Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix

Jessica A Brown1, David Bulkley2, Jimin Wang1, Max I Valenstein1, Therese A Yario3, Thomas A Steitz1–3 & Joan A Steitz1,3

Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is a highly abundant nuclear long noncoding RNA that promotes malignancy. A 3′-stem-loop structure is predicted to confer stability by engaging a downstream A-rich tract in a triple helix, similar to the expression and nuclear retention element (ENE) from the KSHV polyadenylated nuclear RNA. The 3.1-Å resolution crystal structure of the human MALAT1 ENE and A-rich tract reveals a bipartite triple helix containing stacks of five and four U•A-U triples separated by a C•G-C triplet and C-G doublet, extended by two A-minor interactions. In vivo decay assays indicate that this blunt-ended triple helix, with the 3′ nucleotide in a U•A-U triple, inhibits rapid nuclear RNA decay. Interruption of the triple helix by the C-G doublet induces a ‘helical reset’ that explains why triple-helical stacks longer than six do not occur in nature.

Long noncoding RNAs (lncRNAs) function in myriad cellular processes and are associated with various disease states, including cancer1–12. Human MALAT1 is an abundant ~8-kb lncRNA that is upregulated in multiple cancers9. High nuclear levels of MALAT1, with a half-life up to 15 h, promote tumor growth by affecting proliferation, invasion and metastasis, processes associated with altered gene expression in lung cancer3–9. The enhancement of oncogenic processes by MALAT1 in colorectal cancer cells has been localized to an ~1,500-nt fragment near the 3′ end of MALAT1 with a 3′-terminal A-rich tract predicted to engage in a triple-helical ENE structure, on the basis of its similarity to ENE structures discovered in viral lncRNAs and genomic RNAs13,14.

The ENE near the 3′ end of polyadenylated nuclear (PAN) RNA (Fig. 1a), a lncRNA produced by the Kaposi’s sarcoma–associated herpesvirus (KSHV) during the lytic phase of infection, is the best characterized13,15–17. It protects the 3′ end of PAN RNA from rapid deadenylation-dependent nuclear decay, forming a triple helix by sequestration of PAN’s 3′ poly(A) tail within the U-rich internal loop, denoted ENE+A15,17. Structural components important for robust activity of the PAN ENE+A include (i) a triple helix of five consecutive U•A-U triples (where • and ‐ represent interactions along the Hoogsteen and Watson-Crick faces, respectively); (ii) canonical Watson-Crick base pairs in the duplexes flanking the tripod; and (iii) A-minor interactions with three consecutive G-C base pairs adjacent to the triple (Fig. 1a and Supplementary Fig. 1a)16,17. Similar structural features have been predicted for the MALAT1 ENE+A, including two A-minor interactions; however, the predicted triple helix would be markedly different from that of any viral ENE because G and C nucleotides interrupt the U-rich internal loop (Fig. 1a)11,12,14,17.

We set out to determine the crystal structure of the MALAT1 ENE+A, presented here at 3.1-Å resolution. It forms a bipartite triple helix that sequesters the 3′ end of the RNA within a U•A-U triple, conferring resistance to rapid RNA decay. The U•A-U triple is interrupted by a C•G-C triplet (in which ′ denotes protonation) and an adjacent C-G doublet that induces a helical reset, thus suggesting that successive base triples are limited to a finite length. This ENE structure is a major determinant of MALAT1 stability, and it identifies a potential target for reducing MALAT1 levels in cancer cells.

RESULTS
An ENE facilitates accumulation of MALAT1 lncRNA
We tested whether the ENE is responsible for the high levels of cellular MALAT1 (ref. 3). We deleted an ~80-bp region containing the ENE (deletion denoted ΔENE) from plasmids expressing full-length (~6.9-kb) mouse MALAT1 (mMALAT1) or an ~2-kb sequence from the 3′ end of human MALAT1 (Fig. 1b,c). Mouse and human MALAT1 ENE+A sequences are ~90% identical (Supplementary Table 1). Upon transient expression in HEK293T cells, transcripts lacking the ENE accumulated to only ~1.5% of the level of wild-type (WT) mouse or human MALAT1 RNA (Fig. 1d,e). A single U-to-C base substitution on the 5′ side of the U-rich internal loop decreased RNA levels to 23% and 15% for mouse U6612C and human U8275C MALAT1, respectively (Fig. 1a–c and Supplementary Table 2). These

1Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA. 2Department of Chemistry, Yale University, New Haven, Connecticut, USA. 3Howard Hughes Medical Institute, Yale University, New Haven, Connecticut, USA. Correspondence should be addressed to J.A.S. (joan.steitz@yale.edu).

Received 25 March; accepted 21 May; published online 22 June 2014; doi:10.1038/nsmb.2844
and an overall coordinate error based on maximum likelihood of 0.34 Å (Fig. 1f, Table 1 and Supplementary Fig. 3c). Data analyses and figures are based on the most ordered RNA molecule A, whose conformation lies between molecules B and C, with electron density visible for all 76 nts (Supplementary Fig. 2b).

**MALAT1 ENE and A-rich tract form a bipartite triple helix**

The MALAT1 ENE+A core assembles into an intramolecular major-groove RNA triple helix formed by the ENE U-rich internal loop and downstream A-rich tract connected by a single-stranded linker (Figs. 1g and 2a). The triple helix is bipartite, composed of two independent stacks of triples: five consecutive U•A-U triples (denoted with 1–5 in parentheses in triple I; Fig. 2a) capped with one C•G-C triple and four consecutive U•A-U triples (denoted with 6–9 in parentheses in triple II; Fig. 2a). The stacks are separated by one C-G doublet. The bases in the C•G-C and most of the U•A-U triples are within hydrogen-bonding distance (≤3.2 Å) on both the Watson-Crick and Hoogsteen faces (Fig. 2b). The 3’ terminus of the molecule is sequestered within U-A-U(9) of triple II. The bipartite triple is flanked by a GG dinucleotide bulge and two double-helical RNA stems (Fig. 2a). G6-C50 and G5-C51 in stem I interact with A65 and A64 of the A-rich tract to form type I and type II A-minor interactions, respectively (Fig. 2c and Supplementary Fig. 4). Together,
these double- and triple-helical units assemble into a nearly straight rod-like structure similar to the KSHV PAN ENE core structure with oligo(A)₉ (Supplementary Fig. 1)¹⁷.

Although all ten major-groove triples maintain a similar overall layout (Fig. 2b), the observed lengths for potential hydrogen bonds between the 2'‐OH of the Hoogsteen strand and phosphate oxygen (O₂P more often than O₁P) of the A-rich tract or Watson strand vary (Fig. 2d and Supplementary Table 3). For triplexes I and II, the 2'‐OH‐O₂P distances gradually decrease from U•A•U(3) to U•A•U(5) (5.1 to 2.5 Å) and from U•A•U(6) to U•A•U(9) (4.1 to 2.6 Å), creating a zipper-like hydrogen-bonding pattern. These 2'‐OH‐O₂P hydrogen bonds probably stabilize the highly electronegative RNA triple-stranded structure by minimizing potential electrostatic clashes between the O₂ and O₄' of the Hoogsteen strand and O₂P of the Watson strand (Supplementary Table 3)¹⁸,¹⁹. Importantly, an abrupt change in the 2'‐OH‐O₂P distance occurs at the internal triplex junction, thus implicating the C-G doublet in a unique structural role.

Table 1 Data collection and refinement statistics

|                        | Native                  | Iridium derivative | Peak          | Infection     | Remote      |
|------------------------|-------------------------|--------------------|---------------|---------------|-------------|
| **Data collection**    |                         |                    |               |               |             |
| Space group            | P₃₂₁                   | P₃₂₁               | 164.9, 164.9, 64.3 |
| Cell dimensions        |                         |                    | 162.8, 162.8, 65.9 |
| a, b, c (Å)            | 90.0, 90.0, 120.0       | 90.0, 90.0, 120.0  |
| **Resolution (Å)**     | 50–3.1 (3.18–3.10)      | 50–3.4 (3.49–3.40) |
| Rmerge                 | 9.6 (239.7)             | 8.5 (368)          |
| I/σI                   | 17.2 (1.28)             | 15.1 (0.73)        |
| Completeness (%)       | 99.7 (99.8)             | 100 (100)          |
| Redundancy             | 11.2 (11.3)             | 8.7 (8.9)          |
| **Refinement**         |                         |                    |               |               |             |
| Resolution (Å)         | 50–3.1 (I/σI = 1.28)    |                    |
| No. reflections        | 17,565                  | 26,893             |
| Rwork / Rfree          | 22.0 / 25.5             | 26,879             |
| No. atoms              | Protein: 7,072          | 100 (100)          |
|                        | Water: 35               | 100 (100)          |
| B factors              | Protein: 54.8           | 54.8               |
|                        | Water: 31.7             | 31.7               |
| r.m.s. deviations      | Bond lengths (Å)        | 0.0097             |
|                        | Bond angles (°)         | 1.03               |

Data sets were collected from one native crystal and one iridium-soaked crystal.

Values in parentheses are for highest-resolution shell.

Figure 2 Hydrogen-bonding interactions in the triple helix of the MALAT1 ENE+A core RNA. (a) Schematic diagram of the MALAT1 ENE+A core structure, using the notation for hydrogen bonds defined in Figure 1a. The major structural regions are labeled on the right in black; the U•A•U triples are numbered on the left in blue. (b) Potential hydrogen-bonding interactions (≤3.2 Å) for two different major-groove base triples, U•A•U(3) and C+•G-C, shown with blue dashed lines. (c) Stick representation of A-minor interactions with hydrogen bonds represented by blue dashed lines. The A-minor interactions mediated by A65 and A64 are of type I and type II, respectively. (d) Potential hydrogen bonds (blue dashed lines) between the 2'-OH (green spheres) of the Hoogsteen strands and O2P (purple spheres) of the Watson strands (A-rich tract), shown in a close-up view for U11•A70 and C12•G71 (left) and in a view down the helical axis for all triples numbered in a (right). Distances too long for hydrogen-bonding are indicated by black dashed lines.
MALAT1 ENE+A triple helix is interrupted by a C*•G-C and C-G

The disruption of the U•A-U triplex by a central C*•G-C triplet and C-G doublet in the MALAT1 ENE+A is absent from all known viral ENE+A structures. An overlay of C*•G-C, C-G and U•A-U(6) with U•A-U(2–4) revealed that the Hoogsteen strands are structurally different (Fig. 3a). At the C-G doublet, a 1-nt gap in the Hoogsteen strand shifts the backbone inward by ~3 Å, positioning the Hoogsteen C*12 and U13 within hydrogen-bonding distance of the Watson strand. Furthermore, base-stacking interactions are approximately four-fold greater for the internal C*•G-C and C-G than for any U•A-U stack because both the Watson and Crick (3’ side of U-rich loop) strands participate rather than predominantly the Watson strand, for example, A68 and A69 from U•A-U(3) and U•A-U(4) (Fig. 3b). At the C-G doublet, the helical axes of triplexes I and II change their relative orientation by ~15° (Fig. 3c). Likewise, we observed reorientation of ~15° and 21° between the axes of stems I and II, respectively, and the adjacent triplex. These changes in the local helical axes correlate with changes in both major- and minor-groove widths near duplex-triplex transitions (Supplementary Fig. 5 and Supplementary Table 4).

Thus it appears that the C-G doublet enables a ‘reset’ of the helical axis. Because a nucleotide is missing in the Hoogsteen strand,
Figure 5 The MALAT1 ENE+A exhibits a single phase of RNA decay in vivo. (a) Schematic diagrams shown for the βα1,2 constructs containing a tetracycline-responsive promoter (TRP) and bovine growth hormone poly(A) signal (BGH PA). Transcripts containing the mascRNA segment (orange) undergo 3’-end processing via RNase P cleavage, whereas the other transcripts undergo cleavage and polyadenylation. The resulting ENE (green) structures engage a genomically encoded A-rich tract or the poly(A) tail (purple), respectively. (b) Schematic diagrams of the KSHV PAN ENE and MALAT1 ENE structures tested in the decay assays. Hydrogen-bonding interactions are as defined in Figure 1a. (c) Representative northern blots (out of three experiments) probed for β-globin and 7SL RNAs showing the amount of mRNA remaining at different times after the pulse for each construct: TRP-βα1,2-MALAT1 ENE+A+mascRNA WT (green), TRP-βα1,2-MALAT1 ENE+A+mascRNA U8275C (blue), TRP-βα1,2 (black), TRP-βα1,2-KSHV PAN ENE (purple) and TRP-βα1,2-KSHV PAN ENE+A+mascRNA with C•G-C (red). Uncropped blot images are in Supplementary Figure 7. (d) Northern blot data quantitated by normalization of the β-globin signal to the 7SL signal, which serves as a loading control. The end of the transcriptional pulse is labeled as time zero and set at 100%. Values are the average of three biological replicates (independent experiments) ± s.d. Curve colors correspond to the constructs in c. The inset of each plot shows the first 2 h to emphasize the biphasic nature of the curves for βα1,2 (black) and βα1,2-KSHV PAN ENE (purple) transcripts.

we hypothesize that the Hoogsteen strand can interact with the Watson-Crick duplex to form a finite number of successive major-groove triples.

Predicted steric clash between Hoogsteen and Watson strands

A literature survey reveals that experimentally determined tertiary structures of naturally occurring RNA triple helices contain three to six consecutively stacked major-groove triples (Supplementary Table 5)20. Thus, helical reset by the C-G doublet may be a mechanism to circumvent destabilizing interactions occurring with more than six triples. To test this possibility, we computationally extended the Hoogsteen and Watson strands by superimposing a nucleotide at positions n – 2 and n – 1 onto a nucleotide at positions n – 1 and n, respectively, to advance the nucleotide at position n into the extended position n + 1 (for example, U10 overlaid onto U11) for each of the strands in both triple I and II (Fig. 4). This analysis revealed steric clashes between the 2’-OH of the Hoogsteen strands and O2P of the Watson strands in positions C12 + 1 and G71 + 1 for triplex I and U16 + 1 and A76 + 1 for triplex II (Fig. 4a and Supplementary Table 3). Thus, depending on the specific sequence and structural context, successive base triples of more than four to six are destabilizing because of steric hindrance.

MALAT1 ENE+A structure inhibits the rapid phase of RNA decay

Next, we asked whether the MALAT1 ENE+A counteracts the rapid phase of RNA decay similarly to the ENE+A structure of KSHV PAN RNA15–17. We determined the rapid phase of RNA decay for a doxycycline-inducible, introns β-globin transcript (βα1,2) containing the MALAT1 ENE+A+mascRNA sequence in its 3’ untranslated region (Fig. 5a,b). After a transcriptional pulse of 2 h, the transcript (Fig. 5c,d) underwent a single phase of decay with a half-life (t1/2) of 3.4 h (Table 2). Interestingly, the U8275C mutant form of βα1,2-MALAT1 ENE+A+mascRNA, which accumulated to ~15% of WT levels under steady-state conditions (Fig. 1e)13, exhibited a single, but somewhat faster (t1/2 = 1.3 h), phase of decay (Table 2 and Fig. 5d). These profiles sharply contrast with the biphasic decay of βα1,2 and βα1,2-KSHV PAN ENE transcripts (Fig. 5d)15, which undergo rapid (t1/2 = ~7–12 m) followed by slow (t1/2 = ~3–4 h) decay but differ in the percentage of the transcript population undergoing rapid decay (39% for the βα1,2 versus 23% for the βα1,2-KSHV PAN ENE transcript) (Fig. 5 and Table 2).

The distinctly different decay profiles for the KSHV PAN and MALAT1 ENE+A suggest different 3’→5’ exonucleolytic mechanisms. We tested whether the different 3’-end structures of the substrates, a poly(A) overhang for the KSHV PAN ENE+A versus a blunt end for the MALAT1 ENE+A, are responsible by inserting the KSHV PAN ENE upstream of an A-rich tract and mascRNA sequence.

### Table 2 Calculated half-lives and percentage of transcripts rapidly degraded

| Construct | Fast (%) | t1/2 (trans.) | t1/2 (slow) | t1/2 (single) |
|-----------|----------|---------------|-------------|--------------|
| TRP-βα1,2 | 39 ± 8   | 6.5 ± 2.7     | 2.8 ± 0.9   | –            |
| TRP-βα1,2-KSHV PAN ENE | 23 ± 2 | 12 ± 6 | 4.4 ± 0.7 | –            |
| TRP-βα1,2-KSHV PAN | –       | –             | –           | 3.5 ± 0.2    |
| ENE+A+mascRNA | –       | –             | –           | 3.4 ± 0.6    |
| TRP-βα1,2-MALAT1 ENE+A+mascRNA (C•G-C) | – | – | – | 1.3 ± 0.1 |

Values are an average of three biological replicates (independent experiments) ± s.d.
Figure 6 Accumulation levels of βΔ1,2-MALAT1 ENE+A+mascRNA containing mutations in the C•G-C and G-C nucleotides. (a) Schematic of the βΔ1,2-MALAT1 ENE+A+mascRNA construct containing a CMV promoter and BGH pA. Arrowhead denotes 3′-end processing via RNase P cleavage of the transcript containing the MALAT1 ENE (green), A-rich tract (purple) and mascRNA (orange). (b) Schematic of the MALAT1 ENE+A structure with interactions indicated as in Figure 1a. The blue and green boxes highlight mutation sites. (c) Northern blots (top) probed for β-globin and Neomycin resistance (NeoR) mRNAs. Black, WT sequence; red, mutated nucleotides; gray, nucleotide deletion. Bottom, quantification of results by normalisation of the β-globin signal to the NeoR signal, which served as a loading and transfection control. The WT βΔ1,2-MALAT1 ENE+A+mascRNA reporter level was set at an arbitrary value of 1. Relative accumulation is the average of five biological replicates (independent experiments); error bars, s.d. Uncropped blot images are in Supplementary Figure 7. (d) Chemical structures showing the hydrogen-bonding interactions of three different bases triples: U•A-U, C•G-C and U•G-C.

(EP-R), βΔ1,2-KSHV PAN ENE+A+mascRNA C•G-C, which should yield a blunt-ended structure locked into register by the C•G-C triple substitution (Fig. 5b). Indeed, the βΔ1,2-KSHV PAN ENE+A+mascRNA C•G-C transcript decayed with a single slow phase (τ1/2 = 3.5 h), almost identical to that of the βΔ1,2-MALAT1 ENE+A+mascRNA WT transcript (Fig. 5d and Table 2). We conclude that the 3′-blunt-ended MALAT1 ENE+A structure effectively abolishes rapid RNA decay in vivo.

Sequence requirements for MALAT1 ENE+A function

We probing the contributions of the C-G doublet and C•G-C triplet in preventing RNA decay by mutating the βΔ1,2-MALAT1 ENE+A+mascRNA reporter (Fig. 6a,b). First, we replaced the C-G doublet with a G-G mismatch or the three other Watson-Crick base pairs. G-G dramatically reduced stabilization, thus suggesting a requirement for base-pairing at this position (Fig. 6c, lanes 2 and 3). The Watson-Crick base-pair replacement showed relative accumulation levels of C-G (WT) > U-A > G-C = A-U (Fig. 6c, lanes 2 and 4-6), revealing higher accumulation for a pyrimidine-purine pair than purine-pyrimidine. We also inverted the C•G-C and C-G to C-G and C•G-C, and these dropped reporter accumulation to ~39% of WT levels (Fig. 6c, lane 7). Relocating the C•G-C and C-G between U-A-U(2) and U-A-U(3) lowered reporter accumulation to ~76% of WT levels (Fig. 6c, lane 8). We conclude that a pyrimidine-purine arrangement is more effective than the converse for stabilization and that the location of the doublet within the MALAT1 ENE+A triplex affects reporter accumulation.

The C-G doublet stabilizes a protonated C•G-C in vivo

The C•G-C base triple in the MALAT1 ENE+A structure raises the question of whether the Hoogsteen base C8273 is protonated in vivo. Protonation would increase triplex stability by promoting hydrogen-bond formation between N3 of cytosine and N7 of guanine (Fig. 6d). Our X-ray structure shows that the N3 of C12 and N7 of G71 are within hydrogen-bonding distance at 2.7 Å (Fig. 2b). For a protonated C•G-C, two hydrogen bonds are predicted along the Hoogsteen face, whereas the unprotonated C-G triple would form one hydrogen bond, as in a U-G-C triple (Fig. 6d). Thus, we tested a U-G-C mutant in the βΔ1,2-MALAT1 ENE+A+mascRNA reporter and found ~69% accumulation relative to WT (Fig. 6c, lanes 2 and 9). This moderate decrease with the loss of one hydrogen bond contrasts with reporter levels of ≤51% that we observed previously after complete loss of hydrogen-bonding interactions along the Hoogsteen-Watson interface, using the same β-globin reporter system.

In vitro studies of intramolecular DNA tripe helices composed of C-G-C and T-A-T triples by UV melting, iron-affinity cleavage and DNase I footprinting assays have found that adjacent C-G-C triples are destabilizing at neutral pH21-23. In the MALAT1 ENE+A structure, the C-G doublet could stabilize C•G-C by increasing solvent accessibility to satisfy the positive charge or by eliminating electrostatic interference arising from the presence of a third nucleotide. Therefore, we created several βΔ1,2-MALAT1 ENE+A+mascRNA mutants that either delete (Δ) or substitute the C-G doublet with the putative triples C•G-C, U•A-U, C-G-C or U-G-C. Removing the C-G doublet decreased reporter accumulation to ~63% relative to WT (Fig. 6c, lanes 2 and 10), thus indicating that the doublet stabilizes the C•G-C triple more than a U-A-U triple. Likewise, accumulation was reduced to ~34% and ~66% relative to WT when the C-G doublet was replaced with C-G-C and U-A-U, respectively (Fig. 6c, lanes 2, 11 and 12). Because lowered accumulation may be partially due to the loss of the C-G doublet, which confers greater stability than A-U or G-C (Fig. 6c, lanes 2, 4 and 5), we examined reporters with C-G-C and U-G-C triples in the C-G position. We observed ~56% and 88% levels relative to WT, respectively (Fig. 6c, lanes 2, 13 and 14). Thus, the stabilization activity of the MALAT1 ENE+A is lower when the C•G-C triple is adjacent to a C rather than a U in the Hoogsteen strand, in agreement with previous in vitro studies of DNA triplexes21-23. We conclude that the C-G doublet stabilizes the C•G-C triple in the MALAT1 ENE+A in vivo, so that the transcript achieves maximal accumulation when the Hoogsteen C′ base is in a non-base-stacking position.

DISCUSSION

We have determined the structural basis for the stability of the IncRNA MALAT1: its 3′ end forms a triple-helical ENE+A structure whose deletion or mutation greatly reduces MALAT1 levels (Fig. 1). The triple helix involves the U-rich internal loop of the MALAT1 ENE and a genomically encoded downstream A-rich tract interrupted by a C•G-C triplex and C-G doublet (Fig. 1a,g). The C•G-C and C-G were not predicted by computer-generated modeling of a minimal mMALAT1 ENE+A, which proposed a U•A-U triplex interrupted by C-G and C-G.
How the MALAT1 ENE+A counteracts decay

Despite their structural similarity, the MALAT1 and KSHV PAN ENE+A differentially counteract decay in vivo, exhibiting monophasic and biphasic decay, respectively (Fig. 5 and Supplementary Fig. 1). The KSHV PAN RNA exhibits a rapid followed by a slow phase of decay, with the ENE+A structure reducing the percentage of transcripts undergoing rapid decay. Here, we discovered that the MALAT1 ENE+A eliminates the rapid decay phase, instead exhibiting a single phase with a $t_{1/2}$ similar to the slow phase of PAN ENE+A decay (Fig. 5 and Table 2). Perhaps the MALAT1 ENE+A is more effective at preventing rapid 3′→5′ exonucleolytic decay because its 3′-A is engaged in a U-A-U triple to form a blunt-ended structure whose register is determined by the GC dinucleotide, G8350 and C8351, in the A-rich tract of the triple helix (Fig. 6 and Supplementary Fig. 1).

The blunt end of the MALAT1 triplex appears to be a poor substrate for exonucleases compared to the 3′-poly(A) overhang of KSHV PAN RNA (Fig. 5). In a previous study, when an A$_2$ overhang was added onto the 3′ end of the βA1,2-MALAT1 ENE+A+masRNA reporter, transition levels decreased to ~23% relative to WT, indicating accelerated decay. The internal GC in the A-rich tract of the MALAT1 ENE+A may directly slow the rate of deadenylation because purified deadenylases, poly(A)-specific RNase and CCR4, prefer A-containing over G-, C- and U-containing substrates in vitro. On the basis of βA1,2-MALAT1 ENE+A+masRNA reporter assays with Watson-Crick base pairs A-U, G-C, U-A and C-G at the C-G doublet position, the relative substrate preference appears to be A = G > U > C (Fig. 6c, lanes 2 and 4–6). Studies using 3′ rapid amplification of cDNA ends on extracts from cells transfected with the mMALAT1 ENE+A at the 3′ end of a GFP reporter have suggested that the A-rich tract is degraded up to the GC, and this is followed by oligouridylation.

We propose that the MALAT1 ENE+A interferes with RNA decay by reducing enzyme binding to the 3′ end and inhibiting exonucleolytic activity upon encountering the GC dinucleotide. Enzymes targeting RNA 3′ ends may act by engaging directly in 3′→5′ degradation or by synthesizing a single-stranded tail (oligouridylation or polyadenylation) to initiate exonucleolytic decay. The major- and minor-groove widths of the ENE+A are similar to those of ideal A-form double-stranded RNA (Supplementary Fig. 5 and Supplementary Table 4) and thus are suitable for binding decay enzymes or other proteins. The same stabilization mechanism is probably used by the multiple endocrine neoplasia β (MENβ) IncRNA because its proposed 3′-triple-helical ENE+A structure also contains a C•G-C and C-G interruption (Supplementary Table 5).

The MALAT1 ENE+A is further stabilized by strong base-stacking interactions provided by the optimal C•G-C and C-G configuration and the additional hydrogen bond within the apparent C•G-C triple (Fig. 6). We argue that the C•G-C triple is protonated in vivo because the $pK_a$ of an internal C•G-C triple in DNA triplexes is greater than 7, and accumulation of the βA1,2-MALAT1 ENE+A+masRNA reporter is lower when the C•G-C is adjacent to C rather than U in the Hoogsteen strand (Fig. 6c, lanes 2 and 9–14, and refs. 28–30). These conclusions are consistent with studies of DNA triplexes, wherein destabilization is proposed to result from electrostatic repulsion of adjacent positive charges at the N3 position and deprotonation or partial protonation because of a change in the local $pK_a$ to yield fewer hydrogen bonds in a C-G-C versus T-A-T triple.

Implications for RNA triple-helical structures

With ten major-groove triplets, the MALAT1 ENE+A is the most elaborate naturally occurring triple-helical RNA structure known to date (Supplementary Table 5). Its unique bipartite triple helix uses a C-G doublet to reset the helical axis, maintaining alignment between the Hoogsteen and Watson-Crick strands (Figs. 2d and 4a). The helical reset appears necessary; several βA1,2-MALAT1 ENE+A+masRNA reporter mutants designed to form more than six successive canonical base triplets all exhibited decreased accumulation in vivo (Fig. 6c, lanes 8 and 10–12). We speculate that such MALAT1 ENE+A structures are more susceptible to decay because hydrogen bonds along the Hoogsteen face are disrupted. Furthermore, relocating or deleting the C-G doublet, the major determinant of the helical reset, reduces accumulation of the βA1,2-MALAT1 ENE+A+masRNA reporter transcript (Fig. 6c, lanes 2, 7, 8 and 10). The helical reset may also depend on the C•G-C triple because a βA1,2-MALAT1 ENE+A+masRNA reporter with a C•G-C to U-A-U substitution exhibits accumulation of ~60% relative to WT.

Our studies of the MALAT1 ENE+A triplex suggest that triple helices are restricted to a finite length because extension induces a steric clash between the Hoogsteen and Watson strands. We explored whether this structural feature might be common to known RNA triplexes and found that superposition analysis likewise predicted a steric clash between the extended Hoogsteen (U12 + 1) and Watson (A9 + 1) strands in the KSHV PAN ENE core and oligo(A)$_9$ structure (Fig. 4b, Supplementary Table 3 and ref. 17). In contrast, the extended Hoogsteen and Watson strands for the PreQ1-II riboswitch, SAM-II riboswitch, human telomerase and Kluyveromyces lactis telomerase (representative example in Fig. 4c) structures revealed that the 2′-OH-O2P groups at the n + 1 positions do not clash; instead, the bases undergo gross misalignment leading to suboptimal hydrogen-bonding along the Hoogsteen face (Supplementary Table 6 and refs. 31–34). Variations in strand incompatibility may arise from dissimilarities among these triple-helical structures. Superposition analyses of the MALAT1 ENE+A triplex I and other known triplexes yielded r.m.s. deviation values greater than 1 Å except for the KSHV PAN ENE+A (Supplementary Fig. 6 and Supplementary Table 6). Yet the distances between the 2′-OH and O2P groups exhibit a widening-to-narrowing trend for all known RNA triplexes (Supplementary Table 6). However, only the MALAT1 and KSHV PAN ENE+A triplexes (r.m.s. deviation <1 Å), which favor a ‘zipped’ state with more than 50% of the triples poised to form a hydrogen bond between the 2′-OH and O2P groups, undergo a steric clash upon computational strand extension. Other triplexes (r.m.s. deviation >1 Å), which favor an ‘unzipped’ or non-hydrogen-bonded state, show strand misalignment (Fig. 4 and Supplementary Tables 3 and 6). These findings suggest that stacked base triples are restricted to a finite length of three to six, owing to the difficulty of accommodating the third strand.

Unlike base pairs that can form double helices of unlimited length, the irregular geometrical conformation of RNA base triples (Fig. 2 and Supplementary Table 3) indicates that a helical reset is required to correct for the geometrical misalignment and steric clashes that accumulate in consecutively stacked base triples. Such discontinuities may explain why attempts to crystallize DNA or RNA triplexes...
METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 4PLX.

ACKNOWLEDGMENTS

We are grateful for staff assistance at Advanced Photon Source beamline 24-ID and National Synchrotron Light Source beamline X-25, plasmids from K. Prasanth (University of Illinois, Urbana-Champaign, pSV40-mMALAT1) and A. Alexandrov (Yale University, ABA2136) and iodide (III) hexamine trichloride from S. Strobel (Yale University). We thank P. Moore, K. Tycowski and J. Withers for critical review of the manuscript, A. Micheletto for editorial work and all Steitz-laboratory members for thoughtful discussions. This work was supported by US National Institutes of Health grants GM026154 (J.A.S.) and GM022778 (T.A.S.), a laboratory members for thoughtful discussions. This work was supported by US National Institutes of Health grants GM026154 (J.A.S.) and GM022778 (T.A.S.), a Postdoctoral Fellowship (grant 122267-PF-12-077-01-RMC) from the American Cancer Society (J.A.B.) and the Steitz Center for Structural Biology, Gwaguri Institute of Science and Technology, Republic of Korea (J.W.). J.A.S. and T.A.S. are supported as investigators of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

J.A.B. and J.A.S. designed research; J.A.B., D.B., M.L.V. and T.A.Y. performed research; J.A.B., D.B. and J.W. analyzed data; T.A.S. and J.A.S. oversaw research; J.A.B. and J.A.S. wrote the paper; and all authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Qiu, M.T., Hu, J.W., Yin, R. & Xu, L. Long noncoding RNA: an emerging paradigm of cancer research. Tumour Biol. 34, 613–620 (2013).
2. Batista, P.J. & Chang, H.Y. Long noncoding RNAs: cellular address codes in development and disease. Cell 152, 1208–1207 (2013).
3. Gutschner, T., Hammerle, M. & Diederichs, S. MALAT1: a paradigm for long noncoding RNA that supports as investigators of the Howard Hughes Medical Institute.

4. Schmidt, L.H. et al. Upregulated MALAT1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. Mol. Biol. Cell 8, 2289–2294 (2012).
5. Gutschner, T. et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 73, 1180–1189 (2013).
ONLINE METHODS

Plasmids and mutagenesis. The full-length mouse MALAT1 (mMALAT1, nts 1–6982) sequence was generated by PCR with the pSV40-mMALAT1 vector9, a kind gift from K. Prasanth (University of Illinois, Urbana-Champaign), as a template. A fragment of human MALAT1 (MALAT1, nts 6681–8708 from accession No. 002819.2) was created by PCR with human genomic DNA as a template. The mMALAT1 and MALAT1 PCR fragments were subsequently inserted into the XhoI and NotI sites of AVA2136 to generate pCMV-mMALAT1 and pCMV-MALAT1, respectively, by standard molecular biology techniques. AVA2136 is a minimal vector (A. Alexandrov, Yale University) that contains a CMV promoter, a short multiple cloning site, an origin of replication and an ampicillin-resistance gene. QuikChange (Agilent) site-directed mutagenesis was used per the manufacturer's protocol to create mutant forms of pCMV-mMALAT1 (AENE with nts 6591–6672 deleted and U6612c) and pCMV-MALAT1 (AENE with nts 8254–8336 deleted and U8275C). For crystallization studies, the pHDV-MALAT1 ENE+A core (5′-GGAAGGTTTTTCCCTTCTGGGCGAAG TCTCAGGTGTTGCTTTTTGGCCTTTCTTAAAAAAAAAHAAAAAGCAAAA-3′) was created by site-directed mutagenesis to replace the CMV promoter with a tetracycline-responsive promoter (TRP) in the pCMV-Jα1.2 constructs. All mutant forms of pCMV-Jα1.2-MALAT1 ENE+A-rich tract+maskRNA and pTRP-Jα1.2-MALAT1 ENE+A-rich tract+maskRNA were cotransfected into HEK293T cells with pmaxGFP and performed as described for the β-globin reporter assays with the following modifications after RNA isolation: RNA was resolved on a 1% agarose/6.5% formaldehyde gel and 5′-32P-labeled oligonucleotides were used to detect mMALAT1 (5′-TGCCCTCCAAAGTGGTAGAT-3′, 5′-GACTATCGTCCTGCTAGACCCGTTG-3′, 5′-CTGGTCCTGTTGACCTGAG-3′ and 5′-TTTCCTGGAAAGCTGGGGAAAA-3′ (ref. 41)), MALAT1 (5′-GACTTGG AGATGCAGCTTCCGCCAGATGCTAGTGTGGCC AAGTCTGTTATGTT CACC-3′) and GFP (5′-CTGACTTCTCGATGCGGTTTGTTG-3′) on the northern blot. Original images of northern blots used in this study can be found in Supplementary Figure 7.

RNA preparation and crystallization. The 76-nt MALAT1 ENE+A core RNA was prepared from the pHDV-MALAT1 ENE+A core plasmid template, which was linearized with HindIII and transcribed by T7 RNA polymerase; RNA products were gel-purified and exchanged into crystallization buffer (5 mM sodium cacodylate, pH 6.5, 50 mM KCl, 1 mM MgCl2 and 0.1 mM EDTA) as described previously37. RNA (8–mg/ml) was heated at 95 °C for 3 min, snap-cooled on ice for 10 min and allowed to equilibrate at room temperature for at least 1 h before preparation of crystal trays. Crystals grew at 20 °C with the sitting-drop vapor-diffusion method; folded RNA was combined with an equal volume of the reservoir solution (50 mM sodium cacodylate, pH 6.5, 18 mM MgCl2, 2.5 mM spermine and 9% isopropanol). Crystals (final size ~300 μm × 300 μm × 50 μm) appeared within 2 to 14 d and were stabilized by addition of increasing amounts of 2-methyl-2,4-pentanediol (MPD) to a final concentration of 30% with a nonlinear-regression program (KalediaGraph Software), and half-lives (t1/2) were calculated by entering the appropriate rate constant (k, kfast or kslow) from equation (1) or (2) into the equation t1/2 = ln2/k. The RNA population undergoing rapid decay was extrapolated directly from the amplitude, Afast, in equation (2). Curves were fit to the double-exponential equation when the residuals showed a substantial improvement compared to the residuals for a single-exponential fit.
41. Tripathi, V. et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell 39, 925–938 (2010).
42. Walker, S.C., Avis, J.M. & Conn, G.L. General plasmids for producing RNA in vitro transcripts with homogeneous ends. Nucleic Acids Res. 31, e82 (2003).
43. Lykke-Andersen, J., Shu, M.D. & Steitz, J.A. Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. Cell 103, 1121–1131 (2000).
44. Kabsch, W. Xds. Acta Crystallogr. D Biol. Crystallogr. 66, 125–132 (2010).
45. Sheldrick, G.M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. Acta Crystallogr. D Biol. Crystallogr. 66, 479–485 (2010).
46. Winn, M.D. et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242 (2011).
47. DeLaBarre, B. & Brunger, A.T. Considerations for the refinement of low-resolution crystal structures. Acta Crystallogr. D Biol. Crystallogr. 62, 923–932 (2006).
48. Cowtan, K. DM: an automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM newsletter on Protein Crystallography 31, 34–38 (1994).
49. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132 (2004).
50. Vagin, A.A. et al. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr. D Biol. Crystallogr. 60, 2184–2195 (2004).