Published online 22 May 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 2 **1**
https://doi.org/10.1093/nargab/lqad043

TRIPBASE: a database for identifying the human genomic DNA and lncRNA triplexes

Tzu-Chieh Lin †, Yen-Ling Liu†, Yu-Ting Liu†, Wan-Hsin Liu, Zong-Yan Liu, Kai-Li Chang, Chin-Yao Chang, Hung Chih Ni, Jia-Hsin Huang * and Huai-Kuang Tsai *

Institute of Information Science, Academia Sinica, Taipei, 11529, Taiwan

Received November 18, 2022; Revised March 04, 2023; Editorial Decision May 02, 2023; Accepted May 04, 2023

ABSTRACT

Long-non-coding RNAs (lncRNAs) are defined as RNA sequences which are >200 nt with no coding capacity. These lncRNAs participate in various biological mechanisms, and are widely abundant in a diversity of species. There is well-documented evidence that lncRNAs can interact with genomic DNAs by forming triple helices (triplexes). Previously, several computational methods have been designed based on the Hoogsteen base-pair rule to find theoretical RNA–DNA:DNA triplexes. While powerful, these methods suffer from a high false-positive rate between the predicted triplexes and the biological experiments. To address this issue, we first collected the experimental data of genomic RNA-DNA triplexes from antisense oligonucleotide (ASO)-mediated capture assays and used Triplexator, the most widely used tool for lncRNA–DNA interaction, to reveal the intrinsic information on true triplex binding potential. Based on the analysis, we proposed six computational attributes as filters to improve the *in-silico* triplex prediction by removing most false positives. Further, we have built a new database, TRIPBASE, as the first comprehensive collection of genome-wide triplex predictions of human lncRNAs. In TRIPBASE, the user interface allows scientists to apply customized filtering criteria to access the potential triplexes of human lncRNAs in the *cis*-regulatory regions of the human genome. TRIPBASE can be accessed at https://tripbase.iis.sinica.edu.tw/.

INTRODUCTION

Long-non-coding RNAs (lncRNAs) are a type of non-coding RNA with a length greater than 200 nt. With advances in sequencing technology, it has been revealed that lncRNAs can be found in various aspects of the transcriptome and are widely transcribed in species ranging from invertebrates to humans. LncRNAs have gained more attention in recent years and have been implicated in numerous physiological and pathological processes. There is well-documented evidence that lncRNAs can interact with double-stranded DNA (dsDNA) by forming triple helices (triplexes) to mediate gene expression (1,2).

Several experimental methods, such as chromatin isolation by RNA purification (ChIRP) (3) and RNA and DNA interacting complexes ligated and sequenced (RADICL-seq) (4), have been developed to investigate the association between chromatin and non-coding RNAs. These methods use cross-linking to capture chromatin–protein–RNA complexes and enrich the active fractions. However, they are not well-suited to study direct RNA-DNA binding interactions, particularly in the context of lncRNA–DNA:DNA triplex structures. Fortunately, the antisense DNA oligonucleotide (ASO)-based sequencing method (5) provides a genome-wide approach to specifically detect lncRNA–DNA:DNA interactions. This technology employs a cross-link-free approach that eliminates RNA–DNA–protein complexes and R-loop-mediated RNA–DNA interactions, resulting in reliable evidence of RNA–DNA direct interactions experimentally.

However, these biological experiments are expensive and only limited numbers of lncRNAs were investigated in experimental RNA–DNA:DNA interactions. Therefore, an important research topic is the identification of genome-wide lncRNA–DNA:DNA triplexes using computational methods. Most existing methods to date, including Triplexator (6), LongTarget (7) and TRIPLEXES (8), are based on the Hoogsteen base-pairing rules (9) to find the theoretical RNA–DNA:DNA triplexes according to RNA and dsDNA sequences. However, according to the results of experimental assays in wet labs, only a small number of predicted triplex-target sites (TTSs) on DNA sequences using the Triplexator tool can be considered actual interactions (10–12). The low specificity is most probably due to a greedy strategy of applying the Hoogsteen base-pairing rules in enumerating all possible matches.

*To whom correspondence should be addressed. Tel: +886 2 2788 3799 (Ext 1718); Fax: +886 2 2782 4814; Email: hktsai@iis.sinica.edu.tw
Correspondence may also be addressed to Jia-Hsin Huang. Tel: +886 2 2788 3799 (Ext 1475); Fax: +886 2 2651 9574; Email: jiahsin.huang@gmail.com
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

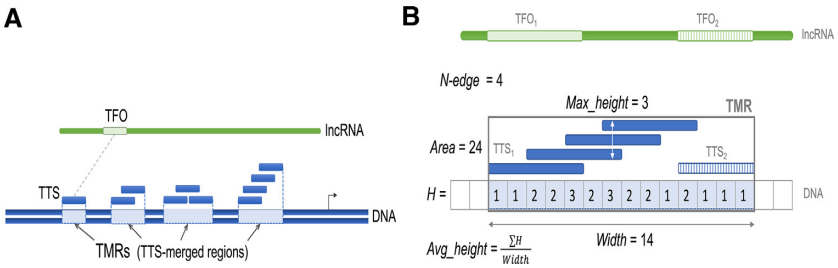

Figure 1. The definitions of TTS, TFO, TMRs and five attributes for filtering false positives. (**A**) The nucleotide sequence of lncRNA that can form hydrogen bonds with the DNA duplex is called a 'triplex-forming oligonucleotide' (TFO). The corresponding position of the triplex on DNA is called the 'triplex-target site' (TTS). The successive TTSs are called TTS-merged regions (TMRs). (**B**) An example to illustrate the five attributes. The four blue blocks stand for TTS_1, which are all predicted by Triplexator and can form a DNA:RNA triplex with TFO_1. The striped block stands for TTS_2, which can form a DNA:RNA triplex with TFO_2. The width stands for the length of the TMR. H is the abbreviation of 'hitting numbers' for each nucleotide sequence. In this TMR, for position 3, where it is covered by two TTSs, the hitting number H is thus 2. The area stands for the sum of the hitting numbers of the TMR. As a result, the area of this TMR is calculated by adding up the hitting number H from position 1 to 14. The area of this TMR is 24. The average height stands for the average of hitting number. Therefore, the average height is the area divided by width, and equals 24/14. The maximal height is the maximal value of the hitting numbers for all the nucleotides of a TMR. Therefore, the maximal value from position 1 to 14 is three, which occurs at positions 5 and 7. The n-edge stands for the number of times triplex formation occurs between a TFO in lncRNA and a TMR in DNA. Because TFO_1 have four TTSs, the n-edge of TFO_1 is 4.

To fill this gap, we first addressed the issue of a massive number of triplex predictions by identifying the optimal filtering combination according to the empirical evaluations using the experimental data of ASO-seq (5). We analyzed the predicted lncRNA–DNA interactions via using Triplexator on ASO-seq to obtain several filters that could remove a large number of predicted lncRNA–DNA interactions (false positives) not present in the ASO-seq peaks. Based on the analysis, we have constructed TRIPBASE, a public database for the *in silico* prediction of human lncRNA–dsDNA triplexes. The current release contains 25 914 protein-coding genes, 7 325 744 enhancer regions and 17 932 lncRNA transcripts. Currently, TRIPBASE covers >1.5 trillion TTS predictions for all lncRNA transcripts and offers 26 982 244 TTS-merged regions (TMRs) for query. In addition, TRIPBASE provides a graphical user interface allowing scientists to search, browse and download the lncRNA triplexes of interest.

MATERIALS AND METHODS

Dataset

We downloaded the human whole genome from the Ensembl database (GRCh38, release 99) and gene annotation and lncRNAs sequences from GENCODE (release 35). Because transcript isoforms share many identical sequences, only the longest isoform of lncRNAs was selected. For *cis*-regulatory DNA sequences, the region from –5000 to +200 of a transcription start site was defined as the promoter region of protein-coding genes, and enhancer regions were collected from EnhancerAtlas 2.0 (13), including enhancer regions from 197 human cell and tissue types. In total, there are 7 325 744 enhancer regions.

Filtering criteria

We employed Triplexator (6) to obtain the putative TTSs of each lncRNA using the following set of parameters: triplex length = 15, maximal error rate = 20 and consecutive error number = 2. For a given lncRNA, as shown in Figure 1A, the nucleotide sequence of the lncRNA that can form hydrogen bonds with the DNA duplex is called the 'triplex-forming oligonucleotide' (TFO). The corresponding position of the triplex on DNA is called a 'triplex-target site' (TTS). The successive TTSs are called TTS-merged regions (TMRs). The adjacent TTSs without overlapped DNA regions are considered independent TMRs. Of note, five TTS attributes within a TMR, i.e. width, area, average height, maximal height and n-edge, are used to set the criteria for filtering TMRs (Figure 1B). Herein, width stands for the length of the TMR, i.e. the number of base pairs. Given that a base pair in a TMR might be 'hit' by multiple TTSs, for each base pair we define 'hitting number' as the number of TTSs that hit that base pair. Area, average height and maximal height stand for the sum, the average and the maximal value of the hitting numbers within a TMR, respectively. The n-edge is the largest number of times that triplex formation can occur by TFOs of the lncRNA. In addition, a previous study has demonstrated that masking the secondary structures of lncRNAs improves the true positive rate (TPR) of the prediction (14). Accordingly, we applied RNAplfold (15) to predict the secondary structure of lncRNAs with a cut-off probability of 0.5. The secondary structure masked as 'N' in the lncRNA transcript sequences is also a filter criterion to remove TMRs which were bound to the secondary structure of the lncRNA region.

Design of TRIPBASE

TRIPBASE consists of seven tables (Figure 2). $t_{triplexator\_result}$ contains predictions from Triplexator. The primary keys in $t_{triplexator\_result}$ are set to ensure every hit in predictions is unique. $t_{chromosome}$ contains chromosome information, including chromosome number and length. t_{assay_peak} is established to record lncRNAs involved in assays, such as ASO-capture-seq. $t_{promoter\_gene\_map}$ records 25 914 protein-coding genes in order to display gene regions on a display panel and also promoter regions regarding protein-coding genes. $t_{homo\_sapients\_enhancer\_gene\_list}$ records 7 325 744 enhancer regions from EnhancerAtlas 2.0, and $t_{homo\_sapients\_enhancer\_gene\_interaction}$

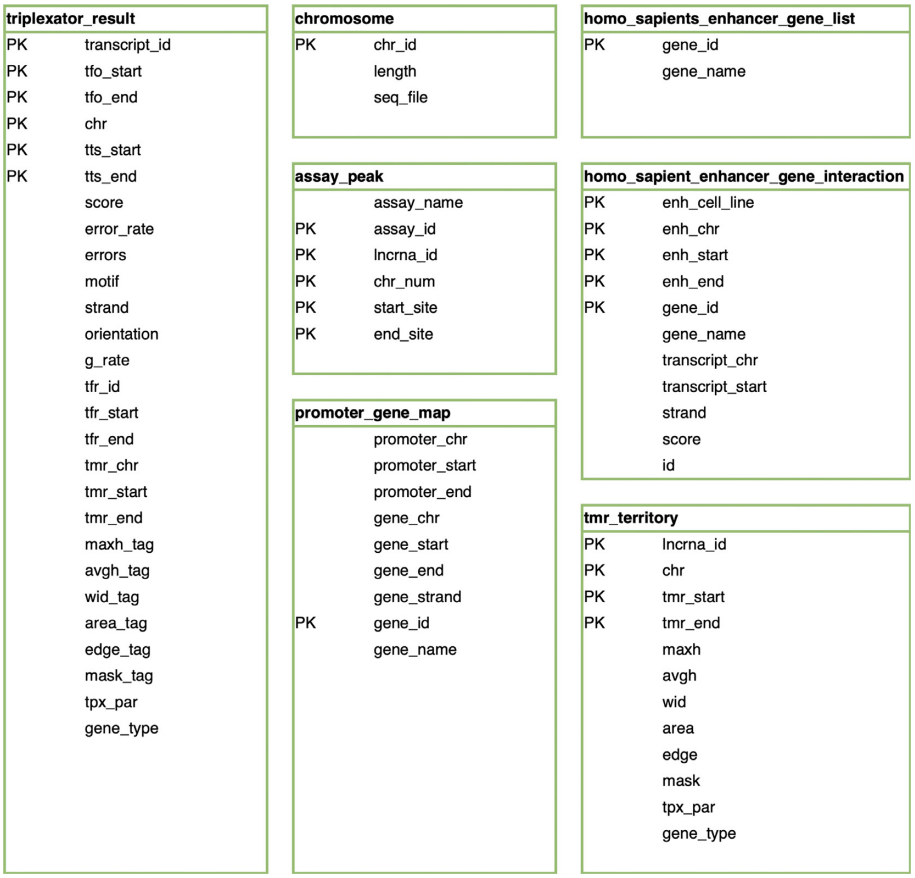

Figure 2. The database relationship diagram of TRIPBASE. TRIPBASE consists of seven tables, namely $t_{triplexator\_result}$ (predictions from Triplexator), $t_{chromosome}$ (chromosome information), t_{assay_peak} (lncRNAs involved in assays), $t_{promoter\_gene\_map}$ (25 914 protein coding genes), $t_{homo\_sapients\_enhancer\_gene\_list}$ and $t_{homo\_sapients\_enhancer\_gene\_interaction}$ (interactions between genes and enhancer regions) and $t_{tmr\_territory}$ (attributes of TMRs).

is set to record interactions between genes and enhancer regions. $t_{tmr\_territory}$ contains TMRs with their different attributes.

RESULTS AND DISCUSSION

Filters for selecting potential lncRNA–dsDNA triplexes

The ASO-capture data of NEAT1-associated DNA downloaded from NCBI's Gene Expression Omnibus (GEO: GSE120850) were used as the ground truth to evaluate the TMRs of the lncRNA NEAT1. The original dataset contains 7 601 780 raw TMRs, of which 18 986 TMRs intersected with assay peaks (positive group) and 7 582 794 TMRs were located outside of peak regions (negative group). To improve the performance of triplex prediction, we employed a brute-force approach that searches for different combinations of the five attributes within the TMRs and secondary mask conditions. As shown in Figure 3, filtering the data using the six criteria improves the TPR but keeps a sensible false-positive rate (FPR). For the unmasked TMRs, using maximum height ≥ 2, average height ≥ 1, width ≥ 15, area ≥ 30 and n-edge ≥ 1 (i.e. the large red dot) reaches the highest TPR of 85%, while the FPR is 54%. For the TMRs with secondary mask conditions, using width ≥ 17 instead (the large blue dot) achieves 64% TPR and 30% FPR. In TRIPBASE, we apply the filtering criteria that reach the

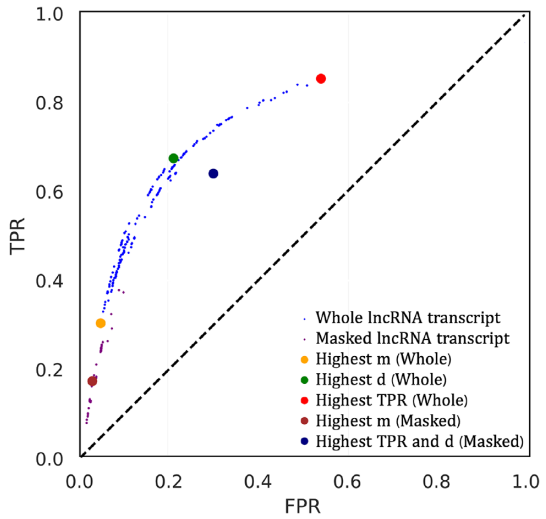

Figure 3. The scatter plot shows FPR and TPR obtained by different combinations of filters, either using whole lncRNA transcript sequences (i.e. small blue dots) or with masked lncRNA transcript sequences with secondary structure masked (i.e. small purple dots). The larger dots denote the filtering combinations that reach the highest TPR/FRP ratio (m), the longest distance from the diagonal line (d) and the highest TPR rate.

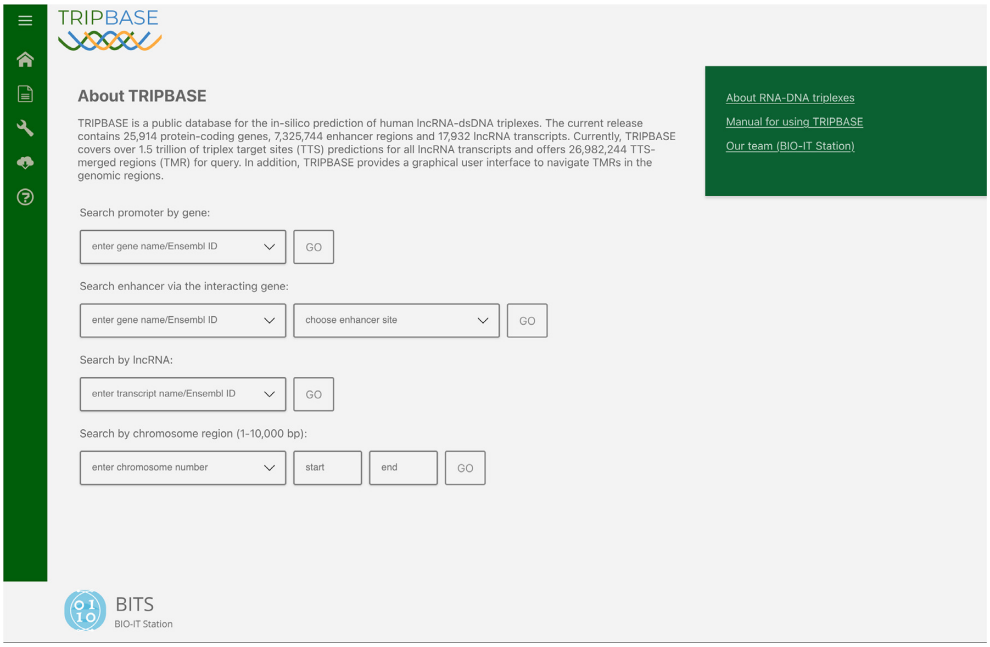

Figure 4. How to use TRIPBASE. TRIPBASE allows a user to search for TMRs in the *cis*-regulatory regions, including the promoter of the protein-coding genes, enhancer regions and a segment of chromosomes, and to query an annotated lncRNA from GENCODE.

highest TPR under the unmasked condition as our recommendation parameters to show the TMR candidates.

TRIPBASE

Based on the above analysis, we have built a new database named TRIPBASE, as the first comprehensive collection of genome-wide triplex predictions of human lncRNAs. TRIPBASE stores and visualizes >1.5 trillion triplex predictions for a total of 17 932 lncRNAs interacting with the *cis*-regulatory elements in the human genome. The homepage of TRIPBASE allows a user to search for TMRs in the *cis*-regulatory regions, including the promoter of protein-coding genes, enhancer regions and a segment of chromosomes, and to query an annotated lncRNA from GENCODE (Figure 4). Additionally, users can search for TMRs by specifying a chromosomal segment of up to 10,000 bp in length.

TRIPBASE displays a graphic browser that easily navigates the TMRs in a specific range of a chromosome. By default, TRIPBASE will show the subset of TMRs with our recommended criteria. The TMRs are ranked according to their area by default. The user can choose another ranking method or refine the filtering criteria according to their preferred parameters to get different subsets of TMRs. By hovering over the TMRs, genes, promoters and enhancers, users can obtain detailed information such as TMR attributes or interacting genes and enhancers. In addition, users can download the complete list of TMRs by clicking the 'download' button. Finally, users can run Triplexator with custom lncRNAs and DNA sequences on the Tools page. Taken together, TRIPBASE offers an integrative platform that will help users to identify functional lncRNA target sites in the *cis*-regulatory elements by evaluating the TMR information.

It is important to note that currently there is only one publicly available ASO-seq experiment. To validate the performance of our filtering parameters on other biological experiments for RNA–dsDNA interaction, we conducted the same analysis on two lncRNAs, MEG3 and HOTAIR, whose RNA–DNA interaction data were obtained by CHOP-seq and CHiRP-seq experiments, respectively. However, CHOP-seq and CHiRP-seq are cross-linking-based methods which might contain interactions other than RNA:DNA triplexes such as those via a protein or R-loop. Our results suggest that the proposed filtering parameters can still reduce the FPR in different types of lncRNA–DNA interaction experiments, but slightly less effectively than by ASO-seq. For example, using the parameters with maximum height ≥ 2, average height ≥ 1, width ≥ 15, area ≥ 30 and n-edge ≥ 1, we obtained a TPR of 71% and an FPR of 67% for MEG3, and a TPR of 67% and an FPR of 80% for HOTAIR.

Due to the computing resource limitations, the current released TRIPBASE only predicted lncRNA–DNA interactions in promoter and enhancer regions. However, we offer a solution by providing users with a programmatic API access, allowing them to analyze their data independently (https://tripbase.iis.sinica.edu.tw/data/). In the future, we plan to expand our computational facilities to enable the prediction of the lncRNA–DNA interactions across the entire human genome. In addition, we hope to include more species, starting with those with smaller genome sizes, such as yeast, Drosophila and Arabidopsis, to provide comparative genomic information on lncRNA–DNA interactions.

DATA AVAILABILITY

TRIPBASE can be accessed at https://tripbase.iis.sinica.edu.tw/.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Dr. Jun-Yi Leu from the Institute of Molecular Biology, Academia Sinica, and Dr. Cheng-Fu Kao from the Institute of Cellular and Organismic Biology, Academia Sinica, for their valuable comments and ideas that greatly contributed to this research. We also thank the Computer Center of the Institute of Information Science, Academia Sinica, for providing computation resources to host the database. Finally, we would like to acknowledge the Ministry of Science and Technology for funding this research.

FUNDING

The Institute of Information Science, Academia Sinica and the Ministry of Science and Technology [MOST 108-2221-E-001-014].

Conflict of interest statement. None declared.

REFERENCES

1. Statello,L., Guo,C.J., Chen,L.L. and Huarte,M. (2021) Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.*, **22**, 96–118.
2. Li,Y., Syed,J. and Sugiyama,H. (2016) RNA–DNA triplex formation by long noncoding RNAs. *Cell Chem. Biol.*, **23**, 1325–1333.
3. Chu,C., Quinn,J. and Chang,H.Y. (2012) Chromatin isolation by RNA purification (ChIRP). *J. Vis. Exp.*,(61), 1868.
4. Bonetti,A., Agostini,F., Suzuki,A.M., Hashimoto,K., Pascarella,G., Gimenez,J., Roos,L., Nash,A.J., Ghilotti,M., Cameron,C.J.F. *et al.* (2020) RADICL-seq identifies general and cell type-specific principles of genome-wide RNA–chromatin interactions. *Nat. Commun.*, **11**, 1018.
5. Senturk Cetin,N., Kuo,C.C., Ribarska,T., Li,R., Costa,I.G. and Grummt,I. (2019) Isolation and genome-wide characterization of cellular DNA:RNA triplex structures. *Nucleic Acids Res.*, **47**, 2306–2321.
6. Buske,F.A., Bauer,D.C., Mattick,J.S. and Bailey,T.L. (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res.*, **22**, 1372–1381.
7. He,S., Zhang,H., Liu,H. and Zhu,H. (2015) LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. *Bioinformatics*, **31**, 178–186.
8. Kuo,C.C., Hanzelmann,S., Senturk Cetin,N., Frank,S., Zajzon,B., Derks,J.P., Akhade,V.S., Ahuja,G., Kanduri,C., Grummt,I. *et al.* (2019) Detection of RNA–DNA binding sites in long noncoding RNAs. *Nucleic Acids Res.*, **47**, e32.
9. 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 nucleotide sequences. *J. Mol. Biol.*, **37**, 63–80.
10. Mondal,T., Subhash,S., Vaid,R., Enroth,S., Uday,S., Reinius,B., Mitra,S., Mohammed,A., James,A.R., Hoberg,E. *et al.* (2015) MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA–DNA triplex structures. *Nat. Commun.*, **6**, 7743.
11. Postepska-Igielska,A., Giwojna,A., Gasri-Plotnitsky,L., Schmitt,N., Dold,A., Ginsberg,D. and Grummt,I. (2015) LncRNA Khps1 regulates expression of the proto-oncogene SPHK1 via triplex-mediated changes in chromatin structure. *Mol. Cell*, **60**, 626–636.
12. O'Leary,V.B., Smida,J., Buske,F.A., Carrascosa,L.G., Azimzadeh,O., Maugg,D., Hain,S., Tapio,S., Heidenreich,W., Kerr,J. *et al.* (2017) PARTICLE triplexes cluster in the tumor suppressor WWOX and may extend throughout the human genome. *Sci. Rep.*, **7**, 7163.
13. Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
14. Matveishina,E., Antonov,I. and Medvedeva,Y.A. (2020) Practical guidance in genome-wide RNA:DNA triple helix prediction. *Int. J. Mol. Sci.*, **21**, 830.
15. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.