BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 178–186
doi:10.1093/bioinformatics/btu643

Genome analysis

Advance Access publication September 26, 2014

LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis

Sha He[1,†], Hai Zhang[2,†], Haihua Liu[1,†] and Hao Zhu[1,*]

[1]Bioinformatics Section, School of Basic Medical Sciences and [2]Network Center, Southern Medical University, Guangzhou 510515, China

Associate Editor: John Hancock

ABSTRACT

Motivation: In mammalian cells, many genes are silenced by genome methylation. DNA methyltransferases and polycomb repressive complexes, which both lack sequence-specific DNA-binding motifs, are recruited by long non-coding RNA (lncRNA) to specific genomic sites to methylate DNA and chromatin. Increasing evidence indicates that many lncRNAs contain DNA-binding motifs that can bind to DNA by forming RNA:DNA triplexes. The identification of lncRNA DNA-binding motifs and binding sites is essential for deciphering lncRNA functions and correct and erroneous genome methylation; however, such identification is challenging because lncRNAs may contain thousands of nucleotides. No computational analysis of typical lncRNAs has been reported. Here, we report a computational method and program (*LongTarget*) to predict lncRNA DNA-binding motifs and binding sites. We used this program to analyse multiple antisense lncRNAs, including those that control well-known imprinting clusters, and obtained results agreeing with experimental observations and epigenetic marks. These results suggest that it is feasible to predict many lncRNA DNA-binding motifs and binding sites genome-wide.

Availability and implementation: Website of *LongTarget*: lncrna.smu.edu.cn, or contact: hao.zhu@ymail.com.

Contact: zhuhao@smu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 1, 2014; revised on September 21, 2014; accepted on September 22, 2014

1 INTRODUCTION

In mammalian cells (especially differentiated somatic cells), many genes are accurately silenced by DNA and chromatin methylation via DNA methyltransferases (DNMTs) and polycomb repressive complexes (PRCs) (Lee, 2009). By contrast, in cancer cells, genomic regions are often erroneously methylated or demethylated (Gupta *et al.*, 2010). Mammals have very limited DNMT and PRC proteins, which lack sequence-specific DNA-binding motifs (Simon and Kingston, 2013). Many long non-coding RNAs (lncRNAs) have DNA- and protein-binding motifs, allowing them to carry DNMTs and PRCs to specific genomic sites (Lee, 2009; Tsai *et al.*, 2010; Zhao *et al.*, 2008), and lncRNAs can bind to DNA by forming RNA:DNA triplexes. However, details of how lncRNAs bind to DNA targets are scarce.

Single-stranded RNAs that are rich in pyrimidines (T and C) can bind to DNA strands that are rich in purines (A and G) by Hoogsteen-type hydrogen binding to form RNA:DNA triplexes (Morgan and Wells, 1968). *In vitro* studies have revealed at least four types of RNA:DNA binding: (i) a pyrimidine-rich RNA strand binds to a polypurine DNA strand by Hoogsteen base-pairing; (ii) a purine-rich RNA strand binds to a polypurine DNA strand by reverse Hoogsteen base-pairing; and (iii and iv) an RNA strand that is rich in both purines and pyrimidines binds to a polypurine DNA strand by either Hoogsteen or reverse Hoogsteen base-pairing (Duca *et al.*, 2008). Triplexes, however, occur widely between RNA:RNA and RNA:DNA. A recent study indicates that there should be additional non-canonical situations of Hoogsteen and reverse Hoogsteen base-pairing (Schmitz *et al.*, 2010). Given the huge number (>14 000) of human lncRNAs (Derrien *et al.*, 2012), the experimental identification of their DNA-binding motifs and binding sites is not feasible.

A computational program was reported to predict RNA:DNA binding based on the above four canonical base-pairing rules and was used to analyse a large set of short (mostly 30–50 bp), chromatin-associated RNAs (Buske *et al.*, 2012). Compared with such short non-coding RNAs, the prediction of lncRNA DNA-binding motifs and binding sites is complicated by two factors and is much more challenging. First, lncRNAs may contain thousands of nucleotides and multiple functional motifs with featured sequences. Second, a considerable number of lncRNAs, such as HOTAIR (Rinn *et al.*, 2007), function *in trans* and regulate remote genome methylation, for which binding to their target regions is also controlled by CTCF (Phillips and Corces, 2009). Here, we introduce a computational method and program (*LongTarget*) to predict lncRNA DNA-binding motifs and binding sites. We used this program to analyse multiple antisense lncRNAs, including some well-characterized lncRNAs that control gene imprinting. The predicted binding motifs and binding sites exhibit significant sensitivity and specificity and agree with experimental observations. These lncRNAs bind not only to promoter regions and CpG sites but also to many transposable elements.

2 METHODS

2.1 Base-pairing rule-sets

In addition to the four canonical base-pairings (Duca *et al.*, 2008), additional Hoogsteen and reverse Hoogsteen base-pairing have been

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

reported, making it necessary to inspect all of the potential base-pairing rules. For example, an RNA strand can bind to a C-rich (instead of purine-rich) DNA strand (Schmitz *et al.*, 2010). Despite being enriched in two nucleotides bound by nucleotides (such as T and C) in binding motifs, most binding sites contain all four types of nucleotides (A, T, G and C). To accurately examine diverse base-pairing between binding motifs in lncRNAs and binding sites in DNA sequences, we collected from intensive literature reviews potential base-pairing rules (Supplementary Table S1) and combined them into 24 rule-sets (Supplementary Table S2). A rule-set, such as 'TA-U, AT-G, CG-G, GC-U', that accurately determines how ATGC in a binding site is bound by U and G in a binding motif, describes RNA:DNA binding more accurately than do two canonical rules such as 'TA-U, CG-G'.

2.2 RNA:DNA triplexes

For a genomic region of interest (Fig. 1A), *LongTarget* constructs two RNA strands for its plus and minus strand on each base-pairing rule-set because a lncRNA may bind to either strand of the DNA duplex (Fig. 1B). Thus, the predicted binding motifs and their binding sites (which are specifically referred to as TFO and TTS hereafter, i.e. triplex-forming oligonucleotides and triplex target sites) as determined by the rule-set in the lncRNA and in the two constructed RNA strands have the same sequences. Because TFOs vary in length but are mutually determined by their TTSs, we use *lalign* in the fasta36 package (Goujon *et al.*, 2010) to align the lncRNA to each constructed RNA strand. *Lalign* aligns a region in the lncRNA to all of the alignable regions in a constructed RNA strand, thus simultaneously identifying all of the TFOs and all of the TTSs of each TFO (Fig. 1C).

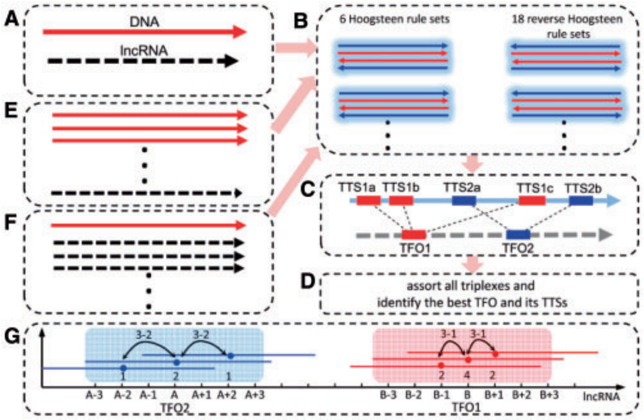
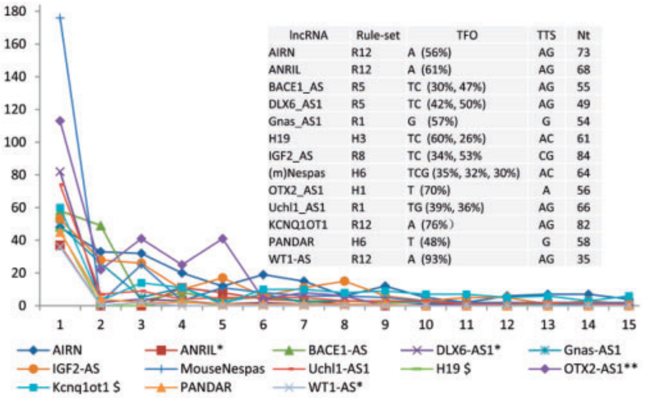

A
DNA
lncRNA
B
6 Hoogsteen rule sets
18 reverse Hoogsteen rule sets
C
TTS1a TTS1b
TTS2a
TTS1c TTS2b
TFO1
TFO2
D
assort all triplexes and identify the best TFO and its TTSs
E
F
G
3-2
3-2
1
2
1
3-1
3-1
2
4
2
A-3 A-2 A-1 A A+1 A+2 A+3
TFO2
B-3 B-2 B-1 B B+1 B+2 B+3
TFO1
lncRNA

Fig. 1. Illustration of the method. (**A**) The input is an lncRNA and a DNA sequence. (**B**) On the 24 rule-sets for the DNA sequence's plus and minus strands, 48 RNA strands (in blue) are constructed. (**C**) The lncRNA is aligned to each constructed RNA strand; the aligned regions in the lncRNA are TFOs and in the constructed RNA strands are TTSs. (**D**) The triplexes that are generated by all of the TFOs and all of the rule-sets are assorted to identify TFO1 (the best TFO) and other TFOs. (**E**) The DNA sequence is shuffled, or (**F**) the lncRNA sequence is shuffled, after which the above steps are repeated for each pair of shuffled sequences. (**G**) Details of triplex assorting. All of the triplexes are handled with their addresses in the lncRNA sequence. In the example, within the range of *offset* (= 3), the midpoints (indicated by red dots) of the triplexes (indicated by red lines) at B – 1, B and B + 1 contribute an *overlap* value (= 3 – 1) to neighbours at B, B – 1, B + 1, and B; the midpoints (indicated by blue dots) of triplexes (indicated by blue lines) at A – 2, A and A + 2 contribute an *overlap* value (= 3 – 2) to neighbours at A, A – 2, A + 2 and A (double arrows). After all of the triplexes are processed, the RNA sequences in overlapping triplexes at close positions form consensus sequences. The position with the highest *overlap* value (at B, value = 4) is chosen first as the midpoint of TFO1, and triplexes with their midpoints within the range of *offset* (shadowed in red) are clustered (and removed after) as TFO1's triplexes. Next, TFO2 and its triplexes (shadowed in blue) are determined on the second highest *overlap* value (at A, value = 2) and removed. The remaining TFOs and their triplexes are determined following the same steps

2.3 Identify the best TFO and its TTSs

The existence of >14 000 lncRNAs in humans (Derrien *et al.*, 2012) suggests that each lncRNA should have a fairly specific function and that, if it binds to DNA, it has a specific DNA-binding motif that may bind to many genomic sites (Chu *et al.*, 2011). The binding sites of a transcription factor share a consensus sequence but show sequence variability. We assume that, due to strong binding affinity, a binding motif can bind to its binding sites with a certain amount of mismatches, thereby generating multiple triplexes at a binding site. Of all of the TFOs that are generated by the 24 base-pairing rule-sets, TFO1 (the best TFO) is determined, without a pre-defined length, by the consensus RNA sequence of triplexes that are distributed most densely with a sufficiently high overlap at a particular position in the lncRNA (triplexes, if in varied sequences, do not reasonably indicate a TFO; Fig. 1G). We find that TFO1 generally (but not always) generates much more triplexes than any other TFO (who are visibly false positives, produce much fewer or poorly overlapped triplexes and simply reflect the quality of TFO1; Figs 1G and 2).

To identify the triplexes that are distributed most densely with a sufficiently high overlap at a particular position in a lncRNA, the *LongTarget* processes all of the triplexes that are generated by all of the rule-sets. For any two triplexes with midpoints within the range of the parameter *offset*, the closer their midpoints are to each other, the higher is the *overlap* value that they contribute to each other. After all of the triplexes are processed, the position in the lncRNA with the highest *overlap* value is chosen as the midpoint of TFO1, and the triplexes with midpoints at both sides within the *offset* are clustered as TFO1's triplexes (Fig. 1G). When, very rarely, TFO1 generates slightly fewer yet more densely distributed triplexes than does TFO2, *LongTarget* still treats TFO1 as the best TFO because the densely distributed overlapping triplexes powerfully indicate a consensus sequence (and thus a true binding motif). Statistically, the fewer the other TFOs and their triplexes, the



lncRNA	Rule-set	TFO	TTS	Nt
AIRN	R12	A (56%)	AG	73
ANRIL	R12	A (61%)	AG	68
BACE1_AS	R5	TC (30%, 47%)	AG	55
DLX6_AS1	R5	TC (42%, 50%)	AG	49
Gnas_AS1	R1	G (57%)	G	54
H19	H3	TC (60%, 26%)	AC	61
IGF2_AS	R8	TC (34%, 53%	CG	84
(m)Nespas	H6	TCG (35%, 32%, 30%)	AC	64
OTX2_AS1	H1	T (70%)	A	56
Uchl1_AS1	R1	TG (39%, 36%)	AG	66
KCNQ1OT1	R12	A (76%)	AG	82
PANDAR	H6	T (48%)	G	58
WT1-AS	R12	A (93%)	AG	35

Fig. 2. Numbers of triplexes that are generated by TFO1. In every case, TFO1 generated many more triplexes than TFO2 did. However, in several cases, TFO3 generated slightly more, instead of fewer, triplexes than TFO2 did, but these triplexes were generated by diverse rule-sets without a clear consensus sequence. In the lncRNA names, '*', '**' and '$' indicate 1/10, 1/20 and 1/100 of their triplex numbers, respectively

better TFO1 is (Fig. 2). If TFO1's triplexes are mainly generated by a highly dominant rule-set, the result is more convincing. The maximum number of overlapped triplexes and the enriched distribution of triplexes in a region are two criteria for prediction of binding sites.

2.4 Stability of triplexes

In vitro RNA:DNA binding is relatively stable. Two datasets of triplex stability, on the stability of different triplets (Supplementary Table S1), are constructed. The default dataset is based on thermal stability (Supplementary Table S1), and the other is based on observed instances of RNA triplets (Abu Almakarem *et al.*, 2012; Leontis *et al.*, 2002). *LongTarget* sums each triplet's ability to obtain triplex stability. The thermal stability of triplets with a mismatch or a gap is assumed zero.

Pyrimidine interruptions (especially successive Ts) in a polypurine region reduce the binding stability (Kukreti *et al.*, 1998; Orson *et al.*, 1999). The parameters *TT penalty* and *CC penalty* penalize TT and CC, respectively. By default, –1000 is set for TT so that TT-containing triplexes generate a negative stability value and are filtered out. Because two sub-triplexes that are separated by TT in a long triplex are still reported if their stability exceeds the parameter *mean stability* and the length exceeds the parameter *minimal TFO length*, *TT penalty* = –1000 simulates the effect of TT breaking a lncRNA:DNA bound as experimentally observed. A small *TT penalty* reduces the TT-containing triplexes' stability. The CC is by default not penalized according to experimental findings (Schmitz *et al.*, 2010).

2.5 Sensitivity and specificity

Because lncRNAs may contain thousands of nucleotides, multiple TFOs are usually reported, especially when lncRNAs are examined against multiple rule-sets. Normally, other TFOs except for TFO1 are obvious false positives because they generate much fewer and poorly overlapped triplexes (Fig. 2) and, therefore, do not affect the prediction. To rigorously evaluate the sensitivity and specificity of TFO1 and its TTSs, *LongTarget* shuffles the lncRNA or DNA sequence and repeats the prediction for each pair of shuffled sequences (Fig. 1E and F). This permutation test examines whether a predicted TFO1 and its TTSs would occur by chance. The program *shuffle* at eyegene.ophthy.med.umich.edu/shuffle/was used (with the default parameters 100 simulations, TRUE_RANDOM) to shuffle the lncRNA and DNA sequences without substituting any nucleotide. This program uses a high-quality random number generator with a period greater than 2×10^{18} and performs better than the program *shuffle* at www.bioinformatics.org. The program *Random DNA Sequence* at www.bioinformatics.org was used to generate random DNA sequences (maximally 10 000 bp).

2.6 Input, output and parameters

LongTarget accepts a pair of lncRNA and DNA sequences as input and outputs nine files, including a list of sorted TFOs and their triplexes, a file reporting permutation test result, a file reporting simple statistics of TFO1, TFO2 and TFO3 and six files describing the TTS positions that are generated by TFO1 and TFO2. There are eight parameters (with default values): *TT penalty* (–1000), *CC penalty* (0), *mean stability* (1.0), *alignment identity* (0.6), *minimal TFO length* (20 bp), *sequence shuffle* (not shuffle), *shuffle times* (0) and *offset* (10). The first five parameters are for filtering triplexes on user-defined conditions, *sequence shuffle* and *shuffle times* are for permutation test, and *offset* constrains distances between the triplex midpoints to control their overlap for sorting triplexes (Fig. 1G). The default parameters *-n -r +5/-4 -f 12 -g 4 -E 10 -m 0* for *lalign* are used. *-n, -r* (+5/-4), *-f* (12), and *–g* (4) indicate nucleotide sequences, substitution scoring matrices, gap open penalty and gap extension penalty, respectively; *-E* controls the maximum number of times that the match is expected to occur by chance; and *-m* controls the output format (www.ebi.ac.uk/Tools/psa/lalign/).

2.7 Gene and genome sequences

For a lncRNA that has multiple transcripts, we used the transcript that contains all of the exons as input, or when such a transcript was lacking, we assembled all of the exons into a sequence as the input. To obtain lncRNA target regions in humans and mice, the human genome hg19 and the mouse genome mm9 were used. Unless specifically mentioned, the lncRNAs and DNA sequences are in the human genome.

3 RESULTS

There are two classes of lncRNAs. One class locates near and antisense to the genes they regulate; a typical example is ANRIL (CDKN2B-AS), which regulates the expression of CDKN2A/ARF/CDKN2B (Kotake *et al.*, 2011). An antisense lncRNA may silence nearby genes if its transcript fully or partially overlaps with these genes, but many lncRNAs, including those controlling imprinted gene clusters, function as long-range *cis*-silencers to regulate non-overlapping genes (Pauler *et al.*, 2012). The second class comprises lncRNAs that are located far from their target genes; the best known is HOTAIR, which is located between HOXC11 and HOXC12 on chromosome 12 and regulates the expression of multiple HOXD genes on chromosome 2 (Rinn *et al.*, 2007). Compared with antisense lncRNAs, the targets of lncRNAs that function *in trans* are poorly understood. The tertiary structures play an important role in target site recognition for some lncRNAs (Engreitz *et al.*, 2013). At this stage, we used *LongTarget* to analyse multiple antisense lncRNAs and obtained results agreeing with experimental observations (Supplementary Results).

3.1 ANRIL

ANRIL, which was initially identified in a melanoma-neural system tumour (Pasmant *et al.*, 2007), regulates the expression of the three cyclin-dependent kinase inhibitors CDKN2A (INK4a/p16), ARF (p14) and CDKN2B (INK4b/p15) and is involved in many tumourigenetic processes (Kotake *et al.*, 2011; Yu *et al.*, 2008). ANRIL has 19 exons and is unique in that it contains 10 transposable elements and exhibits two-stage and clade-specific evolution (He *et al.*, 2013). ANRIL coincides with higher levels of occupancy of EZH2, CBX7 and SUZ12 near the promoter region of CDKN2A (Supplementary Fig. S1; Kotake *et al.*, 2011; Yap *et al.*, 2010), indicating that ANRIL binds to and densely distributes at CDKN2A's promoter region. What remain unknown are ANRIL's DNA-binding motif and binding rules.

First, to predict ANRIL's binding motif and binding sites, we examined the large genomic region at chr9:21900000–22140000 covering CDKN2A/ARF/CDKN2B. The 24 base-pairing rule-sets generated 1236 triplexes that meet the default conditions of '*mean stability*>1.0, *identity*>0.6, *minimal TFO length*>20'. Because only 2 triplexes of the 411 triplexes that were <40 bp were generated by TFO1, we further filtered the triplexes on the condition '*minimal TFO length*>40' and obtained 825 triplexes. Of the 825 triplexes, TFO1 generated 367 triplexes, of which 231 were generated by the dominant rule-set R12. The 367 triplexes

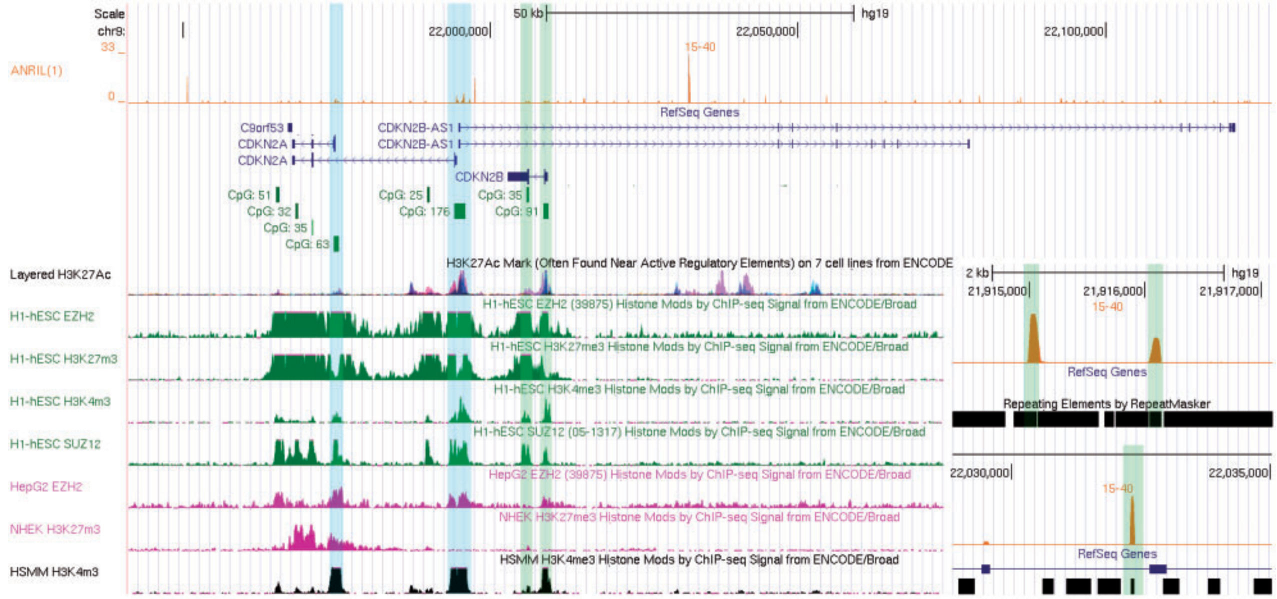


Fig. 3. The distribution of ANRIL's TTSs in the region of CDKN2A/ARF/CDKN2B. Hereafter, TTS distributions are shown as custom tracks of the UCSC Genome Browser. In this and the following figures, right of the lncRNA name, '(1)' indicates TFO1, and the number above '0' indicates the maximal number of overlapped triplexes at an address in the region. In the lncRNA track, the peaks show triplexes, the shadowed blue and green bars mark TTSs at the promoter regions and other sites, and the top central numbers are values of *offset* and *minimal TFO length* (separated by '-'). The others are tracks in the UCSC Genome Browser, including signals of ENCODE Histone Modification (Ram *et al.*, 2011). The two insets in the right-bottom corner show three TTSs with dense triplexes at transposable elements

were, on average, 68 bp and contained an average of 61% A (adenine). By contrast, TFO2 and TFO3 generated only eight and six triplexes, respectively (Fig. 2). TFO1's triplexes were densely distributed at several TTSs in CDKN2A's promoter regions (Fig. 3), agreeing with the experimentally observed H3K27m3 signal and the higher level occupancy of EZH2, CBX7 and SUZ12 at this region (Supplementary Fig. S1) (Kotake *et al.*, 2011; Yap *et al.*, 2010) and with the findings that the methylation of the CDKN2A promoter causes CDKN2A inactivation (Bian *et al.*, 2002; Csepregi *et al.*, 2010). Three TTSs outside the promoter regions and characterized by high peaks fall within transposable elements.

Second, to determine whether the predicted TFO1 (centred at ANRIL's 3403 bp) on the CDKN2A/ARF/CDKN2B region was mainly determined by triplexes at the CDKN2A promoter region, we repeated the prediction using only CDKN2A's promoter region at chr9:21990000–22005000. The same TFO1 and densely distributed triplexes at this region were obtained (Fig. 4A), indicating that these triplexes mainly determined TFO1 in ANRIL.

Third, to evaluate the sensitivity and specificity of TFO1 and its TTSs, we repeated the prediction with ANRIL and the 15 000 bp promoter region that were shuffled 100 times, respectively. When the 15 000 bp promoter region was shuffled, in every case, TFO1 generated few triplexes, indicating the high specificity of the original TFO1's TTSs (Fig. 5A). When ANRIL was shuffled, in most cases, TFO1 generated few triplexes, but in three cases, TFO1 in the 14th, 58th and 65th shuffled ANRIL sequences generated many triplexes (Fig. 5B). Looking at these three cases, we found that the three TFO1s (which were centred at 2465, 3260 and 739 bp) were also highly A-rich, and the dominant rule-set was R12. Moreover, all three of these TFO1s generated triplex distributions that were nearly identical to that generated by the original ANRIL (Fig. 4B–D). Thus, the three putative false positives of TFO1 in shuffled sequences are actually true positives. Two additional rounds of permutation produced similar results (Supplementary Fig. S2). These results indicate that whenever and wherever the A-rich TFO is formed accidentally by shuffling ANRIL, the TFO generates triplexes at the same genomic sites; whenever such an A-rich TFO is not formed, other TFOs fail to generate triplexes at these genomic sites. Thus, only specific A-rich TFOs can generate the experimentally confirmed TTSs at the CDKN2A promoter region.

Fourth, to further evaluate the specificity of ANRIL:TTS binding, we constructed two randomly generated DNA sequences (10 000 bp). At 10 sites with intervals of 1000 bp, we replaced 100 nucleotides with a sequence of 100 bp containing a typical TSS on the minus strand. Running *LongTarget* with ANRIL and the two artificial DNA sequences, we found that ANRIL:DNA binding occurred exactly at the 10 inserted sites on the minus strand (Supplementary Fig. S3), further supporting the specificity of TFO1's TTS.

Finally, we determined whether the parameters of *lalign* would influence the TFO and TTS prediction. The default parameters in the fasta36 package for substitution scoring matrices, gap open penalty and gap extension penalty suit most situations (see Methods); we specifically examined the *-E* parameter. On ANRIL and CDKN2A's promoter region, the examination of

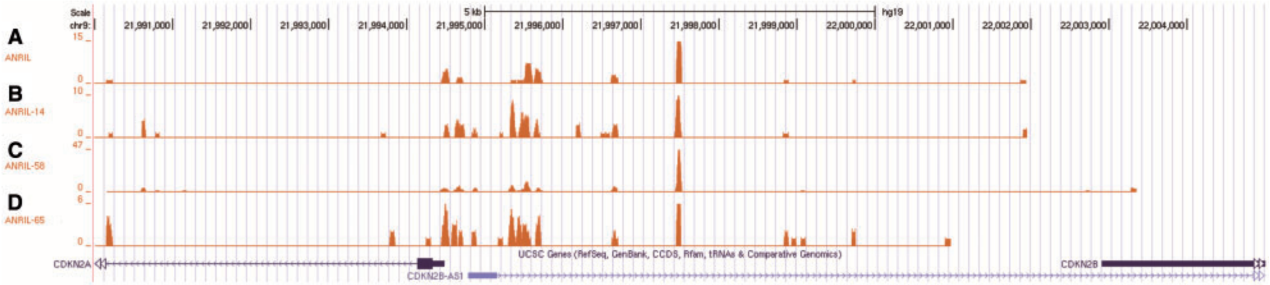


Fig. 4. The distribution of ANRIL's TTSs in CDKN2A's promoter region. The TTSs were generated by the TFO1 in the original ANRIL (**A**) by three TFO1s in three shuffled ANRIL sequences (ANRIL-14: the 14th shuffled sequence, ANRIL-58: the 58th shuffled sequence and ANRIL-65: the 65th shuffled sequence) (**B–D**)

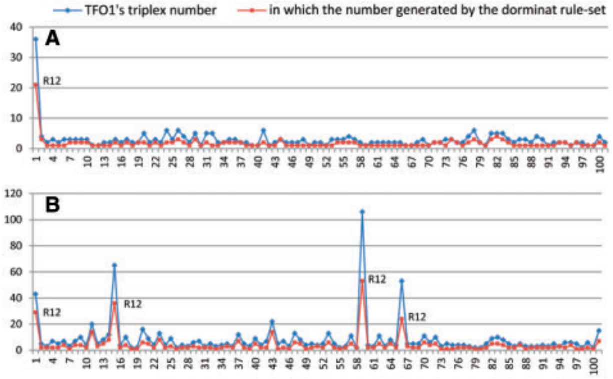


Fig. 5. Sensitivity and specificity of ANRIL's TFO1 and its TTSs. The dots at point 1 denote the original ANRIL sequence + the original CDKN2A promoter [in both (**A**, **B**)], at point 2–101 denote the original ANRIL sequence + 100 shuffled CDKN2A promoter [in (A)] and 100 shuffled ANRIL sequences + the original CDKN2A promoter [in (B)]. At each point the blue and red dots denote the triplex number generated by TFO1, and among which the triplex number generated by rule-set R12. (A) When the CDKN2A promoter was shuffled, in every case TFO1 generated few triplexes. (B) When the ANRIL sequence was shuffled, in three cases (at points 14, 58 and 65) TFO1 generated considerable triplexes, many of which were generated exactly by R12. In other cases TFO1 generated few triplexes

$E = 1$, $E = 10$, $E = 20$ and $E = 50$ indicated that a large E value generated more triplexes (and slightly more TFOs) than did a small E value (Supplementary Fig. S4), and the same E (especially a large one) value may cause slightly different triplex numbers (Fig. 5). Despite these differences, in all the four cases, the same TFO1 was predicted, and its TTSs share the same distribution pattern (Supplementary Fig. S4). The above results suggest that ANRIL's DNA-binding motif and binding sites at the CDKN2A/ARF/CDKN2B region are predicted with satisfactory reliability, sensitivity and specificity.

3.2 H19, Igf2-AS and Airn

LncRNAs regulate the imprinting of many genes (imprinted genes are always expressed from one parental allele). Most imprinted genes are found in clusters that include multiple protein-coding genes and at least one lncRNA (Barlow and Bartolomei, 2014; Ferguson-Smith, 2011). Igf2 is a well-characterized, paternally imprinted gene that is controlled by the maternally expressed lncRNA H19 and a CTCF-binding site (Bell and Felsenfeld, 2000). The human Igf2–H19 locus contains three differentially methylated regions (DMRs): the Igf2 DMR0 at Igf2's exons 2 and 3, the Igf2 DMR2 at Igf2's exons 8 and 9 and the H19 DMR ahead of H19's exon 1 (Supplementary Fig. S6; Boissonnas *et al.*, 2010). Whether and how DMRs are targeted by lncRNAs is unknown.

H19 functions *in cis* to repress the expression of Ins2 and Igf2 (Boissonnas *et al.*, 2010). *LongTarget* predicted that H19's TTSs are located at multiple promoter regions of Igf2 (and of Ins2 and Igf2-AS) (Fig. 6), which agrees well with the experimental findings. Impressively, in addition to Igf2's multiple promoters, H19's TTSs were also predicted largely at the Igf2 DMR0 and the Igf2 DMR2, and the TTSs at the latter position are especially strong (Fig. 6). The Igf2 locus also contains the antisense lncRNA Igf2-AS whose function is unclear (Moore *et al.*, 1997). *LongTarget* predicted that Igf2-AS generates fewer triplexes with a featured distribution at three positions at the Igf2 DMR0, the Igf2 DMR2 and a CpG site, suggesting that Igf2-AS may be involved in the regulation of DMRs instead of the expression of Igf2. If Igf2-AS plays a significant role, it may help explain how the deletion of the H19 alone affects imprinting in only some, but not all, tissues (Barlow and Bartolomei, 2014; Schmidt *et al.*, 1999).

Igf2r (the receptor of Igf2), Slc22a2 and Slc22a3 are imprinted, maternally expressed genes that are controlled by another lncRNA—the imprinted, paternally expressed Airn. Although overlapping only with Igf2r in an antisense orientation, the expression of Airn is correlated with the repression of all the three genes on the paternal allele (Sleutels *et al.*, 2002). Airn accumulates at the Slc22a3 promoter to silence *in cis* Slc22a2 and Slc22a3 by targeting the histone methyltransferase G9a to this region (Nagano *et al.*, 2008). A recent finding suggests that Airn silences the imprinted Slc22a3 and Igf2r genes via different mechanisms (Latos *et al.*, 2012). Agreeing with these findings, *LongTarget* predicted the TTSs of Airn at the promoter regions of Slc22a2, Slc22a3 and Igf2r, overlapping with some CpG sites (Supplementary Fig. S5).

Based on the experimentally uncovered mechanism, that is, the imprinting of Igf2, Igf2r and related genes is controlled by methylation at DMRs and at these genes' promoter regions,

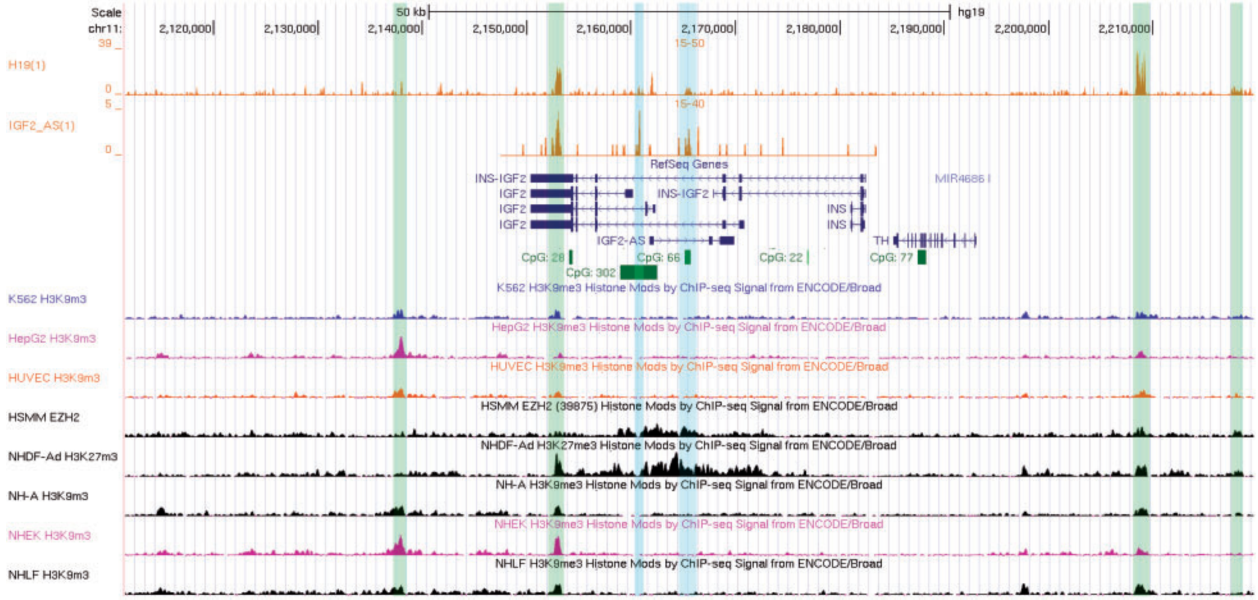


Fig. 6. The distribution of H19's and Igf2-AS's TTSs in the Igf2 imprinting cluster. Two TTSs were at two DMRs, one between Igf2's exon 2 and exon 3 and the other mainly in Igf2's exon 9

we conclude that *LongTarget* predicted the TTSs of H19, Igf2-AS and Airn at reasonable sites, agreeing with the experimentally revealed binding sites of these lncRNAs in the two imprinted clusters.

3.3 Gnas-AS1 (Nespas)

The Gnas cluster is another well-studied case of imprinting genes and includes the protein-coding gene Gnas (Nesp) and the lncRNAs Gnas-AS1 (Nespas). A recent study in mice revealed that when Nespas is present at a low level, the Nesp promoter is unmethylated, indicating that Gnas-AS1 negatively regulates Nesp most likely by interacting with Nesp's promoter (Williamson *et al.*, 2011).

We first predicted human Gnas-AS1 and identified the TTSs exactly at Gnas's multiple promoter regions (Fig. 7). Again, some TTSs impressively match CpG sites and histone methylation marks. This result agrees with experimental findings (Williamson *et al.*, 2011). We next analysed mouse Gnas-AS1. At this cluster in mouse, the TTSs of Gnas-AS were predicted mainly at multiple promoter regions of Gnas, and considerable TTSs matched transposable elements (Supplementary Fig. S7).

3.4 Kcnq1ot1

The extraordinarily long (91 kb) lncRNA Kcnq1ot1 is important for imprinting multiple genes in the large Kcnq1 cluster. Kcnq1ot1 interacts with chromatin, with the H3K9- and H3K27-specific histone methyltransferases G9a and the PRC2 complex (Pandey *et al.*, 2008), and with DNMT (Mohammad *et al.*, 2010). In the large Kcnq1 cluster, impressively, the TTSs of TFO1 are distributed not only at promoter regions of nearby genes, such as Kcnq1, Cdkn1c and Slc22a18, but also at promoter regions of remote genes, such as Ascl2, Tspan32, Cd81 and Phlda2 (Fig. 8; Supplementary Fig. S8), agreeing with experimental findings (Pandey *et al.*, 2008). Moreover, most of these triplexes exceed 70 bp (Fig. 2). The identification of TTSs that were >70 bp and that were at promoter regions of so many imprinted genes strongly indicates that this result is unlikely to have been generated by chance.

Because Kcnq1ot1 is predicted to generate many long triplexes at the promoter regions of eight genes, it is a good example for determining whether the predicted TFO1 and its TTS distribution are specific to parameters (particularly, to the conditions of '*minimal TFO length>50, offset = 15, identity ≥ 0.6*'). We repeated the prediction by separately setting '*minimal TFO length > 70*', '*identity ≥ 0.7*', and '*identity ≥ 0.8*' but maintaining the other two conditions. Under the condition '*minimal TFO length>70*' and the condition '*identity ≥ 0.7*', the same TFO1 and TTS distributions were obtained. Under the condition '*identity ≥ 0.8*', the same TFO1 was obtained, but many TTSs disappeared (Supplementary Fig. S8). We found that the same TFO1 under the conditions '*identity ≤ 0.6*', '*identity ≤ 0.7*' and '*identity ≤ 0.8*' generated 41 625, 5743 and 97 triplexes, respectively. These results demonstrate two indications. First, the abundant relatively shorter triplexes (50 ∼ 70 bp) are not generated randomly at TTSs. As long as *offset* and *minimal TFO length* are within reasonable ranges (not necessarily at the best values), they contribute to identifying the best TFO and its TTSs. Second, setting a very high *identity* does not safely improve the prediction. As EZH2 and CBX7 exhibit higher levels of occupancy at the promoter region of CDKN2A (Yap *et al.*, 2010), dense distributions of triplexes should be more important for correctly identifying TTSs.

4 DISCUSSION

The prediction of lncRNAs' DNA-binding motifs and binding sites is important for uncovering the correct and aberrant genome modification. *LongTarget* is different from the reported

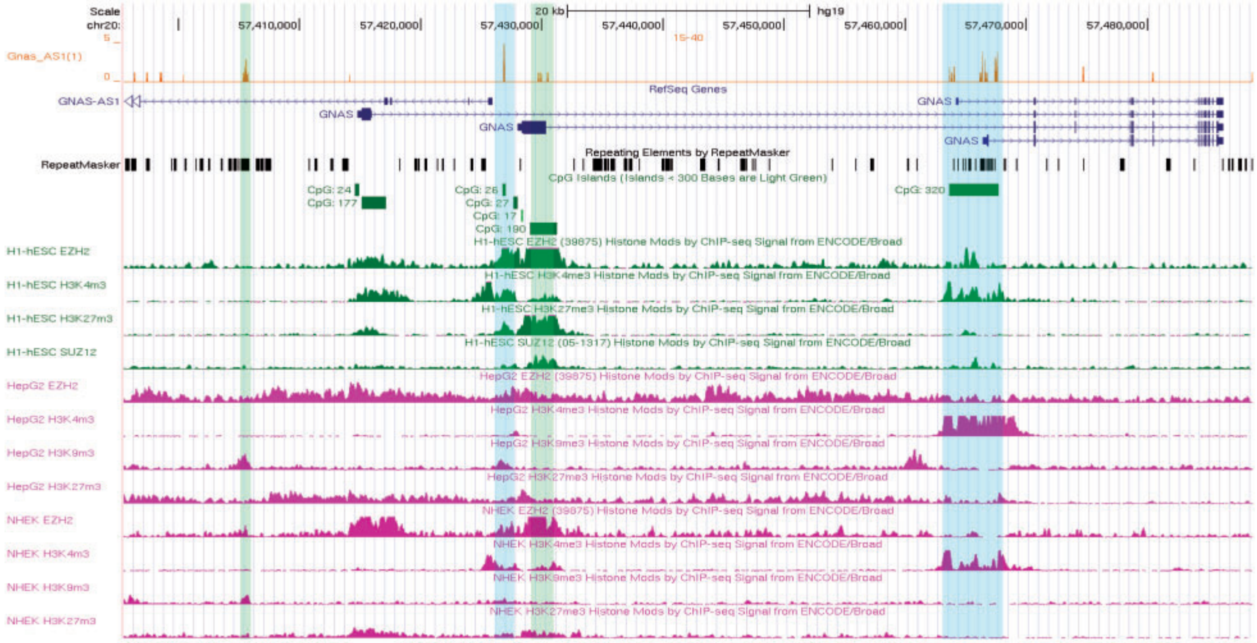


Fig. 7. The distribution of Gnas-AS1's TTSs in the Gnas imprinting cluster. The TTSs were predicted at multiple promoter regions of Gnas

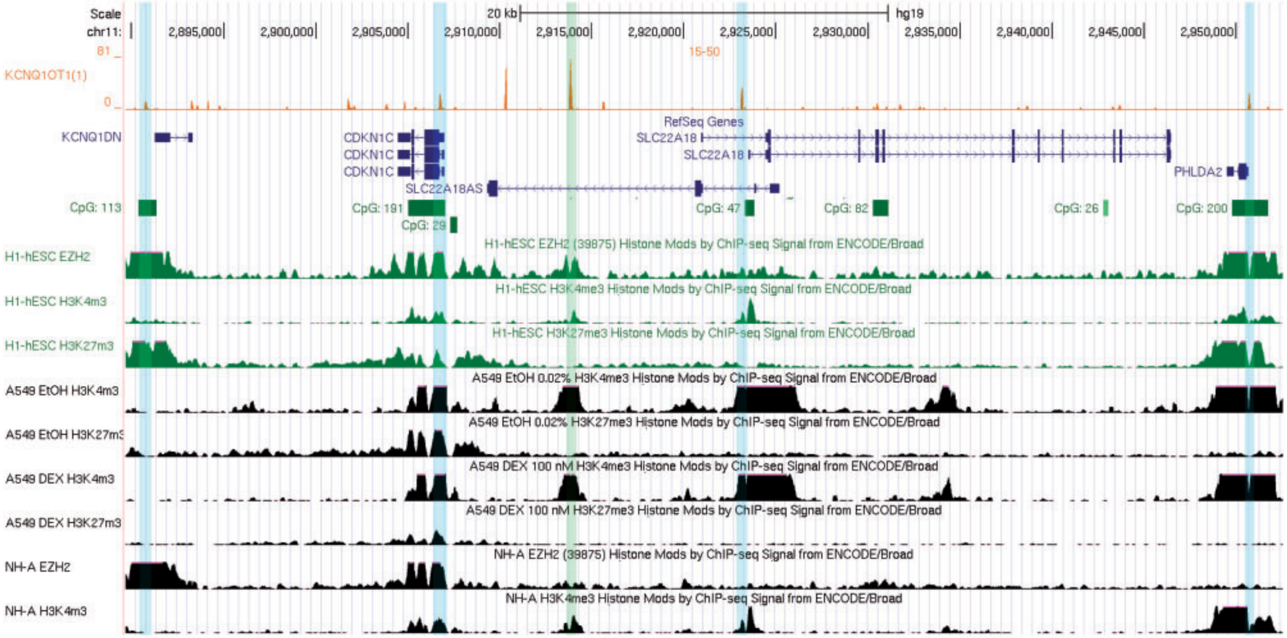


Fig. 8. The distribution of Kcnq1ot1's TTSs in the large Kcnq1 imprinting cluster. TTSs were predicted at the promoter regions of multiple imprinted genes, including Ascl2, Tspan32 and Cd81, which are located far from Kcnq1ot1 (Supplementary Fig. S8). Note that the H3K27me3 signals at this region in the shown datasets are not as strong as the specific report (Pandey *et al.*, 2008)

Triplexator in that it explores a wide range of base-pairing rules (e.g. CG-A and GC-C in the dominant rule-set R12 for multiple lncRNAs) instead of canonical ones, examines base-pairing between all nucleotides instead of enriched ones, determines the single best TFO on the RNA sequences in triplexes that overlap most densely in the lncRNA, determines the best TFO's TTSs upon dense distributions of triplexes in a genome region, and evaluates the triplexes' stability (Buske *et al.*, 2012; Supplementary Table S3; Supplementary Fig. S19). Because binding between ncRNAs and DNA sequences occurs widely

(Schmitz *et al.*, 2010), it is necessary to include non-canonical base-pairing rules. The inclusion of these rules in *LongTarget* is harmless if these rules do not generate significant triplexes but is significant in some case. Because on these rule-sets a nucleotide in a TFO can bind to two different nucleotides in a TTS; TTSs are structurally more complex and informative than are TFOs, which may explain why ANRIL's TTSs are more sensitive to sequence shuffling than its TFO1. The permutation tests of ANRIL and the quality of other lncRNA TFO1 (Fig. 2) indicate that TFO1 and its TTSs have a high sensitivity and specificity. The analysis of these lncRNAs suggests that it is feasible to predict many lncRNA TFOs and TTSs.

Consistent with experimental evidence, most of the predicted TTSs of lncRNAs are at promoter regions of nearby genes and at CpG sites. Examined with signals of ENCODE Histone Modification in the UCSC Genome Browser, these TTSs match well with EZH2, SUZ12 and H3K4m3 signals but poorly with H3K27m3 and H3K9m3 signals in many cells. This poor match can be explained by four reasons. First, the detected H3K27m3 signals may be low in some experiments but high in others (Pandey *et al.*, 2008). Second, lncRNAs, such as Kcnq1ot1, and PRC proteins may have preferred target sites from where they most likely spread to neighbouring imprinted genes (Pandey *et al.*, 2008), making sites of H3K27m3 signals different from and broader than TTSs. Third, the expression levels of lncRNA are low; therefore, the amount of transcripts is likely insufficient for them to bind to the broad H3K27m3 domains. Fourth, in some situations, lncRNAs interact with methylated histone H3 lysine 27 (Yap *et al.*, 2010), suggesting that H3K27m3 signals may exist as a condition for a lncRNA to bind to DNA. It is interesting that many TTSs perfectly match strong H3K4m3 or H3K27Ac signals, which indicate active chromatin and occur in many gene promoter regions, and some TTSs occur at the gap in high level H3K4 and H3K27 signals (Fig. 8, Supplementary Fig. S15). Because in many datasets, both H3K4m3/H3K27Ac signals and H3K9m3/H3K27m3 signals exist at a genomic site, the TTSs at sites with strong H3K4m3 and H3K27Ac signals are sensible, and it is possible that such sites can be competitively bound by lncRNAs of different regulatory functions.

Technically, when predicting a lncRNA TFO1 and TTSs, three issues should be considered. First, the default *minimal TFO length* = 20 is rather small. While this value ensures that no important triplexes would be filtered out, it increases the chance of generating a bad TFO1 because RNA sequences in short triplexes are prone to generate more dense distributions in the lncRNA. A large value (*minimal TFO length*>70) shares the same risk because it filters out too many triplexes. It is advisable to try a small and a large *minimal TFO length* and to determine whether TFO1 is different or whether the number of its triplexes is significantly changed. Normally, *minimal TFO length* = 40 or 50 and *offset* = 10 or 15. Second, we determined whether filtering the triplexes with a large *identity* would improve the quality of TFO1 and found that *identity* = 0.7 may help, but *identity* = 0.8 does not, indicating that lncRNA:DNA triplexes may enjoy a high binding affinity but not accurate base-pairing between all nucleotides. Third, while most lncRNA TFO2s are negligible, some lncRNA TFO2s generate considerable triplexes. Whether some lncRNAs could use two motifs to bind to DNA sequences is an interesting and unresolved issue.

Promoter regions, CpG sites and DMRs are known targets of genome modification (Barlow and Bartolomei, 2014; Ferguson-Smith, 2011); the predicted TTSs of lncRNAs at these positions indicate that they are direct targets of lncRNAs. Transposable elements inserted into the introns of functional genes can reduce the transcription of those genes. Because considerable TTSs are at transposable elements, we postulate that after targeting by lncRNAs, the methylation of these transposable elements may hinder the transcription of genes and that because many lncRNA genes contain multiple transposable elements (He *et al.*, 2013; Kelley and Rinn, 2012), the TTSs at transposable elements in lncRNAs help control the highly tissue-specific expression of lncRNAs. Because most lncRNAs emerge in mammals, beyond the hypothesis that X-inactivation and gene imprinting may have co-evolved in mammals (Reik and Lewis, 2005), imprinting in somatic cells as a genome defence against transposable elements and as a mechanism regulating gene expression may also have an associated origin and evolution.

ACKNOWLEDGEMENTS

We thank Xiaojia Yu, Tao Sun, and Yanfang Guo for their contribution during the early stages.

Funding: This work was supported by the National SuperComputing Center-Guangzhou (2012Y2-00047, 2913Y2-00050).

Conflict of interest: none declared.

REFERENCES

Abu Almakarem,A.S. *et al.* (2012) Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Res.*, **40**, 1407–1423.

Barlow,D.P. and Bartolomei,M.S. (2014) Genomic imprinting in mammals. *Cold Spring Harbor Perspect. Biol.*, **6**, a018382.

Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.

Bian,Y.-S. *et al.* (2002) p16 Inactivation by Methylation of the CDKN2A promoter occurs early during neoplastic progression in Barrett's esophagus. *Gastroenterology*, **122**, 1113–1121.

Boissonnas,C.C. *et al.* (2010) Specific epigenetic alterations of IGF2-H19 locus in spermatozoa from infertile men. *Eur. J. Hum. Genet.*, **18**, 73–80.

Buske,F.A. *et al.* (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res.*, **22**, 1372–1381.

Chu,C. *et al.* (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell*, **44**, 667–678.

Csepregi,A. *et al.* (2010) Promoter methylation of CDKN2A and lack of p16 expression characterize patients with hepatocellular carcinoma. *BMC Cancer*, **10**, 317.

Derrien,T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.

Duca,M. *et al.* (2008) The triple helix: 50 years later, the outcome. *Nucleic Acids Res.*, **36**, 5123–5138.

Engreitz,J.M. *et al.* (2013) The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, **341**, 1237973.

Ferguson-Smith,A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.*, **12**, 565–575.

Goujon,M. *et al.* (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.

Gupta,R.A. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.

He,S. *et al.* (2013) ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. *BMC Evol. Biol.*, **13**, 27.

Kelley,D. and Rinn,J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, **13**, R107.

Kotake,Y. *et al.* (2011) Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene*, **30**, 1956–1962.

Kukreti,S. *et al.* (1998) Triple helices formed at oligopyrimidineoligopurine sequences with base pair inversions: effect of a triplex-specific ligand on stability and selectivity. *Nucleic Acids Res.*, **26**, 2179–2183.

Latos,P.A. *et al.* (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science*, **338**, 1469–1472.

Lee,J.T. (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev.*, **23**, 1831–1842.

Leontis,N.B. *et al.* (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.

Mohammad,F. *et al.* (2010) Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development*, **137**, 2493–2499.

Moore,T. *et al.* (1997) Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. *Proc. Natl Acad. Sci. USA*, **94**, 12509–12514.

Morgan,A.R. and Wells,R.D. (1968) Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotides sequences. *J. Mol. Biol.*, **37**, 63–80.

Nagano,T. *et al.* (2008) The air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science*, **322**, 1717–1720.

Orson,F.M. *et al.* (1999) Triple helix formation: binding avidity of acridine-conjugated AG motif third strands containing natural, modified and surrogate bases opposed to pyrimidine interruptions in a polypurine target. *Nucleic Acids Res.*, **27**, 810–816.

Pandey,R.R. *et al.* (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell*, **32**, 232–246.

Pasmant,E. *et al.* (2007) Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res.*, **67**, 3963–3969.

Pauler,F.M. *et al.* (2012) Mechanisms of long range silencing by imprinted macro non-coding RNAs. *Curr. Opin. Genet. Dev.*, **22**, 283–289.

Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.

Ram,O. *et al.* (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, **147**, 1628–1639.

Reik,W. and Lewis,A. (2005) Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nat. Rev. Genet.*, **6**, 403–410.

Rinn,J.L. *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, **129**, 1311–1323.

Schmidt,J.V. *et al.* (1999) Enhancer competition between H19 and Igf2 does not mediate their imprinting. *Proc. Natl Acad. Sci. USA*, **96**, 9733–9738.

Schmitz,K.-M. *et al.* (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.*, **24**, 2264–2269.

Simon,J.A. and Kingston,R.E. (2013) Occupying chromatin: polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol. Cell*, **49**, 808–824.

Sleutels,F. *et al.* (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810–813.

Tsai,M.C. *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.

Williamson,C.M. *et al.* (2011) Uncoupling antisense-mediated silencing and DNA methylation in the imprinted Gnas cluster. *PLoS Genet.*, **7**, e1001347.

Yap,K.L. *et al.* (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell*, **38**, 662–674.

Yu,W. *et al.* (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature*, **451**, 202–206.

Zhao,J. *et al.* (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, **322**, 750–756.