conservation of coding mRNA regions both at the structure and sequence level, it becomes apparent that the relatively high conservation level of thermo-stable positions in high T_m regions reflects evolutionary pressure to preserve RNA structure. The highest sequence conservation level observed in meltable positions of the low T_m regions is a reflection of the relatively high evolutionary pressure to preserve primary RNA sequence experienced by low T_m regions.

Low T_m regions are under synonymous constraint while high T_m regions exhibit relaxed sequence constraint

It is currently believed that RNA-level functions in coding regions manifest themselves by synonymous constraint (25,26). We therefore compared the synonymous substitution rate (pS) in the high and low T_m regions with the pS values calculated over the entire alignment of yeast paralogous mRNAs. Most low T_m regions exhibit negative ΔpS values while most high T_m regions exhibit positive ΔpS values (chi-squared test, *P*-values < 0.01) (Figure 6), which indicates that the low T_m regions are under synonymous constraint and may harbor functionally important nucleotide sequence motifs, such as ncRNA and protein binding sites (25,26,47). This notion is compatible with the complete absence of thermo-stable nucleotide base pairs in low T_m regions, ensuring good accessibility of binding sites. High concentration of synonymous substitutions in high T_m regions may points to relaxed RNA sequence constraint, which may provide an evolutionary advantage for these regions in terms of accommodating functionally important RNA secondary structure elements (6,25).

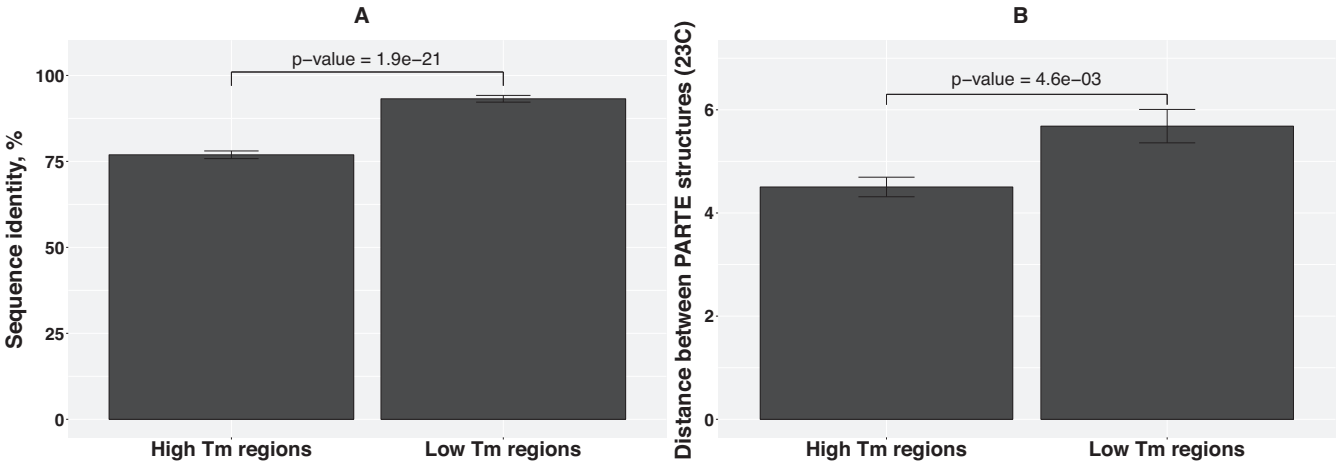

Figure 3. Low T_m regions in paralogous mRNA pairs are more conserved in sequence while high T_m regions are more conserved in RNA secondary structure. (**A**) Sequence identity between mRNAs (Mann–Whitney–Wilcoxon test, *P*-value = 1.9e-21). (**B**) Distance between RNA secondary structures (PARTE structures at 23°C; Mann–Whitney–Wilcoxon test, *P*-value = 4.6e-3). The differences are significant according to Mann–Whitney–Wilcoxon test. Error bars indicate standard error. The investigation of structural distances was effected upon controlling for sequence identity. Only regions with sequence identity 85–95% were considered in this analysis, while the regions with sequence identity < 85%, for which the distance between structures does not differ from randomly selected mRNA pairs as well as the regions with sequence identity > 95%, among which almost no high T_m regions exists, were excluded from consideration.

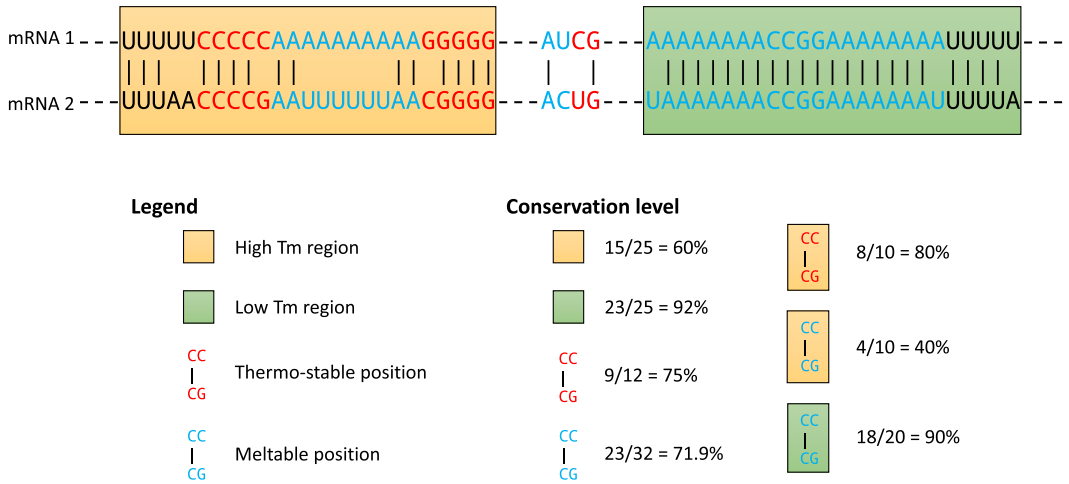

Figure 4. Schematic illustration of the conservation levels of high/low T_m regions and thermo-stable/meltable positions. The alignment of two mRNAs is shown on the top. The low T_m region displays a higher sequence identity than the high T_m region (92 versus 60%). When all thermo-stable and meltable positions are considered together, the thermo-stable positions show a higher conservation level than the meltable positions (75 versus 71.9%). When the thermo-stable and the meltable positions located in high T_m and low T_m regions are considered separately, the meltable positions in the low T_m region are most conserved (90%), followed by the thermo-stable positions in the high T_m region (80%), while the meltable positions in the high T_m region are least conserved (40%).

Functionally important structure elements in the coding regions of yeast mRNAs tend to be thermostable

A typical class of functionally important structure elements in yeast mRNAs is constituted by the so called zipcodes—regions of mRNAs recognized by the RNA-binding protein She2p (5,37,48). Localized mRNAs are transported to the bud tip of the daughter cell by the She protein complex depending on the interaction between She2p and the loop-stem-loop structure of the zipcode (5,37,48). Out of the 12 functional zipcodes in yeast mRNAs identified in a previous study (37), 5 zipcodes (ASH1-E1min, TPO1N, ERG2N, WSC2C and SRL1C) are located in the coding regions and covered by the PARTE melting temperature data. These 5 zipcodes range from 49 (ASH1-E1min) to 178 (SRL1C) nucleotides in length. Another functionally important structure element in yeast mRNA coding regions is the *URE2* IRES (internal ribosome entry site) element, which locates between nucleotides 205 and 309 in the *URE2* coding region and folds into a stem-loop structure (49). This IRES element mediates the cap-independent internal initiation of translation resulting in the expression of an N-terminal truncated form of the Ure2p protein (50). We calculated the T_m of each structure element by averaging the T_m values of every PARTE-probed base within the element. All six structure el-

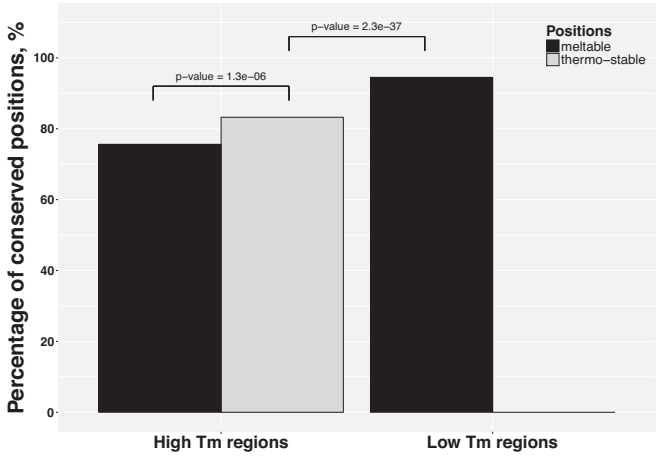

Figure 5. Conservation levels of thermo-stable and meltable positions in high/low T_m regions. Positions in high T_m and low T_m regions are examined separately. In high T_m regions the thermo-stable positions exhibit higher sequence conservation than meltable positions (Z-test for two proportions, P-value = 1.3e-6), while in low T_m regions meltable positions display a very high conservation level (Z-test for two proportions, P-value = 2.3e-37) and thermo-stable position are completely absent.

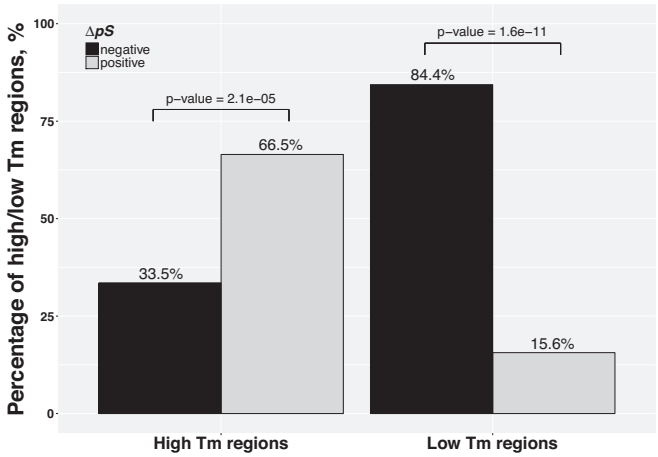

Figure 6. Most low T_m regions exhibit negative ΔpS values while most high T_m regions exhibit positive ΔpS values. The differences are significant according to chi-squared test. Error bars indicate standard error.

ements show high T_m values (ASH1-E1min: 46°C, TPO1N: 51.9°C, ERG2N: 63.8°C, WSC2C: 80°C, SRL1C: 54.7°C and *URE2* IRES: 53.6°C), which fall into the typical range of high T_m regions and thus exhibit high thermostability (Figure 7). This finding supports the hypothesis that high thermostability is indicative of functionally important RNA structure elements in mRNA coding regions.

DISCUSSION

Figure 8 summarizes our findings about high/low T_m regions as well as our inference based on these findings. High T_m regions exhibit a stronger sequence–structure relationship, conserved and thermo-stable RNA secondary structures and relatively divergent nucleotide sequences, while low T_m regions display a weaker sequence–structure

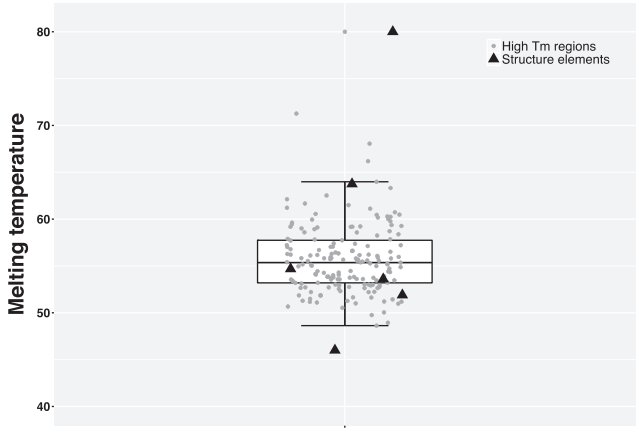

Figure 7. Melting temperatures of high T_m regions (boxplot and grey dots) and experimentally validated structure elements (black triangles).

relationship, divergent and less thermostable RNA secondary structures and highly conserved nucleotide sequences. These findings suggest that high T_m regions are under high evolutionary pressure to preserve RNA secondary structure, whereas low T_m regions are under high evolutionary pressure to preserve primary RNA sequence. We therefore hypothesize that high T_m regions may contain thermo-stable functionally important RNA structure elements (51–56) and thus experience relatively high evolutionary pressure to preserve the RNA structure and a relaxed evolutionary constraint on the nucleotide sequence, as long as the thermo-stable nucleotide base pairs which are crucial for the RNA structure remain intact. Considering the highly conserved nucleotide sequence and low thermostability of low T_m regions, we hypothesize that low T_m regions may contain functionally important RNA sequence elements, for example, binding sites which are conserved in sequence and require a good accessibility to interact with ligands. High and low thermostability is, respectively, indicative of functionally important RNA structures and sequence segments in mRNA coding regions. We therefore speculate that the melting temperature is a crucial parameter for the identification of functionally important RNA structure and sequence elements. We have been able to verify the association of high thermostability with functional importance for two types of RNA structure elements—zipcodes in the yeast mRNA coding regions and *URE2* IRES. The lack of experimentally determined and precisely characterized sequence motifs in the coding regions of yeast mRNAs prevented us from directly assessing the functional implications of low thermostability. While in previous research coding regions carrying RNA-level functions were associated with synonymous constraint elements and relaxed protein structure constraints (25), we find that synonymous constraint is only apparent in functionally important sequence regions (e.g. binding sites) and that functionally important RNA structures are not under synonymous constraint. This finding may prove useful in future investigations of functionally important elements in mRNA coding regions, as the overall attention to mRNAs structure grows. A typical example is constituted by RNA thermometers—temperature-sensitive

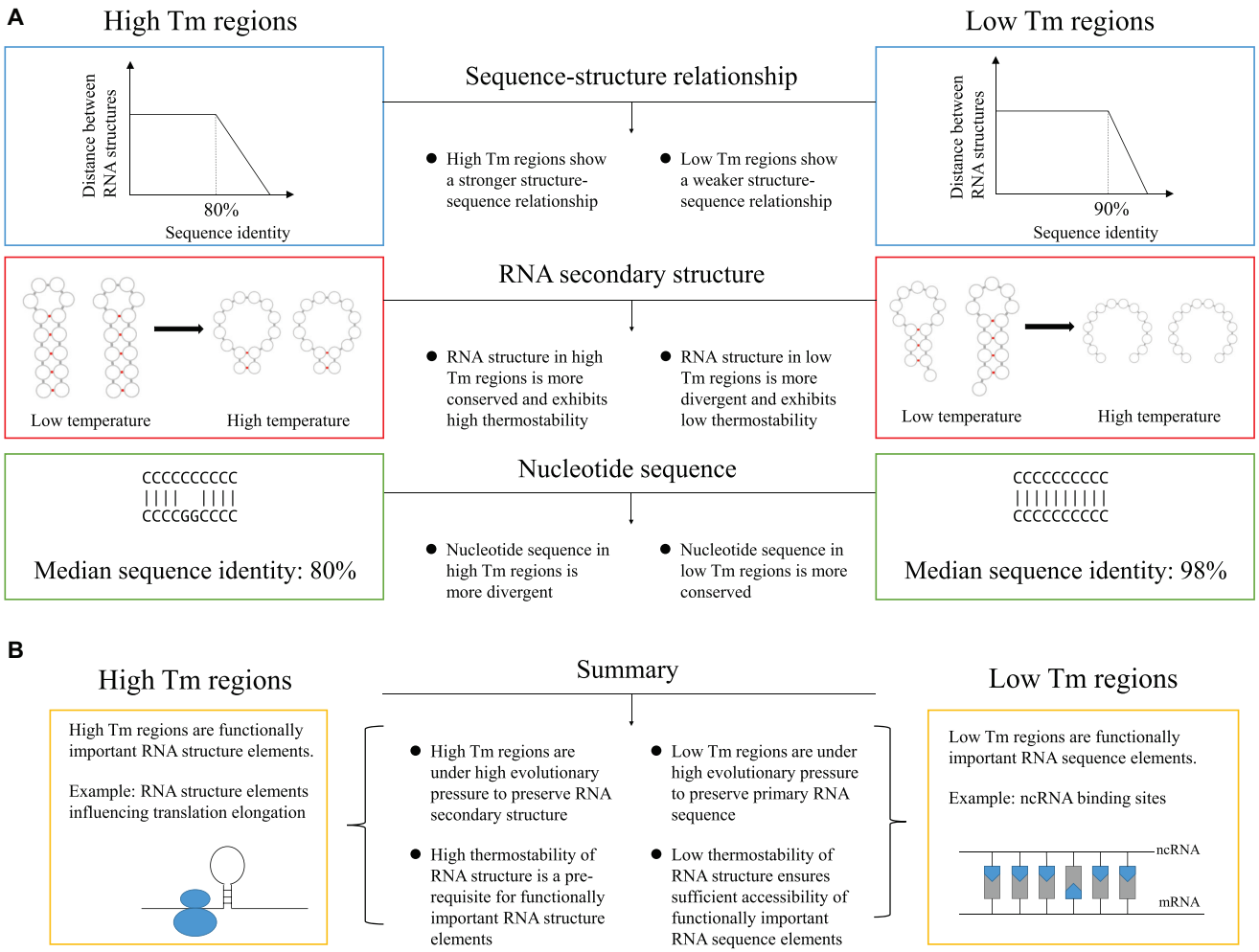

Figure 8. (**A**) Summary of the findings about high T_m and low T_m regions. (**B**) Inferences based on these findings.

RNA structural elements that are typically located in the 5′ UTR of mRNAs and form a secondary structure that traps the ribosome binding site and/or the translation initiation codon (57). In response to temperature changes an RNA thermometer undergoes a conformational transition, which impacts translation efficiency and eventually regulates gene expression (58). RNA thermometers have recently attracted growing attention (59) and efforts have been made to discover new elements of this type (40). Secondary structures of RNA thermometers are more conserved than their primary sequence (60), which is analogous to the high T_m regions described in this work. We therefore speculate that our findings may facilitate the search for new RNA thermometers in mRNA coding regions. The characteristics of high T_m regions, including strong sequence–structure relationships, conservation patterns of thermo-stable and meltable positions, and relaxed sequence constraint could serve as features to narrow down the search space in RNA thermometer discovery.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yue Wan and Howard Y. Chang for kindly providing data on melting temperatures of yeast mRNAs.

FUNDING

China Scholarship Council (No. 201206740006 to F.Q.). This work was supported by the German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.

Conflict of interest statement. None declared.

REFERENCES

1. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
2. Arava,Y., Wang,Y., Storey,J.D., Liu,C.L., Brown,P.O. and Herschlag,D. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 3889–3894.`

3. Wang,Y., Liu,C.L., Storey,J.D., Tibshirani,R.J., Herschlag,D. and Brown,P.O. (2002) Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 5860–5865.
4. Takizawa,P.A., DeRisi,J.L., Wilhelm,J.E. and Vale,R.D. (2000) Plasma membrane compartmentalization in yeast by messenger RNA transport and a septin diffusion barrier. *Science*, **290**, 341–344.
5. Shepard,K.A, Gerber,A.P., Jambhekar,A., Takizawa,P.A., Brown,P.O., Herschlag,D., DeRisi,J.L. and Vale,R.D. (2003) Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 11429–11434.
6. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
7. Wan,Y., Kertesz,M., Spitale,R.C., Segal,E. and Chang,H.Y. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, **12**, 641–655.
8. Garneau,N.L., Wilusz,J. and Wilusz,C.J. (2007) The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.*, **8**, 113–126.
9. Warf,M.B. and Berglund,J.A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.*, **35**, 169–178.
10. Martin,K.C. and Ephrussi,A. (2009) mRNA localization: gene expression in the spatial dimension. *Cell*, **136**, 719–730.
11. Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.
12. Breaker,R.R. (2012) Riboswitches and the RNA World. *Cold Spring Harb. Perspect. Biol.*, **4**, a003566.
13. Wan,Y., Qu,K., Ouyang,Z., Kertesz,M., Li,J., Tibshirani,R., Makino,D.L., Nutter,R.C., Segal,E. and Chang,H.Y. (2012) Genome-wide measurement of RNA folding energies. *Mol. Cell*, **48**, 169–181.
14. Shabalina,S.A. (2006) A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.*, **34**, 2428–2437.
15. Bevilacqua,P.C. and Blose,J.M. (2008) Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.*, **59**, 79–103.
16. Tucker,B.J. and Breaker,R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.
17. Mandal,M. and Breaker,R.R. (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.*, **11**, 29–35.
18. Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11–17.
19. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science*, **324**, 255–258.
20. Nackley,A.G., Shabalina,S.A., Tchivileva,I.E., Satterfield,K., Korchynskyi,O., Makarov,S.S., Maixner,W. and Diatchenko,L. (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, **314**, 1930–1933.
21. Carlini,D.B., Chen,Y. and Stephan,W. (2001) The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics*, **159**, 623–633.
22. Ilyinskii,P.O., Schmidt,T., Lukashev,D., Meriin,A.B., Thoidis,G., Frishman,D. and Shneider,A.M. (2009) Importance of mRNA secondary structural elements for the expression of influenza virus genes. *OMICS*, **13**, 421–430.
23. Duan,J., Wainwright,M.S., Comeron,J.M., Saitou,N., Sanders,A.R., Gelernter,J. and Gejman,P. V (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.*, **12**, 205–216.
24. Goz,E. and Tuller,T. (2015) Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes. *BMC Genomics*, **16**(Suppl. 10), S4.
25. Macossay-Castillo,M., Kosol,S., Tompa,P. and Pancsa,R. (2014) Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments. *PLoS Comput. Biol.*, **10**, e1003607.
26. Lin,M.F., Kheradpour,P., Washietl,S., Parker,B.J., Pedersen,J.S. and Kellis,M. (2011) Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res.*, **21**, 1916–1928.
27. Underwood,J.G., Uzilov,A. V, Katzman,S., Onodera,C.S., Mainzer,J.E., Mathews,D.H., Lowe,T.M., Salama,S.R. and Haussler,D. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.
28. Lucks,J.B., Mortimer,S.A., Trapnell,C., Luo,S., Aviran,S., Schroth,G.P., Pachter,L., Doudna,J.A. and Arkin,A.P. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11063–11068.
29. Mortimer,S.A., Kidwell,M.A. and Doudna,J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, **15**, 469–479.
30. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
31. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A, Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
32. Li,J., Witten,D.M., Johnstone,I.M. and Tibshirani,R. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538.
33. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
34. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
35. Chursov,A., Walter,M.C., Schmidt,T., Mironov,A., Shneider,A. and Frishman,D. (2012) Sequence-structure relationships in yeast mRNAs. *Nucleic Acids Res.*, **40**, 956–962.
36. Korber,B. (2000) HIV signature and sequence variation analysis. In: Rodrigo,AG and Learn,GH Jr (eds). *Computational Analysis of HIV Molecular Sequences*. Kluwer Academic Publishers, Dordrecht, pp. 55–72.
37. Jambhekar,A., McDermott,K., Sorber,K., Shepard,K.A, Vale,R.D., Takizawa,P.A and DeRisi,J.L. (2005) Unbiased selection of localization elements reveals cis-acting determinants of mRNA bud localization in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 18005–18010.
38. Mathews,D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**, 526–532.
39. Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.
40. Churkin,A., Avihoo,A., Shapira,M. and Barash,D. (2014) RNAthermsw: direct temperature simulations for predicting the location of RNA thermometers. *PLoS One*, **9**, e94340.
41. Chursov,A., Kopetzky,S.J., Bocharov,G., Frishman,D. and Shneider,A. (2013) RNAtips: analysis of temperature-induced changes of RNA secondary structure. *Nucleic Acids Res.*, **41**, W486–W4891.
42. Clote,P., Ferré,F., Kranakis,E. and Krizanc,D. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.
43. Higgs,P.G. (2000) RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.*, **33**, 199–253.
44. Ringnér,M. and Krogh,M. (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, e72.
45. Li,X., Quon,G., Lipshitz,H.D. and Morris,Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–1107.
46. Hiller,M., Pudimat,R., Busch,A. and Backofen,R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**, e117.
47. Steigele,S., Huber,W., Stocsits,C., Stadler,P.F. and Nieselt,K. (2007) Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol.*, **5**, 25.

48. Olivier,C., Poirier,G., Gendron,P., Boisgontier,A., Major,F. and Chartrand,P. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol.*, **25**, 4752–4766.
49. Reineke,L.C., Komar,A.A., Caprara,M.G. and Merrick,W.C. (2008) A small stem loop element directs internal initiation of the URE2 internal ribosome entry site in Saccharomyces cerevisiae. *J. Biol. Chem.*, **283**, 19011–19025.
50. Komar,A.A., Lesnik,T., Cullin,C., Merrick,W.C., Trachsel,H. and Altmann,M. (2003) Internal initiation drives the synthesis of Ure2 protein lacking the prion domain and affects [URE3] propagation in yeast cells. *EMBO J.*, **22**, 1199–1209.
51. Bentley,D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, **15**, 163–175.
52. Kanhere,A., Viiri,K., Araújo,C.C., Rasaiyaah,J., Bouwman,R.D., Whyte,W.A., Pereira,C.F., Brookes,E., Walker,K., Bell,G.W. *et al.* (2010) Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol. Cell*, **38**, 675–688.
53. Vargas,D.Y., Shah,K., Batish,M., Levandoski,M., Sinha,S., Marras,S.A.E, Schedl,P. and Tyagi,S. (2011) Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, **147**, 1054–1065.
54. Meyer,M., Plass,M., Pérez-Valle,J., Eyras,E. and Vilardell,J. (2011) Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol. Cell*, **43**, 1033–1039.
55. Zamft,B., Bintu,L., Ishibashi,T. and Bustamante,C. (2012) Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 8948–8953.
56. Eperon,L.P., Graham,I.R., Griffiths,A.D. and Eperon,I.C. (1988) Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell*, **54**, 393–401.
57. Righetti,F. and Narberhaus,F. (2014) How to find RNA thermometers. *Front. Cell. Infect. Microbiol.*, **4**, 132.
58. Narberhaus,F., Waldminghaus,T. and Chowdhury,S. (2006) RNA thermometers. *FEMS Microbiol. Rev.*, **30**, 3–16.
59. Kortmann,J. and Narberhaus,F. (2012) Bacterial RNA thermometers: molecular zippers and switches. *Nat. Rev. Microbiol.*, **10**, 255–265.
60. Waldminghaus,T., Gaubig,L.C. and Narberhaus,F. (2007) Genome-wide bioinformatic prediction and experimental evaluation of potential RNA thermometers. *Mol. Genet. Genomics*, **278**, 555–564.