RESEARCH NOTE

REVISED Purine-rich low complexity regions are potential RNA binding hubs in the human genome [version 2; peer review: 3 approved]

Previously titled: Triplex target sites of MEG3 RNA-chromatin interactions

Ivan Antonov [1,2], Yulia A. Medvedeva[1-3]

[1]Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Moscow, Russian Federation
[2]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russian Federation
[3]Vavilov institute of General Genetics, Russian Academy of Sciences, Moscow, Russian Federation

v2 **First published:** 17 Jan 2018, **7**:76 (https://doi.org/10.12688/f1000research.13522.1)
Latest published: 09 May 2019, **7**:76 (https://doi.org/10.12688/f1000research.13522.2)

Abstract

Many long noncoding RNAs are bound to the chromatin and some of these interactions are mediated by triple helices. It is usually assumed that a transcript can form triplexes with a distinct set of genomic loci also known as triplex target sites (TTSs). Here we performed computational analyses of the TTSs that have been experimentally identified for particular RNAs. To assess the ability of these TTSs to bind other transcripts we developed a method to estimate the statistical significance of the predicted number of triplexes for a given RNA-DNA pair. We demonstrated that each DNA set included a subset of sequences that have a potential to form a statistically significant (adjusted *p*-value < 0.01) number of triplexes with the majority (>90%) of the analyzed transcripts. Due to the predicted ability of these DNA sequences to interact with a wide range of different RNAs, we called them "universal TTSs". While the universal TTSs were quite rare in the human genome (around 0.5%), they were more frequent (>15%) among the MEG3 binding sites (ChOP-seq peaks) and especially among the shared Capture-seq peaks (40%). The universal TTSs were enriched with the purine-rich low complexity regions. Nowadays, the role of the chromatin bound RNAs in the formation of 3D chromatin structure is actively discussed. We speculated that such universal TTSs may contribute to establishing long-distance chromosomal contacts and may facilitate distal enhancer-promoter interactions. All the scripts and the data files related to this study are available at: https://github.com/vanya-antonov/universal_tts

Keywords

MEG3 lncRNA, triple helix, triplex target sites (TTS)

Open Peer Review

Reviewer Status ✓ ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
REVISED **version 2** published 09 May 2019	✓ report		✓ report
version 1 published 17 Jan 2018	✗ report	✓ report	✗ report

1 **Ingrid Grummt**, German Cancer Research Center (DKFZ), Heidelberg, Germany

2 **Andrey A. Mironov**, Moscow State University, Moscow, Russian Federation
RAS (Russian Academy of Sciences), Moscow, Russian Federation

3 **Hao Zhu**, Southern Medical University, Guangzhou, China

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Yulia A. Medvedeva (ju.medvedeva@gmail.com)

Author roles: Antonov I: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Medvedeva YA**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Russian Science Foundation [grant 14-15-30002].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Antonov I and Medvedeva YA. **Purine-rich low complexity regions are potential RNA binding hubs in the human genome [version 2; peer review: 3 approved]** F1000Research 2019, **7**:76 (https://doi.org/10.12688/f1000research.13522.2)

First published: 17 Jan 2018, **7**:76 (https://doi.org/10.12688/f1000research.13522.1)

REVISED **Amendments from Version 1**

As it was suggested by the reviewers, we significantly increased the number of the RNA sequences (from 6 to 306) that were used for the analysis and used the exact locations for all the peaks (instead of 3kb bins that were originally used). Additionally, we developed a probabilistic approach to estimate the statistical significance of each RNA-DNA interaction. Finally, we added the analysis of the recently published data obtained using the Capture-seq experiment.

The data has been expanded and uploaded to Zenodo and GitHub.

We have updated the title of the article to: 'Purine-rich low complexity regions are potential RNA binding hubs in the human genome'.

Nevertheless, with all these changes our main conclusion remained the same. Namely, we proposed that the purine rich low complexity genomic regions may have an ability to interact with various different RNAs.

We also speculate about possible biological role of these special genomic loci.

See referee reports

Introduction

Many human long noncoding RNAs are localized in the nucleus and can potentially participate in chromatin formation and remodeling[1]. Recently, technologies such as ChIRP[2], ChRIP[3], ChOP[4], CHART[5], RAP[6], MARGI[7] and GRID[8] have been developed to map the genomic interacting sites of various lncRNAs. RNA can interact with chromatin by associating with DNA binding proteins, nascent transcripts, single-stranded or double-stranded DNA, forming R-loops or triple helices, respectively. Growing body of evidence shows that RNA-DNA triplex formation based on the Hoogsteen[9] base pairing rules plays important role in RNA-chromatin interactions. Several studies have provided *in vitro* and *in vivo* evidence for the existence and biological relevance of triplexes, including pRNA[10], Fendrr[11], Khps1[12], PARTICLE[13], and MEG3[4].

Computational analyses have revealed that a large population of triplex-forming motifs is present across the human genome with the majority of annotated genes containing at least one triplex target site (TTS), preferentially in regulatory gene regions[14,15]. Considering the large number of purine-rich sequences in the genome, triplex-mediated targeting of lncRNAs and associated proteins to distinct genomic loci is very likely a commonly used mechanism of gene regulation. Still, there are only a few bioinformatic studies of triplex-based RNA-DNA interactions on the genome-wide scale[16].

Here we analyze the genomic regions that are known to interact with the MEG3 lncRNA (the ChOP-seq peaks) or three different short oligos, corresponding to the DNA binding domains (DBDs) of MEG3 and GATA6-AS lncRNAs (the Capture-seq peaks). The current literature usually assumes that the triplex-based interactions have high sequence specificity and each triplex forming oligonucleotide (a transcript region) has a distinct set of genomic binding sites ("triplexome"). We investigate whether all the DNA sites capable of triplex formation are specific enough to be regulated by one particular RNA only or whether different transcripts may have shared TTSs. Our computational analysis revealed a group of genomic regions that may have a very high propensity for triplex formation with a wide range of different RNAs. Therefore, we named such DNA sequences "universal TTSs". We also attempted to reveal the features of these sequences that may be responsible for the observed phenomenon.

Methods

The genomic coordinates of the 6837 MEG3 binding sites[4] (ChOP-seq peaks) were mapped from the hg19 to the hg38 human genome version using liftOver[17] (Nov 7, 2017 version). Next, from the 6800 successfully converted peaks we removed two cases corresponding to the genomic regions with ambiguous base-pairs (N) keeping the 6798 ChOP-seq peaks for the analysis. Additionally, to simulate the genomic background 6798 control regions with the lengths matching the selected ChOP-seq peaks were randomly sampled from the human genome using the bedtools[18] (version 2.27.1, see Supplementary Figure 1).

Triplex-based interactions were predicted by the Triplexator[14] (version 1.3.2) with the following parameters: `-fr off -l 10 -e 10`. These values were optimized so that the tool could predict binding between all three RNA-DNA sequence pairs that have been validated *in vitro* in the original study[4] (Supplementary Table 1). To detect the statistically significant RNA-DNA interactions we developed a method to estimate a p-value from the number of predicted triple helices. Since the MEG3 peaks have different lengths, the expected number of triplexes (i.e. the parameter of the Poisson distribution) is computed based on the lengths of the input RNA and DNA sequences (see below).

The MEG3 binding sites have been identified in the triple negative breast cancer cell line BT-549[4]. To identify all the genes expressed in this cell line we used the RNA-seq data from the control knockdown experiment (ERR652847). The reads were aligned to the human genome (hg38) using HISAT2[19] (version 2.1.0) and the number of reads corresponding to each GENCODE[20] (version 28) transcript was calculated by the HTSeq-count tool[21] (version 0.10.0). Next, the RPKM values were computed as RPKM = $C/(N \times L)$, where C is the number of reads aligned to all the transcript exons, N is the total number of mapped reads (in millions) and L is the transcript length (in kilobases). The most highly expressed (in terms of RPKM) isoform of each gene was considered only. Next, 153 expressed transcripts with the length and GC content similar to the MEG3 lncRNA (NR_002766.2) were selected using the RANN (version 2.6) R package (Supplementary Figure 2). Additionally, 153 random RNA sequences were obtained by di-nucleotide shuffling the original MEG3 transcript using the uShuffle tool[22].

All the heatmaps were generated using the Complex-Heatmap[23] R package. The alignment of the RNA oligo sequences was obtained by the MUSCLE[24] (version 3.8). The locations of the RepeatMasker repeats in the human genome were downloaded from the UCSC Genome Browser[17].

Calculation of the statistical significance of the predicted triplexes

For a given pair of RNA and DNA sequences, Triplexator outputs all the possible triple helices that satisfy the user-defined thresholds. Notably, the number of the predicted triplexes increases with the lengths of input sequences (Supplementary Figure 3A). To account for this dependence the normalized number of triplexes (i.e. the "triplex potential" or t_{pot}) is also computed by the Triplexator. Although this allows to compare triplexes predicted for RNA-DNA pairs with different lengths, it does not provide information about significance of these interactions.

To estimate the probability to observe a particular number of triplexes by chance (e.g. from the random sequences with the same lengths) we analyzed the average number of predicted triplexes between random sequences of various lengths. Namely, we considered four different RNA lengths (L_{RNA} = {500, 1000, 1500, 2000}) and ten different DNA lengths (L_{DNA} = {200, 400, ..., 1800, 2000}). For each of the 40 combinations of (L_{RNA}, L_{DNA}), 100 random sequences with the length of L_{RNA} and 100 random sequences with the length of L_{DNA} were generated (with the equal frequencies for all the four nucleotides).

For each RNA-DNA pair triple helices were predicted by the Triplexator with the parameters optimized for MEG3 lncRNA (`-fr off -l 10 -e 10`). Next, for every combination of (L_{RNA}, L_{DNA}) the average number of predicted triplexes (λ) was computed from all the 10000 predictions (Supplementary Table 2). Finally, a linear regression model for λ was fitted to all the obtained values (adjusted R^2 = 87%, see Supplementary Figure 3B):

$$\lambda(L_{RNA}, L_{DNA}) = \theta_0 + \theta_1 \times L_{RNA} + \theta_2 \times L_{DNA} \qquad (1)$$

where $\theta_0 = -0.688$, $\theta_1 = 5.37 \times 10^{-4}$ and $\theta_2 = 6.03 \times 10^{-4}$.

Thus, the statistical significance of the number of triple helices N_{tpx}, predicted between RNA of length L_{RNA} and DNA of length L_{DNA}, can be estimated as follows. First, the expected average number of predicted triplexes (λ) is computed from the equation (1). Next, the expected distribution of the number of predicted triplexes (H_0) is simulated by the Poisson distribution with the obtained value of λ (Supplementary Figure 3C). Finally, the p-value of the observed number of triple helices (N_{tpx}) can be estimated as follows:

$$\text{P-value}(N_{tpx}) = P(X \geq N_{tpx} | X \sim \text{Pois}(\lambda)) \qquad (2)$$

Importantly, the N_{tpx} value is taken from the "Total (abs)" column of the `triplex_search.summary` file. The same value is used by the Triplexator to compute its "triplex potential" (t_{pot}). The 'Total (abs)' is the total number of *all possible* triplexes that satisfy the user-defined thresholds (overlaps are allowed). Thus, for a single triplex longer than the minimal length (10 in our settings), the 'Total (abs)' value may be greater than 1. For example, 11 bp DNA fragment `5'-GAGAGAGAGAG-3'` and 11 nt RNA oligo `5'-GAGAGAGAGAG-3'` can interact with each other forming one long anti-parallel triplex without mismatches. However, with the minimal allowed triplex length set to 10, the N_{tpx} is equal to 3. This includes the long triplex of length 11 as well as the two triplexes of length 10 without the first or the last position of the long triplex. Therefore, a single long triplex is likely to produce a large N_{tpx} value and, consequently, a statistically significant p-value.

Calculation of purine and poly-purine contents

Due to the properties of the Watson-Crick base pairing model, the GC content of a sequence corresponding to the forward (+) DNA strand is equal to the GC content of the reverse (-) DNA strand. However, the GA content is more important for triplex based interactions because RNA can only form triple helices with the purines in the DNA. In contrast to the GC content, the GA content can be different between the DNA strands. Moreover, if one strand is purine rich, the other strand is automatically purine poor. For example, for the DNA sequence `5'-GGGGGAGA-3'` the purine content of the direct strand is 100%, while the other strand (`3'-CCCCCTCT-5'`) has no purines at all (i.e. its purine content is 0%).

Thus, we define the purine content of a given DNA fragment as the maximum value between the two strands, i.e.:

$$\text{GA-content} = \max_{s=\{+,-\}} \frac{\text{NumPurines}(\text{DNA}^s)}{\text{Length}(\text{DNA})} \qquad (3)$$

where DNA^+ (DNA^-) denotes the sequence corresponding to the forward (reverse) DNA strand and $NumPurines(DNA^s)$ is the total number of G or A nucleotides present in the DNA strand s.

It should be noted that the purine content computed by the formula (3) is always ≥ 50%. To work with a measure that is defined from 0% to 100%, we introduce the *poly-purine content*. We define a poly-purine element P^s as a continuous stretch of 10 or more purines located on the DNA strand s (where s = {+, −}) . For a given DNA sequence that has N^+ poly-purine elements on the forward strand and N^- poly-purine elements on the reverse strand, the poly-purine content is computed as follows:

$$\text{Poly-GA content} = \max_{s=\{+,-\}} \frac{\sum_{i=1}^{N^s} \text{Length}(P_i^s)}{\text{Length}(\text{DNA})} \qquad (4)$$

where $Length(P_i^s)$ is the length of the poly-purine element i on the DNA strand s, i.e. according to the above definition $\text{Length}(P_i^s) \geq 10$.

Results

We used Triplexator to predict possible triple helices between the full length MEG3 transcript and the 6798 experimentally identified ChOP-seq peaks as well as 6798 control DNA regions (see Methods). The genomic sites with a statistically significant number of triplexes were identified in each DNA set by our custom probabilistic approach (see Methods). As anticipated, more statistically significant (Bonferroni adjusted p-value < 0.01) interactions with the MEG3 lncRNA were predicted for the ChOP-seq peaks than for the control regions. Namely, the interactions with the 3825 (56.3%) ChOP-seq peaks were classified as statistically significant (Figure 1A, left) while there were only 617 (9.1%, odds ratio test p-value < 2.2×10^{-16}) such cases among the control DNA regions

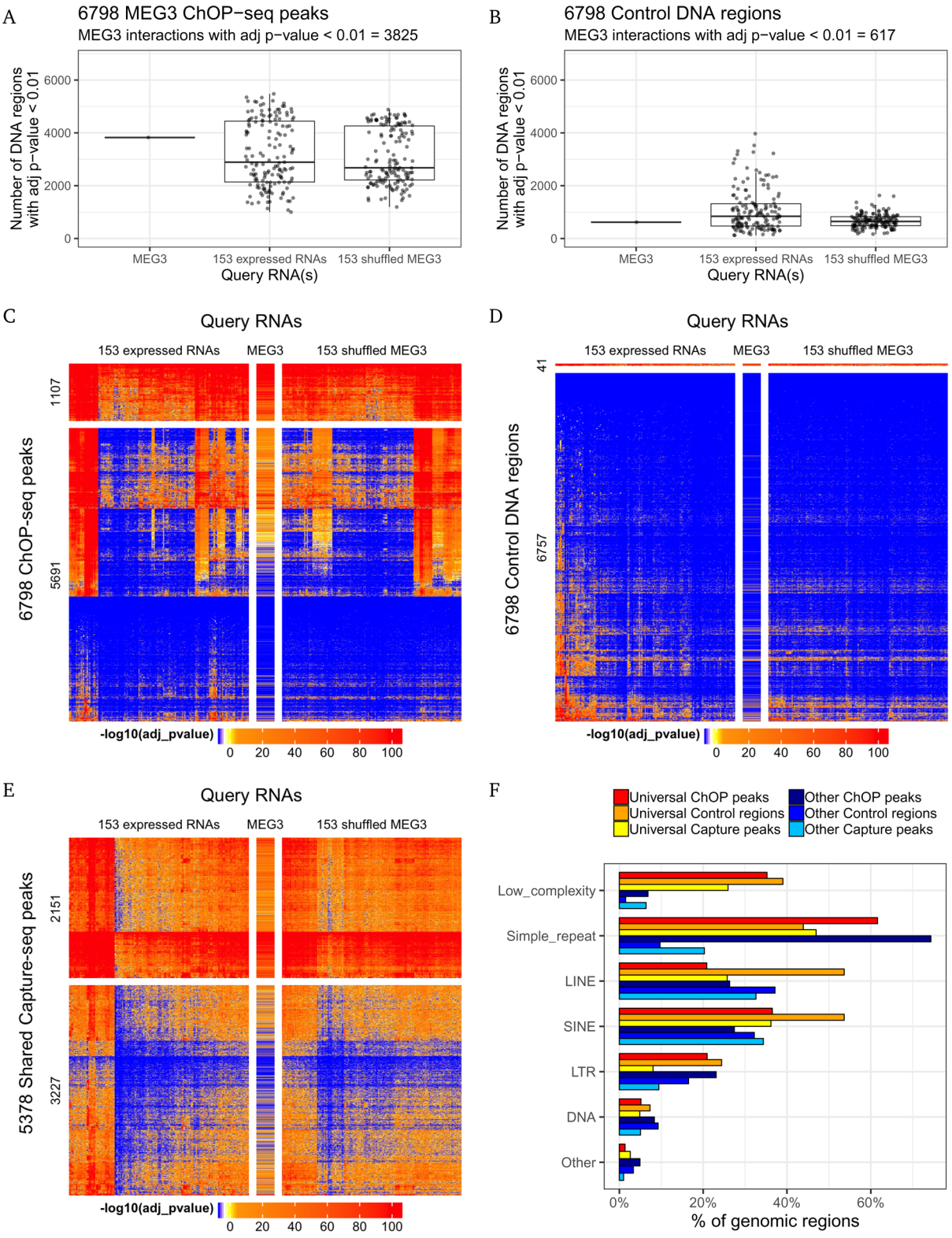

Figure 1. (**A**,**B**) The number of the DNA sequences from (**A**) the ChOP-seq or (**B**) the control genomi set with the statistically significant number of predicted triplexes for different query RNAs (the black dots). (**C**,**D**,**E**) The heat maps of the $-\log_{10}$ (adjusted p-value) corresponding to the predicted triplexes between the 307 different query RNAs (columns) and (**C**) all the ChOP-seq peaks, (**D**) the control genomic sites or (**E**) the Shared Capture-seq peaks (rows). The universal TTSs were identified based on their interactions with the 153 expressed transcripts (left part of each heat map) and visualized as a separate (top) cluster. The MEG3 column was intentionally drawn wider. The blue color corresponds to the RNA-DNA pairs with adjusted p-value = 1 (including cases where no triplexes were predicted). (**F**) Repeat classes present in different sets of genomic regions.

(Figure 1B, left). Since the ChOP-seq method has detected RNA contacts with the chromatin (and not the naked DNA) the obtained binding sites can correspond to several different interaction mechanisms including direct RNA-DNA interactions via triple helices or R-loops, RNA-RNA hybridization with nascent transcripts as well as bindings to nuclear proteins. This may be the reason that many MEG3 binding sites did not produce statistically significant predictions with the MEG3 lncRNA. Therefore, these results supported the original conclusion that the MEG3 lncRNA is able to directly interact with the genomic DNA via triple helices.

To check the ability of other RNAs to form triplexes with MEG3 binding sites, we applied Triplexator to a set of 153 expressed transcripts (see Methods). Surprisingly, 65 analyzed RNAs showed results similar to MEG3 lncRNA – they were predicted to form statistically significant interactions with the majority (> 50%) of the ChOP-seq peaks (Figure 1A, middle). To further investigate possible interactions with the ChOP-seq peaks, 153 artificial sequences were generated by di-nucleotide shuffling of the MEG3 transcript (see Methods). Strikingly, these random "RNAs" produced statistically significant number of triplexes with 39% and 9% of the ChOP-seq peaks and control DNA regions, respectively (Figure 1A, B, right). These results indicated that the set of the ChOP-seq peaks was different from the randomly sampled genomic sites in that it contained a number of DNA sequences that may be able to interact not only with the MEG3 lncRNA, but with other RNAs as well.

Based on these observations we hypothesized that some of the MEG3-bound genomic sites may be 'universal', i.e. they may have a potential to form multiple triplexes with a number of different RNAs. Analysis of the Triplexator predictions obtained for the 153 expressed RNAs revealed 1107 (16.3%) ChOP-seq peaks that were predicted to form statistically significant number of triplexes with more than 90% of the analyzed transcripts (Supplementary Figure 4A). In contrast, the genomic background contained only 41 (0.6%) such sites (Supplementary Figure 4B). Due to the predicted ability of these genomic regions to form triple helices with various RNAs, we called them "universal triplex target sites (TTSs)". Notably, the identified universal TTSs produced strong p-values for the MEG3 lncRNA as well as for the 153 MEG3 shuffled sequences (Figure 1C, D). Thus, according to our predictions some of the DNA sequences were more prone to formation of triple helices with different long RNAs and a number of such genomic regions were present among the experimentally identified MEG3 binding sites (ChOP-seq peaks).

To further investigate the predicted ability of the universal TTSs to bind various RNAs we analyzed the results of the recent Capture-seq experiment[25]. This *in vitro* study has determined genomic binding sites of three different short RNA oligos that corresponded to the DNA-binding domains (DBDs) of the MEG3 and GATA6-AS lncRNAs. Namely, the *MEG3_13_41*, *GATA6_AS_78_118* and *MEG3_839_890* oligos were 28, 40 and 48 nt long, respectively. Since the experiment has been performed on the RNA- and protein-free ("naked") genomic DNA, the majority of the identified interactions have been assumed to be direct and mediated by triple helices. Comparison of the genomic coordinates corresponding to the identified target DNA fragments demonstrated that most of the interactions were specific to one oligo only (Supplementary Figure 5). This can be explained by the fact that the oligo sequences had limited similarity with each other (the identities between the *MEG3_13_41-GATA6_AS_78_118*, *MEG3_839_890-GATA6_AS_78_118* and *MEG3_13_41-MEG3_839_890* oligo pairs were 33%, 40% and 25%, respectively – Supplementary Figure 6). Still, 5379 genomic regions were captured by each of the three oligos (Supplementary Figure 5). Thus, we expected that this set of 'shared Capture-seq peaks' can be enriched with the potential universal TTSs. To check this we predicted their possible interactions with the 153 RNA sequences that were used in the analysis of the ChOP-seq peaks (see above). Indeed, 2151 (40%) shared Capture-seq peaks were classified as universal TTSs – they had statistically significant number of triplexes with most (> 90%) of the analyzed transcripts (Figure 1E and Supplementary Figure 4C). Therefore, the fact that the experimentally identified set of shared Capture-seq peaks contained such a high fraction of the universal TTSs indirectly confirmed the predicted property of these special genomic loci.

Finally, we attempted to reveal the features of the universal TTSs that may allow them to interact with several different RNAs. For this purpose we compared sequence composition of the universal and all the other (i.e. non-universal) genomic regions from each set. While the GC content of the universal and non-universal DNA sequences were similar, the universal TTSs had higher purine (G or A) and, especially, poly-purine content (see Supplementary Figure 7 and Methods for the definitions). To find out the origin of these poly-purine elements we analyzed the classes of the overlapping genomic repeats. All three sets of the universal TTSs were enriched with the purine-rich low complexity regions, LCRs (Figure 1F and Supplementary Figure 8). Additional analysis of several universal TTSs confirmed that these LCRs were predicted to form multiple triple helices with the majority of the analyzed transcripts (see Supplementary Figure 9 for representative cases). Therefore, the presence of the purine-rich low complexity elements was the characteristic property of the universal TTSs that potentially allowed them to interact with a wide range of different RNAs. All together our results suggested the existence of a special type of genomic loci that may function as RNA-binding hubs.

Discussion

The importance of triplex-dependent gene regulation in the genomes of higher organisms is becoming a generally accepted concept. Here we performed a large-scale bioinformatic analysis of the genomic regions (ChOP-seq and Capture-seq peaks) that have been shown experimentally to interact with particular RNAs (MEG3 lncRNA or short oligos). To filter out not significant Triplexator predictions, the statistical significance of every RNA-DNA interaction was estimated from the Poisson distribution. To our surprise for some genomic regions (that we called "universal triplex target sites") Triplexator predicted statistically

significant (adjusted p-value < 0.01) number of triplexes with the majority (> 90%) of the analyzed transcripts. According to our analysis universal TTSs are quite rare in the human genome – there were only 0.6% of them among the 6798 randomly sampled regions. On the other hand, 16.3% of the experimentally identified MEG3 binding sites (ChOP-seq peaks) were classified as universal TTSs. Additionally, genomic regions that have been shown to form triplexes with three different oligos (shared Capture-seq peaks) contained 40% of the universal TTSs. All three sets of the identified universal TTSs were enriched with the purine rich low complexity regions.

The theoretical possibility of the universal TTS existence comes from the degeneracy of the Hoogsteen rules[9]. In fact, the triplex-based interaction can be formed in both orientations (parallel and anti-parallel) and it involves only purines (G or A) in the DNA. Additionally, the DNA guanine and adenine can bind to RNA guanine and uracil, respectively, in both orientations while the A::A pairing occurs in the anti-parallel orientation only. This makes the long poly-purine elements a possible targets for a number of different RNA oligos.

One of the possible and actively discussed roles of the chromatin bound RNAs (including lncRNAs) is to bring different chromosomal parts together to enable the remote DNA-DNA contacts[8]. Moreover, it has recently been shown that RNAs originating from super-enhancers form triplexes at distant regions[26]. Therefore, it is possible that universal TTSs may facilitate distal enhancer-promoter interactions via engagement with the same enhancer RNA. In line with this hypothesis, we observed the statistical significant enrichment of the universal Capture-seq peaks near (< 1 kb) the transcription start sites (TSSs) of the annotated genes (Supplementary Figure 10A). However, the computationally predicted universal ChOP-seq and background TTSs did not have such trend (Supplementary Figure 10B,C). Thus, the experimentally identified shared Capture-seq peaks may be more suitable for subsequent functional validation of the universal TTSs in living cells.

Importantly, the current computational analysis has a number of limitations. Namely, the triplex-based interactions of the full length transcripts were predicted without taking their secondary structure into account. We are not aware of any bioinformatics tools that would be able to produce such predictions. Moreover, cellular localization of the 153 selected expressed transcripts as well as DNA binding proteins and chromatin compaction were not considered. Therefore, our simulations are more similar to the *in vitro* Capture-seq experiments with short oligos than to the interactions of long transcripts with the chromatin inside the nucleus. Comprehensive identification of all the RNA-DNA interactions obtained by high throughput experimental methods may clarify the predicted functionality of the universal TTSs in the cell. Although a few methods for this task have recently been developed, the length of the sequencing reads (e.g. about 40 bp of DNA in case of GRID-seq) does not allow to reliably determine interactions with the long low complexity regions (including universal TTSs). We are looking forward to the new high quality experimental data to gain further insight into the triplex-based RNA-chromatin interactions *in vivo*.

Data availability

Underlying data

Zenodo: vanya-antonov/universal_tts: The initial release of the code, data files and images related to universal TTSs. http://doi.org/10.5281/zenodo.2654800[27]

This project contains the following underlying data:

- universal_tts-v1.0.0.zip?download=1.zip
  - data (folder containing underlying data, description of individual files can be found in Supplementary Table 3)

Extended data

Zenodo: vanya-antonov/universal_tts: The initial release of the code, data files and images related to universal TTSs. http://doi.org/10.5281/zenodo.2654800[27]

This project contains the following extended data:

- universal_tts-v1.0.0.zip?download=1.zip
  - images_R (folder containing R scripts to generate figures)
  - scripts (folder containing scripts to compute p-values based on the Triplexator predictions)

Grant information

This work was partially supported by Russian Science Foundation (grant 14-15-30002 to YAM).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are thankful to Dr. Chandrasekhar Kanduri (University of Gothenburg, Sweden) for providing the original coordinates of the ChOP-seq peaks.

Supplementary material

Supplementary File 1. File containing Supplementary figure 1–Supplementary figure 10, and Supplementary table 1–Supplementary table 3.

Click here to access the data

References

1. Khalil AM, Guttman M, Huarte M, *et al.*: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.** *Proc Natl Acad Sci U S A.* 2009; **106**(28): 11667–72.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Chu C, Qu K, Zhong FL, *et al.*: **Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions.** *Mol Cell.* 2011; **44**(4): 667–78.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Pandey RR, Mondal T, Mohammad F, *et al.*: **Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation.** *Mol Cell.* 2008; **32**(2): 232–246.
PubMed Abstract | Publisher Full Text

4. Mondal T, Subhash S, Vaid R, *et al.*: **MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures.** *Nat Commun.* 2015; **6**: 7743.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Simon MD, Wang CI, Kharchenko PV, *et al.*: **The genomic binding sites of a noncoding RNA.** *Proc Natl Acad Sci U S A.* 2011; **108**(51): 20497–502.
Publisher Full Text

6. Engreitz JM, Pandya-Jones A, McDonel P, *et al.*: **The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome.** *Science.* 2013; **341**(6147): 1237973.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Sridhar B, Rivas-Astroza M, Nguyen TC, *et al.*: **Systematic Mapping of RNA-Chromatin Interactions In Vivo.** *Curr Biol.* 2017; **27**(4): 602–609.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Li X, Zhou B, Chen L, *et al.*: **GRID-seq reveals the global RNA-chromatin interactome.** *Nat Biotechnol.* 2017; **35**(10): 940–950.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Hoogsteen K: **The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine.** *Acta Cryst.* 1963; **16**(9): 907–916.
Publisher Full Text

10. Schmitz K-M, Mayer C, Postepska A, *et al.*: **Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes.** *Genes Dev.* 2010; **24**(20): 2264–2269.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Grote P, Wittler L, Hendrix D, *et al.*: **The tissue-specific lncrna *fendrr* is an essential regulator of heart and body wall development in the mouse.** *Dev Cell.* 2013; **24**(2): 206–214.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Postepska-Igielska A, Giwojna A, Gasri-Plotnitsky L, *et al.*: **LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure.** *Mol Cell.* 2015; **60**(4): 626–636.
PubMed Abstract | Publisher Full Text

13. O'Leary VB, Ovsepian SV, Carrascosa LG, *et al.*: **PARTICLE, a Triplex-Forming Long ncRNA, Regulates Locus-Specific Methylation in Response to Low-Dose Irradiation.** *Cell Rep.* 2015; **11**(3): 474–485.
PubMed Abstract | Publisher Full Text

14. Buske FA, Bauer DC, Mattick JS, *et al.*: **Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data.** *Genome Res.* 2012; **22**(7): 1372–81.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Goñi JR, De La Cruz X, Orozco M: **Triplex-forming oligonucleotide target sequences in the human genome.** *Nucleic Acids Res.* 2004; **32**(1): 354–360.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Soibam B: **Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation.** *RNA.* 2017; **23**(11): 1729–1742.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Kent WJ, Sugnet CW, Furey TS, *et al.*: **The human genome browser at UCSC.** *Genome Res.* 2002; **12**(6): 996–1006.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods.* 2015; **12**(4): 357–60.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Harrow J, Frankish A, Gonzalez JM, *et al.*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res.* 2012; **22**(9): 1760–1774.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Anders S, Pyl PT, Huber W: **Htseq--a python framework to work with high-throughput sequencing data.** *Bioinformatics.* 2015; **31**(2): 166–169.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Jiang M, Anderson J, Gillespie J, *et al.*: **uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts.** *BMC Bioinformatics.* 2008; **9**: 192.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Gu Z, Eils R, Schlesner M: **Complex heatmaps reveal patterns and correlations in multidimensional genomic data.** *Bioinformatics.* 2016; **32**(18): 2847–2849.
PubMed Abstract | Publisher Full Text

24. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res.* 2004; **32**(5): 1792–1797.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Kuo CC, Hänzelmann S, Sentürk Cetin N, *et al.*: **Detection of RNA-DNA binding sites in long noncoding RNAs.** *Nucleic Acids Res.* 2019; **47**(6): e32.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Sentürk Cetin N, Kuo C-C, Ribarska T, *et al.*: **Isolation and genome-wide characterization of cellular DNA:RNA triplex structures.** *Nucleic Acids Res.* 2019; **47**(5): 2306–2321.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Antonov I: **vanyaantonov/universal_tts: The initial release of the code, data files and images related to universal TTSs (Version v1.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2654800

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 21 May 2019

https://doi.org/10.5256/f1000research.21026.r48244

✔ **Hao Zhu**
Bioinformatics Section, Southern Medical University, Guangzhou, China

The manuscript is improved and well-written, and the idea of "universal TTSs" is interesting, but there are still several weak points. These points make some of my concerns remain but are not as serious as before. So, I think it is better to let readers make their own judgement.

First, the authors analyzed basically only the data of MEG3, which makes the basis of the conclusion weak. The authors say, "To check the ability of other RNAs to form triplexes with MEG3 binding sites, we applied Triplexator to a set of 153 expressed transcripts". The specific 153 ones (given that there are ~20,000 annotated lncRNA genes) make the following "65 analyzed RNAs show results similar to MEG3 lncRNA" quite specific.
Second, both the ChOP-seq data and Triplexator results may have defects. For example, the authors say, "the number of the predicted triplexes increases with the lengths of input sequences". Why? Theoretically, if the longer sequences do not contain more TFO, the number of the triplexes should not increase.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 13 May 2019

https://doi.org/10.5256/f1000research.21026.r48245

✔

Ingrid Grummt
Division of Molecular Biology of the Cell II , DKFZ-ZMBH-Allianz, German Cancer Research Center (DKFZ), Heidelberg, Germany

The authors have responded to most of my concerns. They increased the number of RNA sequences and included published data in their analysis. Though some concerns are left, the overall conclusion of this study – that is, purine-rich sequences interact with various different RNAs – supports the notion that regulatory RNAs may target transcriptional co-activators to distinct genomic loci via Hoogsteen interactions with purine-rich gene sequences.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Regulation of gene expression by noncoding RNA

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 February 2018

https://doi.org/10.5256/f1000research.14683.r29953

Hao Zhu
Bioinformatics Section, Southern Medical University, Guangzhou, China

Many lncRNAs can bind to DNA sequences by forming triplexes (the binding sites are often called TTS, triplex-targeting sites). Whether there are “universal TTS” as described here is interesting and unreported, and deserves a careful investigation. But I have a major concern about the work: the authors reach the conclusion upon too few examples. Also, why these lncRNAs (BE2L6, LILRA3, HMOX1) were chosen (randomly or selected for some reasons)?

A few others issues should also be addressed. First, what is the relationship between the “universal TTSs” and base-pairing rules is untouched. For example, do the universal TTSs allow many lncRNAs to bind to them using the same rules? If very different rules are involved, what does this mean? To some extent, binding upon different rules indicates lncRNA specific TTSs, instead of universal TTSs. Second, I feel that the genomic regions used to sum scores are unreasonably long (3000 bp). Finally, it is said that “the median Triplexator SumScores were 48 and 25, respectively (p-value=5.2e-100)”. Statistically, the difference is significance, but biologically might not. I think 48 is not that large and 25 is not that small.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
No

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genome analysis, lncRNA analysis, molecular evolution

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 10 May 2019

Ivan Antonov, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russian Federation

Many lncRNAs can bind to DNA sequences by forming triplexes (the binding sites are often called TTS, triplex-targeting sites). Whether there are “universal TTS” as described here is interesting and unreported, and deserves a careful investigation. But I have a major concern about the work: the authors reach the conclusion upon too few examples. Also, why these lncRNAs (BE2L6, LILRA3, HMOX1) were chosen (randomly or selected for some reasons)?
In the current version we have significantly increased the number of query RNAs (i.e. 153 expressed transcripts and 153 random sequences obtained by di-nucleotide shuffling of MEG3 lncRNA). The 153 transcripts that we use now were chosen so that their lengths and GC contents were similar to the MEG3 lncRNA.

1) A few others issues should also be addressed. First, what is the relationship between the “universal TTSs” and base-pairing rules is untouched. For example, do the universal TTSs allow many lncRNAs to bind to them using the same rules? If very different rules are involved, what does this mean? To some extent, binding upon different rules indicates lncRNA specific TTSs, instead of universal TTSs.
Our preliminary analysis indicated that different RNAs interacted with universal TTSs via mixed (G or U) motif a little bit more frequently than via the purine or pyrimidine motifs. Importantly, all the analyzed transcripts were predicted to form a lot of triple helices (using different RNA motifs) with the universal TTSs. This is what makes uTTS special rather than specific motifs .

2) Second, I feel that the genomic regions used to sum scores are unreasonably long (3000 bp).
In the current version we decreased the genomic region lengths by considering the exact

ChOP-seq and Capture-seq peaks.

3) Finally, it is said that “the median Triplexator SumScores were 48 and 25, respectively (p-value=5.2e-100)”. Statistically, the difference is significance, but biologically might not. I think 48 is not that large and 25 is not that small.
We now use p-values instead of SumScore to estimate the statistical significance of each RNA-DNA interaction.

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 January 2018

https://doi.org/10.5256/f1000research.14683.r29955

Andrey A. Mironov
Department of Bioengineering and Bioinformatics, Moscow Technological University, Moscow, Russian Federation

The manuscript describes an application of the Triplexator software for search possible binding sites of the imprinting related MEG3 linc RNA on the human genome. The authors give a good example of the statistical analysis of the results. The main paper about this software has 62 references (google scholar data). Most of them have only reference to the software and only a few of them used the Triplexator. Only a few reports show a success story about the application of the Triplexator software and comparison the results with an experiment. In some papers, a significant enrichment of triplex targets on regions of interest was found. But they did not analyze the specificity of the predicted triplex formation. The current paper focused on a specificity of the Triplexator predictions. The authors got unexpected results that the Triplexator gives many non-specific hits for the case.

Comments:

1. Description of similar transcripts and the parameters of the Triplexator software should be rearranged because the appearance of some RNA names before definition sounds strange. The parameters of the Triplexator software contains a reference to the BE2L6 RNA while the description of the control RNA set has a reference on UBE2L6.

2. The di-nucleotide shuffling seems more adequate for RNA analysis.

3. In one paper[1] the Triplexator software also was used for analysis of MEG3 RNA-DNA contacts. The comparison of the obtained results with the results of given manuscript should be provided. Seems in current manuscript a more accurate analysis with good controls was provided.

4. It would be good to look at the practice of using the program on literature and make sure that the program has a sufficiently low specificity.

References
1. Mondal T, Subhash S, Vaid R, Enroth S, Uday S, Reinius B, Mitra S, Mohammed A, James AR, Hoberg E, Moustakas A, Gyllensten U, Jones SJ, Gustafsson CM, Sims AH, Westerlund F, Gorab E, Kanduri C: MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures.*Nat Commun*. 2015; **6**: 7743 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 10 May 2019

Ivan Antonov, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russian Federation

Dear reviewer,

We would like to apologise for a significant delay with the reply to all the comments. To implement the changes suggested by the reviewers we had to completely redesign our study and significantly increase the number of analyzed RNAs. Particularly, we developed a new method to estimate the statistical significance of the number of triplexes predicted for each RNA-DNA pair. Moreover, we analyzed the results of the recently published Capture-seq experiment that identified interactions of three different RNA oligos with "naked" DNA. We hope that all these analyses improved our study.

1) Description of similar transcripts and the parameters of the Triplexator software should be rearranged because the appearance of some RNA names before definition sounds strange. The parameters of the Triplexator software contains a reference to the BE2L6 RNA while the

description of the control RNA set has a reference on UBE2L6.
We have made the appropriate corrections in the text

2) The di-nucleotide shuffling seems more adequate for RNA analysis.
We now use the di-nucleotide shuffling to generate random RNA sequences.

3) In one paper1 the Triplexator software also was used for analysis of MEG3 RNA-DNA contacts. The comparison of the obtained results with the results of given manuscript should be provided. Seems in current manuscript a more accurate analysis with good controls was provided.
In this original paper the authors focused on the triplex-based interactions of a single RNA (MEG3) with the chromatin. In the present study we are interested whether other transcripts may have a potential to interact with the same genomic regions. Taking into account the different aims (and approaches) of the studies we are not sure if it is reasonable to compare their results.

4) It would be good to look at the practice of using the program on literature and make sure that the program has a sufficiently low specificity.
In our recent benchmarking study [PMID:29697742] we showed that Triplexator was the most accurate tool as of 2018. This is why we used it in the present study.

Competing Interests: No competing interests were disclosed.

Reviewer Report 23 January 2018

https://doi.org/10.5256/f1000research.14683.r29954

Ingrid Grummt
Division of Molecular Biology of the Cell II , DKFZ-ZMBH-Allianz, German Cancer Research Center (DKFZ), Heidelberg, Germany

Long noncoding RNAs (lncRNA) can regulate gene expression by targeting specific DNA sequences via Hoogsteen base paring, forming RNA-DNA triple helical structures. Computational analyses have revealed that a large population of triplex-forming motifs is present across the genome, the majority of annotated human genes containing at least one unique and high-affinity triplex target site, preferentially in regulatory gene regions (Goni et al. 2004, Buske et al. 2012). Moreover, several studies have provided in vitro and in vivo evidence for the existence and biological relevance of RNA-DNA triplexes, including pRNA (Schmitz et al. 2010), Fendrr (Grote et al. 2013), Khps1 (Postepska-Igielska et al. 2015), PARTICLE (O'Leary et al. 2015), and MEG3 (Mondal et al. 2015). MEG3 has been shown to associate with AG-rich DNA motifs and facilitate recruitment of PRC2 to target sites. Considering the large number of purine-rich sequences in the genome, triplex-mediated targeting of lncRNAs and associated proteins to distinct genomic loci is very likely a commonly used mechanism of gene regulation.

Given the importance and emerging acceptance of the concept of triplex-dependent gene regulation, it is more than surprising, if not irritating, that the authors challenge this concept feeding the 'Triplexator' only with a few RNAs and a subset of MEG3-interacting regions rather than providing any experimental data and/or more global bioinformatic analysis.

Just some specific comments:

- In the abstract they claim '*these triplex interactions might contribute to establishing long-distance chromosomal contact'* without providing any information or bioinformatic analyses.
- They use the term 'hybridization' for the interaction between RNA and dsDNA. This is wrong as hybridization refers to Watson-Crick base-pairing between RNA and ssDNA and not to Hoogsteen bonding.
- They took MEG3-interacting DNA peaks shorter than 1000 bp, then selected 3000 bp bins centering these regions and used these bins for analysis. There is no rationale for this selection which of course determines the final outcome of the analysis. Accordingly, the majority of these bin regions did not coincide with regions determined by ChOP-seq. Probably, a shorter binning would be more reliable to analyze the available data.
- They focused on bins that overlap genes. Even if partial overlapping was accepted, they might have missed some promoters. Intergenic regions containing regulatory sequences (e.g. enhancers) were excluded.
- Why were only peaks overlapping with annotated genes considered to be significant (or "real")? Genomic regions that do not harbor annotated genes, such as enhancers, are important regulatory elements that are targeted by lncRNAs and as such are functional RNA-binding sites, highly relevant for this study. In addition, since it is known that the base composition of genic and intergenic regions is different, exclusion of intergenic regions introduces a considerable bias to the analysis.
- Selection of just three additional RNAs is certainly not adequate for the far-reaching conclusion: '*TTSs are able to hybridize with various different RNAs almost irrespectively of their sequence'*. It would be more convincing to show the results from scanning more RNAs, irrespective of their length and GC-content. Also, there is no attention given to the expression profiles of selected MEG3-mimicking RNAs. This is important because transcription of MEG3 is highly tissue-specific.
- The sum scores from the Triplexator analysis are shown which does not mean that the same regions are involved in triplex formation. It would be much more convincing to show similarities (or differences) of triplex-forming RNAs for a given TTS in a given bin.
- The terms 'universal TTS' and 'universal bins' are not synonymous and interchangeable!! One bin (3000 bp) can contain many putative TTSs.
- If there are only 18 'universal bins' out of 3620 bins among seven RNA analyses, this small number is not sufficient for claiming that there is no specificity in RNA targeting.
- They hypothesize that *'the universal TTS can be viewed as the anchor point which can be bound by various nuclear RNAs to provide long-distance chromosomal contacts'*. Even if this might be true, without any supportive data this is pure speculation.

Altogether, the authors´ claim that triplex formation occurs almost sequence-independent is not justified but is based solely on *in silico* analyses. At least another available bioinformatics tool should have been used and standard *in vitro* assays (e.g. EMSA experiments) should have been performed to validate that the candidate RNAs are indeed capable to form triplexes. The authors do not even mention that the *in vivo* situation might be completely different than algorithm-based predictions and that there might be additional factors/constraints involved in triplex formation and stability.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?
No

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Regulation of gene expression by noncoding RNA

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 10 May 2019

Ivan Antonov, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russian Federation

We would like to thank Dr. Grummt for the extended comments to our work. They have helped us improve the design of our study and obtain additional results. We hope that they made our conclusions more reliable and reproducible.

Given the importance and emerging acceptance of the concept of triplex-dependent gene regulation, it is more than surprising, if not irritating, that the authors challenge this concept feeding the 'Triplexator' only with a few RNAs and a subset of MEG3-interacting regions rather than providing any experimental data and/or more global bioinformatic analysis.
We do not challenge the possibility of triplex-dependant regulation. We simply claim that RNA interactions with some genomic regions have low sequence specificity because many other RNAs may be able to bind the same loci. We investigated features of such regions and found them to be enriched in purine rich low complexity repeats. In the current version, we completely redesigned the study and incorporated the analysis of 306 RNA sequences to confirm our findings. We also modified the text so the main conclusions are clear and non-misleading.

Just some specific comments:
1) In the abstract they claim 'these triplex interactions might contribute to establishing long-distance chromosomal contact' without providing any information or bioinformatic analyses.

We believe it is reasonable to speculate about this possibility in the discussion for the following reasons. First, it has recently been shown that "RNAs originating from super-enhancers form triplexes at distant regions" [PMID: 30605520]. Second, we showed that the predicted Capture-seq universal TTSs were highly enriched in gene promoters (Supplementary Figure 10). Together these observations indicate that the same eRNA may be able to interact with several different universal TTSs and therefore contribute to the long-distance chromosomal (i.e. enhancer-promoter) contacts. However, additional experimental verification of this hypothesis is required.
We modified the text in the abstract as follows: "We speculated that such universal TTSs may contribute to establishing long-distance chromosomal contacts and may facilitate distal enhancer-promoter interactions."

2) They use the term ‘hybridization’ for the interaction between RNA and dsDNA. This is wrong as hybridization refers to Watson-Crick base-pairing between RNA and ssDNA and not to Hoogsteen bonding.
We have corrected the terminology used in the manuscript.

3) They took MEG3-interacting DNA peaks shorter than 1000 bp, then selected 3000 bp bins centering these regions and used these bins for analysis. There is no rationale for this selection which of course determines the final outcome of the analysis. Accordingly, the majority of these bin regions did not coincide with regions determined by ChOP-seq. Probably, a shorter binning would be more reliable to analyze the available data.
We now use the exact locations for all the ChOP-seq peaks. To compensate for the peak length variability we developed a method that estimates the statistical significance of the number of triplexes predicted for a RNA-DNA pair taking into account lengths of both sequences.

4) They focused on bins that overlap genes. Even if partial overlapping was accepted, they might have missed some promoters. Intergenic regions containing regulatory sequences (e.g. enhancers) were excluded.
We now analyze all the ChOP-seq peaks without considering their overlaps with the annotated genes.

5) Why were only peaks overlapping with annotated genes considered to be significant (or ”real“)? Genomic regions that do not harbor annotated genes, such as enhancers, are important regulatory elements that are targeted by lncRNAs and as such are functional RNA-binding sites, highly relevant for this study. In addition, since it is known that the base composition of genic and intergenic regions is different, exclusion of intergenic regions introduces a considerable bias to the analysis.
We now analyze all the ChOP-seq peaks.

6) Selection of just three additional RNAs is certainly not adequate for the far-reaching conclusion: ‘TTSs are able to hybridize with various different RNAs almost irrespectively of their sequence’. It would be more convincing to show the results from scanning more RNAs, irrespective of their length and GC-content. Also, there is no attention given to the expression profiles of selected MEG3-mimicking RNAs. This is important because transcription of MEG3 is highly tissue-specific.
We have increased the number of the considered query RNAs to 306 and used the expressed transcripts only.

7) The sum scores from the Triplexator analysis are shown which does not mean that the same regions are involved in triplex formation. It would be much more convincing to show similarities (or differences) of triplex-forming RNAs for a given TTS in a given bin.
We no longer use sum scores as a measure of triplex-based interaction. Instead, we estimate the statistical significance (p-value) of each RNA-DNA interaction based on the number of predicted triplexes.
Our work was focused on the properties of the DNA sequences that may allow them to interact with various different RNAs. We therefore analyzed the parts of the ChOP-seq/Capture-seq peaks universal TTSs that allowed them to interact with various different RNAs. Our analysis indicates that such triplex-forming hot-spots frequently coincide with the purine-rich low complexity genomic regions.

8) The terms ‘universal TTS’ and ‘universal bins’ are not synonymous and interchangeable!! One bin (3000 bp) can contain many putative TTSs.
We do not use the concept of bins and ‘universal bins’ in the current version of the manuscript. However, we kept the term ‘universal TTS’.

9) If there are only 18 ‘universal bins’ out of 3620 bins among seven RNA analyses, this small number is not sufficient for claiming that there is no specificity in RNA targeting.
We would like to emphasize that our paper don't question the concept of the sequence specific triplex-dependent gene regulation (moreover, we support this idea and conduct research in this direction). We claim that some genomic regions may have a potential to form triple helices with a variety of different long RNAs forming "universal" triplex target sites (TTSs). At the same time, we do not challenge the sequence specificificity of the other triplex-based RNA-DNA interactions.

10) They hypothesize that ‘the universal TTS can be viewed as the anchor point which can be bound by various nuclear RNAs to provide long-distance chromosomal contacts’. Even if this might be true, without any supportive data this is pure speculation.
We agree that at the moment our claim is a hypothesis/speculation. Nevertheless, the recent published results [PMID: 30605520] as well our own indicate the possibility of such mechanism. By discussing it in the current manuscript we hope to attract attention of experimental biologists to further study this topic.
We modified the text in the paper to clarify the issue:
"One of the possible and actively discussed roles of the chromatin bound RNAs (including lncRNAs) is to bring different chromosomal parts together to enable the remote DNA-DNA contacts. Moreover, it has recently been shown that RNAs originating from super-enhancers form triplexes at distant regions. Therefore, it is possible that universal TTSs may facilitate distal enhancer-promoter interactions via engagement with the same enhancer RNA. In line with this hypothesis, we observed the statistical significant enrichment of the universal Capture-seq peaks near (< 1 kb) the transcription start sites (TSSs) of the annotated genes (Supplementary Figure 10C)."

11) Altogether, the authors´ claim that triplex formation occurs almost sequence-independent is not justified but is based solely on in silico analyses. At least another available bioinformatics tool should have been used and standard in vitro assays (e.g. EMSA experiments) should have been performed to validate that the candidate RNAs are indeed capable to form triplexes. The authors do not even mention that the in vivo situation might be completely different than algorithm-based

predictions and that there might be additional factors/constraints involved in triplex formation and stability.

We added analysis of the recent in vitro data obtained by the Capture-seq method. According to this experimental data some genomic fragments can interact with all three RNA oligos used in the original study. This supports the observations obtained in the analysis of the ChOP-seq peaks. Our computational analysis classified almost 40% of these shared Capture-seq peaks as universal TTSs. Moreover, the ChOP-seq and the Capture-seq universal TTSs were similar in that they were enriched with the purine rich low complexity regions. We believe that these results indirectly support the existence of universal TTS . Yet, experimental validation of these results is beyond the scope of the current paper.

Still, at the end of the manuscript, we discuss the limitations of our computational approach and mention that the obtained results resemble mostly the situation with the naked DNA in vitro, than the interactions with the chromatin in vivo.

We added the following text to the discussion:

"Importantly, the current computational analysis has a number of limitations. Namely, the triplex-based interactions of the full length transcripts were predicted without taking their secondary structure into account. We are not aware of any bioinformatics tools that would be able to produce such predictions. Moreover, cellular localization of the 153 selected expressed transcripts as well as DNA binding proteins and chromatin compaction were not considered. Therefore, our simulations are more similar to the in vitro Capture-seq experiments with short oligos than to the interactions of long transcripts with the chromatin inside the nucleus."

Competing Interests: No competing interests were disclosed.