LETTER

doi:10.1038/nature12756

In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features

Yiliang Ding[1,2,3]*, Yin Tang[1,3,4]*, Chun Kit Kwok[2,3]*, Yu Zhang[4,5], Philip C. Bevilacqua[2,3,6] & Sarah M. Assmann[1,3,4,6]

RNA structure has critical roles in processes ranging from ligand sensing to the regulation of translation, polyadenylation and splicing[1–4]. However, a lack of genome-wide *in vivo* RNA structural data has limited our understanding of how RNA structure regulates gene expression in living cells. Here we present a high-throughput, genome-wide *in vivo* RNA structure probing method, structure-seq, in which dimethyl sulphate methylation of unprotected adenines and cytosines is identified by next-generation sequencing. Application of this method to *Arabidopsis thaliana* seedlings yielded the first *in vivo* genome-wide RNA structure map at nucleotide resolution for any organism, with quantitative structural information across more than 10,000 transcripts. Our analysis reveals a three-nucleotide periodic repeat pattern in the structure of coding regions, as well as a less-structured region immediately upstream of the start codon, and shows that these features are strongly correlated with translation efficiency. We also find patterns of strong and weak secondary structure at sites of alternative polyadenylation, as well as strong secondary structure at 5′ splice sites that correlates with unspliced events. Notably, *in vivo* structures of messenger RNAs annotated for stress responses are poorly predicted *in silico*, whereas mRNA structures of genes related to cell function maintenance are well predicted. Global comparison of several structural features between these two categories shows that the mRNAs associated with stress responses tend to have more single-strandedness, longer maximal loop length and higher free energy per nucleotide, features that may allow these RNAs to undergo conformational changes in response to environmental conditions. Structure-seq allows the RNA structurome and its biological roles to be interrogated on a genome-wide scale and should be applicable to any organism.

Most existing RNA structure mapping has been performed *in vitro*[5–8]. Among RNA structure probing reagents, dimethyl sulphate (DMS) can penetrate cells and has been used to map structures of high-abundance RNAs *in vivo* in various organisms[9–12]. DMS methylates the base-pairing faces of A and C of RNA in loops, bulges, mismatches and joining regions. The base-pairing status of U and G nucleotides can be inferred from structural mapping of As and Cs, because constraining even some nucleotides substantially improves predictions of other regions[13]. However, a method for genome-wide study of RNA structure *in vivo* has been lacking. Here we combine DMS methylation with next-generation sequencing to establish structure-seq, an *in vivo* quantitative measurement of genome-wide RNA secondary structure at nucleotide resolution.

We optimized DMS treatment conditions for *Arabidopsis* etiolated seedlings (Extended Data Fig. 1a), and then generated two independent biological replicates of (+)DMS and (−)DMS libraries (Fig. 1). DMS-induced methylation sites were highly reproducible (Pearson correlation coefficient (PCC) of 0.91 for the two (+)DMS libraries (Extended Data Table 1a)). Nucleotide modification in the (+)DMS library was specific to As and Cs (Extended Data Fig. 1b). Notably, 98% of the combined 206 million sequence reads were mappable to the *Arabidopsis* genome; these reads include diverse classes of RNAs, with a predominance of mRNAs and ribosomal RNAs (Extended Data Fig. 1c and Extended Data Table 1b, c). The reverse transcriptase stops are evenly distributed along the transcripts, with no 3′ bias (Extended Data Fig. 1d). In particular, 10,781 transcripts had sufficient coverage at nucleotide resolution to obtain secondary-structure constraints (Extended

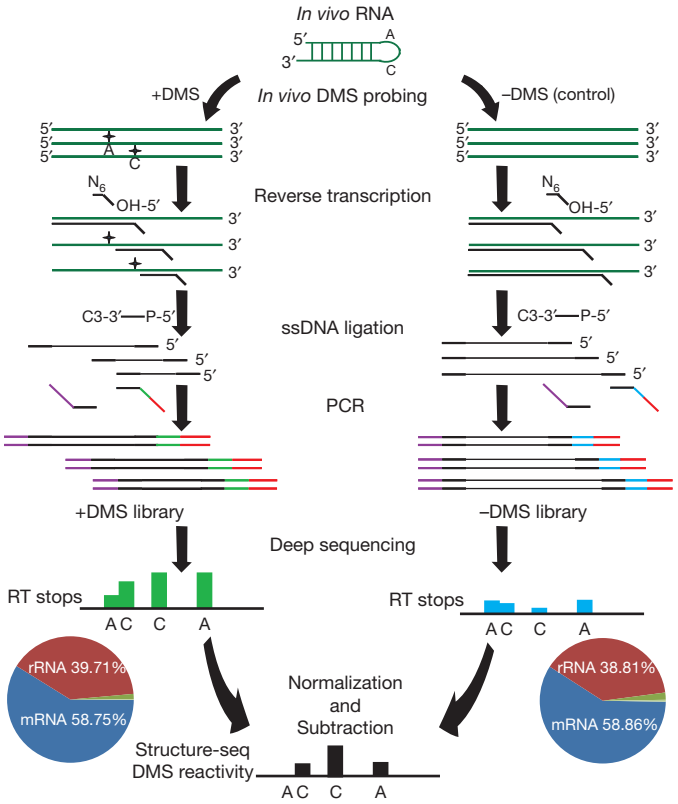


Figure 1 | Overview of structure-seq. *Arabidopsis* seedlings are treated with DMS. Reverse transcription is performed using random hexamers (N_6) with adaptors (thicker black lines). Reverse transcriptase stalls one nucleotide before DMS-modified As and Cs[11] (black crosses). Single-stranded (ss) DNA ligation attaches a single-stranded DNA linker (thicker black line) to the 3′ end. Double-stranded DNA is generated by PCR (purple line, forward primer; green–red line, unique index (green) and universal portion (red) of reverse primer). A (−)DMS library is prepared in parallel. Deep sequencing is performed with different indices for (+)DMS and (−)DMS libraries. Counts of the reverse transcriptase (RT) stops are normalized and subtracted. Pie charts depict percentages of RNA types for the (+)DMS (left) and (–)DMS (right) libraries. Green portions represent other RNA types plus unmappable reads (see Extended Data Table 1b, c).

[1]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [2]Department of Chemistry, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [3]Center for RNA Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [4]Bioinformatics and Genomics Graduate Program, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [5]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [6]Plant Biology Graduate Program, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
*These authors contributed equally to this work.

Data Fig. 2a). Abundance of individual mRNAs in structure-seq correlated well with mRNA abundance from RNA-seq analyses[14] (Extended Data Fig. 2b, c).

To validate *in vivo* structure-seq, we mapped DMS reactivities of 18S rRNA (Fig. 2a and Extended Data Fig. 3). Overall, the reactivities are consistent (Extended Data Fig. 3) with structure mapping of 30S subunit-bound 16S rRNA[15] and with the phylogenetically derived structures[16], which are the evolutionarily conserved structures and are the closest models of *in vivo*, protein-associated structure[17]. Further, comparison of DMS modifications from structure-seq with those from conventional gel-based *in vivo* structure probing yielded strong agreement for all regions of 18S rRNA tested (PCCs of 0.78 (Fig. 2b, c), 0.71 (Extended Data Fig. 4a, b) and 0.68 (Extended Data Fig. 4c, d)), as well as for a randomly chosen mRNA, *CAB1* (At1g29930) (Extended Data Fig. 4e, f, g). We thus conclude that structure-seq accurately probes RNA structures *in vivo* on a genome-wide basis. Importantly, complete coverage can be provided in a single experiment even for long transcripts, which is not the case for conventional gel-based methods.

We accordingly investigated global features and discovered several notable genome-wide *in vivo* RNA structural properties of *Arabidopsis* mRNAs (Fig. 3). We found that the average DMS reactivity of untranslated regions (UTRs) is significantly higher than that of coding sequences (CDS) (Extended Data Fig. 5a). The ~5 nucleotides (nt) immediately upstream of the start codon show particularly high DMS reactivity, which indicates less structure (Extended Data Fig. 5a). These findings agree with previous findings in yeast and *Arabidopsis* UTRs *in vitro*[5,6] and with *in silico* predictions in mouse and human[18]. Unstructured regions near start codons may facilitate ribosome binding and translation initiation. To evaluate this hypothesis, we ranked our mRNAs according to their polyribosome association on the basis of previous *in vivo* polyribosome profiling in *Arabidopsis* seedlings[19]. The unstructured region upstream of the start codon was enriched in high translation efficiency mRNAs and was absent in low translation efficiency mRNAs (Fig. 3a). Although a related observation was made *in vitro* for yeast[5], our data demonstrate that this is a genuine *in vivo* phenomenon, and extend these results to the plant kingdom.

When DMS reactivity along the CDS was averaged across mRNAs in our data set (see Methods for details), a periodic trend was revealed. A discrete Fourier transform applied to the CDS gave a period of 3, whereas periodicity was absent in UTR regions (Fig. 3a insets and

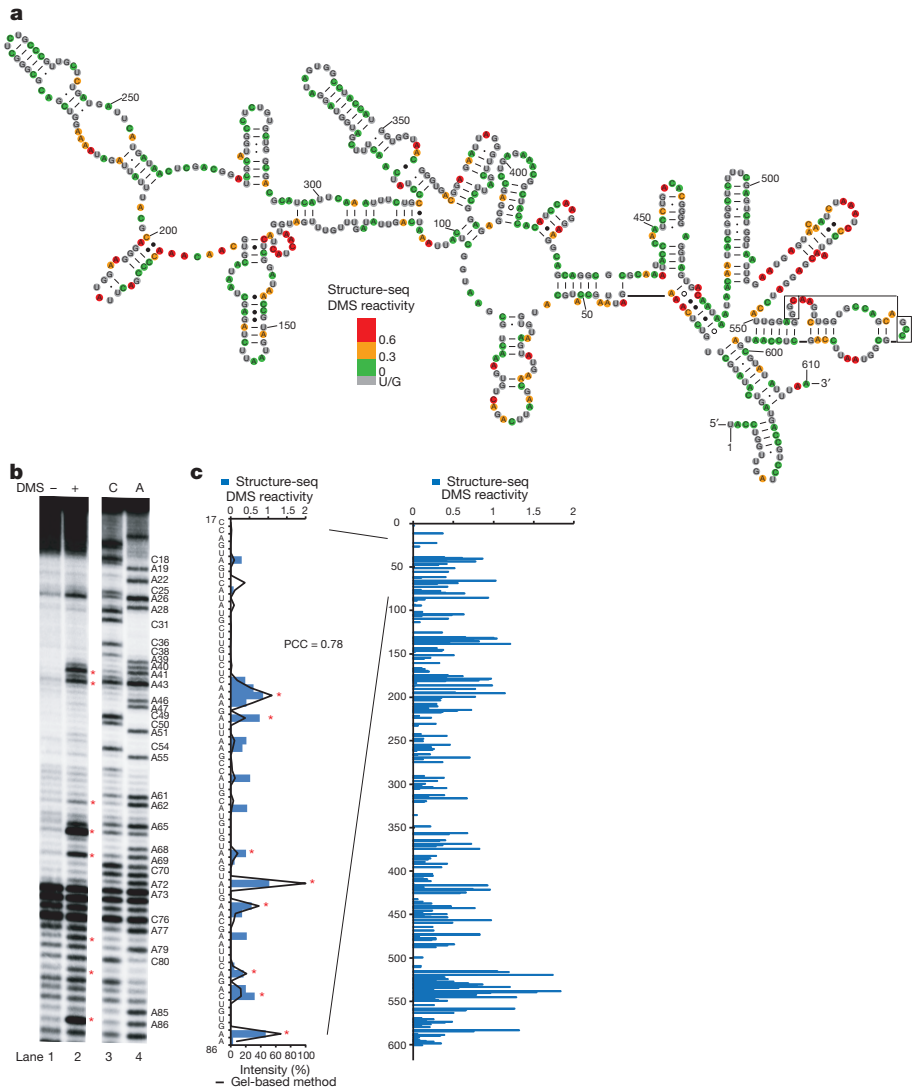


Figure 2 | Structure-seq accurately maps 18S rRNA and agrees with gel-based *in vivo* structure probing. **a**, Nucleotides 1–610 of the phylogenetic 18S rRNA structure[16], colour-coded according to structure-seq DMS reactivity. **b**, Nucleotides 17–86 of 18S rRNA structure-mapped by gel-based probing. Lanes 1–2, (−)DMS and (+)DMS treatments; lanes 3–4, C/A sequencing. **c**, Comparison of structure-seq (blue bars) and gel-based probing (black line, normalized to 0–100%) yields a PCC of 0.78. Structure-seq reactivity for nucleotides 1−610 is shown on the right. The red asterisks indicate nucleotides that have significant DMS modifications from both methods, and are also shown in panel **b**.

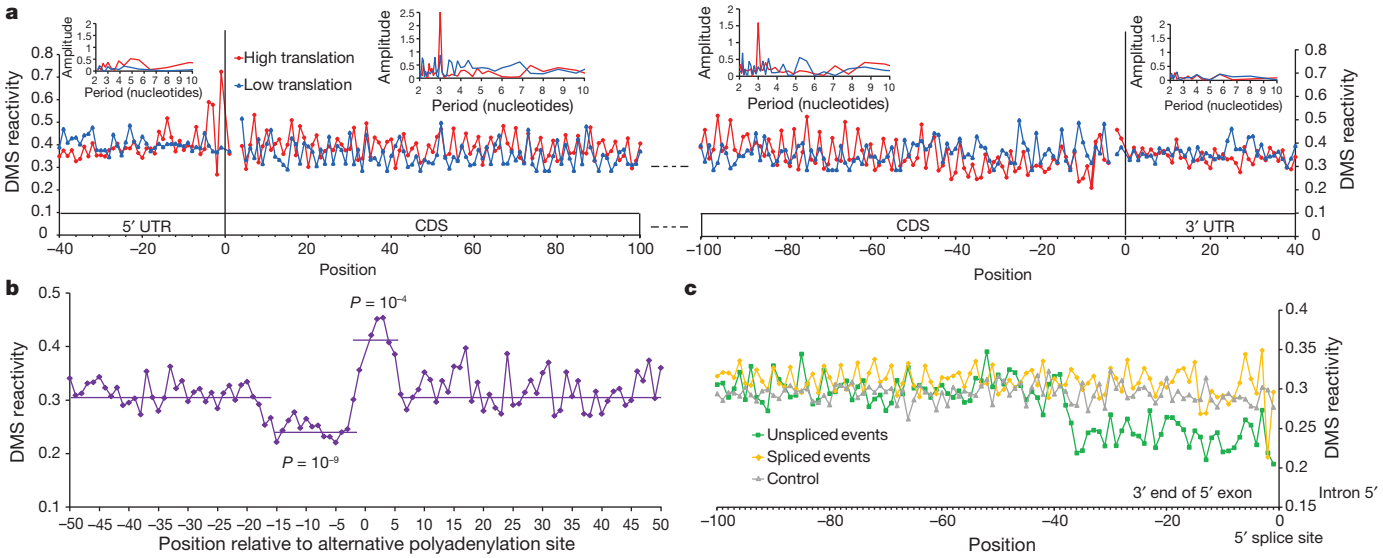


Figure 3 | Structure-seq reveals new features of mRNA secondary structures that prevail *in vivo*. **a**, RNA structure associated with translation. DMS reactivities of selected regions (5′ UTR, 40 nt upstream of the start codon; CDS, 100 nt downstream of the start codon and 100 nt upstream of the stop codon; 3′ UTR, 40 nt downstream of the stop codon) across high (red) or low (blue) translation efficiency mRNAs were averaged. mRNAs were aligned by their start/stop codons (vertical black lines). Discrete Fourier transforms (insets) of average DMS reactivity of selected regions across high (red) or low (blue) translation efficiency mRNAs shows structural periodicity only in high translation efficiency CDS. **b**, RNA structures associated with alternative polyadenylation. DMS reactivities 50 nt upstream and downstream of alternative polyadenylation sites (indicated by 0) were averaged (violet). One region with significantly lower DMS reactivity (−15 to −2 nt, $P = 10^{-9}$, Student's *t*-test) and one region with significantly higher DMS reactivity (−1 to 5 nt, $P = 10^{-4}$, Student's *t*-test) are highlighted. **c**, RNA structure associated with alternative splicing. DMS reactivities along 100 nt of the 3′ end of the 5′ exon were averaged from each of unspliced (green) and spliced (yellow) events. For unspliced events, the significance of the difference in average DMS reactivity between the 40 nt upstream of the 5′ splice site and the remaining 60-nt upstream region was $P = 10^{-25}$ (Student's *t*-test). For spliced events, the *P* value was > 0.05. Absence of structure in a nucleotide composition control is in grey.

Extended Data Fig. 5b, c). This represents the first *in vivo* demonstration of triplet periodicity in the structure of the CDS in a multicellular organism. Observation of an *in vivo* triplet periodicity in CDS structure in plants, as well as its presence in both *in vivo* (from ribosome profiling)[20] and *in vitro* (ribosome-free)[5] yeast data sets, and its proposed presence in mammals[18], suggests that periodic structure may have evolved as a universal regulatory feature of translated portions of mRNAs.

Our genome-wide *in vivo* structurome allowed us to evaluate the hypothesis that robustness of the periodic structure signal might influence translation. Notably, the periodic signal was intensified in high translation efficiency transcripts and absent from low translation efficiency transcripts[5] (Fig. 3a insets and Extended Data Fig. 5d). Further analysis revealed that differential presence of periodic structure between these two mRNA populations did not arise from differential codon usage or differential nucleotide bias in any of the three codon positions (Extended Data Fig. 5e). Our results thus reveal a hidden code in *in vivo* RNA structure that influences polyribosome association and, by inference, translation[21].

Alternative polyadenylation has been observed for ~60% of *Arabidopsis* mRNAs[22]. We assessed DMS modification 50 nt upstream and downstream of the known[22] alternative polyadenylation cleavage sites for the corresponding 5,959 mRNAs in our RNA structurome. For alternative polyadenylation, RNA secondary structure upstream of the cleavage site from nt −15 to −2 showed significantly lower DMS reactivity than the average reactivity throughout the 100-nt region, indicating more structure *in vivo* in the U- and A-rich upstream region (Fig. 3b and Extended Data Fig. 6a). This finding provides genome-wide support for a regulatory role of RNA structure in this region, in line with an early mutagenesis study of polyadenylation efficiency on one selected RNA assayed *in vitro*[23]. We also found that nt −1 to 5 had significantly higher DMS reactivity than average (Fig. 3b). This leads to a structured–unstructured pattern (Fig. 3b) that is not simply due to nucleotide composition (Extended Data Fig. 6b, c). These results, newly revealed by structure-seq, suggest that structural elements near the cleavage site may help to regulate alternative polyadenylation.

Alternative splicing has been proposed to be regulated by RNA secondary structure[24,25]. We considered a previous compilation of alternative splicing events in *Arabidopsis* seedlings[26] and identified, for each mRNA in our data set, whether introns were spliced out or whether alternative splicing (including exon skipping and intron retention) occurred. Notably, we found significantly lower DMS reactivity in the region ~40 nt upstream of the 5′ splice site for the unspliced events (Fig. 3c). This structural pattern was not found in the spliced events or in a nucleotide composition control (Fig. 3c), nor was it apparent at the 3′ splice site (Extended Data Fig. 6d). Secondary structure at the 5′ splice site appears to disfavour the first step of splicing, providing a regulatory mechanism for alternative splicing.

Current *in silico* structure prediction based on thermodynamics estimates a set of probable RNA structures, but constraints from experimental data significantly improve predictions[13,27]. Individual nucleotide DMS reactivities for each of the 10,623 mRNAs with ≥ 1 reverse transcriptase stop/nucleotide provided a rich data set (Fig. 4a) to compare RNA structure predictions with and without inclusion of *in vivo* DMS-guided structural constraints. First, we compared *in silico*-predicted structures and our *in vivo* structures with available *in vitro* structures[6]. We find that *in vitro* and *in vivo* structures differ, and that *in vitro* structures are more similar to *in silico* structures than are *in vivo* structures (Extended Data Fig. 7a). Next, using RNAstructure[27], we calculated for each of the 10,623 mRNAs the positive predictive value (PPV)[28], which indicates the proportion of base pairs in the *in vivo* DMS-constrained RNA structure that also appear in the *in silico*-predicted RNA structure. Most mRNAs did not fold *in vivo* according to *in silico*-predicted structures, as is evident from the broad PPV distribution (Fig. 4b). Such poor correlation could, in theory, be explained by mRNA association with proteins that block DMS reactivity *in vivo*. This hypothesis was not supported, however, as low

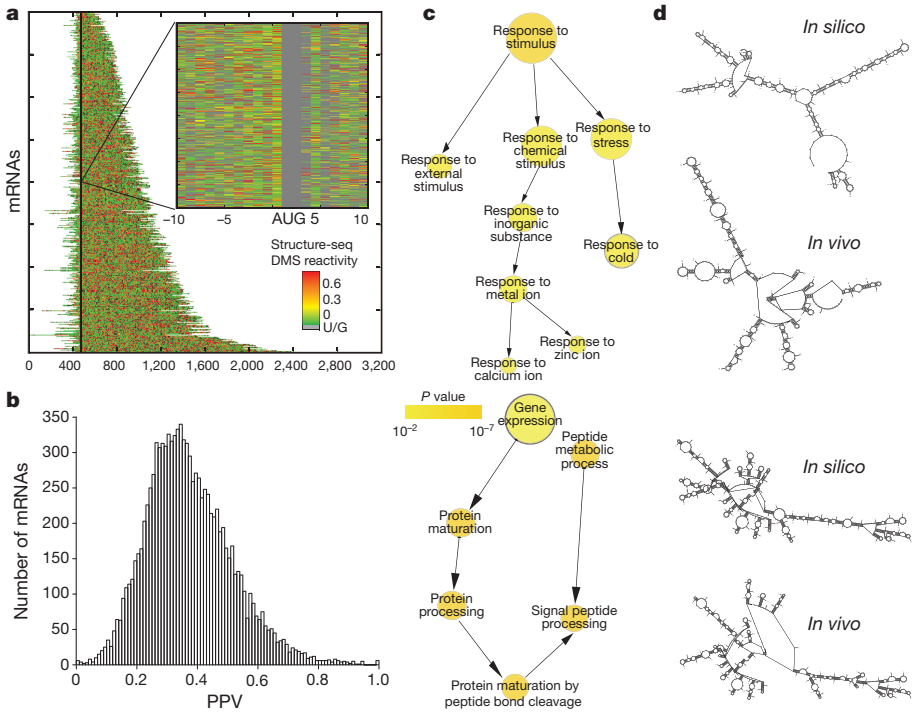


Figure 4 | Structure-seq provides *in vivo* RNA structure information at nucleotide resolution across 10,623 mRNAs and reveals correlations between RNA structure and biological function. **a**, DMS reactivity of each of 10,623 mRNAs. **b**, PPV distribution for *in vivo* versus *in silico* structures of 10,623 mRNAs; a higher PPV value indicates less difference. **c**, mRNAs with low PPV are enriched in functional annotations related to stress and stimulus responses; mRNAs with high PPV are enriched in basic biological functions. Gene Ontology categories over-represented in the 5% of 10,623 mRNAs with lowest and highest PPV are shown at the top and bottom, respectively. **d**, *In silico* and *in vivo* structures of one illustrative low PPV mRNA (top), *RCI2A* (At3g05880), are highly dissimilar, whereas such structures for one illustrative high PPV mRNA (bottom), *S24* peptidase (At1g52600), are highly similar.

reactivity did not correlate with low PPV, nor was PPV correlated with mRNA length (Extended Data Fig. 7b, c). The results of Fig. 4b and Extended Data Fig. 7a demonstrate the critical contribution of *in vivo* constraints in prediction of the RNA secondary structures that prevail in living cells. This is also illustrated by an improvement in predicting the phylogenetic structure of 18S rRNA when *in vivo* constraints are used (Extended Data Table 2).

We next asked whether genome-wide relationships exist between *in vivo* mRNA structures and biological functions of the encoded proteins. Intriguingly, the Gene Ontology annotations of those transcripts in the lowest 5% of the PPV distribution are enriched in annotations of biological functions related to stress and stimulus responses[29] (Fig. 4c and Extended Data Fig. 8a, b). For example, mRNAs of cold and metal ion stress-response genes folded significantly differently *in vivo* from their unconstrained *in silico* predictions (Fig. 4c, d and Extended Data Fig. 8a, b). Interestingly, these stresses are known to affect RNA structure and thermostability[27,30]. By contrast, genes involved in basic biological functions such as gene expression, protein maturation and processing, and peptide metabolic processes show little difference in their *in vivo*-constrained and *in silico*-predicted RNA secondary structures, as indicated by their enrichment in the highest 5% of the PPV distribution (Fig. 4c, d and Extended Data Fig. 8a, b). Speculatively, mRNAs related to cell maintenance and showing high PPV may have evolved to resist large conformational changes in order to maintain homeostasis.

Table 1 | RNA structural features differ between high and low PPV mRNAs

	Single-strand percentage	Maximum loop length of structure	Free energy per nucleotide
In silico	0.99	3.7×10^{-2}	7.73×10^{-3}
In vivo	5.80×10^{-19}	4.7×10^{-7}	3.07×10^{-34}

The significance of the difference for several RNA structural features was assessed between high PPV mRNAs and low PPV mRNAs. Each entry is the *P* value of a Student's *t*-test between the 5% of mRNAs with highest PPV and the 5% of mRNAs with lowest PPV. The comparisons were performed on *in silico*-predicted (without in *vivo* constraints) and *in vivo* (*in silico* prediction with constraints from our *in vivo* structure-seq data) structures. Small *P* values confirm that there are significant differences in RNA structural features between high- and low-PPV mRNAs.
(Pseudoknots are uncommon (~1 pseudoknot per 1,000 nt) in both high- and low-PPV mRNA data sets (calculated from the 1% mRNAs with highest PPV and the 1% mRNAs with lowest PPV). The *P* values for comparison of pseudoknot prevalence between these two groups are 0.48 and 0.31 for *in silico*-predicted and *in vivo* structures, respectively.)

We next compared several structural features between low and high PPV mRNAs. We found that the fraction of a mRNA's nucleotides with DMS reactivity greater than a 0.6 threshold is significantly higher in the low than in the high PPV mRNAs ($P = \sim 2 \times 10^{-42}$; two sample *t*-test), which provides experimental support independent of computational structure prediction that the low PPV mRNAs exist in multiple conformations and/or are less structured. The low PPV mRNAs, enriched in functions related to stress, also tend to have more single-stranded regions (consistent with higher average reactivity per nucleotide; $P = \sim 10^{-85}$; Student's *t*-test), longer maximum loop length and higher free energy per nucleotide when assessed *in vivo* (Table 1 and Extended Data Fig. 8b). These features might favour change in RNA structure in response to, for example, cold or metal ions, stress conditions with which these mRNAs are associated (Fig. 4c). In other words, stress-response RNAs may be more plastic, changing their structure in response to changing cellular conditions. As sessile organisms, plants face extreme environmental stresses; it will be of interest to ascertain whether the RNA structure–function relationships revealed in Fig. 4c, d prevail in other kingdoms.

In summary, we have established a high throughput, genome-wide method that profiles RNA secondary structure with high accuracy and nucleotide resolution *in vivo*. Our comprehensive study reveals new insights into how global native RNA structural characteristics regulate RNA processing and translation, and associates mRNA structural characteristics with functions of the encoded proteins. These trends are not discernible by studies on just one or a few RNAs, nor are they necessarily found in *in vitro* genome-wide studies. Structure-seq provides a broadly applicable method for the investigation of RNA structure–function relationships in living systems.

METHODS SUMMARY

Five-day-old *Arabidopsis thaliana* etiolated seedlings were treated with DMS, followed by dithiothreitol quench. Extracted poly(A)-selected RNA was reverse transcribed. First-strand complementary DNAs were ligated at their 3′ ends to a DNA linker and PCR was performed. Different barcode indices were used for the (+)DMS and (−)DMS libraries, which were subjected to Illumina sequencing. Two independent biological replicates were performed. Reads were mapped to the *Arabidopsis* transcriptome and genome using Bowtie (v.0.12.8). The natural log (ln) was taken of reverse transcriptase stops in both (+) and (−) DMS libraries, followed by normalization for abundance and length. Raw DMS reactivity was

calculated by subtracting from the normalized (+)DMS library values the normalized number of reverse transcriptase stops in the (−)DMS library, and further normalized (2–8% normalization) to obtain the final DMS reactivity of each nucleotide. PPVs were used to compare *in vivo*- and *in silico*-predicted structures for each mRNA. mRNAs with PPV values in the top and bottom 5% were subjected to Gene Ontology analysis using the hypergeometric test ($P < 0.01$ as significant).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 April; accepted 9 October 2013.
Published online 24 November 2013.

1. Buratti, E. *et al.* RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. *Mol. Cell. Biol.* **24,** 1387–1400 (2004).
2. Cruz, J. A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136,** 604–609 (2009).
3. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361,** 13–37 (2005).
4. Sharp, P. A. The centrality of RNA. *Cell* **136,** 577–580 (2009).
5. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
6. Li, F. *et al.* Regulatory impact of RNA secondary structure across the *Arabidopsis* transcriptome. *Plant Cell* **24,** 4346–4359 (2012).
7. Zheng, Q. *et al.* Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. *PLoS Genet.* **6,** e1001141 (2010).
8. Wan, Y. *et al.* Genome-wide measurement of RNA folding energies. *Mol. Cell* **48,** 169–181 (2012).
9. Senecoff, J. F. & Meagher, R. B. *In vivo* analysis of plant RNA structure: soybean 18S ribosomal and ribulose-1,5-bisphosphate carboxylase small subunit RNAs. *Plant Mol. Biol.* **18,** 219–234 (1992).
10. Wells, S. E., Hughes, J. M. X., Igel, A. H. & Ares, M. Use of dimethyl sulfate to probe RNA structure *in vivo*. *Methods Enzymol.* **318,** 479–493 (2000).
11. Zaug, A. J. & Cech, T. R. Analysis of the structure of Tetrahymena nuclear RNAs *in vivo*: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA* **1,** 363–374 (1995).
12. Zemora, G. & Waldsich, C. RNA folding in living cells. *RNA Biol.* **7,** 634–641 (2010).
13. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101,** 7287–7292 (2004).
14. Oh, E., Zhu, J. Y. & Wang, Z. Y. Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. *Nature Cell Biol.* **14,** 802–809 (2012).
15. Moazed, D., Stern, S. & Noller, H. F. Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension. *J. Mol. Biol.* **187,** 399–416 (1986).
16. Cannone, J. J. *et al.* The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3,** 2 (2002).
17. Gutell, R. R., Lee, J. C. & Cannone, J. J. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* **12,** 301–310 (2002).
18. Shabalina, S. A., Ogurtsov, A. Y. & Spiridonov, N. A. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* **34,** 2428–2437 (2006).
19. Branco-Price, C., Kawaguchi, R., Ferreira, R. B. & Bailey-Serres, J. Genome-wide analysis of transcript abundance and translation in *Arabidopsis* seedlings subjected to oxygen deprivation. *Ann. Bot. (Lond.)* **96,** 647–660 (2005).
20. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
21. Branco-Price, C., Kaiser, K. A., Jang, C. J., Larive, C. K. & Bailey-Serres, J. Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in *Arabidopsis thaliana*. *Plant J.* **56,** 743–755 (2008).
22. Shen, Y. *et al.* Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing. *Genome Res.* **21,** 1478–1486 (2011).
23. Loke, J. C. *et al.* Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures. *Plant Physiol.* **138,** 1457–1468 (2005).
24. Solnick, D. Alternative splicing caused by RNA secondary structure. *Cell* **43,** 667–676 (1985).
25. Jin, Y., Yang, Y. & Zhang, P. New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. *RNA Biol.* **8,** 450–457 (2011).
26. Filichkin, S. A. *et al.* Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20,** 45–58 (2010).
27. Lu, Z. J., Gloor, J. W. & Mathews, D. H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **15,** 1805–1813 (2009).
28. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl Acad. Sci. USA* **106,** 97–102 (2009).
29. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
30. Misra, V. K. & Draper, D. E. The linkage between magnesium binding and RNA folding. *J. Mol. Biol.* **317,** 507–521 (2002).

Acknowledgements This research is supported by Human Frontier Science Program (HFSP) grant RGP0002/2009-C, the Penn State Eberly College of Science, and a Penn State Huck Institutes HITS grant to P.C.B. and S.M.A. We thank F. Pugh, Y. Li, A. Chan and K. Yen for help with Illumina sequencing; D. Mathews and A. Spasic for advice on RNA structure analysis; M. Axtell for reading of the manuscript; and P. Raghavan for access to the CyberSTAR server, funded by the National Science Foundation through grant OCI–0821527. We also thank L. Song, D. Chadalavada and S. Ghosh for discussions.

Author Contributions Y.D. and C.K.K. performed the experiments. Y.D., Y.T. and C.K.K. performed data analysis. Statistical analyses were designed by Y.Z. and Y.T., with input from all authors. Y.D., Y.T. and C.K.K. contributed equally to this work. All authors contributed ideas, discussed the results and wrote the manuscript.

Author Information Sequencing data are deposited in the Sequence Read Archive (SRA) on the NCBI website under the accession number SRP027216. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M.A. (sma3@psu.edu) or P.C.B. (pcb5@psu.edu).

METHODS

Plant materials and growth conditions. *Arabidopsis thaliana* seeds of the Columbia (Col-0) accession were sterilized with 70% (v/v) ethanol and plated on half-strength Murashige and Skoog medium. The plates were wrapped in foil and stratified at 4 °C for 3–4 days and then grown in a 22–24 °C growth chamber for 5 days.

***In vivo* DMS chemical probing.** All manipulations involving DMS were conducted in a chemical fume hood.

Five-day-old *A. thaliana* etiolated seedlings grown as described above were suspended intact and completely covered in 20 ml 1 × DMS reaction buffer in a 50 ml Falcon tube that contained 100 mM KCl, 40 mM HEPES (pH 7.5) and 0.5 mM $MgCl_2$. DMS was added to a final concentration of 0.75% (~75 mM) and allowed to react for 15 min at room temperature (~22 °C) with periodic swirling. This DMS concentration and reaction time allowed DMS to penetrate plant cells and modify the RNA *in vivo* with single-hit kinetics conditions. Single-hit kinetics conditions can be directly observed in the (+)DMS lanes of Fig. 2b and Extended Data Figs 1a and 4e, in which an intense full-length peak is observed for both rRNA and mRNA, and is confirmed in structure-seq data by the presence of transcripts with no internal reverse transcriptase stops. To quench the reaction, freshly prepared dithiothreitol (DTT) was added to a final concentration of 0.5 M, and after swirling for 2 min the reaction mixture was decanted and the seedlings were washed with ~2 × 50 ml deionized water. The seedlings were immediately frozen with liquid N_2 and ground into powder using a mortar and pestle pre-cleaned with RNase Zap (Ambion). Lysis buffer was added to the powder, and then the sample was subjected to total RNA extraction, following the protocol described in the RNeasy Plant Mini Kit (Qiagen).

Illumina library construction. *In vivo* total RNA isolation was followed by one round of poly(A) selection using the Poly(A) purist Kit (Ambion). The poly(A)-selected RNA (2 μg) was then treated with TURBO DNase (Ambion) following the manufacturer's protocol, followed by phenol chloroform extraction and ethanol precipitation. The RNA was re-suspended in RNase-free water and subjected to reverse transcription using the SuperScript III First Strand Kit (Invitrogen) and random hexamers fused with an Illumina TruSeq Adapter (5′-CAGACGTGT GCTCTTCCGATCNNNNNN-3′). The resultant first-strand cDNAs were then ligated at their 3′ ends to a ssDNA linker (5′-pNNNAGATCGGAAGAGCGTC GTGTAG-3′-Spacer, where '5′ p' is a 5′ phosphate and '3′-Spacer' is a 3-carbon linker) using CircLigase ssDNA Ligase (Epicentre), with slight modifications to the manufacturer's and literature procedures[31], as follows. In brief, the cDNA was re-dissolved in RNase-free water and reagents were added to yield the following final concentrations in a total volume of 20 μl: 70 μM ssDNA linker, 50 mM MOPS (pH 7.5), 10 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 0.05 mM ATP, 2.5 mM $MnCl_2$ and 200 U total CircLigase. The ligation was performed at 65 °C for 12 h and then the sample was heated at 85 °C for 15 min to deactivate the CircLigase. PCR amplification was performed on the ligated cDNA using Illumina TruSeq Primers (Illumina TruSeq forward primer, 5′-AATGATACGGCGACCACCGA GATCTACACTCTTTCCCTACACGACGCTCTTCCGATC T-3′; Illumina TruSeq reverse primer index 1, 5′-CAAGCAGAAGACGGCATACGAGATTGGTCAGT GACTGGAGTTCAGACGTGTGCTCTTCCGATC-3′; Illumina TruSeq reverse primer index 2, 5′-CAAGCAGAAGACGGCATACGAGATGATCTGGTGACT GGAGTTCAGACGTGTGCTCTTCCGATC-3′). Three rounds of gel purification were performed to remove adapters and achieve a uniform size distribution of PCR products between 150 and 650 base pairs (bp) using both a 50-bp DNA Ladder and a 1 Kb Plus DNA Ladder (Invitrogen) as references. This, together with carefully measured loading DNA concentration, allowed an optimized cluster density to reduce unmappable reads (c.f. the manufacturer's protocol (Illumina)). Different barcode indices were used for the (+)DMS library and (−)DMS libraries. The dsDNA libraries were subjected to next-generation sequencing on an Illumina HiSeq 2000. An independent biological replicate was prepared in the same way and separately subjected to next-generation sequencing.

Illumina sequence mapping. Illumina sequencing read lengths of 37 nt were obtained and mapped to the *Arabidopsis thaliana* transcriptome and genome (TAIR v10 release 2010). Twenty-one nucleotides was determined to be the threshold length required for unique mapping of a sequencing read after the reads were linker trimmed at their 3′ ends. Up to three mismatches without any insertions or deletions were allowed to account for PCR and sequencing errors. Reads that could not be mapped or uniquely mapped to the genome were designated as 'not mappable'. Mapping of the reads was performed using Bowtie[32] (v0.12.8) (http://bowtie-bio.sourceforge.net/index.shtml).

As shown in Extended Data Table 1a, there is high correlation between the two (+)DMS libraries and between the two (−)DMS libraries from the biological replicates. Therefore, biological replicates were combined for further analysis.

Determination and normalization of DMS reactivity. To compare the (+)DMS and (−)DMS data sets and derive the final DMS reactivity for each nucleotide, the following three-step procedure was used:

Step 1. For a transcript, suppose $P_r(i)$ and $M_r(i)$ are the raw numbers of reverse transcriptase stops for nucleotide i (including all four bases) on the transcript in the (+)DMS and (−)DMS libraries (P and M, respectively), l is the length of the transcript and $P_r(0)$ and $M_r(0)$ are the raw numbers of full-length reverse transcriptase reads on the transcript in the (+)DMS and (−)DMS libraries, respectively.

For each nucleotide on each transcript, take the natural log (ln) of the number of reverse transcriptase stops mapped to that nucleotide position ($\ln[P_r(i)]$ or $\ln[M_r(i)]$) and divide the number by the average of the ln of reverse transcriptase stops per position, yielding equations (1) and (2). The average of the ln of reverse transcriptase stops per position is calculated as the sum of the ln of reverse transcriptase stops at each position (including all four bases (as random reverse transcriptase stalling can occur at any base) and full length reverse transcriptase reads) of the entire transcript, divided by the length of the transcript, as provided in the denominators of equations (1) and (2).

$$P(i) = \frac{\ln[P_r(i)]}{(\sum_{i=0}^{l} \ln[P_r(i)])/l} \quad (1)$$

Equation (1) is the normalized number of reverse transcriptase stops for nucleotide i in the (+)DMS library.

$$M(i) = \frac{\ln[M_r(i)]}{(\sum_{i=0}^{l} \ln[M_r(i)])/l} \quad (2)$$

Equation (2) is the normalized number of reverse transcriptase stops for nucleotide i in (−)DMS library.

Step 2. For each nucleotide, the raw DMS reactivity is calculated by subtracting the normalized number of reverse transcriptase stops for the nucleotide between (+) and (−)DMS libraries. All negative values are taken as 0 for the raw DMS reactivity.

$$\theta(i) = \max((P(i) - M(i)), 0) \quad (3)$$

Equation (3) gives the raw DMS reactivity for nucleotide i.

Step 3. Normalization (2–8% normalization[28]) is then performed on the raw DMS reactivity, $\theta(i)$, of all the nucleotides on all the transcripts to obtain the final DMS reactivity of each nucleotide. The reactivity is capped at seven[5].

In all of the figures in which the average DMS reactivity of a region is given, it is the average of the DMS reactivity of all adenine and cytosine nucleotides in that region, for all of the transcripts under consideration. Those transcripts that have no reverse transcriptase stops for any of the nucleotides are not used in further structure analyses, as they provide no structure information.

***In vivo* RNA structure analysis of the genome-wide transcriptome using DMS reactivity.** Global *in vivo* mRNA structure trends. We determined global transcriptome trends in mRNA structure by averaging DMS reactivity from selected regions of mRNAs: the 5′ UTR region (the first 40 nt upstream of the start codon); the CDS-beginning region (the first 100 nt downstream of the start codon); the CDS-ending region (the 100 nt upstream of the stop codon); and the 3′ UTR region (the first 40 nt downstream of the stop codon). There were 22,721 unique mRNAs (including splice variants) that had at least 40 nt in both the 5′ UTR region and the 3′ UTR region and at least 200 nt in the CDS; these mRNAs were analysed for global trends (Extended Data Fig. 5a).

We analysed the global mRNA structure of polyribosome-associated mRNAs defined in a previous study[21], ranking the transcripts according to their polyribosome-associated mRNA abundance relative to their mRNA abundance. We selected the top 5% (1,136 mRNAs) and the bottom 5% (1,136 mRNAs) of mRNAs from the ranking. We defined the top 5% as the 'high translation efficiency mRNAs' and the bottom 5% as the 'low translation efficiency mRNAs'. We analysed the global transcriptome trends of DMS reactivity of the 5′ UTR, CDS and 3′ UTR for both the high translation efficiency mRNAs and the low translation efficiency mRNAs (Fig. 3a).

Codon periodicity and codon position signature. We assessed the codon periodicity by applying a discrete Fourier transform. We collected the DMS reactivity data from the Fourier-transformed patterns of the 40-nt 5′ UTR, the first 100 nt of the CDS, the last 100 nt of the CDS and the 40-nt 3′ UTR regions (Fig. 3a and Extended Data Fig. 5b). We also computed the average DMS reactivity for each codon position, collected from the entire CDS across 22,721 unique mRNAs (see above for explanation of mRNAs chosen). We applied the Student's *t*-test to assess

the significance of the difference between the average DMS reactivity for different codon positions ($P < 0.01$ as significant) (Extended Data Fig. 5c). The same methodology was applied to the high and low translation efficiency mRNA subsets (Fig. 3a and Extended Data Fig. 5d).

Alternative polyadenylation structural patterns. Alternative polyadenylation sites were defined on the basis of a previous genome-wide study of alternative polyadenylation in *A. thaliana*[22]. First we computed and plotted the nucleotide occurrence 50 nt upstream and 50 nt downstream of the alternative polyadenylation cleavage site for all alternatively polyadenylated mRNAs represented in our RNA structurome (Extended Data Fig. 6a). There were 5,959 mRNAs in our data set with alternative polyadenylation cleavage sites. Then we mapped the average DMS reactivity of these upstream and downstream regions. We applied the Student's *t*-test to analyse the significance of the difference in the average DMS reactivity between the structured region (−15 nt to −2 nt uracil- and adenine- rich region upstream of the alternative polyadenylation cleavage sites) and the average DMS reactivity of the whole 100 nt (Fig. 3b). We also did the same analysis for the significance of the difference in the average DMS reactivity between the unstructured region (−1 nt to 5 nt adenine-rich region of the alternative polyadenylation cleavage sites) and the average DMS reactivity of the whole 100 nt (Fig. 3b).

Structure across alternative splice sites. On the basis of a previous study of genome-wide alternative splicing in *Arabidopsis* seedlings[26], we identified, for each mRNA in our data set, whether all introns were spliced out or whether alternative splicing (including exon skipping and intron retention) occurred. This yielded a data set of 15,441 mRNAs with alternative splicing events. We then examined average DMS reactivity of the 100 nt at the 3′ end of the 5′ exon and compared this parameter in unspliced versus spliced events (Fig. 3c). For the unspliced events, we applied the Student's *t*-test to analyse the significance of the difference in the average DMS reactivity between the 40 nt upstream of the 5′ splice site and the remaining 60 nt of the 100-nt region upstream of the 5′ splice site. The same analysis was performed for the spliced events. As a nucleotide composition control for the unspliced events, the identical nucleotide composition of the 40 nt upstream of the 5′ splice site in the unspliced events was shuffled and remapped to find regions on the mRNAs in the TAIR *Arabidopsis* cDNA library that were not located at 5′ splice site junctions. For all of the resulting regions that were also present in our data set, the average DMS reactivity for each nucleotide along the 40-nt regions plus the additional 60 nt upstream of these regions was collected as a total 100-nt control, and the resulting average DMS reactivity was compared to that of the unspliced events (Fig. 3c). The above set of analyses was also applied to the 100-nt regions of the 3′ splice site except that the nucleotide composition control was performed with a 100-nt shuffle (Extended Data Fig. 6d).

All global structure trends in mRNA regions and periodicity, alternative polyadenylation and alternative splicing that we describe (Fig. 3 and Extended Data Figs 5 and 6) remained significant when global analyses were redone on the smaller, 10,623 mRNA subset with ≥ 1 average reverse transcriptase stop per (A+C) nucleotide.

Comparison between *in vivo* constrained RNA structures and *in silico* predicted RNA structures. *In vivo* DMS-constrained RNA structuromes were graphed with nucleotide resolution. A total of 10,623 mRNAs with ≥ 1 average reverse transcriptase stop per (A+C) nucleotide were analysed (see below). We used the criterion of ≥ 1 average reverse transcriptase stop per (A+C) nucleotide because PPV and Gene Ontology analyses rely on nucleotide resolution throughout the entire mRNA. All 10,623 mRNAs (including all splice variants) were aligned by their start codon. Colour scales were applied to indicate the DMS reactivity. Each row in Fig. 4a represents the DMS-guided RNA structurome information of one mRNA. mRNAs were organized by transcript length. The figure was constructed using Python matplotlib module (http://matplotlib.org/).

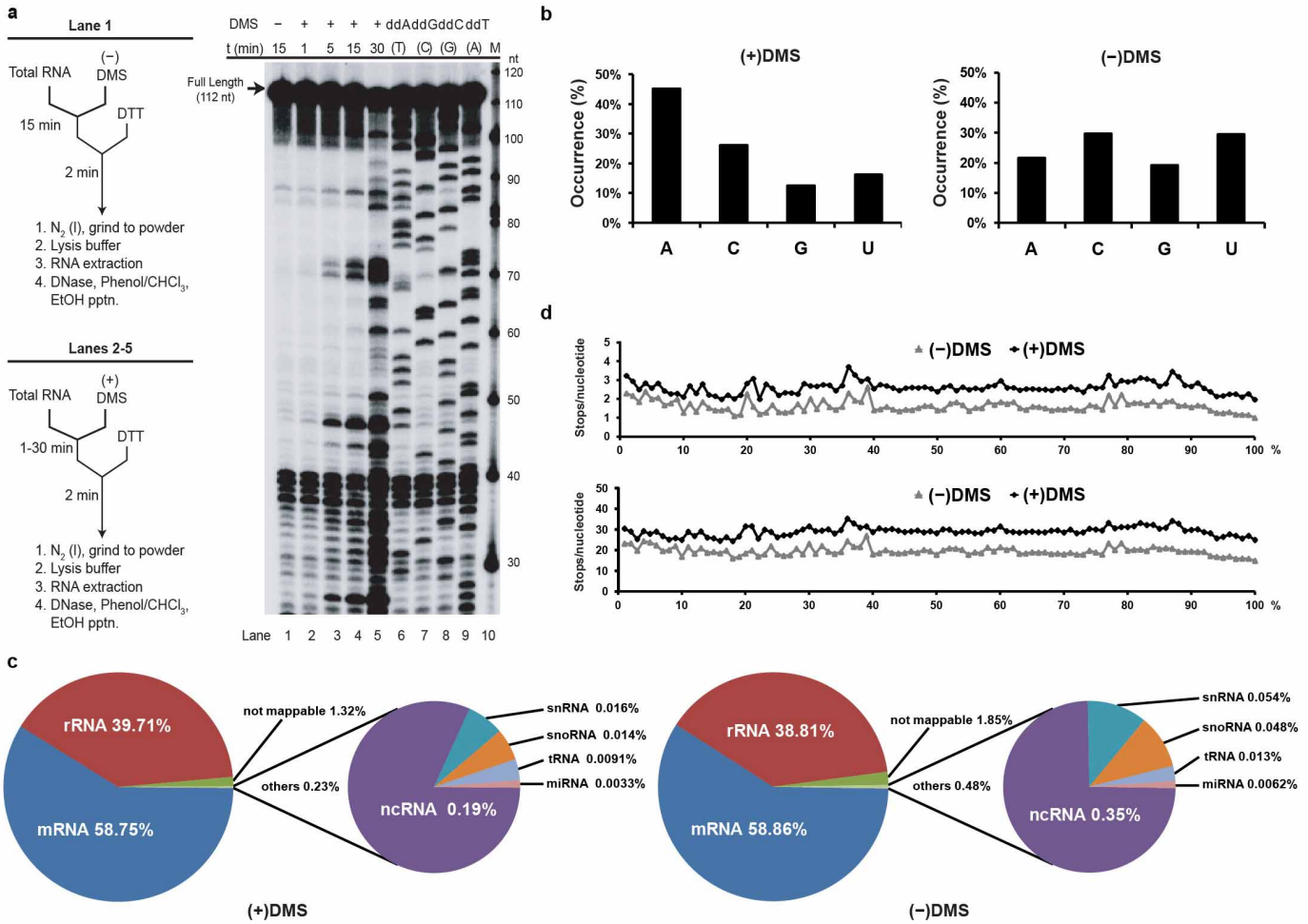
To obtain predicted RNA structures, we folded each of the 10,623 *A. thaliana* mRNAs with ≥ 1 average reverse transcriptase stop per (A+C) nucleotide using the program RNAstructure[27] (http://rna.urmc.rochester.edu/RNAstructure.html) with slope (1.8) and intercept (−0.6) for the pseudo-free energy function and either with or without our *in vivo* DMS constraints. (After testing on several protein-free regions of 18S rRNA, we concluded that for the pseudo-free energy function used by RNAstructure[27] the intercept and slope as defined in ref. 33 were adequate.) We compared *in vivo* DMS-constrained RNA structure with *in silico*-predicted RNA structure (that is, without constraints) for each mRNA by examining the PPV and sensitivity of base pairs[28]. Simply, when comparing two structures, PPV implies the proportion of base pairs in the *in vivo* DMS-constrained RNA structure that also appear in the *in silico*-predicted RNA structure[28]. The sensitivity indicates the proportion of base pair coverage *in silico* that also appears *in vivo*[28]. These criteria indicate the extent of divergence of *in vivo* constrained and *in silico* structures[28]. In our data, the PPV and sensitivity for the mRNA population are highly correlated (PCC = 0.99), thus we use PPV to represent the difference between the *in vivo* and *in silico* structures (Fig. 4b). Negative predictive value (NPV)[34] implies the proportion of single-stranded nucleotides common to both structures. The PPV and NPV are also highly correlated in our data set (PCC = 0.90), and so PPV was used for subsequent analyses. We plotted the PPV values for each transcript across the 10,623 mRNAs (Fig. 4b). We then took the mRNAs with PPV values in the top 5% and those with PPV values in the bottom 5% and performed Gene Ontology annotation analysis[29] for these two groups using the hypergeometric test ($P < 0.01$ as significant) (Fig. 4c). For Gene Ontology analysis of mRNAs with splice variants, we defined the PPV value as the average of the PPV values of all the splice variants of that mRNA present in our data set. Structure prediction with inclusion of pseudoknot prediction was performed for the top 1% and bottom 1% of mRNAs in the PPV distribution using RNAstructure (ShapeKnots command)[33,35].

Comparison of high and low PPV mRNAs. To better understand the underlying mechanisms causing the variation of PPV among mRNAs, we selected the mRNAs in the top 5% and bottom 5% of the PPV distribution and performed two sample *t*-tests to assess whether there was significant difference between the two groups for several RNA structural features for both *in silico* structures and *in vivo* structures: single-strand percentage, maximum loop length and free energy per nucleotide within an mRNA, and DMS reactivity per nucleotide. We similarly compared the prevalence of pseudoknots (pseudoknots per 100 nt of structure) in the top and bottom 1% of the PPV distribution.

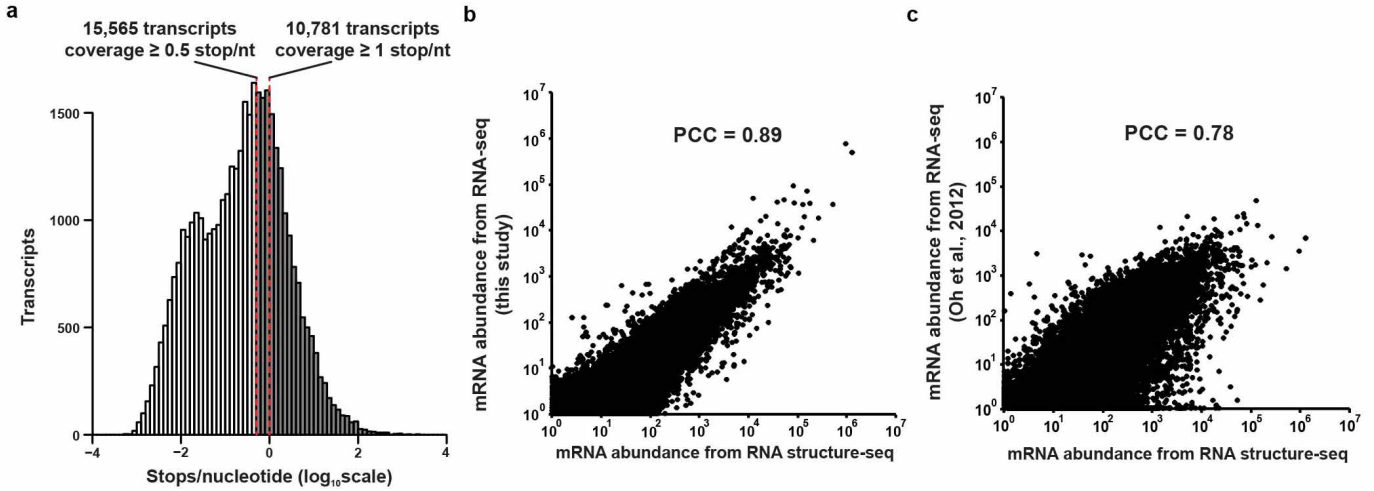
Gel-based method data collection and quantification. The gel-based method of structure probing used the same *in vivo* total RNA pools from the same (+)DMS and (−)DMS plant material as for high-throughput RNA structure-seq. To accomplish gel-based structure probing, reverse transcription was performed using gene-specific ^{32}P-radiolabelled DNA primers (18S reverse primer for region 1 for gel-based method, 5′- AACTGATTTAATGAGCCATTCGCAG-3′; 18S reverse primer for region 2 for gel-based method, 5′-GAGCCCGCGTCGACCTTTTATC-3′; 18S reverse primer for region 3 for gel-based method, 5′-GGTAATTTGCGCG CCTGCT-3′; *CAB1* mRNA (At1g29930) reverse outer primer for gel-based method, 5′-TTCCAAGGACTTCAGATGCC-3′; *CAB1* mRNA (At1g29930) reverse inner primer for gel-based method, 5′-GGAAAGCTTGACGGCCTTAC-3′; ssDNA adaptor for gel-based method, 5′-pNNNCTGCTGATCACCGACTGCCCATAG AG-3′−Spacer; adaptor forward primer for gel-based method, 5′-CTCTATGGG CAGTCGGTGAT-3′). The cDNA samples were then size fractionated on 8.3 M urea 8% polyacrylamide gels for DNA size separation. The power was maintained at 90–100 W throughout the 1.5–2 h run, and the surface temperature was ~55–65 °C, which helps to ensure denaturation of the DNA. Each gel was dried and exposed using a PhosphorImager (Molecular Dynamics) cassette.

Gel images were collected with a Typhoon PhosphorImager 9410, and bands were quantified using ImageQuant 5.2. The differences in band intensity between (+)DMS and (−)DMS samples were calculated. The most intense peak was normalized as 100% intensity[5]. As DMS specifically targets the Watson–Crick position of A and C nucleotides, the G and U nucleotides were not included during signal processing.

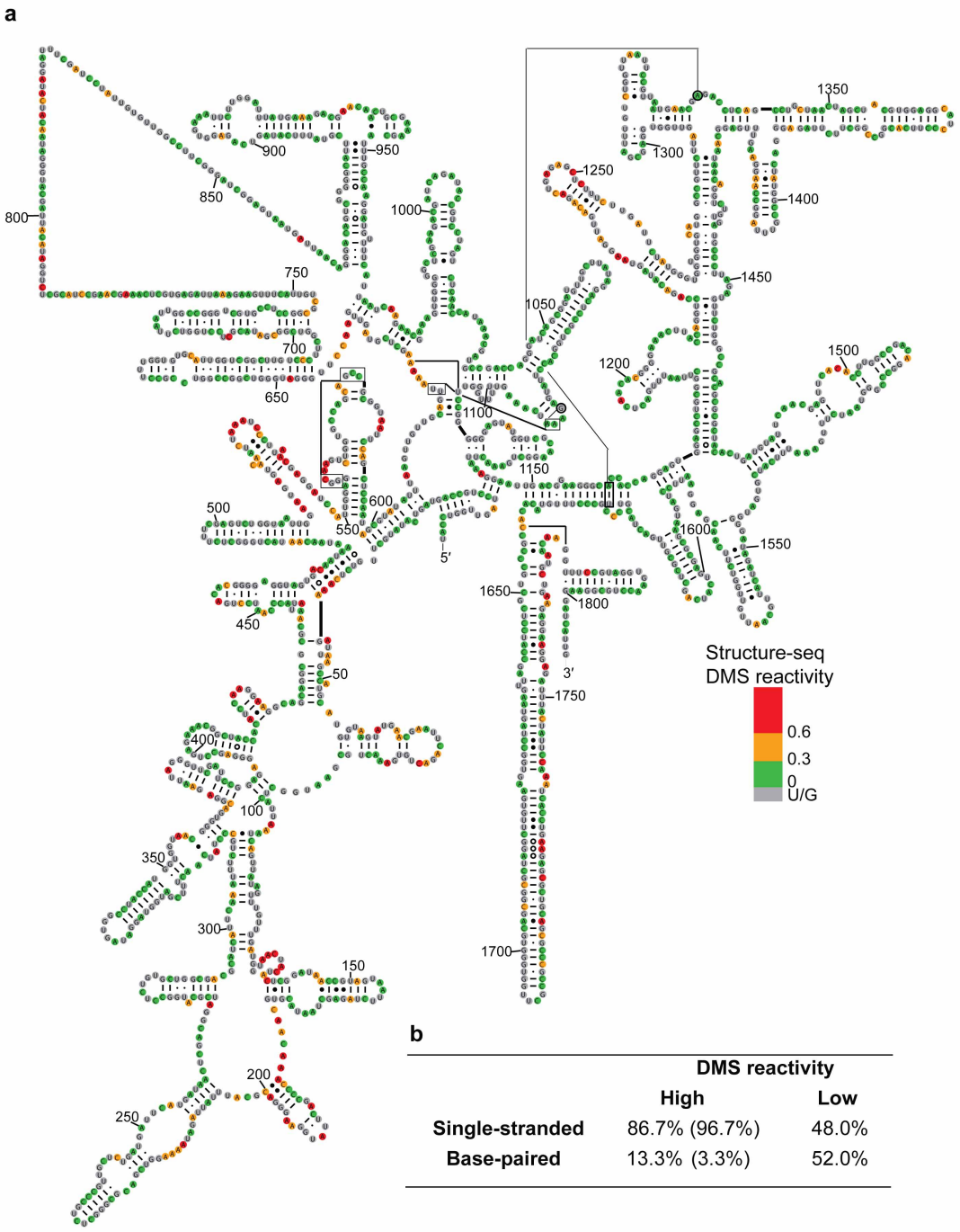
31. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl Acad. Sci. USA* **108,** 11063–11068 (2011).
32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
33. Hajdin, C. E. *et al.* Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl Acad. Sci. USA* **110,** 5498–5503 (2013).
34. Smith, C. J. Diagnostic tests (2) – positive and negative predictive values. *Phlebology* **27,** 305–306 (2012).
35. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11,** 129 (2010).
36. Lawley, P. D. & Brookes, P. Further studies on the alkylation of nucleic acids and their constituent nucleotides. *Biochem. J.* **89,** 127–138 (1963).
37. Weeks, K. M. & Crothers, D. M. RNA recognition by Tat-derived peptides: interaction in the major groove? *Cell* **66,** 577–588 (1991).



Extended Data Figure 1 | Time course of DMS modification and overview of structure-seq libraries. **a**, Time course of *in vivo* DMS modification of 18S rRNA in *Arabidopsis* etiolated seedlings. Five-day-old *Arabidopsis* etiolated seedlings were DMS-treated for different durations (1 min, 5 min, 15 min and 30 min; lanes 2–5, respectively). In all cases the final DMS concentration was 0.75% (~75 mM). The 18S rRNA DMS modification read-out was assessed by gel-based probing, which was done here near the 5′ end to afford a view of the full-length RNA band. The 15-min time point is the optimal duration for DMS modification, as it is the longest time point for which single-hit kinetics still occur as revealed by the intense full-length band. The 30-min time point is too long, as revealed by significant loss of the full-length band and increase of shorter length bands. Lanes 6–9 show the dideoxy sequencing of 18S rRNA. Lane 1 is the (−)DMS control. Lane 10 is a DNA marker (M) that was size fractionated to confirm the size of the full-length band (112 nt). **b**, DMS modification is RNA nucleotide specific. Nucleotide occurrence of RNA bases one nucleotide upstream of the position of reverse transcriptase stalling on the (+)DMS library and (−)DMS library, respectively. The (+)DMS library shows higher occurrence of A and C than of U and G (A is more than 1 standard deviation higher compared to C, G and U, and C is more than 1 standard deviation higher compared to G and U if leaving out A), consistent with the properties of DMS modification of nucleobases[36]. The percentages of each RNA base in the (−)DMS library are also indicated and are found to be similar (within 1 standard deviation). This figure combines results from both biological replicates. **c**, The total number of reads was classified into different classes of RNAs on a percentage basis from a total number of 121,258,873 reads for the (+)DMS library and 85,371,519 reads for the (−)DMS library. This figure combines results from both biological replicates. **d**, Structure-seq reads coverage. RNA structure information from structure-seq is distributed evenly across transcripts, with no 3′ bias. Each of the 37,558 transcripts (all transcripts with ≥ 1 internal reverse transcriptase stop and length ≥ 100 nt) was divided into 100 bins to normalize the transcript length. The reverse transcriptase stops per each A and C nucleotide (top) and the reverse transcriptase stops per each A and C nucleotide with ≥ 1 reverse transcriptase stop (bottom) from both the (+)DMS library (black diamonds) and the (−)DMS library (grey triangles) were averaged within each bin and plotted. The reverse transcriptase stops are well distributed over the entire transcript length.

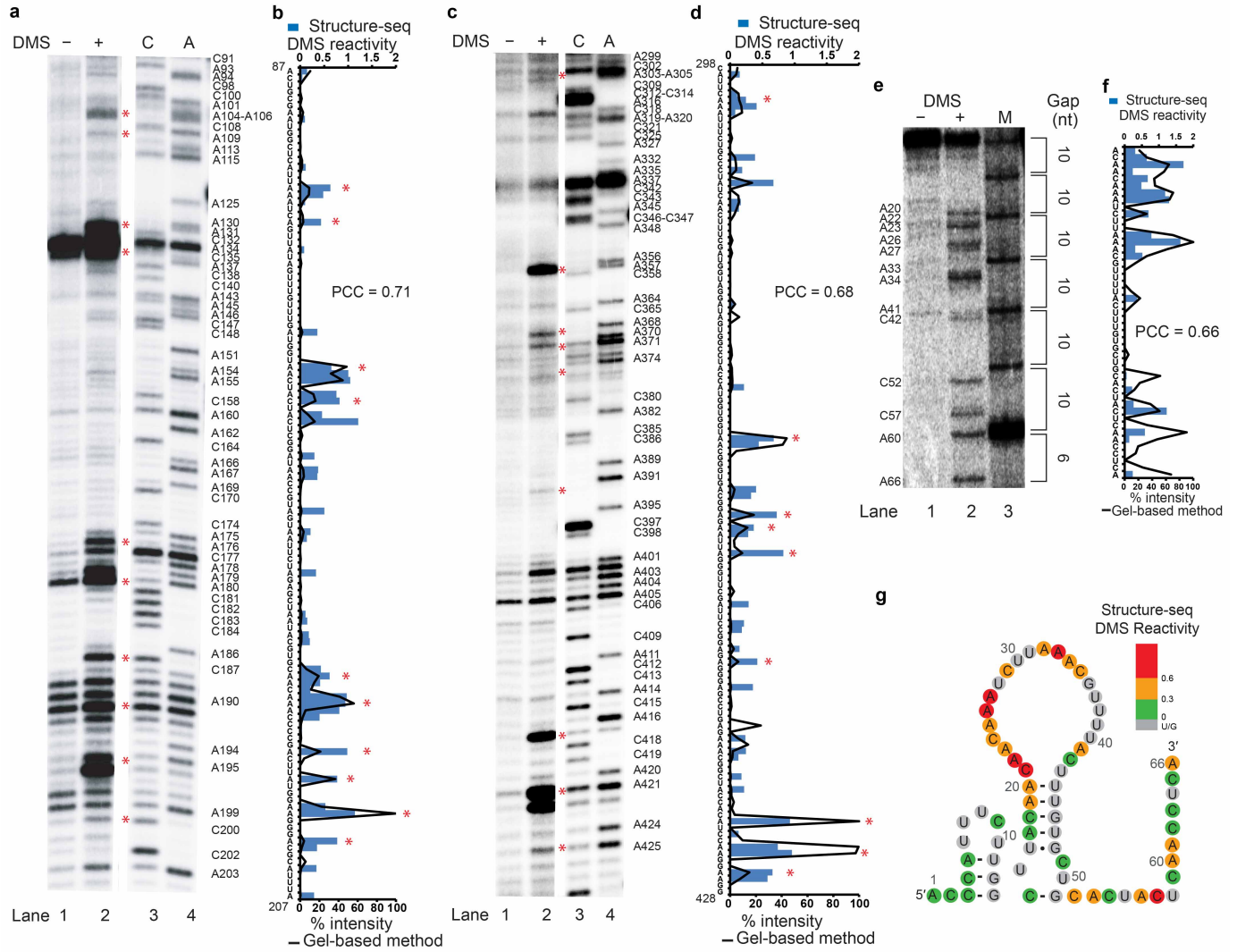


Extended Data Figure 2 | Structure-seq reveals *in vivo* RNA secondary structures for over 10,000 transcripts and correlates with mRNA abundance. **a**, Structure-seq reveals *in vivo* RNA secondary structures for over 10,000 transcripts. The histogram shows the number of transcripts as a function of the average reverse transcriptase stops associated with A + C nucleotides of a transcript, divided by the total number of the A + C nucleotides of that transcript, calculated for all individual transcripts in our data set. (Note that it is expected that not all As and Cs of a transcript will be DMS-modified and associated with an reverse transcriptase stop, because some As and Cs will be protected, for example, by base-pairing, tertiary structure or protein binding.) There are 10,781 transcripts with ≥ 1 average read per A + C nucleotides (dark shading and to the right of the right-most dashed red line). With a threshold of 0.5 average reads per A + C nucleotides, there are 15,565 transcripts (to the right of the left-most dashed red line). It is of interest to compare structure-seq, which provides the first high-throughput *in vivo* RNA structurome, with previous high-throughput studies of RNA structures conducted *in vitro*[5–8]. We have coverage with ≥ 1 average reverse transcriptase stop per nucleotide across 10,623 mRNAs, which compares favourably with ~3,000 mRNAs with load (number of reads per nucleotide) > 1 from an *in vitro* study of yeast[5]. In comparison with 3.9×10^5 reads (0.0078 RNase One cleavages per nucleotide on average) on mRNAs in the single-stranded RNA-seq library of an *in vitro* study of RNA structure in *Arabidopsis*[6], we have much improved coverage with 7.1×10^7 reads (1.4 reverse transcriptase stops per nucleotide on average) on mRNAs in our (+)DMS *in vivo* library. **b**, **c**, Structure-seq queries *in vivo* RNA structures in proportion to their abundance in the transcriptome. mRNA abundance within our structure-seq data set is highly correlated with mRNA abundance from RNA-seq analysis in this study (**b**) and with RNA-seq analysis from a previous study (**c**)[14]. Correlation of mRNA abundance is based on average sequencing reads per mRNA between structure-seq and RNA-seq. The RNA-seq data set in our study was generated in parallel with the structure-seq data set from seedlings under the identical growth conditions but without DMS; that is, the RNA-seq data are extracted from the (−)DMS library. The RNA-seq data set from ref. 14 was generated from five-day-old etiolated seedlings. The PCCs of 0.89 and 0.78, respectively, indicate that more abundant mRNAs are more likely to have sufficient coverage available for structure-seq analysis.

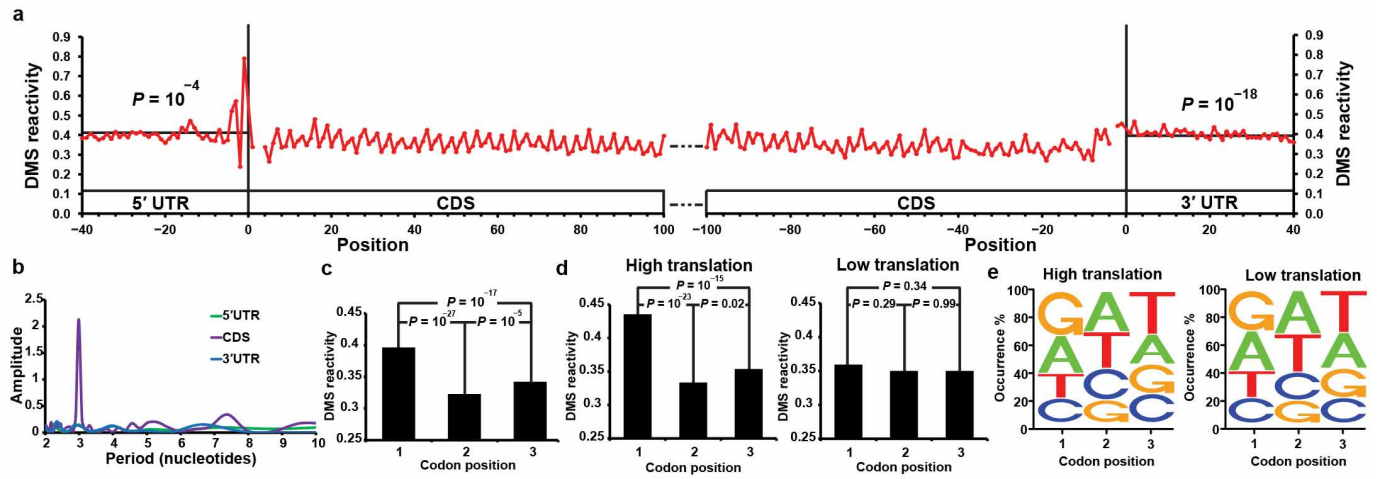


	DMS reactivity	
	High	Low
Single-stranded	86.7% (96.7%)	48.0%
Base-paired	13.3% (3.3%)	52.0%

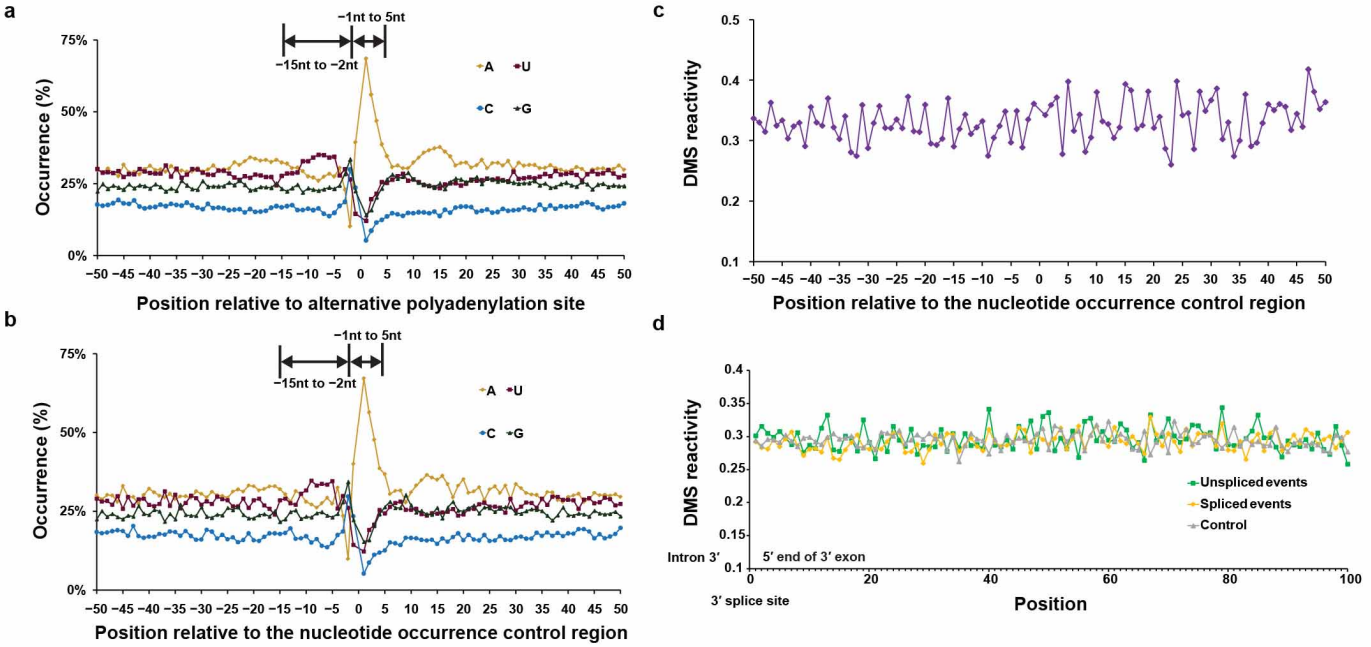
Extended Data Figure 3 | Structure-seq provides the complete map of the 18S rRNA *in vivo* structure at nucleotide resolution. **a**, Structure-seq provides the complete map of the 18S rRNA *in vivo* structure at nucleotide resolution. The complete 18S rRNA phylogenetic structure[16] is colour-coded according to the DMS reactivity generated from structure-seq (DMS reactivity ≥0.6 marked in red; DMS reactivity 0.3–0.6 marked in yellow; DMS reactivity 0–0.3 marked in green; and U/G bases marked in grey). **b**, High correlation between structure-seq and 18S rRNA phylogenetic structure. In the entire 18S rRNA (length = 1,808 nt), 86.7% (true positive) of the As and Cs that show high *in vivo* DMS reactivity (defined as ≥0.6) in our data set correspond to single-stranded regions in the phylogenetic structure[16], whereas 52.0% (true negative) of the As and Cs that show low *in vivo* DMS reactivity (defined as ≤ 0.3) in our data set correspond to base-paired regions in the phylogenetic structure. The 48.0% (false negative) of the As and Cs that show low *in vivo* DMS reactivity in our data set but correspond to single-stranded regions in the phylogenetic structure presumably are protected by either ribosomal proteins or non-base-pairing tertiary RNA structure. Of the 13.3% (false positive) reactive nucleotides (defined as ≥ 0.6 from structure-seq) that are annotated as base-paired in the phylogenetic structure, 75% of these nucleotides are positioned either at the end of a helix or adjacent to a helical defect such as a bulge or loop, locations that are known to lead to flexibility[37]. Values in parentheses, corrected for this positioning, show higher true positive and lower false positive percentages.



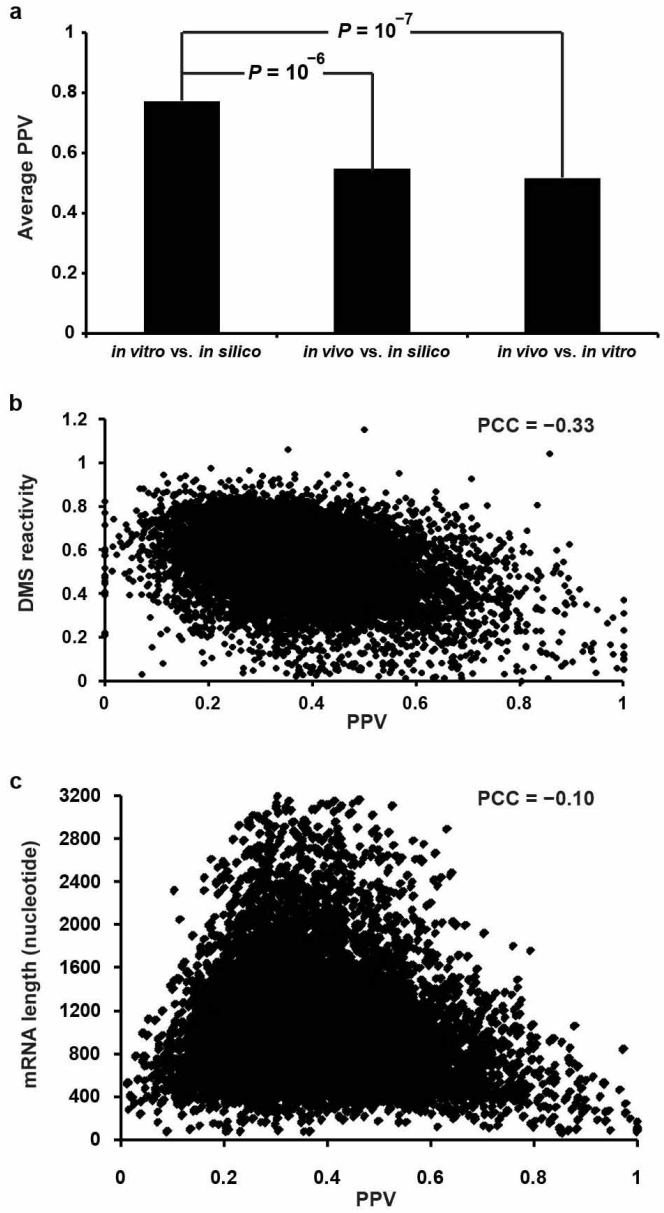
Extended Data Figure 4 | Structure-seq results are strongly correlated with results from the conventional gel-based RNA structure probing method. **a**, Nucleotides 87–207 of 18S rRNA were probed by the conventional gel-based method. Lanes 1–2 show the (−)DMS and (+)DMS results on the region of interest. Lanes 3–4 show C and A dideoxy sequencing. For both this panel and structure-seq, the starting material was the same total population of *in vivo* DMS-modified RNA. **b**, The results from structure-seq (blue bars) are compared to results from the conventional gel-based method, presented as normalized band intensity (black lines), with the highest intensity normalized to 100%[5]. The red asterisks indicate nucleotides that have significant DMS modifications from both methods, and are also shown in panel **a**. Structure-seq results are strongly correlated with results from the conventional gel-based RNA structure probing method: the PCC between the two methods is 0.71. **c**, **d**, Nucleotides 298–428 of 18S rRNA as probed by structure-seq and also analysed by the conventional gel-based method. The PCC is 0.68. **e**–**g**, Structure-seq results are also strongly correlated with results from the conventional gel-based RNA structure probing method for an individual mRNA, *CAB1* (At1g29930). The 5′ UTR of *CAB1* was probed by structure-seq and analysed by the gel-based method; in both cases, the starting material was the same total population of *in vivo* DMS-modified RNA. **e**, Lanes 1–2 show the (−)DMS and (+)DMS results on the region of interest as analysed by the conventional gel-based method. A 10-nt marker (M) was size fractionated (lane 3) to allow nucleotide assignment based on spacing. **f**, DMS reactivity from structure-seq is plotted with nucleotide resolution (blue bars). Results from the gel-based RNA structure probing method are presented as normalized quantified band intensity (black lines), with the highest intensity normalized to 100%[5]. For the gel-based method, the nucleotides near the 5′ end cannot be confidently quantified and assigned due to band compression at the top of the gel and proximity to the full-length band. The PCC between the two methods is 0.66. **g**, The secondary structure of the 5′ UTR of *CAB1* mRNA (At1g29930) was determined using the i*n vivo* DMS constraints obtained from structure-seq. (DMS reactivity ≥ 0.6 marked in red; DMS reactivity 0.3–0.6 marked in yellow; DMS reactivity 0–0.3 marked in green; and U/G bases marked in grey).



Extended Data Figure 5 | Structure-seq reveals global trends in mRNA secondary structure *in vivo* that correlate with translation efficiency. **a**, Average DMS reactivity on an A + C nucleotide basis in selected regions of 22,721 mRNAs (including all splice variants) that have 5′ and 3′ UTR regions longer than 40 nt: 5′ UTR region (40 nt upstream of the start codon); CDS initial region (100 nt downstream of the start codon); CDS final region (100 nt upstream of the stop codon); and 3′ UTR region (40 nt downstream of the stop codon) are depicted. The transcripts were aligned by their start codon and stop codon (vertical lines). (Us and Gs in the start codon and the stop codon were not counted, marked by a break in the red line.) The 40-nt 5′ UTR and 3′ UTR regions show significantly higher average DMS reactivity than the flanking 100 nt of the CDS region, with *P* values of 10^{-4} and 10^{-18}, respectively (Student's *t*-tests). The first 5 nt immediately upstream of the start codon show significantly higher reactivity than the average DMS reactivity across the first 100 nt of the CDS with *P* value of 10^{-112} (Student's *t*-test). **b**, Discrete Fourier transform of average DMS reactivity on a nucleotide basis was performed on the 40-nt 5′ UTR (green line), the first 100 nt of the CDS (purple line) and the 40-nt 3′ UTR (blue line) regions. Only the CDS shows the periodic signal. For the analysis, the 40-nt 5′ UTRs and 3′ UTRs were compared to the first 100 nt of the CDS regions. **c**, The average DMS reactivity of the three positions in each codon was computed from the entire CDS regions of all 22,721 mRNAs. The first position of each codon shows significantly higher average DMS reactivity compared with the second position of each codon ($P = 10^{-27}$). The third position of each codon shows significantly higher average DMS reactivity compared with the second position ($P = 10^{-5}$) but significantly lower average DMS reactivity compared with the first position of each codon ($P = 10^{-17}$) (Student's *t*-tests). **d**, Structure-seq reveals significantly stronger periodic signal in the coding regions of high translation efficiency mRNAs (1,136 mRNAs) as compared to low translation efficiency mRNAs (1,136 mRNAs). We analysed the polyribosome-associated mRNA populations defined in a previous study[21], ranking the mRNAs according to their polyribosome-associated mRNA abundance[21]. We defined the top 5% ($n = 1{,}136$ mRNAs) as the 'high translation efficiency mRNAs' and the bottom 5% ($n = 1{,}136$ mRNAs) as the 'low translation efficiency mRNAs'. The average DMS reactivity of the three positions of each codon was computed along the entire CDS for the high translation efficiency mRNAs and the low translation efficiency mRNAs. The difference in average DMS reactivity between the three nucleotides is significantly greater in the high translation efficiency transcripts (nt 1–2, $P = 10^{-23}$; nt 2–3, $P = 0.02$; nt 1–3, $P = 10^{-15}$) than in the low translation efficiency transcripts (nt 1–2, $P = 0.29$; nt 2–3, $P = 0.99$; nt 1–3, $P = 0.34$) (Student's *t*-tests). **e**, No nucleotide or codon bias in high versus low translation efficiency mRNAs occurs in any of the three positions of the codon. There is no difference between high translation efficiency mRNAs (1,136 mRNAs) and low translation efficiency mRNAs (1,136 mRNAs) in the frequency of nucleotide occurrence at each codon position. The correlation between the codon usage of the high translation efficiency mRNAs and low translation efficiency mRNAs is very high (PCC = 0.90).

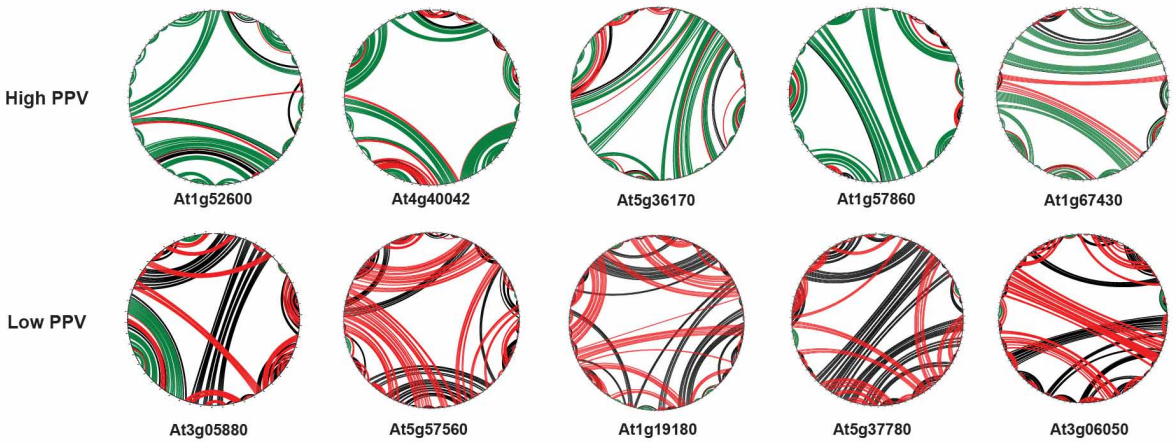


Extended Data Figure 6 | Control analyses for alternative polyadenylation and alternative splicing. **a**, The percentages of nucleotide occurrence around the site of alternative polyadenylation show a U/A rich region from -15 nt to -2 nt ($P = 10^{-16}$ Student's *t*-test), and the region from 1 nt upstream to 5 nt downstream (nt -1 to 5) of the cleavage site is A-rich ($P = 10^{-5}$ Student's *t*-test). This pattern is not unlike that reported for a combined data set of all polyadenylation sites[23]. The percentages of nucleotide occurrence are plotted relative to the alternative polyadenylation site position collected from a previous study[22], indicated by 0: (A (orange diamonds); U (dark red squares); C (blue circles); and G (green triangles)). **b**-**c**, Nucleotide composition and sequence alone cannot account for the RNA structural pattern of the alternative polyadenylation site. **b**, We identified 20 nt regions in our structure-seq mRNA data set that are not alternative polyadenylation cleavage sites but contain the same exact nucleotide sequence as the region 15 nt upstream and 5 nt downstream of each alternative polyadenylation cleavage site that we analysed. The percentages of nucleotide occurrence are plotted relative to the position corresponding to where the alternative polyadenylation site (designated as position zero) would be situated: (A (orange diamonds); U (dark red squares); C (blue circles); and G (green triangles)). **c**, For the selected control region from panel **b**, DMS reactivity of these selected 20 nt control regions as well as the regions upstream (35 nt) and downstream (45 nt) was averaged on a nucleotide basis and plotted, revealing absence of any structural features (violet line). **d**, Extensive RNA secondary structure was not apparent at the 3′ splice site. A previous genome-wide study of alternative splicing (AS) in *Arabidopsis* seedlings[26] was used to identify for each mRNA in our data set, whether all introns were spliced out or whether AS (including exon skipping and intron retention) occurred. DMS reactivity along 100 nt in the exons upstream of the 3′ splice site was averaged on a nucleotide basis from the unspliced events, including both exon skipping and intron retention (green lines), and the spliced events (yellow lines). The same nucleotide composition of the 100 nt in the unspliced AS events was shuffled and remapped to regions in our structure-seq mRNA data set that were not located at the junction of a 3′ splice site. The averaged DMS reactivity collected from the control regions with the same nucleotide composition served as the control (grey lines).



Extended Data Figure 7 | *In vitro* structures differ from *in vivo* structures; PPV does not correlate with average DMS reactivity or with mRNA length. **a**, *In vitro* structures differ from *in vivo* structures, and *in vitro* structures are more similar to *in silico* structures than are *in vivo* structures. The 61 *Arabidopsis* mRNAs with coverage ≥ 0.5 cleavages per nucleotide from Li *et al.*'s *in vitro* data were compared among the *in silico* structure (from RNAstructure), the *in vitro* structure (*in silico* structures from RNAstructure constrained by Li *et al.*'s *in vitro* data)[6], and the *in vivo* structure (*in silico* structures from RNAstructure constrained by our *in vivo* data). PPV (the base pairs in one structure that are also present in another structure, as a proportion) was averaged across these 61 mRNAs. The PPV between *in vitro* structures and *in silico* structures is 0.77, which is significantly higher than the PPV between *in vivo* structures and *in silico* structures and is also significantly higher than the PPV between *in vivo* and *in vitro* structures, according to two sample *t*-tests with *P* values as shown in the figure. *In vivo* structures are different from both *in vitro* structures (PPV = 0.51) and *in silico* structures (PPV = 0.55). **b**, PPV does not correlate with average DMS reactivity per nucleotide. For each of 10,623 mRNAs in our structure-seq data set, the corresponding PPV of each mRNA was plotted, revealing an absence of correlation between PPV and average DMS reactivity per nucleotide (PCC = −0.33). **c**, PPV does not correlate with mRNA length. For each of 10,623 mRNAs, the corresponding PPV of each mRNA was plotted as a function of mRNA length, revealing an absence of correlation between these two variables (PCC = −0.10).

a



b

High PPV

ID	Annotation	PPV	Single strandedness (*in silico*)	Single strandedness (*in vivo*)	Max. loop length (*in silico*)	Max. loop length (*in vivo*)	Free energy per nucleotide (*in silico*)	Free energy per nucleotide (*in vivo*)
At1g52600	Peptidase S24/S26	0.87	0.35	0.35	11	11	−0.26	−0.46
At4g40042	SPC12 Microsomal signal peptidase 12 kDa subunit	0.78	0.35	0.34	12	12	−0.26	−0.51
At5g36170	*ATPRFB* Required for normal processing of polycistronic plastidial transcripts	0.78	0.37	0.34	15	13	−0.28	−0.53
At1g57860	Translation protein SH3-like family protein	0.77	0.37	0.40	8	7	−0.28	−0.45
At1g67430	Ribosomal protein L22p/L17e	0.78	0.39	0.39	10	10	−0.27	−0.39

Low PPV

ID	Annotation	PPV	Single strandedness (*in silico*)	Single strandedness (*in vivo*)	Max. loop length (*in silico*)	Max. loop length (*in vivo*)	Free energy per nucleotide (*in silico*)	Free energy per nucleotide (*in vivo*)
At3g05880	*RCI2A* (*RARE COLD-INDUCIBLE 2A*)	0.19	0.37	0.42	11	12	−0.21	−0.31
At5g57560	*TCH4* (*TOUCH 4*), endotransglucosylase/ hydrolase rapidly upregulated in response to environmental stimuli	0.07	0.39	0.43	11	39	−0.25	−0.40
At1g19180	*JAZ1* (*JASMONATE-ZIM-DOMAIN PROTEIN 1*), involved in jasmonate signaling	0.15	0.38	0.41	17	21	−0.22	−0.39
At5g37780	*CAM1* (*CALMODULIN 1*), detection of calcium ion	0.09	0.38	0.43	9	11	−0.24	−0.33
At3g06050	*ATPRXIIF*, *PEROXIREDOXIN IIF* involved in redox homeostasis under oxidative stress	0.09	0.37	0.41	13	19	−0.29	−0.41

Extended Data Figure 8 | Examples for *in vivo* and *in silico* structural feature comparison of high and low PPV mRNAs. **a**, Ten examples for *in vivo* and *in silico* structural comparison of high and low PPV mRNAs. Five examples from the high PPV mRNA group (top) and five examples from the low PPV mRNA group (bottom). At1g52600 and At3g05880 mRNA structures were given in Fig. 4d. Base pair predictions are indicated with coloured lines: red, uniquely *in vivo* base pair; black, uniquely *in silico* base pair; green, base pair present in both the *in vivo* and the *in silico* structure. Plots were generated using the CircleCompare program in the RNAstructure package[35]. Low PPV mRNAs show more extensive differences between *in vivo* and *in silico* structures than do high PPV mRNAs. **b**, Characteristics of *in vivo* and *in silico* structural features in the ten high and low PPV mRNAs. The same five examples from both high PPV and low PPV mRNAs as in **a** were assessed for RNA structural features in both *in silico*-predicted (without *in vivo* constraints) and *in vivo* (*in silico* prediction with constraints from our *in vivo* structure-seq data) structures. *In vivo* structures of low PPV mRNAs show more single stranded regions, longer maximum loop length, and higher (that is, less favourable) free energy per nucleotide as compared to high PPV mRNAs. By contrast, *in silico*-predicted structures do not show such major differences between low and high PPV mRNAs.

Extended Data Table 1 | Statistical analysis of structure-seq libraries

a

	Between Two Biological Replicates		Biological Replicate I	Biological Replicate II
Library	(+)DMS/(+)DMS	(−)DMS/(−)DMS	(−)DMS/(+)DMS	(−)DMS/(+)DMS
Correlation	0.91	0.74	0.49	0.61

b

Library	Total reads	uniquely mapped reads	uniquely mapped %	not mappable reads	not mappable %
(−)DMS Biological Replicate 1	3.93×10^7	3.86×10^7	98.24	6.94×10^5	1.76
(+)DMS Biological Replicate 1	3.25×10^7	3.22×10^7	98.88	3.65×10^5	1.12
(−)DMS Biological Replicate 2	4.60×10^7	4.52×10^7	98.07	8.87×10^5	1.93
(+)DMS Biological Replicate 2	8.87×10^7	8.75×10^7	98.61	1.23×10^6	1.39

c

RNA type	(+)DMS library (reads)	(+)DMS library (% reads)	(−)DMS library (reads)	(−)DMS library (% reads)
mRNA	7.12×10^7	58.75	5.03×10^7	58.86
rRNA	4.81×10^7	39.71	3.31×10^7	38.81
ncRNA	2.29×10^5	0.19	3.02×10^5	0.35
snRNA	1.95×10^4	0.016	4.57×10^4	0.054
tRNA	1.11×10^4	0.0091	1.15×10^4	0.013
miRNA	3.96×10^3	0.0033	5.31×10^3	0.0062
snoRNA	1.68×10^4	0.014	4.12×10^4	0.048
Total	1.21×10^8	100	8.53×10^7	100

a, High correlation (PCC) between biological replicates for (+) and (−)DMS libraries, and low correlation between the (+)DMS and (−)DMS libraries for each biological replicate. **b**, High read number and mappability of our (+)DMS and (−)DMS libraries. **c**, mRNAs and rRNAs predominate among different classes of RNAs in (+)DMS and (−)DMS libraries (combined data from two biological replicates).

Extended Data Table 2 | *In vivo* constraints improve the prediction of structure in 18S rRNA

Row		PPV/Sensitivity
1	*in silico* vs. phylogenetic structure	0.27/0.31
2	*in vivo* vs. phylogenetic structure	0.41/0.45
3	*in vivo* vs. phylogenetic structure, omitting false negatives	0.50/0.52
4	ideal A/C constraint vs. phylogenetic structure	0.63/0.63
5	ideal A/C/U/G constraint vs. phylogenetic structure	0.68/0.65
6	*in vivo* vs. *in silico*	0.48/0.46

We calculated the PPV/sensitivity between *in silico* and phylogenetic structure, *in vivo* and phylogenetic structure, and *in vivo* and *in silico* structure in 18S rRNA. (Sensitivity is defined as the proportion of base pairs occurring *in silico* that also appear *in vivo*.) We also compared the *in vivo* structure with the phylogenetic structure upon omission of false negatives (i.e., we did not apply a pseudo-free energy constraint to the false negative data), because false negatives presumably result from protection by either ribosomal proteins or non-base-pairing tertiary RNA structure rather than base pairing. In addition, we folded the RNAs with the constraints generated from ideal A/C or ideal A/C/U/G base-pairing information (the predicted structure with the A/C or A/C/U/G constraints as generated directly from the phylogenetic structure), and compared the resultant structure predictions with actual phylogenetic structures.