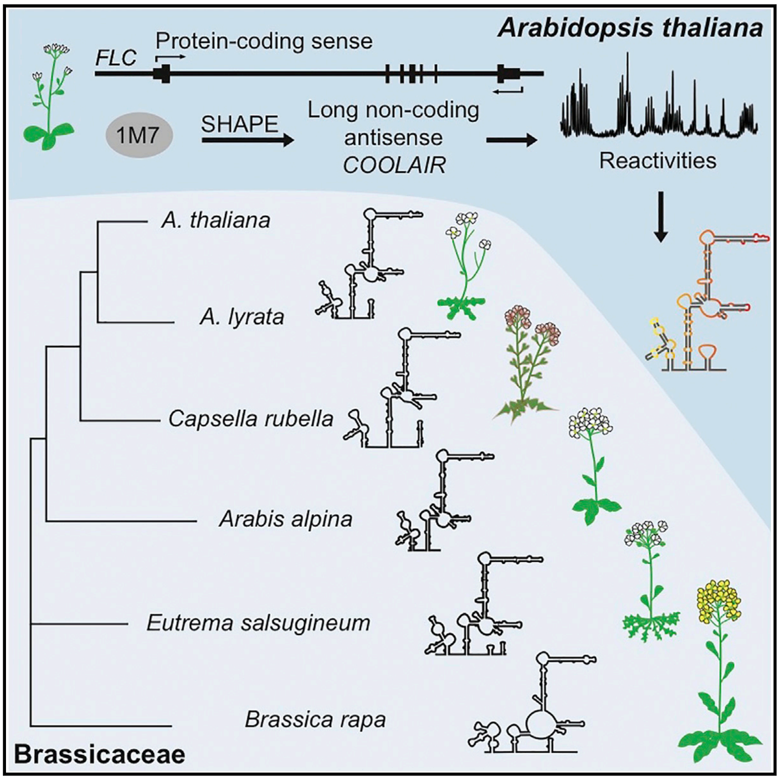
Cell Reports

Report

COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures

Graphical Abstract



Authors

Emily J. Hawkes, Scott P. Hennelly, Irina V. Novikova, Judith A. Irwin, Caroline Dean, Karissa Y. Sanbonmatsu

Correspondence

kys@lanl.gov

In Brief

Hawkes et al. chemically probed the *Arabidopsis thaliana* long non-coding RNA, *COOLAIR*, to reveal its complex secondary structure. The evolutionary conservation of this secondary structure across Brassicaceae species, despite low sequence similarity, indicates a functional role. This is supported by variation in secondary structure in a phenotypically significant *COOLAIR* splicing isoform.

Highlights

- Long non-coding (lnc)RNA, COOLAIR, has a complex structure with multi-helix junctions
- The distally polyadenylated lncRNA contains two right-hand turn (r-turn) motifs
- A phenotypically significant lncRNA isoform has an altered secondary structure
- The *COOLAIR* secondary structure has been evolutionarily conserved across species

Hawkes et al., 2016, Cell Reports *16*, 3087–3096
September 20, 2016
http://dx.doi.org/10.1016/j.celrep.2016.08.045

COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures

Emily J. Hawkes,[1,5] Scott P. Hennelly,[2,3,5] Irina V. Novikova,[2,4] Judith A. Irwin,[1] Caroline Dean,[1] and Karissa Y. Sanbonmatsu[2,3,6,7,*]
[1]John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK
[2]Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[3]New Mexico Consortium, Los Alamos, NM 87544, USA
[4]Pacific Northwest National Laboratory, Environmental Molecular Sciences Laboratory, Richland, WA 99354, USA
[5]Co-first author
[6]Lead Contact
[7]Twitter: @karissascience
*Correspondence: kys@lanl.gov
http://dx.doi.org/10.1016/j.celrep.2016.08.045

SUMMARY

There is considerable debate about the functionality of long non-coding RNAs (lncRNAs). Lack of sequence conservation has been used to argue against functional relevance. We investigated antisense lncRNAs, called *COOLAIR*, at the *A. thaliana FLC* locus and experimentally determined their secondary structure. The major *COOLAIR* variants are highly structured, organized by exon. The distally polyadenylated transcript has a complex multi-domain structure, altered by a single non-coding SNP defining a functionally distinct *A. thaliana FLC* haplotype. The *A. thaliana COOLAIR* secondary structure was used to predict *COOLAIR* exons in evolutionarily divergent Brassicaceae species. These predictions were validated through chemical probing and cloning. Despite the relatively low nucleotide sequence identity, the structures, including multi-helix junctions, show remarkable evolutionary conservation. In a number of places, the structure is conserved through covariation of a non-contiguous DNA sequence. This structural conservation supports a functional role for *COOLAIR* transcripts rather than, or in addition to, antisense transcription.

INTRODUCTION

Long non-coding RNAs (lncRNAs) have emerged as potentially important players in the epigenetic regulation of development and disease in many organisms. These RNAs are typically 1–10 kb in length, polyadenylated, capped, and alternatively spliced (Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). They can be *cis*- or *trans*-acting and have been associated with gene regulation in mechanisms including chromatin scaffolding, Polycomb complex (PRC2) recruitment to chromatin, mRNA decay, and decoys for proteins and microRNAs (miRNAs). Specific functional studies have shown lncRNAs to be essential for Xist regulation, paraspeckle formation, lineage commitment, stem cell development, cancer-associated effects, coactivation of hormone response, and brain development (Klattenhoff et al., 2013; Novikova et al., 2012; Sauvageau et al., 2013).

While the functional importance of lncRNAs such as Xist is well accepted, more general roles are still controversial, especially in light of low primary sequence conservation through evolution (Graur et al., 2013; Nitsche et al., 2015). A conserved RNA secondary structure can occur despite weak conservation of the primary sequence. For example, riboswitches (regulatory RNAs with exquisite control over metabolism in bacteria) typically have nucleic acid sequence identities of only 50%–65% but secondary structures conserved across thousands of sequences (Mandal and Breaker, 2004; Nawrocki et al., 2015; Roth and Breaker, 2009). Likewise, the U2 and U4 spliceosomal RNAs, 5S rRNA, and group I introns have low sequence identities but highly conserved structures (Nawrocki et al., 2015). In such cases of low sequence identity, sequence-based search algorithms (e.g., BLAST) are not generally productive; however, a strategy that aligns syntenic sequences according to chemically probed structures has proved successful for identifying riboswitch RNAs (Cheah et al., 2007; Weinberg et al., 2009). While few in vivo chemical probing studies on riboswitches have been performed, nearly every in vitro chemical probing-derived riboswitch structure has been validated with high-resolution crystallographic structures (Roth and Breaker, 2009). Nature often evolves structural RNAs with changes in helical length or with addition or subtraction of entire helices, presenting formidable challenges for computation. We therefore adopt an integrative approach, proven to be accurate for riboswitches and ribosomes, of time-consuming iteration among chemical probing, secondary structure refinement, sequence alignment refinement, and functional studies.

Over the past few years, researchers have been laying the groundwork for lncRNA structure-function studies. Genome-wide studies suggest lncRNAs are more structured than mRNAs but less structured than rRNAs (Ding et al., 2014; Quinn et al., 2016; Wan et al., 2012). Studies of MALAT1 and related RNAs show the 3′ end forms a triple helix protecting it from RNase degradation (Brown et al., 2014). Other pioneering studies have examined stem-loop-related structures (Quinn et al., 2016) and lncRNA-protein interactions (Davidovich et al., 2013). However, few have attempted to determine the secondary structure of complete, intact, single lncRNA systems. Those studies that have done so revealed hierarchically structured RNAs with sub-domains containing modular RNA secondary structure motifs (Ilik et al., 2013; Novikova et al., 2012; Somarowthu et al., 2015).

We chose to investigate *Arabidopsis thaliana* antisense lncRNAs, named *COOLAIR*, which are important in the regulation of a major plant developmental gene *FLOWERING LOCUS C* (*FLC*). These initiate just downstream of the protein-coding sense transcript poly(A) site and are alternatively spliced and polyadenylated, either at a proximal site to give ~400 nt class I transcripts or at a distal site within the *FLC* promoter region to give ~750 nt class II transcripts (Figure 1A). These transcripts act in a feedback mechanism linking *COOLAIR* processing to *FLC* gene body histone demethylation, reduced *FLC* transcription, and earlier flowering (Liu et al., 2010). *COOLAIR* is upregulated during prolonged cold, contributing to a Polycomb-mediated epigenetic switch between opposing chromatin states (Csorba et al., 2014). While the *COOLAIR* promoter region is evolutionarily conserved, sequence conservation is low in regions corresponding to *FLC* 5′ and 3′ untranslated and intronic sequences (Castaings et al., 2014; Li et al., 2016). Whether it is *COOLAIR* transcription, *COOLAIR* transcripts, or both that are functionally important is not yet known.

Here, we apply the riboswitch strategy supplemented with the shotgun secondary structure (3S) determination method to determine the secondary structure of the *COOLAIR* transcripts (Novikova et al., 2013; Weinberg et al., 2007). We find the distal *COOLAIR* transcript is highly structured in *A. thaliana*, with numerous secondary structure motifs, an intricate multi-way junction, and two unusual asymmetric 5′ internal loops (right-hand turn [r-turn] motifs). Part of this structure is altered by a single non-coding SNP that has been shown to confer functional *cis*-regulatory variation to a naturally occurring *FLC* haplotype. The secondary structure was used to predict *COOLAIR* exonic sequences in a range of evolutionarily distinct Brassicaceae species, including *Arabidopsis lyrata*, *Capsella rubella*, and *Brassica rapa*, which were then validated in vivo.

RESULTS

3S Chemical Probing of *A. thaliana COOLAIR* Transcripts

COOLAIR transcripts were probed in vitro using selective 2′-OH acylation analyzed by primer extension (SHAPE) (Merino et al., 2005). In addition, to isolate modularly folded regions within *COOLAIR*, fragments of the full-length distal transcript were probed using 3S (Novikova et al., 2013). We divided the distal class II.i isoform into three segments of ~200–250 nt (positions 1–235, 211–433, and 403–658). In the first fragment, the SHAPE reactivity profile of the 5′ region had significant overlap with the SHAPE reactivity profile of the full RNA (positions ~1–125), suggesting this region possesses an autonomous, modular fold in the context of the full *COOLAIR* with a well-defined three-way junction (Figure 1). While the relative ratio in reactivity differed slightly, the positions of base-paired nucleotides remained the same. The reactivity profile of the 3′ half of fragment 1 (positions ~125–235) differed significantly, suggesting that this region forms interactions outside of fragment 1 positions. The reactivity profile of most of fragment 2 agreed with the full *COOLAIR* profile, suggesting a modular fold with two well-defined helices joined by a large internal loop. Combining 3S fold information from fragments 1 and 2 with SHAPE probing data from the full-length transcript allowed us to produce the secondary structure for the distal *COOLAIR* II.i transcript (Figure S1), confirmed by CMCT (1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluene sulfonate) probing data (Figure 2A).

The Distal *COOLAIR* Transcript Has a Complex Structural Architecture Organized into Three Distinct Domains

The distal *COOLAIR* lncRNA structure is arranged into 12 helices, seven stem loops, a three-way junction, a five-way junction, and two rare r-turns (Figure 2A). Nucleotides that exhibit high SHAPE reactivities are mainly located in the terminal loops, internal loops, and junction regions, e.g., the terminal loops of helix 3 (H3) and H12, the internal loop separating H7 and H8, and the multi-way junction connecting H5, H6, H7, H10, and H11. Many of these single-stranded regions are purine rich. This is consistent with the secondary structures of rRNAs, riboswitch RNAs, the steroid receptor RNA activator (*SRA1*), and *Braveheart*, each of which shows a similar propensity for purine-rich single-stranded locations. Nucleotides restrained by base-pairing interactions generally show a much lower tendency toward modification. There are a few select instances in which nucleotides involved in base pairing, located close to the single-stranded regions or bulges, can also be reactive toward the SHAPE reagent, such as in H8 and H9. This was observed from SHAPE probing of the 16S rRNA, whose secondary structure is well known (Noller and Woese, 1981). Minor instances of SHAPE-reactive nucleotides positioned in the central part of helices have been previously observed in rRNA (Deigan et al., 2009).

COOLAIR appears to be organized into three major domains: the 5′ domain in exon 1, characterized by a three-way junction; the 3′ major domain (3′ M or central domain) in exon 2, containing the long H4, r-turn, and five-way junction; and the 3′ minor domain (3′ m or stalk) also in exon 2 and containing the two long helices H8 and H9 connected by the second r-turn. Most distal *COOLAIR* structural features do not correspond to *FLC* protein exonic regions apart from the stalk domain, which is formed from sequences within exon 1 of the sense transcript. The extensive distal H4 corresponds to sense intronic regions. The sequence underlying the first exon (H1–H3) of both the proximal and the distal transcripts also corresponds with the non-coding sequence.

The r-Turn Motif

The secondary structure motifs of *COOLAIR* are found in many instances of rRNAs and RNase P RNAs (Table S3), with the exception of the two r-turn motifs. These are internal loop

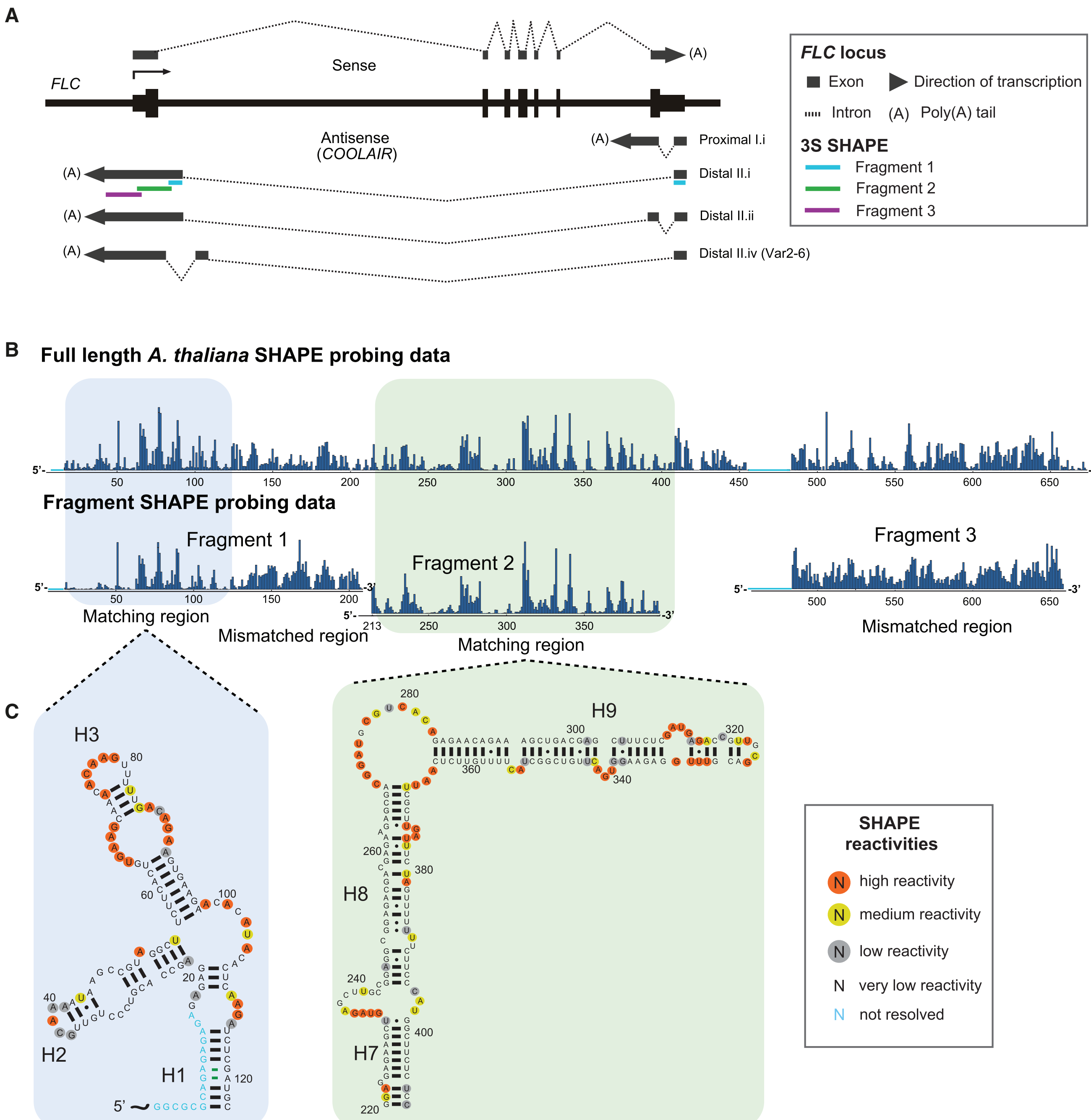


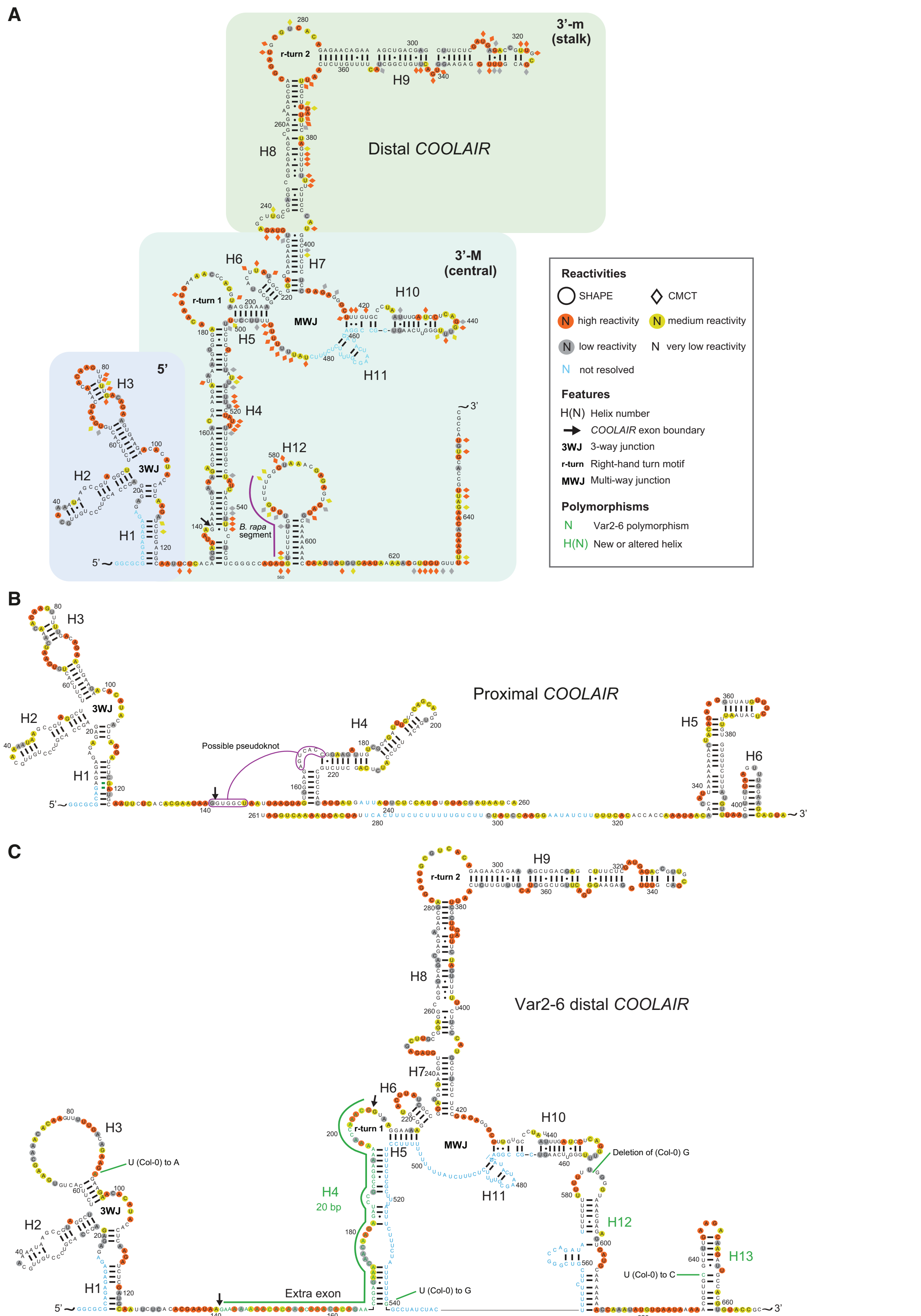
Figure 1. 3S Determination via SHAPE Probing of the Distal *A. thaliana COOLAIR* lncRNA

(A) Schematic representation of *FLC* and *COOLAIR* transcripts at the *FLC* locus, with 3S fragment positions mapped.

(B) SHAPE reactivities for the full-length *A. thaliana* distal *COOLAIR* (class II.i) transcript are compared with shorter fragments 1–3 for 3S determination.

(C) Modular secondary structure corresponding to reactivity data of the boxed regions in (B).

structures consisting of a large single-stranded region (19 and 13 nt) on the 5′ side and a very short single-stranded region (2 and 3 nt) on the 3′ side, corresponding to type 1 and type 2 r-turns, respectively. Internally, this motif contains two adjacent pairs consisting of potential non-canonical GA or of canonical Watson-Crick (WC) or GU base pairs. We have followed the definition of "motif" used by Moore and coworkers (Klein et al., 2001). Because the r-turn is well defined and recurrent, it may play a role in function either directly, through binding protein or ligand, or indirectly, through positioning helices or engaging in tertiary contacts (Yesselman and Das, 2015). Two crystallographic studies revealed similar motifs (Figure S1) in the U6 small nuclear



(legend on next page)

ribonucleoprotein and pistol ribozyme (Montemayor et al., 2014; Ren et al., 2016). For U6, the r-turn is a receptor for a protein forming an extensive interface with multiple RNA recognition motif (RRM) regions of the Prp24 protein. In pistol, the r-turn is a receptor for a pseudo-knot interaction. The r-turn occurs in two other lncRNA systems: *SRA1* and in the *Braveheart* asymmetric G-rich internal loop (AGIL) (Novikova et al., 2012; Xue et al., 2016).

The Proximal Isoform Shares the Distal 5′ Domain Structure

We also performed SHAPE probing on the proximally polyadenylated *COOLAIR* I.i transcript (Figure 2B). This transcript was substantially disordered (high reactivity), with three localized regions of secondary structure. The secondary structure of the 5′ domain for the proximal transcript was identical to that of the distal transcript (H1–H3 in both), because they share a common first exon. The 3′ domain of the proximal transcript consists of three helical structures (H4–H6), each capped by a stem loop, with H4 underlying exon 7 of the protein-coding sense transcript. H4 contains three internal loops and an eight-member stem loop. The potential for a pseudo-knot interaction, consistent with the probing data, exists between a stretch of sequence (5′-GGUGGCU-3′) spanning the exon 1/exon 2 splice junction and the first internal loop (5′-AGUCACC-3′) of H4.

Functionally Important Natural *cis* Polymorphism Influences *COOLAIR* Secondary Structure

To investigate the functional significance of the *COOLAIR* secondary structure, we took advantage of natural variation at *FLC*. In the preceding experiments, we probed *COOLAIR* RNA from the widely used Columbia (Col) accession. Other functionally distinct *FLC* haplotypes exist in *A. thaliana* accessions from different parts of the world. Haplotype 11, characterized in the Var2-6 accession from northern Sweden, contains a SNP that changes the splicing pattern of *COOLAIR*, causing a shift to a downstream distal splice acceptor site and inclusion of an internal exon (Li et al., 2015). This distal isoform (class II.iv) (Figure 1A) cotranscriptionally increases transcription of the *FLC* nascent transcript, thus delaying flowering. We compared the secondary structure of the functionally distinct Var2-6 distal transcript with the Col transcript using SHAPE analysis (cf. Figures 2A and 2C). The structure was nearly identical, including the 5′ domain, the first r-turn and multi-way junction, and the stalk. However, there were several significant differences. In the Var2-6 isoform, the 3′ end of the additional exon forms half of H4; thus, a shorter H4 is maintained in a similar position but composed of an entirely different sequence. This supports the need for H4 to be maintained; without it, the first r-turn would not form. In addition to its shorter length (17 versus 37 bp), H4 contains a large highly reactive internal loop (14 bases), making it likely to be less stable. The 5′ end of the additional exon is apparently unstructured, with many highly reactive bases creating a longer distance between H1 and H4. The 3′ end is more structured, with H12 bifurcated and an additional H13. While four additional polymorphisms found within the Var2-6 haplotype group have been highlighted in Figure 2C, the most significant alteration to the structure is caused by the splice site shift. The U-A SNP in H3 disrupts base pairing and promotes one large terminal loop, in contrast to the internal and smaller terminal loop in Col. Although structurally interesting, this SNP is found in a large number of accessions and is not responsible for the Var2-6 phenotype. The changed functionality of the Var2-6 *COOLAIR* transcript is therefore most likely to be due to the structural changes associated with H4.

Use of the Secondary Structure to Identify *COOLAIR* in Other Plant Species

Although distal and/or proximal isoforms of *COOLAIR* have been identified in *A. lyrata*, *A. alpina*, and *B. rapa*, low sequence conservation complicated the identification of all isoforms (Castaings et al., 2014; Li et al., 2016). We derived *COOLAIR* secondary structures for five Brassicaceae species, *A. lyrata*, *A. alpina*, *C. rubella*, *E. salsugineum*, and *B. rapa*, representing ca. 13 million to 43 million years divergence from *A. thaliana* (Beilstein et al., 2010; Koch and Kiefer, 2005). Following the strategy of Weinberg et al. (2007), we scanned syntenic regions for stretches of sequence identity and then improved the alignment with the chemically determined *A. thaliana* structure, using covariant base pairs to help validate helices.

COOLAIR H8–H9 are antisense to a highly conserved coding region of *FLC* (containing the MADS box motif) and therefore were used to align homologous sequences across the five species. Next, stretches of sequence flanking H8–H9 were shifted to improve alignment with helices in *A. thaliana*. This was repeated and iterated outward toward the 5′ and 3′ ends. The resulting secondary structures show a high degree of similarity with *A. thaliana*, maintaining most structural elements (Figure 3). We find the 5′ domain, the two r-turns, the stalk, and the terminal region of H4 to be conserved across all six species. Each contains covariant base pair flips in the helices and greater variation in the single-stranded regions, supporting the conservation of the secondary structure.

Looking at the consensus structure in Figure 3F, five species contain a five-way junction, while one species (*B. rapa*, SHAPE probed to produce Figure 3C) contains a four-way junction in the central domain due to lack of H6. H7, H8, and H10 are conserved across six species but exhibit length variation. H11

Figure 2. Secondary Structure of the Distal and Proximal *A. thaliana COOLAIR* lncRNAs

(A) Secondary structure of the distal class II.i lncRNA from the *A. thaliana* Col accession, based on SHAPE and CMCT probing experiments. Normalized SHAPE reactivity is represented as colored circles, and normalized CMCT reactivity is represented as colored diamonds. A short segment of the sequence was replaced with the *B. rapa* sequence to improve reactivity data read and to confirm the predicted fold in Figure S1. For the rarity of the structural motifs, see Figure S1 and Table S3.

(B) Secondary structure of the proximal class I.i lncRNA from the *A. thaliana* Col accession, based on SHAPE probing experiments. The potential pseudo-knot may be conserved across species (Figure S4).

(C) Secondary structure of the distal class II.iv lncRNA from the *A. thaliana* Var2-6 accession, based on SHAPE probing experiments.

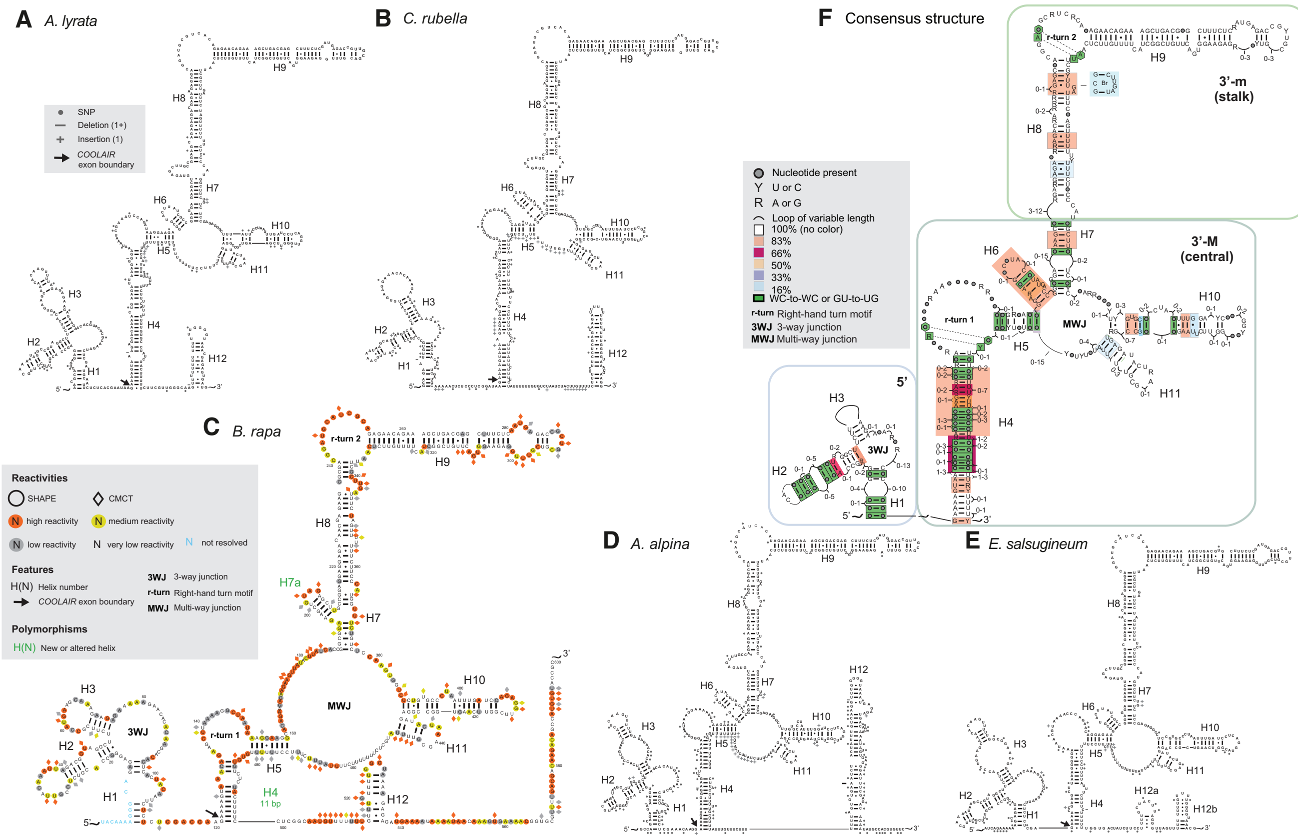


Figure 3. Predicted *COOLAIR* Distal lncRNA Secondary Structures for Five Brassicaceae Species

(A–E) Predicted secondary structure of the distal class II.i transcript of *COOLAIR* for (A) *A. lyrata* (*FLC1*), (B) *C. rubella*, (C) *B. rapa* (*FLC3*), (D) *A. alpina*, and (E) *E. salsugineum*. (C) is annotated with SHAPE and CMCT data, and a variant structure for a different accession is given in Figure S3. Polymorphisms are mapped from pairwise alignment with the *A. thaliana COOLAIR* distal class II.i transcript, and sequence divergence is represented in more detail in Figure S2.

(F) Consensus diagram combining structural information from the five species plus *A. thaliana* shows conservation of the secondary structure, where colored boxes represent the percentage of conservation across species, i.e., pink box = 83% = 5/6 species conserve that structural element. Dots in a looping region signify that the length of the loop is conserved, but the sequence varies. Dots paired to dots in helices signify that a base pair is always present but the sequence varies (i.e., a covariant base pair).

exists in all six species. Covariant base pairs were found in all helices apart from H3, H8, H9 and H11. The terminal four base pairs of H3 are conserved across all six species. Although this helix does not exhibit covariant base pairing from species to species, the helix length and loop length vary, supporting the existence of the helix. As H8 and H9 overlap with a coding region in the sense transcript, they exhibit minimal sequence variation and therefore have almost no opportunity for base pair covariance. While H11 does not exhibit covariant base pairing, *C. rubella* H11 has an extended length, with three extra bases on either side of the loop, forming three additional base pairs.

Conservation of key structural features, despite low sequence similarity in non-protein-coding regions (Figure S2), strongly supports a functional role. In effect, nature has maintained these structural features even from the sequence of largely, or (in the case of Var2-6 H4) completely, different composition. Although a role for the distal *COOLAIR* transcript in the cold-induced epigenetic silencing of *FLC* is perhaps less likely because the two perennial species (*A. alpina* and *A. lyrata*) do not exhibit distinct structural features, the Var2-6 data are supportive of a role in setting initial levels of *FLC* expression in the warm.

Validation of *COOLAIR* Spliced Transcripts in Evolutionarily Diverse Species

Primers designed from the predicted secondary structures in Figure 3 confirmed the in vivo presence of the proximal and distal *COOLAIR* isoforms in *A. lyrata*, *C. rubella*, and *B. rapa* (Figure 4A). Three major splice variants were identified and classified according to their similarity to *A. thaliana*: the proximal class I.i and the distal II.i and II.ii transcripts (Figure 1A). Splice sites are largely conserved, with the exception of the proximal 3′ acceptor splice site in *B. rapa* and the distal class II.ii terminal exon 3′ acceptor site in *C. rubella* (Figure 4B).

Whereas the same proximal isoform is conserved in all four species, two distal isoforms were identified. Differential distal splicing in *A. thaliana* accessions (Var2-6 versus Col) resulted in changes in *FLC* expression and thus may be equally important across species. Comparison of the loci is complicated by ancient polyploidization and tandem duplication events creating multiple

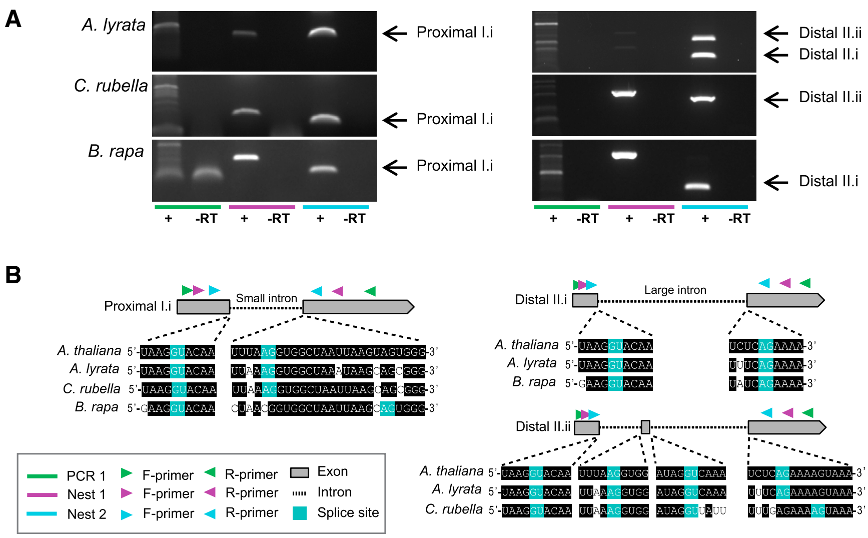


Figure 4. Experimental Validation of *COOLAIR* Transcripts in *A. lyrata*, *C. rubella*, and *B. rapa*

(A) RT-PCR experiments probing for the proximal (left) and distal (right) forms of *COOLAIR* from non-vernalized *A. lyrata* and *C. rubella*, and 2-week-vernalized *B. rapa* leaf tissue. Initial RT-PCR (green line) was followed by two rounds of nested PCR (purple and blue lines) to amplify a specific band, where the + column is the cDNA sample and the −RT column is the DNA contamination control. Different splice variants have been labeled according to *A. thaliana* classes in Figure 1A.

(B) Sequencing the RT-PCR products revealed the major *COOLAIR* splicing isoforms, with gray boxes in the schematic representing exon positions and triangles representing primer positions. Sequences were aligned with *A. thaliana* to compare splice sites, highlighted in blue.

copies of *FLC* in *B. rapa* and *A. lyrata*. We analyzed one of four *B. rapa* copies (*FLC3*) and one of two *A. lyrata* copies (*FLC1*). As loci diverge independently over time, it may be that each expresses unique splicing isoforms. A distal *COOLAIR* isoform with an alternate 3′ acceptor site was identified at the *FLC2* locus in *B. rapa* (Li et al., 2016).

Detection of *COOLAIR* isoforms with similar architecture to *A. thaliana* supported the predicted conservation of the secondary structure. To validate this, we performed SHAPE and CMCT analysis on the more diverged *B. rapa* distal *COOLAIR* (Figure 3C). This class II.i transcript is spliced in the same way as the *A. thaliana* Col isoform but contains multiple polymorphisms. We know from Var2-6 that a single SNP can significantly alter the secondary structure, but covariance analysis predicted the *B. rapa* structure would be maintained. We found that the 5′ domain, including the three-way junction, and stalk were conserved, with covariant base pair flips in the helices and greater variation in the single-stranded regions. Strong sequence conservation of the protein-coding exon (H8 and H9) retains the second r-turn. Similar to the *A. thaliana* Var2-6 isoform and the *A. alpina* and *E. salsugineum* transcripts, H4 was significantly shorter, partially due to an 11 bp deletion disrupting its 5′ side. The 17 bp stem of H4 in Var2-6 could be responsible for its altered behavior and late-flowering phenotype; the *B. rapa* distal *COOLAIR*, with its even shorter stem, may therefore behave more similarly to Var2-6 than Col. *B. rapa* genotypes exhibit a range of morphological and flowering phenotypes; this could be partly a consequence of sequence polymorphism between *COOLAIR* transcripts. We have identified a SNP within H4 between two *B. rapa FLC3* alleles that correlates with differences in *FLC* sense expression and flowering time (Figure S3). Maintenance of even a short H4 preserves the first r-turn, connecting H4 and H5. The multi-way junction is present but contains one less helix (H6) relative to the other species. In addition, H1 and H7 are less stable than they are for other species, and some helices have shifted or changed length. H3 has a large terminal loop and no internal loop, reminiscent of the Var2-6 structure. Potential base pairing between the loop of the multi-way junction (which contains 13 nt with low SHAPE reactivities) and the nucleotides forming the 3′ side of H12 could affect tertiary folding.

We have confirmed that covariation of physically separated regions of the primary *COOLAIR* sequence has maintained *COOLAIR* secondary structures over evolutionary time. The conserved H8/H9 structure, flanked by a robust and complicated secondary structure unit that shows covariance (first r-turn plus multi-way junction) suggests an important functional role, reinforced by the finding that this region associates with *FLC* chromatin in chromatin isolation by RNA purification (ChIRP) experiments (Csorba et al., 2014).

DISCUSSION

COOLAIR is a set of antisense RNAs expressed from the *A. thaliana FLC* locus, different components of which have been shown to regulate expression of *FLC*. To further investigate *COOLAIR* function, we determined the secondary structure of the *COOLAIR* transcripts using chemical probing experiments. The transcripts were found to be highly modular and organized by exon, suggesting a mix-and-match strategy for lncRNA structure that was also observed in the *SRA* and *HOTAIR* lncRNA structures conserved throughout mammals (Novikova et al., 2012; Somarowthu et al., 2015). The first exon of both the proximal and the distal transcripts of *COOLAIR* is shared, while their distinct second exons display a conserved structural core with variations in certain structural elements.

Overall, we find intricate secondary structures (e.g., multi-way junctions, as opposed to single stem loops) to be conserved despite low sequence conservation. Similar phenomena occur in domain IV of *SRA1* across vertebrates (Sanbonmatsu, 2016). Although commonplace in other RNA systems, this may have implications for lncRNAs, a large number of which have been dismissed as non-conserved. The Bartel lab identified lncRNAs of more than 2 kb (megamind and cyrano) that were functionally conserved from zebrafish to human, despite only a 26 nt conserved stretch of sequence (Ulitsky et al., 2011). Likewise, human and mouse local repeats within the mammalian functional intergenic repeating RNA element (FIRRE) lncRNA have only 68% nucleic acid sequence identity and yet share protein-binding functions (Hacisuleyman et al., 2016), while orthologs of the *Drosophila melanogaster* RNA-on-the-X (roX) system have low sequence homology but conserved structure and function (Quinn et al., 2016). We have shown that the 3S method finds conserved secondary structures when faced with lncRNAs containing short patches of conserved sequence surrounded by regions with much lower sequence conservation. In light of the large number of such low sequence identity syntenic lncRNAs recently identified, this approach might be useful for other systems (Hezroni et al., 2015).

COOLAIR exons largely correspond to non-coding sequences from the sense strand and are relatively poorly conserved by sequence in evolutionarily distant plant relatives. We characterized *COOLAIR* from a range of species within the family Brassicaceae, using the Weinberg et al. (2007) strategy of experimentally probing an RNA of one species to determine its secondary structure and then using this to find *COOLAIR* in other species. We then validated the RNAs through cloning and chemical structure probing of the most evolutionarily distant species analyzed. In vivo chemical probing will be an essential tool to complement methods used in the present study (Ding et al., 2014). However, in vivo, it is difficult to assign protected bases to RNA helices, because protein binding can give similar protection. While in vivo chemical probing will help to validate these in vitro structures, we emphasize that many in vitro-determined structures have been proved in vivo and in crystallographic studies (Noller and Woese, 1981; Roth and Breaker, 2009). The in vitro secondary structure is also a critical step for cryo electron microscopy (cryo-EM) and crystal structure determination. Modular domains of the in vitro-determined secondary structure of human *SRA1* were validated via binding studies (Arieti et al., 2014; Huet et al., 2014). NMR studies demonstrate that helices H12 and H13 of *SRA1* (also known as structure 7) produce a shift in the ^{15}N transverse relaxation-optimized spectroscopy (TROSY) spectrum of SHARP RRM domain 1, supporting specific interactions between H12-H13 on the RNA and the ribonucleoprotein (RNP)1 and RNP2 motifs on the β sheet surface of the protein (Bilinovich et al., 2014). In addition, the in vitro secondary structure of the well-characterized mammalian *HOTAIR* lncRNA was determined to gain insight into how it functions on a molecular level (Somarowthu et al., 2015).

The conservation of *COOLAIR* structural features, from *A. thaliana* to *B. rapa*, suggests they may be involved in *FLC* regulation. The proximal transcripts are functional in the autonomous pathway mechanism that results in restraint of *FLC* expression. In addition, an R loop formed over the *COOLAIR* promoter represses *COOLAIR* and *FLC* expression (Sun et al., 2013). H1–H3, combined with proximal H4–H6, could be involved in these mechanisms. From our functional (Li et al., 2015) and structural analysis, distal H4 appears to be an important component of the regulation of *FLC* transcription in the warm. Its length and stability are significantly altered by the SNP responsible for the Var2-6 late-flowering phenotype. This helix is also shorter in *A. alpina*, *E. salsugineum*, and *B. rapa*, the more distant species in our study. We propose that the changed functionality of the Var2-6 *COOLAIR* transcript therefore results from the structural changes associated with H4. Identification of the *COOLAIR* interacting protein complex or complexes will help us to determine whether this is correct. H4, plus other structures in the distal *COOLAIR* transcript, including the multi-way junction, may also play a role during vernalization, the process in which prolonged cold epigenetically silences *FLC.* Distal *COOLAIR* associates with the *FLC* locus near the nucleation region, where chromatin modifications switch from an active H3K36me3 state to an inactive H3K27me3 state (Csorba et al., 2014). By analogy, one of the only large RNA crystal structures solved to date (the ribosome) possesses a highly conserved core, along with separate variable structures that allow for adaptation. Further *COOLAIR* studies, including motif deletion and compensatory mutations, will aid in interrogating structure-function relationships, including roles in temperature perception. Identifying *COOLAIR* in more species will allow more iterations of consensus secondary structure refinement.

In summary, the central domain and stalk of *COOLAIR* have withstood evolutionary selection, while the variation in H4 length, linked to trait variation, has varied, potentially allowing adaptation to a changing environment. By solving the in vitro secondary structure of *COOLAIR*, we move a step closer to understanding its role in establishing expression levels of the floral repressor *FLC*. Clarifying the role of *COOLAIR* in monitoring long-term exposure to fluctuating temperatures experienced by plants during winter, and how this function has evolved during adaptation, will provide an important paradigm for lncRNA studies.

EXPERIMENTAL PROCEDURES

RNA Synthesis, Chemical Probing, and Capillary Electrophoresis Analysis

RNA was synthesized using the Standard RNA IVT kit (CELLSCRIPT) for runoff transcription. For SHAPE probing, folded RNA was probed using 1M7. Parallel RNA samples were treated with DMSO as a blank. For CMCT, 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (Sigma-Aldrich) was added to 50 mM. Both were reacted for 5 min at 22°C and precipitated. The modified sites of RNA were analyzed by reverse transcription using site-specific 5′-fluorophore-labeled primers and SuperScript III reverse transcriptase (Life Technologies). The samples, supplemented with the dideoxy terminate sequencing products of Cy3-labeled primer extension, were denatured and loaded on an ABI PRISM 3100-Avant genetic analyzer. Capillary electrophoresis traces will be deposited online in the repository of RNA structure probing (RNA Mapping Database, http://rmdb.stanford.edu) (Cordero et al., 2012).

3S Determination Analysis

In combination with full-length lncRNA analysis, three overlapping fragments covering the *COOLAIR* distal RNA were probed as in Novikova et al. (2013). Modular regions were determined by comparison to the full-length RNA; non-modular regions were searched for long-range interactions.

Conservation and Covariance of the Secondary Structure across Species

A. lyrata MN47, *C. rubella* Monte Gargano, and *E. salsugineum* Pall. *FLC* sequences were obtained from Phytozome, *A. alpina* FJ543377.1 (GenBank: FJ543377.1), and *B. rapa* R018 was obtained from in-house sequencing. Multiple sequence alignments for a conserved ~250 nt region of the sense coding region of *FLC* were used as an initial alignment and improved manually using 3S of *A. thaliana COOLAIR*, according to Weinberg et al. (2007) and Griffiths-Jones (2005). For the consensus structure (Figure 3F), a conservative approach was used. Only WC base pairs and GU wobble base pairs are reported as pairs. They are not defined as base pairs in the consensus structure if any mutation in any of the six species causes a pair to break (i.e., no bar between bases). Covariant base pairs were reported in which at least one base pair flip occurs (WC to WC or GU to UG; we do not count GU-AU, UG-UA, CG-UG, and GC-GU transitions). Only one covariant pair included a GU to UG flip; all others were WC to WC.

RT-PCR Analysis of *COOLAIR* in Three Species

Total RNA was extracted from non-vernalized *A. lyrata* MN47 and *C. rubella* Cr22.5 and from vernalized (for 2 weeks at 4°C) *B. rapa* R018 leaf tissue as in Box et al. (2011). DNA was removed with the TURBO DNA-free kit (Ambion), and RNA was reverse transcribed with SuperScript III (Invitrogen) and gene-specific primers. cDNA was amplified by touchdown PCR using GoTaq DNA Polymerase (Promega), followed by two nested PCRs (Tables S1 and S2). RT-PCR products were gel purified, cloned, and sequenced.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.08.045.

AUTHOR CONTRIBUTIONS

C.D., K.Y.S., I.V.N., S.P.H., and E.J.H. conceived the project. I.V.N. and S.P.H. performed the chemical probing experiments. I.V.N., K.Y.S., S.P.H., and E.J.H. produced the structures. K.Y.S. performed the covariance analysis. E.J.H. conducted the in vivo analysis. All authors discussed the results and wrote and commented on the manuscript.

ACKNOWLEDGMENTS

K.Y.S. acknowledges funding from US DOE LANL LDRD. C.D. acknowledges funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme. Additional funding was provided by the UK Biotechnology and Biological Sciences Research Council (BBSRC) Institute Strategic Program (grant BB/C517633/1) to C.D. and J.A.I. and the BBSRC iCASE Studentship (grant BB/K012371/1) with Bayer CropScience to E.J.H. We thank Julia Questa, Rebecca Bloomer, and Yiliang Ding for useful contributions and editing of the manuscript and Nicola Stacey for providing *C. rubella* leaf tissue. We also thank Peijin Li, Katie Sanbonmatsu, Bette Korber, Zasha Weinberg, and Ron Breaker for useful discussions on the Var2-6 *Arabidopsis* accession, covariance, and phylogenetic analysis.

Received: January 21, 2016
Revised: June 3, 2016
Accepted: August 12, 2016
Published: September 20, 2016

REFERENCES

Arieti, F., Gabus, C., Tambalo, M., Huet, T., Round, A., and Thore, S. (2014). The crystal structure of the Split End protein SHARP adds a new layer of complexity to proteins containing RNA recognition motifs. Nucleic Acids Res. *42*, 6742–6752.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA *107*, 18724–18728.

Bilinovich, S.M., Davis, C.M., Morris, D.L., Ray, L.A., Prokop, J.W., Buchan, G.J., and Leeper, T.C. (2014). The C-terminal domain of SRA1p has a fold more similar to PRP18 than to an RRM and does not directly bind to the *SRA1* RNA STR7 region. J. Mol. Biol *426*, 1753–1765.

Box, M.S., Coustham, V., Dean, C., and Mylne, J.S. (2011). Protocol: a simple phenol-based method for 96-well extraction of high quality RNA from *Arabidopsis*. Plant Methods *7*, 7.

Brown, J.A., Bulkley, D., Wang, J., Valenstein, M.L., Yario, T.A., Steitz, T.A., and Steitz, J.A. (2014). Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. Nat. Struct. Mol. Biol. *21*, 633–640.

Castaings, L., Bergonzi, S., Albani, M.C., Kemi, U., Savolainen, O., and Coupland, G. (2014). Evolutionary conservation of cold-induced antisense RNAs of *FLOWERING LOCUS C* in *Arabidopsis thaliana* perennial relatives. Nat. Commun. *5*, 4457.

Cheah, M.T., Wachter, A., Sudarsan, N., and Breaker, R.R. (2007). Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. Nature *447*, 497–500.

Cordero, P., Lucks, J.B., and Das, R. (2012). An RNA mapping database for curating RNA structure mapping experiments. Bioinformatics *28*, 3006–3008.

Csorba, T., Questa, J.I., Sun, Q., and Dean, C. (2014). Antisense *COOLAIR* mediates the coordinated switching of chromatin states at *FLC* during vernalization. Proc. Natl. Acad. Sci. USA *111*, 16160–16165.

Davidovich, C., Zheng, L., Goodrich, K.J., and Cech, T.R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nat. Struct. Mol. Biol. *20*, 1250–1257.

Deigan, K.E., Li, T.W., Mathews, D.H., and Weeks, K.M. (2009). Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. USA *106*, 97–102.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696–700.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B., Zufall, R.A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol. *5*, 578–590.

Griffiths-Jones, S. (2005). RALEE—RNA alignment editor in Emacs. Bioinformatics *21*, 257–259.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large noncoding RNAs. Nature *482*, 339–346.

Hacisuleyman, E., Shukla, C.J., Weiner, C.L., and Rinn, J.L. (2016). Function and evolution of local repeats in the *Firre* locus. Nat. Commun. *7*, 11021.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. *11*, 1110–1122.

Huet, T., Miannay, F.A., Patton, J.R., and Thore, S. (2014). Steroid receptor RNA activator (SRA) modification by the human pseudouridine synthase 1 (hPus1p): RNA binding, activity, and atomic model. PLoS ONE *9*, e94610.

Ilik, I.A., Quinn, J.J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R.C., Luscombe, N., Backofen, R., et al. (2013). Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in *Drosophila*. Mol. Cell *51*, 156–173.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S., et al. (2013). *Braveheart*, a long noncoding RNA required for cardiovascular lineage commitment. Cell *152*, 570–583.

Klein, D.J., Schmeing, T.M., Moore, P.B., and Steitz, T.A. (2001). The kink-turn: a new RNA secondary structure motif. EMBO J. *20*, 4214–4221.

Koch, M.A., and Kiefer, M. (2005). Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species—*Capsella rubella*, *Arabidopsis lyrata* subsp. *petraea*, and *A. thaliana*. Am. J. Bot. *92*, 761–767.

Li, P., Tao, Z., and Dean, C. (2015). Phenotypic evolution through variation in splicing of the noncoding RNA *COOLAIR*. Genes Dev. *29*, 696–701.

Li, X., Zhang, S., Bai, J., and He, Y. (2016). Tuning growth cycles of *Brassica* crops via natural antisense transcripts of *BrFLC*. Plant Biotechnol. J. *14*, 905–914.

Liu, F., Marquardt, S., Lister, C., Swiezewski, S., and Dean, C. (2010). Targeted 3′ processing of antisense transcripts triggers *Arabidopsis FLC* chromatin silencing. Science *327*, 94–97.

Mandal, M., and Breaker, R.R. (2004). Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol. *5*, 451–463.

Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. (2005). RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). J. Am. Chem. Soc. *127*, 4223–4231.

Montemayor, E.J., Curran, E.C., Liao, H.H., Andrews, K.L., Treba, C.N., Butcher, S.E., and Brow, D.A. (2014). Core structure of the U6 small nuclear ribonucleoprotein at 1.7-Å resolution. Nat. Struct. Mol. Biol. *21*, 544–551.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., and Finn, R.D. (2015). Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. *43*, D130–D137.

Nitsche, A., Rose, D., Fasold, M., Reiche, K., and Stadler, P.F. (2015). Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. RNA *21*, 801–812.

Noller, H.F., and Woese, C.R. (1981). Secondary structure of 16S ribosomal RNA. Science *212*, 403–411.

Novikova, I.V., Hennelly, S.P., and Sanbonmatsu, K.Y. (2012). Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res. *40*, 5034–5051.

Novikova, I.V., Dharap, A., Hennelly, S.P., and Sanbonmatsu, K.Y. (2013). 3S: shotgun secondary structure determination of long non-coding RNAs. Methods *63*, 170–177.

Quinn, J.J., Zhang, Q.C., Georgiev, P., Ilik, I.A., Akhtar, A., and Chang, H.Y. (2016). Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. Genes Dev. *30*, 191–207.

Ren, A., Vušurović, N., Gebetsberger, J., Gao, P., Juen, M., Kreutz, C., Micura, R., and Patel, D.J. (2016). Pistol ribozyme adopts a pseudoknot fold facilitating site-specific in-line cleavage. Nat. Chem. Biol. Published online July 11, 2016. http://dx.doi.org/10.1038/nchembio.2125.

Roth, A., and Breaker, R.R. (2009). The structural and functional diversity of metabolite-binding riboswitches. Annu. Rev. Biochem. *78*, 305–334.

Sanbonmatsu, K.Y. (2016). Towards structural classification of long non-coding RNAs. Biochim. Biophys. Acta *1859*, 41–45.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F., and Pyle, A.M. (2015). HOTAIR forms an intricate and modular secondary structure. Mol. Cell *58*, 353–361.

Sun, Q., Csorba, T., Skourti-Stathaki, K., Proudfoot, N.J., and Dean, C. (2013). R-loop stabilization represses antisense transcription at the *Arabidopsis FLC* locus. Science *340*, 619–621.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

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.

Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., Wang, J.X., Lee, E.R., Block, K.F., Sudarsan, N., et al. (2007). Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res. *35*, 4809–4819.

Weinberg, Z., Perreault, J., Meyer, M.M., and Breaker, R.R. (2009). Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. Nature *462*, 656–659.

Xue, Z., Hennelly, S.P., Doyle, A.T., Gulati, A.A., Novikova, I.V., Sanbonmatsu, K.Y., and Boyer, L.A. (2016). A G-rich motif in the lncRNA Braveheart interacts with a zinc finger transcription factor to specify the cardiovascular lineage. Mol Cell.

Yesselman, J.D., and Das, R. (2015). Modeling small non-canonical RNA motifs with the Rosetta FARFAR Server. bioRxiv, Published online November 11, 2015. http://dx.doi.org/10.1101/031278.