Computational and Structural Biotechnology Journal 21 (2023) 3091–3102

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

3plex enables deep computational investigation of triplex forming lncRNAs

Chiara Cicconetti [a,b], Andrea Lauria [a,b], Valentina Proserpio [a,b], Marco Masera [a], Annalaura Tamburrini [a,b], Mara Maldotti [a,b], Salvatore Oliviero [a,b,*], Ivan Molineris [a,b,*]

[a] Dipartimento di Scienze della Vita e Biologia dei Sistemi and MBC, Università di Torino, Via Nizza 52, 10126 Torino, Italy
[b] Italian Institute for Genomic Medicine (IIGM), Sp142 Km 3.95, Candiolo 10060 (Torino), Italy

ARTICLE INFO

Article history:
Received 6 February 2023
Received in revised form 15 May 2023
Accepted 15 May 2023
Available online 17 May 2023

Keywords:
LncRNA
Triplex
Bioinformatics

ABSTRACT

Long non-coding RNAs (lncRNAs) regulate gene expression through different molecular mechanisms, including DNA binding via the formation of RNA:DNA:DNA triple helices (TPXs). Despite the increasing amount of experimental evidence, TPXs investigation remains challenging. Here we present *3plex*, a software able to predict TPX interactions *in silico*. Given an RNA sequence and a set of DNA sequences, *3plex* integrates 1) Hoogsteen pairing rules that describe the biochemical interactions between RNA and DNA nucleotides, 2) RNA secondary structure prediction and 3) determination of the TPX thermal stability derived from a collection of TPX experimental evidences. We systematically collected and uniformly re-analysed published experimental lncRNA binding sites on human and mouse genomes. We used these data to evaluate *3plex* performance and showed that its specific features allow a reliable identification of TPX interactions. We compared *3plex* with the other available software and obtained comparable or even better accuracy at a fraction of the computation time. Interestingly, by inspecting collected data with *3plex* we found that TPXs tend to be shorter and more degenerated than previously expected and that the majority of analysed lncRNAs can directly bind to the genome by TPX formation. Those results suggest that an important fraction of lncRNAs can exert its biological function through this mechanism. The software is available at https://github.com/molinerisLab/3plex.

1. Introduction

LncRNAs are a class of RNA molecules longer than 200 nucleotides, mainly transcribed by the RNA polymerase II and often spliced and polyadenylated. They can fold in peculiar structures and may show highly specific tissue and development expression levels. Notably, many lncRNAs have been shown to regulate gene expression both transcriptionally and post-transcriptionally by interacting with proteins, other RNA molecules or with the DNA [1,2].

One of the molecular mechanisms that enables lncRNAs to bind specific DNA sequences *in cis* or *in trans* is the formation of RNA:DNA:DNA triple helices (TPXs). These structures are formed by an RNA molecule that, laying on the DNA major groove, establishes hydrogen bonds with DNA purine nucleotides already paired in the double helix. Those RNA-DNA hydrogen bonds do not follow the usual Watson-Crick laws, but the Hoogsteen rules [3,4]. The bound portion of the lncRNA is named Triplex Forming Oligonucleotide (TFO) and the cognate portion in the DNA molecule is named Triplex Target Site (TTS). The RNA molecule can be directed in parallel or antiparallel orientation with respect to the bound DNA strand. These configurations require different RNA nucleotide compositions: RNA pyrimidine motifs (C and U) can form only parallel Hoogsteen triplexes, RNA purine motifs (G and A) exclusively antiparallel meanwhile mixed motifs (composed of G and U), can form triplexes in both the orientations [3,4].

Some lncRNAs are known to regulate gene expression of specific target genes through binding to specific TTS in their DNA regulatory regions, acting analogously to transcription factors (TFs). For

Abbreviations: dsDNA, double stranded DNA; IDR, Irreproducible Discovery Rate; lncRNA, long non-coding RNA; ssRNA, single stranded RNA; TF, Transcription Factor; TFO, Triplex Forming Oligonucleotide; TPX, Triple Helix (Triplex); TTS, Triplex Target Site

* Corresponding author at: Dipartimento di Scienze della Vita e Biologia dei Sistemi and MBC, Università di Torino, Via Nizza 52, 10126 Torino, Italy.
E-mail addresses: salvatore.oliviero@unito.it (S. Oliviero), ivan.molineris@unito.it (I. Molineris).

https://doi.org/10.1016/j.csbj.2023.05.016

example, the murine lncRNA EPR binds on Arrdc3 promoter, thus activating its expression and modulating the epithelial to mesenchymal transition [5]. Similarly, LncSmad7 binds and recruits p300 to enhancer regions *in trans*, triggering their acetylation and transcriptional activation of their target genes, controlling the expression of key stemness regulators [6]. Other noteworthy examples can be found in recent reviews [7,8].

In order to investigate these lncRNA:DNA interactions at genome-wide level, several experimental methods have been developed. The first one was the Chromatin Isolation by RNA purification (ChIRP) technique [9], which isolates the genomic regions bound by a specific RNA using biotinylated oligonucleotides on cross-linked nuclei. The purified DNA is then sequenced using a next generation sequencing system. A few variants of this method exist: ChOP-seq [10], CHART-seq [11] and RAP-seq [12]. We can define this group of methodologies as "one-to-all", in contrast with the "all-to-all'' experimental techniques which enable the identification of all the RNA transcripts interactions with the chromatin exploiting proximity ligation: MARGI [13], GRID-seq [14], RADICL-seq [15], RedC [16], RedChIP [17] and CHAR-seq [18]. Of particular interest for TPX investigation is the RADICL-seq technique, which provides a non-protein mediated (NPM) RNA:DNA interactions dataset. In fact, it is worth to note that none of the other mentioned all-to-all techniques are able to discriminate between TPX and other RNA:DNA interactions (e.g. protein mediated binding or R-loops [19]). Nevertheless, all-to-all methods present different limitations, mostly related to the low expression levels of the lncRNAs and to the complexity of the experimental procedures. Indeed, most of the detected contacts involve nascent transcripts or highly expressed lncRNAs. For instance, Li et al. reported that lncRNAs constituted only the 10% of RADICL-seq isolated transcripts, the majority of them having a modest DNA interaction count (median around 10). Moreover, most of the *trans* interactions were lost because of the experimental complexity (0.6% in NPM dataset) [15]. A similar behaviour is observed in RedC data, where the RNA showed the highest interaction frequency in the vicinity of the gene and then along the same chromosome [16]. Apart from one-to-all and all-to-all techniques, Sentürk Cetin et al. developed a method that retrieves RNA and DNA sequences specifically involved in TPX structures. However, this technique does not permit the identification of direct RNA:DNA pairs [20].

From the computational point of view, a few bioinformatic software have been developed to predict the ability of a lncRNA to form TPXs identifying TFO and TTS couples. The first published TPX prediction tool was *Triplexator* [21], whose algorithm is based on the set of canonical Hoogsteen pairing rules to find all the TFOs, TTSs and their TPX matches. *Triplex Domain Finder (TDF)* is one of the most used TPX prediction tools and is based on the same logic of *Triplexator*, with a further statistical evaluation of the TPX forming potential of the RNA [22]. Another approach is implemented in *LongTarget* [23] and its recently improved version *fasim-LongTarget* [24]. Those tools aim to search for genome-wide TPX interactions considering non-canonical Hoogsteen rules and implement an estimate of the TPX stability based on denaturation experiments and frequency of observed nucleotides in TPXs. More recently, two methods that take advantage of machine learning algorithms have been developed. *TriplexFPP* uses the feature extracted from a convolutional neural network (CNN) trained with experimental TPX data to predict if a RNA sequence can form TPXs [25]. *TriplexAligner* instead uses probabilistic nucleotide pairing models learned by expectation-maximisation from training experimental TPX data [26]. Warwick et al. recently reviewed these computational methods [27].

Despite the effort in the development of prediction algorithms, TPX investigation remains challenging. We particularly investigated the following aspects: 1) What is the typical length of a TPX? 2) Can a TPX stability estimate improve the prediction? 3) Can we increase the prediction accuracy by integrating the RNA secondary structure information?

An important parameter that affects the performance of TPX prediction software is the minimum number of nucleotides involved in the binding. For instance, this parameter is set by default at 16 in the most used approaches (*Triplexator* and *TDF*). *LongTarget* and *fasim-LongTarget* focus on long TPXs in general and consider a default minimal length of 20. Two are the principal reasons to focus on long TPXs: 1) the identification of small TPXs is computationally demanding, 2) longer sequences are required for a stronger and more specific match. For these reasons, shorter TPXs have been so far mainly neglected in this type of analysis. As an example, Jalaly et al. [17] performed a genomic survey of potential TPXs in the human genome by using a cut-off of 35 for the TPX length. Yet, a rigorous discussion on TPX minimum length is required in order to improve the understanding of this molecular interaction and its biological role. Indeed, different investigations of TF binding have shown that transient binding to low-affinity sequences plays an important role [28,29] and in principle the same could be true for TPX-forming lncRNAs. Some recent observations made by Matveishina at al. point in this direction [30].

Another key aspect of predictive algorithms is the choice of a proper method to score the results. *Triplexator* and *TDF* rank the identified TPXs by counting the number of nucleotides involved in the interaction (matches). *Triplexator* additionally computes the TPX potential of a DNA region, considering all the possible TPXs it can form given an RNA sequence. *TDF* instead performs randomisation tests to associate a p-value to each set of overlapping TFOs. *LongTarget* takes advantage of in vitro studies of TPX structures (e.g. thermal stability derived from denaturation experiments and frequency of observed nucleotides in TPXs) summarised in [23]. Lastly, *TriplexAligner* scores the TPXs by computing the E-value of the sequence alignment produced with the codes learned by expectation-maximization on training data. Here, we evaluated different scoring strategies and we observed a better performance using stability derived measures compared to those based on simple match counts.

Many lncRNAs fold in peculiar secondary RNA structures involving hydrogen bonds between its nucleotides [31]. Theoretically an RNA region can only form Hoogsteen hydrogen bonds in a TPX structure if its nucleotides are not already bound in Watson-Crick double strand [32], it is reasonable to expect that TPXs could originate from single-stranded RNA regions only. Indeed, the validated TFOs of the human lncRNA *MEG3* correspond mainly to unpaired RNA regions, as identified in the experimentally determined secondary structure model [33]. Moreover, a previous work suggests that RNA secondary structure can improve the TPX prediction specificity [30]. Despite this evidence, there is no TPX prediction software that integrates RNA structure.

Here we present *3plex*, a TPX prediction software developed by integrating Hoogseten pairing rules as implemented in *Triplexator*, computational evaluation of thermal stability derived from *LongTarget* and RNA secondary structure prediction.

2. Methods

2.1. Collection of high throughput RNA:DNA interaction data

We collected high-throughput RNA:DNA interaction data primarily through manual mining of literature published from 2011 (when the first ChIRP-seq was published) to 2022. Gene Expression Omnibus (GEO) [34], PubMed and European Nucleotide Archive (ENA) [35] databases were queried using as keywords each "one-to-all" experimental techniques that map RNA-binding sites at the whole genome level including: ChIRP-seq [9], ChOP-seq [10], CHART-seq [11] and RAP-seq [12]. RNA molecules were uniformly annotated using GENCODE version 32 and M25 gene symbols [36] and the

lncRNA sequences were retrieved from the same database. When more than one transcript per gene is reported we selected the longest one. As the lncRNA *lncSmad7* transcript is not present in GENCODE, we took the sequence from the original publication [6]. We excluded from the analysis experiments without available raw data or with unclear annotation of the replicates. *Xist* has been excluded because the experimental data for this lncRNA were limited to the X chromosome [12]. The data related to some recent papers have not been analysed because they were published after the data collection was completed [37,38].

2.2. Data uniformation and replicate handling

We downloaded raw fastq files from GEO or ENA and analysed them using the ENCODE ChIP-seq pipeline (https://www.encode-project.org/chip-seq/transcription_factor/ version v1.6.1) [39]. The experimental designs were different across the collected data. The majority of the experiments used a couple of even/odd sets of probes that we defined as replicate 1 and 2 in the ENCODE pipeline. When more then one even/odd set of probes were available, peak calling was applied to each even/odd set considered as replicate 1 and replicate 2 and then the intersection of the called peaks is the final set of peaks.

In the ENCODE pipeline, after SPP [40] peak calling with low stringency, replicates are handled by defining overlapping peaks or by measuring the Irreproducible Discovery Rate (IDR) [41]. Both these methods can be applied to true replicates only (conservative procedure) or considering the pseudoreplicates obtained by subsampling pooled replicates (optimal procedure). The optimal set is more sensitive, in particular when one of the replicates has substantially lower data quality than the other. We additionally filtered the Overlap set by considering only those peaks having a SPP q-value below 0.05. We defined all these replicate handling procedures as “peak filtering methods”. Apart from IDR (conservative and optimal) and Overlap (conservative and optimal) peaks sets, we created a selection of the top 1000 peaks (Top1000) obtained as follows: 1) preliminary selection of IDR conservative peaks, 2) if the number of selected peaks is below 1000, inclusion of the best IDR optimal peaks, not overlapping with IDR conservative ones, 3) if the number of selected peaks is still below 1000, addition of the best Overlap conservative peaks that do not overlap with any of the IDR peaks.

ENCODE guidelines suggest to evaluate the reproducibility of the data by computing:

- N1: Replicate 1 self-consistent peaks (comparing two pseudoreplicates generated by subsampling Rep1 reads)
- N2: Replicate 2 self-consistent peaks (comparing two pseudoreplicates generated by subsampling Rep2 reads)
- Nt: True Replicate consistent peaks (comparing true replicates Rep1 vs Rep2)
- Np: Pooled-pseudoreplicate consistent peaks (comparing two pseudoreplicates generated by subsampling pooled reads from Rep1 and Rep2)
- Self-consistency Ratio: max(N1,N2) / min(N1,N2)
- Rescue Ratio: max(Np,Nt) / min(Np,Nt)

Then the reproducibility test returns “ideal” if Self-consistency Ratio > 2 and Rescue Ratio > 2, “acceptable” if one of the two ratios is < 2, “concerning” if both the ratios are < 2.

2.3. 3plex scoring method based on thermal stability

Given a TPX found by running *Triplexator* algorithm, a triplet is defined as a group of two Watson-Crick interacting nucleotides in the dsDNA and one Hoogsteen interacting nucleotide in the ssRNA. *3plex* implements a new scoring method based on experimentally in vitro determined triplets stability data derived from He et al. [23] (Supplementary Table 5). We defined the Normalised stability as the length normalised sum of thermal stability of all the TPXs predicted on the entire considered DNA sequence. In particular, we used *bedtools merge* [42] to avoid overestimation of stability. Since different TPXs can overlap, for each set of overlapping TPXs we considered only the one with maximal stability. Hence, we selected the maximal stability value among all merged TTSs for each dsDNA and we divided it by the dsDNA length. This scoring method undercounts the stability for each dsDNA, indeed the contribution to the stability of some dsDNA portion is discarded because of partial overlap. Using experimentally determined binding sites of 7 TPX-validated lncRNAs, we compared the performance of this scoring in terms of AUC with an alternative that overcounts the stability by ignoring overlaps. We found that the undercounting performs similarly or slightly better than the overcounting.

2.4. RNA secondary structure prediction

RNAplfold from the *ViennaRNA* package [43] used with -u option measures the probability that stretches of n sequential nucleotides on the RNA sequence are unpaired, which is useful for predicting possible binding sites. Given an RNA sequence, we associated to each nucleotide the probability that a window of length 8 centred on that position is single stranded. Subsequently, we masked nucleotides from the RNA sequence which report a probability value lower than a certain cut-off. The selected cut-offs enable masking the 10%, the 20% or the 50% of the sequence (ss10, ss20, ss50). The unmasked sequence is defined as ss0.

2.5. Positive and negative regions selection for 3plex parameter evaluation

Positive regions are peaks identified with the 5 different peak filtering methods available as described in Section 3.2. If more than one even/odd set were provided, the intersection of the called peaks is considered.

Negative regions are randomly selected from the genome maintaining the same length distribution of the positive ones using *bedtools shuffle*. Positive regions, genome gaps and ENCODE blacklist regions are excluded from the selection. When the Top1000 peaks were considered as positive regions, peaks found in other datasets (e.g. IDR optimal) for the same lncRNA were excluded from the random selection.

We obtained ROC curves and AUC values for different scoring methods using *pROC* package version 1.17.0.1 [44]. We compared different AUC using *roc.test* of the same package (two sided Delong’s test). One sided Mann-Whitney test was performed using the R function *wilcox.test* in order to test the significance of the difference between the scores of the positive and negative regions. Benjamini-Hochberg correction on p-values was computed using the R function *p.adjust* setting method to “BH”.

2.6. Linear models for parameters relevance evaluation

To evaluate the relevance of parameters in the TPX prediction, we explored the following parameter space:

- Peak filtering method, described in Section 3.2, with possible values:
  - o IDR conservative
  - o IDR optimal
  - o Overlap conservative, filtered at q < 0.05
  - o Top1000 as described in Section 3.2
  - o The set of peaks provided by the authors of the experiment

- ssRNA, a factorial variable indicating the specific lncRNA
- the number of peaks for the given lncRNA and peak filtering method
- single strandedness, as described in Section 3.4, with possible values:
  - ss0
  - ss10
  - ss20
  - ss50
- the minimal length of TPX to be considered, with possible values:
  - 16
  - 12
  - 10
  - 8
- the minimal error rate allowed in a TPX, with possible values:
  - 10%
  - 20%
- the maximum number of consecutive errors allowed in TPX, with possible values:
  - 1
  - 3
- the minimal guanine rate allowed in TTS, with possible values:
  - 10%
  - 40%
  - 70%
- repeat filter, indicating if low complexity regions in ssRNA should be masked or not
- the TPX scoring method as described in Sections 3.3 and 4.2.3, with possible values:
  - *Triplexator* potential
  - Normalised stability
  - *Triplexator* best score
  - Best stability

We considered 2 different models.

The first model aims to specifically investigate the effect of the minimal TPX length parameter on TPX prediction performance. The parameters already discussed in Matveishina et al. [30] were fixed at the value suggested by the authors (namely: minimal length=10, error rate=20%, guanine rate=40%), meanwhile the others parameters were kept at default level. For this specific investigation, we excluded 3plex specific parameters (thermal stability scoring methods and secondary structure). We computed AUC for the 7 TPX-validated lncRNAs for each combination of peak filtering methods and minimal TPX length taking into consideration the Triplexator potential score. These AUC values are represented in Fig. 2, panel "minimal length". Subsequently, we fitted a linear model (lm function on R) to study the dependency of AUC on different parameters controlling for the ssRNA as covariate (formula: AUC ~ peak filtering method + ssRNA + n peaks + minimal length) model. Moreover, we fitted an analogous nested model with the same formula but excluding the minimal length parameter. By doing this, we could evaluate the difference in performance between the two models (ANOVA function of R) obtaining a p-value $P_{anoval} < 2.2e\text{-}16$. Since the different observations are not independent, the analytically estimated P_{anova} considers inflated degrees of freedom and is not reliable. To obtain a statistically correct estimation of the difference in performance between the full and nested model, we performed a randomisation test. We computed 10000 permutations of the AUC versus parameter association and fitted the full and nested model on randomised data as before. Noticeably, randomised data maintained the correlation structure of real data and only the association between the AUC and the independent variables was lost. For each permutation we computed the p-value P_{anova_random} comparing full and nested models. To measure the relevance of the minimal length parameter, we counted the number of permutations resulting in $P_{anova\_random} <= P_{anova}$. Since we did not obtain any random p-value less than the real one, we can say that the minimal length parameter is relevant and we can estimate its probability as less than 1e-4.

Subsequently, we considered the full parameter space of *3plex* (formula: AUC ~ peak filtering method + ssRNA + n peaks + single strandedness + minimal length + error rate + guanine rate + repeat filter + consecutive errors + scoring method), still focusing on the 7 lncRNAs having experimental evidence of functional TPX. To reduce the computational time we considered only the values 8 and 10 for the minimal length. These AUC values are represented in Fig. 2 (all panels but "minimal length"). We performed a permutation test as previously explained for each parameter and we obtained a randomised p-value lower than 1e-4 for all of them except for minimal length (p-value < 1e-3) and consecutive error (not significant).

To efficiently explore the parameter space considered we leveraged the Paramspace functionality of snakemake [45].

2.7. Positive and negative regions selection for 3plex comparison with other TPX prediction software

For TPX software comparison, we considered as a positive dataset the Top1000 peaks set for the 7 TPX-validated lncRNAs (see Methods). We set the parameters of *3plex*, *TriplexAligner* and *fasimLongTarget* to default. We associated each peak with the following scores: Best stability for *3plex*, best -log10E for *TriplexAligner* and best MeanStability for *fasim-LongTarget*. If no interaction was predicted with a DNA sequence, we assigned the score 0. Several *TriplexAligner* interactions scored infinite, in those cases we converted the value to 1000 for the AUC computation. In order to compare the performance of the TPX prediction tools, we generated two different analysis strategies:

1) In the first strategy, we ran the software on a random selection of genomic regions that maintains the same length distribution of the positive ones by using *bedtools shuffle*. We excluded all the positive regions, genome gaps and ENCODE blacklist regions. We repeated the negative region sampling 1000 times in order to estimate the AUC standard deviation.

2) In the second strategy, the negative control cases were generated by running the TPX software on the positive dataset and shuffling the transcripts with *fasta-shuffle-letters* from the *MEME* suite with *-kmer* 1 [46].

2.8. TF binding sites prediction evaluation

A set of TFs for binding sites prediction evaluation was selected from Jayaram et al. [47] in the human K562 cell line (CTCF, E2F1, GATA2, IRF1, MAX, NFYA, TAL1, YY1). We downloaded the ChIP-seq IDR conservative thresholded peaks from the ENCODE database as true positive binding sites. Hence, we selected genomic random regions as negative controls, maintaining the same length distribution of the positive ones using *bedtools shuffle* as for lncRNAs binding sites. We downloaded the PWM models of binding sites for those TFs from *JASPAR* [48] (respectively MA0139, MA0024, MA0036, MA0050, MA0058, MA0060, MA0091, MA0095) and used *FIMO* [49] from the *MEME* suite [46] with standard parameters in order to associate a score to each genomic region. The obtained values were employed to evaluate the performance of the TFs binding sites predictor using ROC and AUC values, computed with the *pROC* r package [44] as for lncRNAs binding sites.

2.9. TriplexFPP

We ran *TriplexFPP* [25] as described by the authors (https://github.com/yuuuuzhang/TriplexFPP). We used the lncRNAs sequences as input and executed the "triplex lncRNA prediction" section in TriplexFPP.ipynb.

2.10. Computation time testing

The tests were performed on a dedicated physical cluster computing node, equipped with AMD EPYC 7313 processor with 16 cores, 250 GB of RAM and NVMe storage. *Triplexatror* and *3plex* have the capability of using many computing cores per single process, but for this test we fixed them to 1. We used as input two ssRNA sequences, the shorter (*TERC* transcript, 541 nucleotides) and longer (*NEAT1* transcript, 22743 nucleotides) among the TPX-validated lncRNAs, thus we ran 10 times each software with the short sequence and 10 time with the long one. Input DNA sequences were the Top1000 peaks for both ssRNA plus the negative random controls built as described in the section 3.7.

To evaluate the computation cost of *3plex* specific features we run *Triplexator* using two settings, 1) default parameters and 2) the optimal parameters setting we described in section 4.2.7 (excluding the *3plex* specific parameters) that is the default setting for *3plex*. In the second case *Triplexator* is slower for two principal reason 1) lower minimal length require many more putative TPXs to be evaluated, 2) the combination of error-rate and minimum length does not allow for efficient filtering with q-grams (in this case *Triplexator* is forced to use brute-force approach).

The Fig. 6 represents average wall-clock times, standard deviations are always < 5% and not represented.

3. Results

3.1. 3plex implementation

We developed *3plex*, a software for the RNA:DNA:DNA TPXs prediction. *3plex* is built on top of *Triplexator*, which is the fastest and most used tool in the field [21]. We further integrated TPX evaluation methods based on thermal stability derived by *LongTarget* [23] and RNA secondary structure prediction with the *ViennaRNA* package (Fig. 1).

The input of the *3plex* algorithm is a single stranded RNA sequence (ssRNA) and a set of double stranded DNA sequences (dsDNAs). The output consists of 1) all possible TPXs that satisfy a set of constraints derived from *Triplexator* with an associated thermal stability evaluation, and 2) a score for each dsDNA that predicts the capability of a ssRNA to bind to it.

3.2. 3plex performance evaluation

In order to evaluate the performance of *3plex*, we collected the experimentally determined lncRNA:DNA binding events at genomic level in humans and mice resulting in approximately 800 GB of data (see Methods). We uniformly re-analysed the available data by employing the ENCODE ChIP-seq pipeline. Among the lncRNAs for which the DNA binding sites have been determined, we identified 7 (in human: *TERC*, *NEAT1*, *HOTAIR*, *MEG3*, *ANRIL*; in mouse: *Meg3*, *lncSmad7*) for which the TPX binding mechanism have been experimentally validated and we called them TPX-validated lncRNAs (Supplementary Table 1). For these 7 TPX-validated lncRNAs that are known to form TPXs, we expected a good fraction of RNA binding sites with high predicted TPX scores. Consequently, we could use those experimentally identified lncRNA binding sites as a benchmark to evaluate the goodness of TPX prediction approaches.

3.2.1. Minimal TPX length

In order to investigate the influence of the minimal TPX length on the prediction, we started by comparing the *Triplexator* default minimal TPX length of 16 with shorter lengths (12, 10 and 8). Indeed, previously reported cases of TPX interactions between lncRNAs and genomic regions showed that even few nucleotides are sufficient to establish a functional recognition (see Introduction). To carry out this analysis, we fixed the previously investigated *Triplexator* parameters to the values suggested by Matveishina et al. [30] and we excluded *3plex* specific parameters (thermal stability and secondary structure). We computed the AUC for each minimal TPX length and then we compared a multivariate linear model that investigates the dependency of the AUC on different parameters (minimal TPX length, number of peaks, peak filtering method and different ssRNA) with a similar nested model where the dependency on minimal TPX length is removed. We observed a significant influence of this parameter (p-value < 1e-04, permutation test), in particular the best minimal TPX length values were 8 and 10 (equally). The default cut-off of 16 resulted to be the worst, producing an average reduction in AUC of 0.1, meanwhile 12 gave an intermediate performance (Fig. 2).

Considering the obtained results on the minimal TPX length we fixed that parameter to 8 or 10 and we further investigated *3plex* parameter space by introducing other parameters in the linear model (single strandedness cut-off, minimal error rate, guanine rate, repeat filter, maximal number of consecutive error, scoring method). Again, we compared the dependency of the AUC from these parameters with a similar nested model where the dependency from minimal TPX length is removed. We obtained a moderate, but significant preference for a value of 8 over 10 (average AUC difference of 0.005 p-value < 9e-4, permutation test).

Noteworthy, 8 is below the minimum available threshold of *Triplexator*. Therefore, in order to allow for a more accurate TPX prediction, we modified the source code of *Triplexator* to permit the identification of small TFOs and TTSs.

Our observations highlight the fact that typical TPXs are shorter than previously expected. We evaluated the minimal TPX length of 6 in a subsample of the parameter space and did not obtain any further improvement (Supplementary Figure 1).

3.2.2. Error rate and number of consecutive errors

With the same strategy used to evaluate the relevance of the minimal TPX length, we also evaluated the relevance of the error rate. In this context, an error is a non-proper pairing in the TTS and TFO that forms a TPX, as well as deviation from the rules that sequences forming TSSs and TFOs should satisfy. The linear model shows that this parameter is relevant (p-value < 1e-4 permutation test). Setting the error rate to 20% resulted in an increase of average AUC of 0.02 compared to 10%. Moreover, we tested if the number of consecutive errors was relevant and we observed no significant differences in average AUC, allowing for just 1 or up to 3 consecutive errors.

These data suggest that TPXs are generally relatively small and degenerated. This observation could be surprising, since it seems that such models account for too little information to be biologically relevant. Actually, these binding models do not differ much from many famous TF binding models considering their typical length and degeneration (see Discussion). We quantified this consideration by comparing ROC curves describing the prediction of RNA binding sites and those describing the prediction of TF binding sites (Fig. 3).

3.2.3. Effect of the thermal stability computation

3plex implements a thermal stability evaluation of the predicted TPXs similar to the one developed in *LongTarget* [11]. In particular, given a couple of ssRNA and a dsDNA sequences, *3plex* finds all the TPXs that satisfy the constraints as in *Triplexator*, then reports the best and the average stability over all the identified TPXs. To evaluate

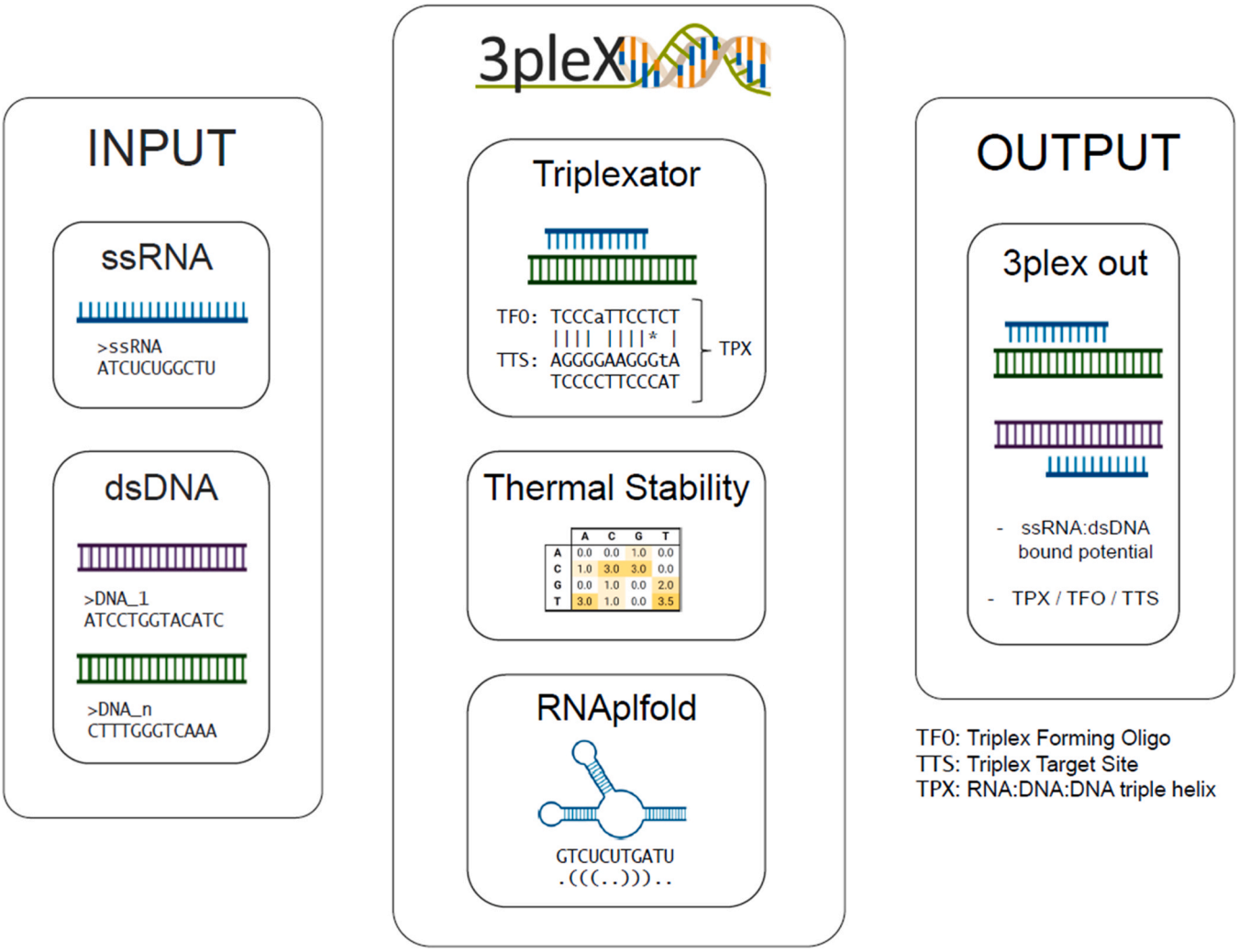

Fig. 1. *3plex* analysis workflow. *3plex* is a TPX prediction software that integrates methods from *Triplexator*, *RNAplfold* and *LongTarget*. *Triplexator* algorithm scans a couple of ssRNA and dsDNA sequences considering the Hoogsteen hydrogen bonding set of rules, and a score computed from a collection of in vitro denaturation experiments is assigned to each TPX. *RNAplfold* secondary structure prediction can be used in order to mask ssRNA regions. *3plex* outputs the list of all the possible TPXs, TFOs and TTSs and calculates a TPX bound potential of the target regions.

the probability that a ssRNA may form a TPXon a given dsDNA sequence, we compared 4 alternative TPX scoring methods:

1. The Best stability (the stability score of the more stable among the predicted TPXs).
2. The Normalised stability (the length normalised average thermal stability of the predicted TPXs).
3. The *Triplexator* best score (the highest *Tripelxator* score among predicted TPXs, where the score is computed as the TPX length minus errors, such as mismatches or nucleotides not suitable for the TFO and TTS model).
4. The *Triplexator* potential (the commonly used TPX potential returned by *Triplexator*, that is the length normalised number of TPXs [21].

The *Triplexator* best score and the Best stability score collectively correspond to “maximum metrics” as they summarise numerous small predicted TPXs with the highest scoring one. Instead, the *Triplexator* potential and the Normalised stability take into account the contribution of all the TPXs in a certain DNA sequence.

We considered a full multivariate linear model including all studied parameters as covariates (peak filtering method, single strandedness cut-off, minimal error rate, guanine rate, repeat filter, maximal number of consecutive error, scoring method) and evaluated the effect of the TPX scoring method by comparing the full model with a similar nested model where the dependency from TPX scoring method is removed. We found that the AUC depends on this parameter (p-value < 1e-4, permutation test). In particular, the Best stability method outperformed the other scoring functions and produced an increase in average AUC of 0.01 if compared with the usual *Triplexator* potential (Fig. 2).

Considering each lncRNA individually, the parameter sets that produced the best AUC required a stability based scoring method (Best stability or Normalised stability) for 4 out of 7 lncRNAs. By changing the scoring method while maintaining all the other parameters fixed, we found that there was a statistically significant reduction in AUC when the scoring method was not based on stability (*Triplexator* best score or *Triplexator* potential). On the contrary, 3 lncRNAs showed the best AUC with *Triplexator* best score or *Triplexator* potential. It should be noted that in these cases the difference in AUC with respect to the one computed using a stability based scoring method is not significant (see Supplementary Table 2).

3.2.4. Peak filtering method

The use of different peak filtering methods (see Section 3.2) is one of the covariates included in the multivariate model. By

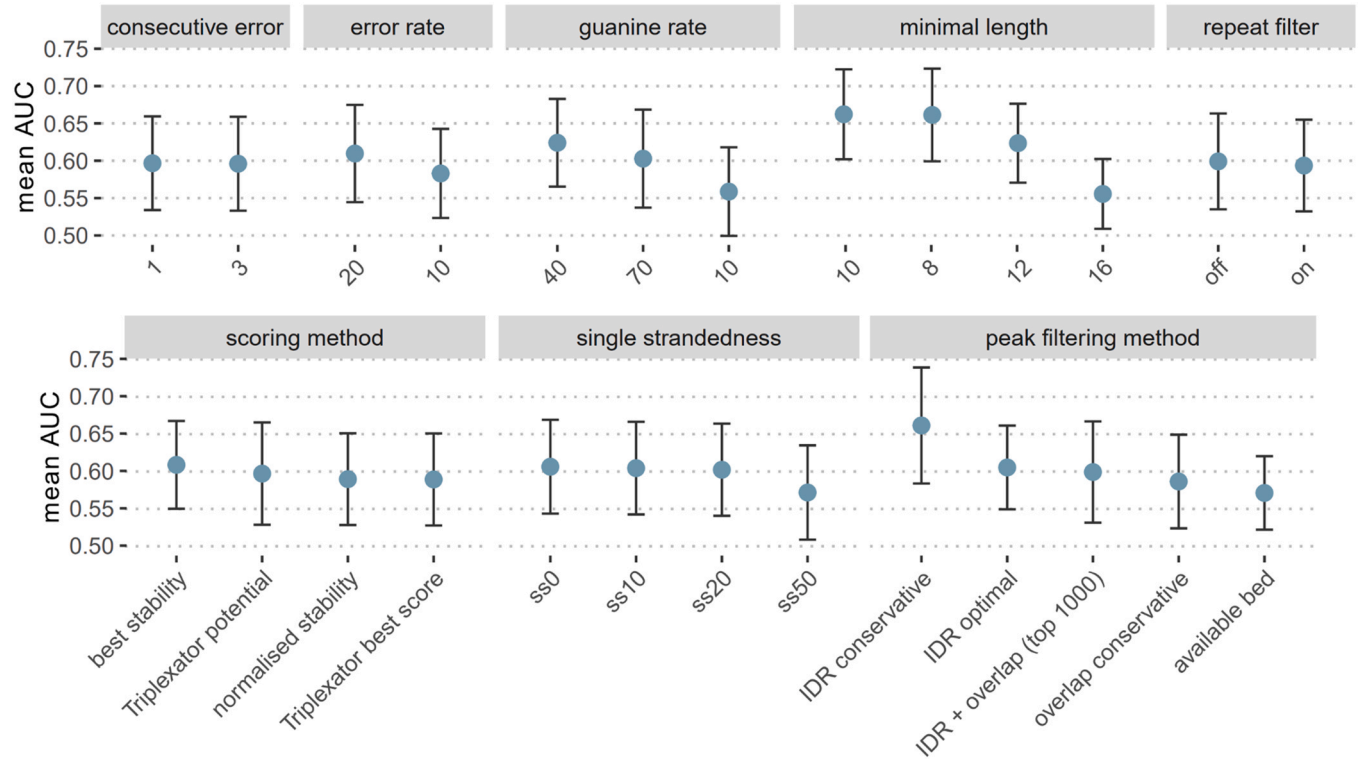

Fig. 2. Impact of various parameters in TPX prediction accuracy. *3plex* prediction performance is expressed as AUC values obtained by comparing TPX scores of experimentally determined RNA binding sites with randomised negative genomic regions. Every tested parameter combination produces an AUC value for each specific lncRNA. For all the levels of the investigated parameters, the average AUC is represented as a blue dot, error bars represent standard deviation.

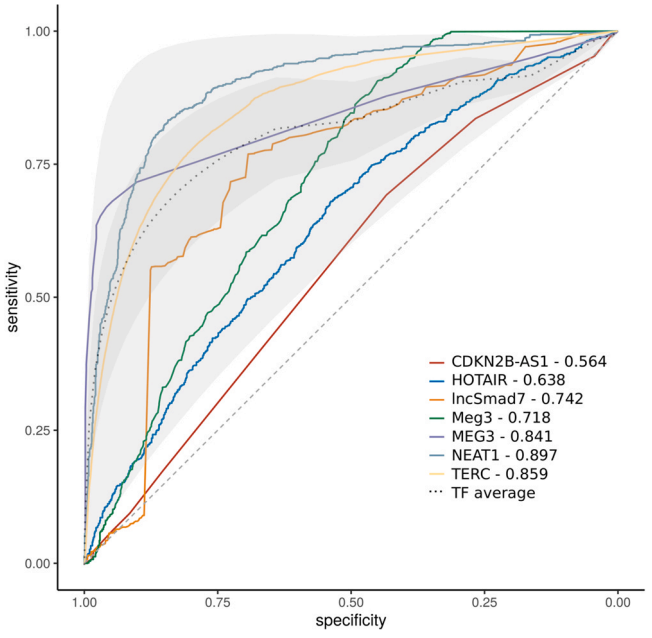

Fig. 3. ROC curves showing the performance of TPX prediction applied to lncRNAs compared with the performance of DNA binding sites prediction applied to TFs. The coloured continuous lines are the ROC curves obtained by applying *3plex* to each lncRNA with experimental evidence of functional TPX. *3plex* parameters combination is the one that produced the best AUC value (see Supplementary Table 2). The grey dashed line represents the performance of a random classifier. The grey dotted line illustrates the average ROC curve computed from the performance evaluation of *FIMO* TFs binding sites predictions. The dark grey shade represents the standard deviation of ROC curves for TFs, the light grey shade boundaries correspond to the worst and best TF ROC curves. AUC values of TPX prediction are reported in the figure legend.

comparing the full model with the corresponding nested one and excluding this parameter, we found that it is relevant for the AUC (p-value < 1e-4, permutation test). Interestingly, we observed that the RNA binding sites redefined using the ENCODE pipeline showed on average higher AUC value than the sets of peaks provided by the authors. The best method was the IDR conservative one, producing an increase in average AUC of 0.09 compared with published peaks sets. Only 2 out of 7 lncRNA considered in this analysis had a defined set of IDR conservative peaks, while the other 5 showed insufficient reproducibility among replicates to apply this method. Anyway, we controlled for different lncRNAs as covariate in our model. Moreover, considering IDR optimal peaks that are available for 5 out of 7 lncRNA, we still obtained a significant increase in average AUC of 0.015 (Fig. 2).

These results highlight the importance of a uniform and standard practice of RNA:DNA interaction data. Indeed, the peaks derived from custom analysis by the authors of single papers (available for 6 out of the 7 lncRNAs) showed the worst performance.

3.2.5. RNA secondary structure

In *3plex* we implemented a parameter that determines the importance of secondary structure in the prediction of TPX by masking the RNA sequence according to the probability of base pairing (see Methods).

Among the available tools implemented in the *ViennaRNA* package [43], we focused on *RNAplfold* because theoretically a TPX interaction requires the stretch of nucleotides on the RNA transcript not to be involved in any further hydrogen bonds. *RNAplfold* specifically computes the probability that a stretch of consecutive nucleotides is unpaired in order to predict possible binding sites. Additionally, Matveishina et al. reported an improvement in TPX prediction accuracy for some lncRNAs by masking RNA regions according to this tool's predictions [30].

By testing different extents of sequence masking (0%, 10%, 20% or 50%), we found that 3 out of 7 TPX-validated lncRNAs (*MEG3* in human and *Meg3* and *lncSmad7* in mouse) had an improvement in the AUC using secondary structure filtering (Supplementary Table 2). Thus these 3 lncRNAs may preferentially harbour the TFOs in regions likely to be single stranded, while the others may have the TFOs in regions predicted to be in a double strand conformation.

These observations suggest that RNA secondary structure has an impact on TPX prediction that may be further investigated and *3plex* allows the users to consider this information in the investigation of lncRNAs TPXs.

3.2.6. Other parameters

According to constraints due to Hoogsteen hydrogen bonding rules, a minimum guanine rate in the TTS should be required in TPX prediction. Our analysis showed that this parameter is relevant (p-value < 1e-4, permutation test) and in general a minimum rate of 40% is preferable with respect to 70% or 10%, producing an increase in average AUC of 0.06 compared to the baseline 10%.

The possibility of filtering out the dsDNA repetitive elements from the TPX search is implemented in *Triplexator*. The choice of this parameter significantly impacts the AUC in general (p-value < 1e-4, permutation test) and it is remarkable that avoiding filtering improves the TPXprediction.

3.2.7. Default parameter settings

Taken together, our *3plex* parameter space exploration resulted in a combination of optimal default settings we suggest using for TPX investigation. In particular, we recommend setting minimum TPX length 8, minimum error rate 20%, minimum guanine rate 40%, number of consecutive errors 1, repeat filter off, single strandedness 0 and Best stability as scoring metrics. *3plex* runs with this combination of parameters as default, returning all the scoring metrics.

3.3. 3plex comparison with other TPX prediction software

In the previous section we showed that *3plex* improves the TPX prediction accuracy with respect to *Triplexator.* Afterwards, we compared *3plex* with the other available TPX prediction software, *TriplexAligner* [26] and *fasim-LongTarget* [24]. We excluded *TDF* from this evaluation since its algorithm is built on the same logic of *Triplexator*. In order to compare the different methods, we chose to set the parameters to default for all the software. In particular, we considered maximum metrics for TPX interaction scoring (*TriplexAligner*: -log10E; *fasim-LongTarget*: best meanStability; *3plex*: Best stability). We selected the Top1000 RNA:DNA interaction peaks for the 7 TPX-validated lncRNAs as a positive dataset and randomised genomic regions as a negative dataset (see Methods). In this strategy, we found that *3plex* outperforms the other software. We furthermore confirmed these results by performing a resampling procedure and obtaining the following AUC: *TriplexAligner:* 0.519 ± 0.004; *fasim-LongTarget:* 0.595 ± 0.004; *3plex:* 0.660 ± 0.002 (Fig. 4 A). We noticed that the AUC value of *TriplexAligner* in our strategy was considerably different with respect to the one reported by the authors. We reasoned that this discrepancy may be related to differences in our approaches. (Fig. 5).

Firstly, we considered TPX-validated lncRNAs from both human and mouse with an associated experimental one-to-all RNA:DNA interaction dataset. Warwick et al., instead, tested nuclear expressed RNAs whose genomic interaction was detected through RADICL-seq or Red-C techniques. One-to-all techniques typically result in thousands of peaks for a given RNA, while all-to-all RNA:DNA interaction techniques report a median of 10 interactions for a given RNA (see Introduction). Notably, these techniques are highly susceptible to the level of expression of different lncRNAs, indeed few outlier lncRNAs account for a large fraction of the total interactions found. On the contrary, the Top1000 peaks set used in our testing strategy avoids biases towards highly expressed lncRNAs. Other differences between one-to-all and all-to-all techniques are discussed in the Introduction. For these reasons, we decided to focus on one-to-all data which are more robust, balanced and therefore still to be considered as the gold standard for the identification of RNA:DNA interactions.

Secondly, the randomization procedure used to generate negative controls differs: we chose to use random genomic regions as negative controls, while Warwick et al. shuffled the RNA sequences. Accordingly, we set up another testing strategy, similar to the one used by *TriplexAligner* authors. We defined the negative controls by RNA transcripts shuffling and ran the TPX prediction tools on the Top1000 peaks set of the TPX-validated lncRNAs. In this testing strategy the AUC of all the software increased and *TriplexAligner* resulted in the best performance (Fig. 4).

We reasoned that the first testing strategy was more stringent, as suggested by the lower AUC values reported for all the tested TPX prediction tools. This observation can be explained by the fact that negative random regions preserve properties of genomic sequences that can be lost in the sequence shuffling procedure. We developed *3plex* to specifically discriminate true lncRNAs binding sites from random genomic regions, notably, the first testing strategy directly and more strictly evaluated this ability and *3plex* outperformed the other tools in this case.

In terms of computation time *3plex* is slightly slower than *Triplexator* but greatly outperforms *TriplexAliner* and *fasim-Longtarget* (even avoiding the available internal parallelization available in *3plex* and not in *TriplexAligner* and *fasim-Longtarget*), retaining a comparable or superior accuracy (Fig. 6).

3.4. Evidence of TPX potential for other lncRNAs

From our RNA:DNA interaction data collection, we considered the remaining 20 human and murine lncRNAs with no experimental evidence of functional TPX mechanism of action. For two of them, the capability of forming TPXs *in silico* was tested by the authors (*AC087482.1*, *Eprn*) [5,50], while for *Tug1* the TPX mechanism of binding was only speculated [51].

We wondered for how many lncRNAs we could observe statistical evidence supporting a mechanism of DNA binding involving TPX formation, therefore we used *3plex* to predict TPXs on the available binding sites. We considered those sites as positive controls and created a randomised selection of genomic regions as negative control (see Methods), that we used to investigate the same parameter space as before.

Interestingly, we found 3 lncRNAs having an AUC above 0.8 (*AK156552.1*, *HAND2-AS1*, *MALAT1*), 4 lncRNAs above 0.7 (*AL109615.3*, *7SK*, *1200007C13Rik*, *CPEB2-DT*) and 6 lncRNAs above 0.6 (*Rn7sk*, *AC087482.1, Tug1*, *Eprn, AC018781.1, 1200007C13Rik@NDL, SRA1, MIR503HG, Bloodlinc*). All the mentioned AUC values had a corresponding highly significant p-value, even considering the high number of tests due to the parameter space exploration (FDR < e-10, Mann-Whitney test with Benjamini-Hochberg correction), suggesting that they may directly bind the DNA forming TPX structures (Fig. 4, supplementary table 3).

We compared these results with scores produced by *TriplexFPP* [25]. *TriplexFPP* uses a machine learning approach to face a different problem with respect to *3plex*: the classification of lncRNAs that are (or are not) prone to bind the DNA through TPX formation. For this reason, it cannot be used to predict where a specific lncRNA binds on the genome which is the specific purpose of *3plex*. Nevertheless, in this section we evaluated the capability of a lncRNA to form TPXs in general, thus we found it useful to

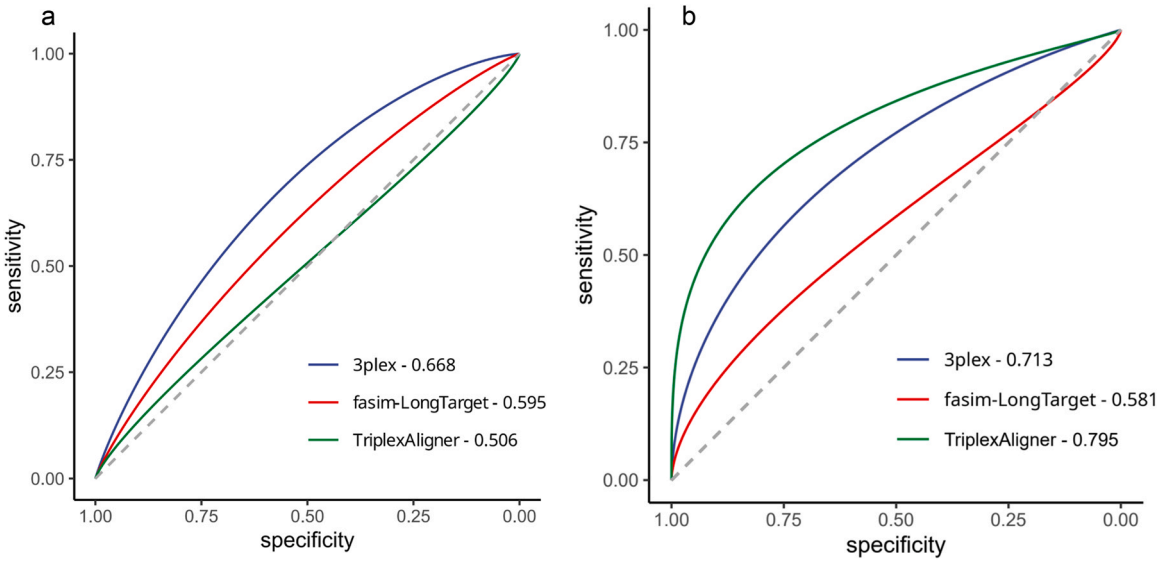

Fig. 4. *3plex* performance comparison with *TriplexAligner* and *fasim-LongTarget*. (A) ROC curves produced from the performance evaluation of the TPX prediction tools in the genomic region shuffling strategy. (B) ROC curves produced from the performance evaluation of the TPX prediction in the transcript shuffling strategy. The legends report the AUC computed for each software.

compare *3plex* results with those obtained by running *TriplexFPP*. Unfortunately, we did not find a good agreement between the two predictions. Out of 23 evaluated lncRNAs, only two were predicted to be TPX forming by TriplexFPP (*HOTAIR* and *MIR503HG*, see Supplementary Table 4). Notably, numerous lncRNAs experimentally known to form TPXs were not correctly identified by TriplexFPP, including *CDKN2B-AS1*, *lncSmad7*, *MEG3*, *NEAT1* and *TERC*. Our evaluation required experimental binding data for each lncRNA considered, and then data were interpreted according to a model of TPX formation. This approach is more powerful yet more demanding with respect to the machine learning approach of *TriplexFPP*, that does not require experimental binding data. A possible explanation for the poor agreement of the two predictions could be related to the training and testing datasets of *TriplexFPP* that are strictly related to HELA cells, while our collection of RNA:DNA interactions spans various cell lines and tissues.

4. Discussion

Numerous papers published in the last ten years have pointed out the RNA:DNA:DNA TPX interaction as one of the mechanisms that enable lncRNAs to specifically regulate gene expression. In order to investigate the functional role of a lncRNA it is necessary to be able to predict its putative TPX interactions. Various algorithms already exist for the identification and ranking of TPXs given an RNA and a DNA sequence. Nevertheless, this computational task remains challenging. Here, we presented *3plex*, a software for RNA:DNA:DNA TPX interaction prediction. We applied *3plex* to a collection of experimentally determined RNA binding sites in order to test the relevance of the available parameters and to compare *3plex* accuracy with respect to the other algorithms.

Interestingly, we showed that lncRNAs tend to form TPXs that are short (around 8–10 nucleotides) and degenerate (around 10%–20% of

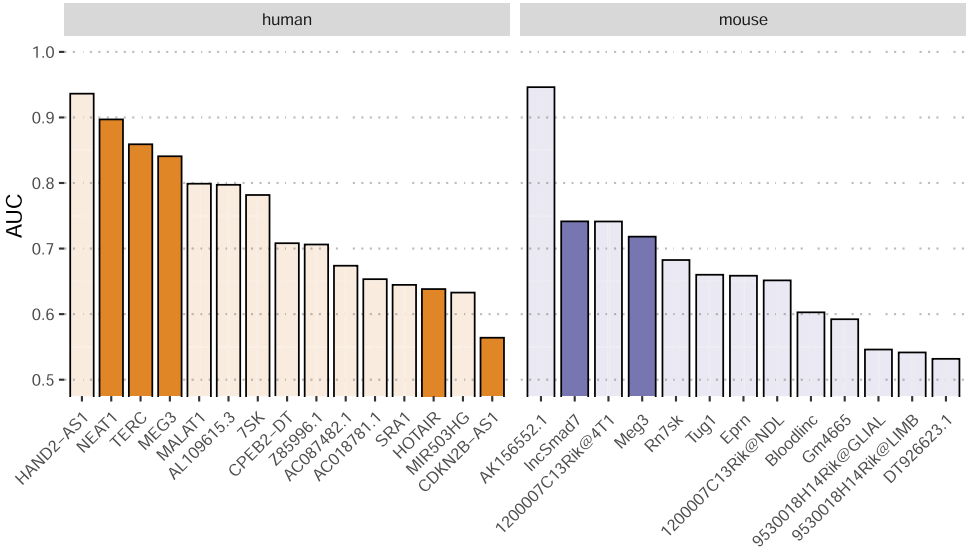

Fig. 5. AUC values of all the analysed lncRNAs. The AUC values are reported from the best parameter set for each lncRNA (see Supplementary Table 3). Darker bars highlight lncRNAs with experimental TPX evidence. If data from different tissues are available for a single lncRNA, the label reports the name of the lncRNA and the specific tissue separated by the symbol @.

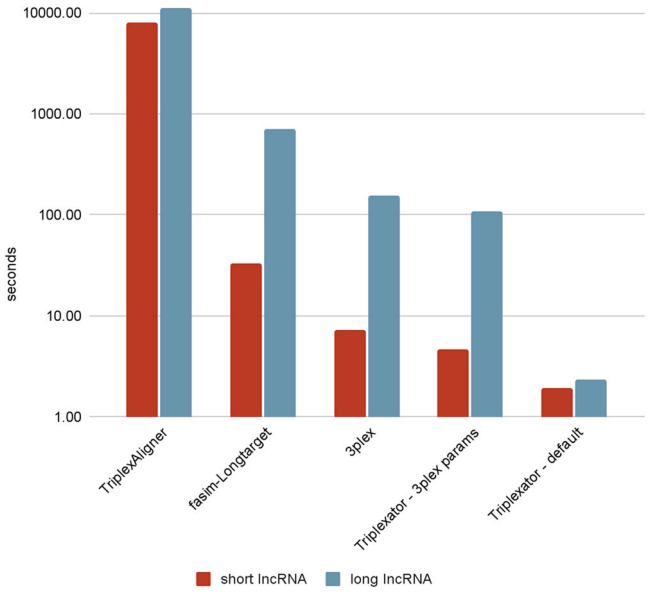

Fig. 6. Computational time required by different TPX prediction tools. The bars represent average computation time required by different software to predict TPX for two lncRNA: short (TERC transcript, 541 nucleotides) and long (NEAT1 transcript, 22743 nucleotides). To evaluate the computational cost of *3plex* specific features, the *3plex* values should be compared with *Triplexator* with *3plex* params as discussed in Methods.

errors). This observation could be surprising, since it may be argued that such models account for too poor information to be biologically relevant. Notably, TPXs typical length and degeneration do not differ much from the classical DNA binding sites of many important TFs. This is reflected by the comparison of ROC curves describing the prediction of RNA binding sites and those describing the prediction of TF binding sites (Fig. 4). Indeed, the information content of a typical TF binding site is paradoxically low when compared to the complex regulatory programs they orchestrate in eukaryotic gene regulation [52]. These observations suggest that lncRNAs (as well as TFs) may not work alone, reinforcing the common idea that eukaryotic gene regulation is a complex process involving cooperation and competition of many different molecules. Another interesting possibility is that multiple short TPXs close to each other in the same genomic region act cooperatively to enhance the specificity of the binding, a phenomenon actually observed in the X chromosome inactivation by *Xist* [53].

Moreover, Modal et al. showed, with in vitro biochemical experiments, that TXPs as small as 12 nucleotides can be functional [10], and in a previous work we experimentally validated functional specificity of some short TPXs formed by *lncSmad7* in vivo [6]. In 2020 Matveishina et al. considered the experimentally validated binding sites of 4 lncRNA to elaborate a practical guidance in genome-wide RNA:DNA TPX prediction. Consistent with our observations, they found that the lowest available cut-off in *Triplexator* minimal TPX length produced the best prediction [30]. We confirmed these findings on a wider dataset and for even smaller lengths.

It should be noted that in a recent in vitro study Kunkler et al. suggested that a longer minimal length should be used in TPX prediction [54]. Our results do not contradict these findings in vitro but suggest that despite low stability short TPX could be relevant in vivo by possibly relying on cooperative effects. Additionally, the high error rate tolerance we observed in TPX predictions may be explained by non-canonical Hoogsteen bonds the authors proved to be relevant in the same paper. Those non-canonical pairing are not yet considered by *3plex*, suggesting that there is room for improvement in the algorithm.

Recently, Matveishina et al. verified that RNA secondary structure can improve the TPX prediction specificity for some lncRNAs [30]. Starting from this observation, we implemented the first TPX prediction software that integrates the RNA secondary structure information. By comparing *3plex* predictions with gold standard experimental RNA:DNA binding data, we confirmed the relevance of the RNA folding on a wider set of lncRNAs. Importantly, we observed that masking the paired nucleotides in RNA secondary structure improved the TPX prediction only on a subset of the investigated lncRNAs. This suggests that TFOs can reside in RNA regions likely to be in a double strand conformation. Further studies would be required to identify possible secondary structure characteristics of these lncRNAs. Moreover, it is to be considered that the RNA secondary structure is highly dynamic. This implies that nucleotides predicted to be paired when considering the RNA sequence alone can be in an unpaired state in particular cellular contexts. It may also be reasonable that the unbound portion of a TFO initiates the TPX binding, thus triggering the conformational changes in the lncRNA necessary for a stable interaction with the DNA. Lastly, RNA secondary structure predictions present accuracy limitations that may be improved by integrating experimental information such as SHAPE data [55]. Unfortunately, experimentally determined structures are scarce on lncRNAs because of their typical low expression level.

Testing and comparing prediction software is a complex and delicate matter in general, as the choice of data and methodologies can greatly influence the outcome. To compare *3plex* with *TriplexAligner* and *fasim-LongTarget,* we devised two testing strategies. In the first one we designed negative controls by selecting random regions in the genome, in the second one we shuffled the sequences as done in the *TriplexAligner* paper. The best performing tool resulted in *3plex* for the first one but in *TriplexAligner* for the second one. We showed that the first strategy is more stringent as the performance of all thested software decreased, moreover in this case the AUC of *3plex* does not drop considerably. Because of these observations, we consider that *3plex* could be more robust than *TriplexAligner* even if the accuracy of the two is roughly comparable. We hence encourage the researchers to experiment with both of them.

In terms of computation time, *3plex* is slightly slower than *Triplexator* but greatly outperforms *TriplexAliner* and *fasim-Longtarget*, retaining a comparable or superior accuracy.

Finally, we investigated the TPX potential of 20 lncRNAs not previously reported to bind DNA using this mechanism but having publicly available DNA binding sites. We found a strong evidence of TPX formation for 3 lncRNAs (*AK156552.1, HAND2-AS1, MALAT1*), a good evidence for other 4 lncRNAs (*AL109615.3, 7SK, 1200007C13Rik, CPEB2-DT*), and a moderate evidence for other 8 (*Rn7sk, AC087482.1, Tug1, Eprn, AC018781.1, SRA1, MIR503HG, Bloodlinc*). These results suggest that DNA binding through TPX formation could be a widespread mechanism adopted by lncRNAs. This idea was previously reported in literature [56,57] using purely computational methods. Our RNA:DNA binding sites collection allowed the first large-scale evaluation of this hypothesis leveraging high-throughput experimental data.

Funding

V.P. was supported by Fondazione Umberto Veronesi (FUV). S.O. was supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC) IG 2022 ID 27155, PRIN 2018, and IIGM Institutional Funds.

PNRR CN3 - National Center for Gene Therapy and Drugs based on RNA Technology, Spoke 3.

CRediT authorship contribution statement

Chiara Cicconetti: Methodology, Software, Writing – original draft. **Andrea Lauria:** Software, Validation. **Valentina Proserpio:** Visualization, Investigation, Writing – review & editing. **Marco Masera:** Software. **Annalaura Tamburrini:** Software, Validation. **Mara Maldotti:** Visualization, Investigation. **Salvatore Oliviero:** Supervision, Funding acquisition. **Ivan Molineris:** Conceptualization, Methodology, Software, Project administration, Writing – review & editing.

Data Availability

3plex source code is available at https://github.com/molinerisLab/3plex.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank Danny Incarnato, Department of Molecular Genetics, University of Groningen and Paolo Provero, Dipartimento di Neuroscienze, University of Torino for the helpful discussion of the results.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.05.016.

References

[1] Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 2021;22(2):96–118.
[2] Mirzadeh Azad F, Polignano IL, Proserpio V, Oliviero S. Long noncoding RNAs in human stemness and differentiation. Trends Cell Biol 2021;31(7):542–55.
[3] Sun JS, Garestier T, Hélène C. Oligonucleotide directed triple helix formation. Curr Opin Struct Biol 1996;6(3):327–33.
[4] Duca M, Vekhoff P, Oussedik K, Halby L, Arimondo PB. The triple helix: 50 years later, the outcome. Nucleic Acids Res 2008;36(16):5123–38.
[5] Zapparoli E, Briata P, Rossi M, Brondolo L, Bucci G, Gherzi R. Comprehensive multi-omics analysis uncovers a group of TGF-β-regulated genes among lncRNA EPR direct transcriptional targets. Nucleic Acids Res 2020;48(16):9053–66.
[6] Maldotti M, Lauria A, Anselmi F, Molineris I, Tamburrini A, Meng G, et al. The acetyltransferase p300 is recruited in trans to multiple enhancer sites by lncSmad7. Nucleic Acids Res 2022;50(5):2587–602.
[7] Brown JA. Unraveling the structure and biological functions of RNA triple helices. Wiley Inter Rev RNA 2020;11(6):e1598.
[8] Li Y, Syed J, Sugiyama H. RNA-DNA triplex formation by long noncoding RNAs. Cell Chem Biol 2016;23(11):1325–33.
[9] Chu C, Quinn J, Chang HY. Chromatin Isolation by RNA Purification (ChIRP). J Vis Exp 2012;61:3912.
[10] Mondal T, Subhash S, Vaid R, Enroth S, Uday S, Reinius B, et al. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. Nat Commun 2015:6.
[11] Simon M.D. Capture hybridization analysis of RNA Targets (CHART). Curr Protoc Mol Biol. 2013;Chapter 21:Unit 21.25.
[12] Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 2013;341(6147):1237973.
[13] Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, et al. Systematic mapping of RNA-chromatin interactions in vivo. Curr Biol 2017;27(4):602–9.
[14] Li X, Zhou B, Chen L, Gou LT, Li H, Fu XD. GRID-seq reveals the global RNA-chromatin interactome. Nat Biotechnol 2017;35(10):940–50.
[15] Bonetti A, Agostini F, Suzuki AM, Hashimoto K, Pascarella G, Gimenez J, et al. RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. Nat Commun 2020;11(1):1018.
[16] Gavrilov AA, Zharikova AA, Galitsyna AA, Luzhin AV, Rubanova NM, Golov AK, et al. Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. Nucleic Acids Res 2020;48(12):6699–714.
[17] Gavrilov AA, Sultanov RI, Magnitov MD, Galitsyna AA, Dashinimaev EB, Lieberman Aiden E, et al. RedChIP identifies noncoding RNAs associated with genomic sites occupied by Polycomb and CTCF proteins. Proc Natl Acad Sci USA 2022;119(1):e2116222119.
[18] Bell JC, Jukam D, Teran NA, Risca VI, Smith OK, Johnson WL, et al. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. eLife 2018;7:e27024.
[19] Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 2012 30;45(6):814–25.
[20] Cetin NS, Kuo CC, Ribarska T, Li R, Costa IG, Grummt I. Isolation and genome-wide characterization of cellular DNA:RNA triplex structures. Nucleic Acids Res 2019;47(5):2306–21.
[21] Buske FA, Bauer DC, Mattick JS, Bailey TL. Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome Res 2012;22(7):1372–81.
[22] Kuo CC, Hänzelmann S, Sentürk Cetin N, Frank S, Zajzon B, Derks JP, et al. Detection of RNA-DNA binding sites in long noncoding RNAs. Nucleic Acids Res 2019;47(6):e32.
[23] He S, Zhang H, Liu H, Zhu H. LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. Bioinformatics 2015;31(2):178–86.
[24] Wen Y, Wu Y, Xu B, Lin J, Zhu H. Fasim-LongTarget enables fast and accurate genome-wide lncRNA/DNA binding prediction. Comput Struct Biotechnol J 2022;20:3347–50.
[25] Zhang Y, Long Y, Kwoh CK. Deep learning based DNA:RNA triplex forming potential prediction. BMC Bioinforma 2020;21(1):1–13.
[26] Warwick T, Seredinski S, Krause NM, Bains JK, Althaus L, Oo JA, et al. A universal model of RNA.DNA:DNA triplex formation accurately predicts genome-wide RNA-DNA interactions. Brief Bioinf 2022;23(6). bbac445.
[27] Warwick T, Brandes RP, Leisegang MS. Computational methods to study DNA:DNA:RNA triplex formation by lncRNAs. Non-Coding RNA 2023;9(1):10.
[28] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 2008;451(7178):535–40.
[29] Grassi E, Zapparoli E, Molineris I, Provero P. Total Binding Affinity profiles of regulatory regions predict transcription factor binding and gene expression in human cells. PLoS One 2015:1–13.
[30] Matveishina E, Antonov I, Medvedeva YA. Practical guidance in genome-wide RNA:DNA triple helix prediction. Int J Mol Sci 2020;21(3):830.
[31] Bugnon LA, Edera AA, Prochetto S, Gerard M, Raad J, Fenoy E, et al. Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches. Brief Bioinf 2022;23(4). bbac205.
[32] Takahashi S, Sugimoto N. Watson–crick versus hoogsteen base pairs: chemical strategy to encode and express genetic information in life. Acc Chem Res 2021;54(9):2110–20.
[33] Sherpa C, Rausch JW, Le, Grice SF. Structural characterization of maternally expressed gene 3 RNA reveals conserved motifs and potential sites of interaction with polycomb repressive complex 2. Nucleic Acids Res 2018;46(19):10432–47.
[34] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10.
[35] Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, Gupta D, et al. The european nucleotide archive in 2022. Nucleic Acids Res 2022. gkac1051.
[36] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019 8;47(D1):D766–73.
[37] Bezzecchi E, Pagani G, Forte B, Percio S, Zaffaroni N, Dolfini D, et al. MIR205HG/LEADR long noncoding RNA binds to primed proximal regulatory regions in prostate basal cells through a triplex- and alu-mediated mechanism. Front Cell Dev Biol 2022;10:909097.
[38] Ducoli L, Agrawal S, Sibler E, Kouno T, Tacconi C, Hon CC, et al. LETR1 is a lymphatic endothelial-specific lncRNA governing cell proliferation and migration through KLF4 and SEMA3C. Nat Commun 2021;12(1):925.
[39] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.
[40] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 2008;26(12):1351–9.
[41] Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat 2011;5(3):1752–79.
[42] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl 2010;26(6):841–2.
[43] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol AMB 2011;6:26.
[44] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinforma 2011;12:77.
[45] Köster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinforma Oxf Engl 2012;28(19):2520–2.
[46] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009;37:W202–8. (Web Server issue).
[47] Jayaram N, Usvyat D, Martin AC. Evaluating tools for transcription factor binding site prediction. BMC Bioinforma 2016;17(1):1–12.

[48] Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2022;50(D1):D165–73.
[49] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinforma Oxf Engl 2011;27(7):1017–8.
[50] Merry CR, Forrest ME, Sabers JN, Beard L, Gao XH, Hatzoglou M, et al. DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer. Hum Mol Genet 2015;24(21):6240–53.
[51] Long J, Badal SS, Ye Z, Wang Y, Ayanga BA, Galvan DL, et al. Long noncoding RNA Tug1 regulates mitochondrial bioenergetics in diabetic nephropathy. J Clin Invest 2016;126(11):4205–18.
[52] Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annu Rev Cell Dev Biol 2019;35:357–79.
[53] Matsuno Y, Yamashita T, Wagatsuma M, Yamakage H. Convergence in LINE-1 nucleotide variations can benefit redundantly forming triplexes with lncRNA in mammalian X-chromosome inactivation. Mob DNA 2019;10:33.
[54] Kunkler CN, Hulewicz JP, Hickman SC, Wang MC, McCown PJ, Brown JA. Stability of an RNA•DNA-DNA triple helix depends on base triplet composition and length of the RNA third strand. Nucleic Acids Res 2019;47(14):7213–22.
[55] Li P, Zhou X, Xu K, Zhang QC. RASP: an atlas of transcriptome-wide RNA secondary structure probing data. Nucleic Acids Res 2021;49(D1):D183–91.
[56] Jalali S, Singh A, Maiti S, Scaria V. Genome-wide computational analysis of potential long noncoding RNA mediated DNA: DNA: RNA triplexes in the human genome. J Transl Med 2017;15(1):1–17.
[57] Soibam B. Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation. RNA 2017;23(11):1729–42.