OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–12

https://doi.org/10.1093/bib/bbac445

Advance access publication date 13 October 2022

Problem Solving Protocol

A universal model of RNA·DNA:DNA triplex formation accurately predicts genome-wide RNA–DNA interactions

Timothy Warwick, Sandra Seredinski, Nina M. Krause, Jasleen Kaur Bains, Lara Althaus, James A. Oo, Alessandro Bonetti, Anne Dueck, Stefan Engelhardt, Harald Schwalbe, Matthias S. Leisegang, Marcel H. Schulz and Ralf P. Brandes

Corresponding authors. Ralf P. Brandes, Institute for Cardiovascular Physiology, Goethe University, Theodor-Stern-Kai 7, D-60590, Frankfurt am Main, Germany. E-mail: brandes@vrc.uni-frankfurt.de; Marcel H. Schulz, Institute for Cardiovascular Physiology, Goethe University, Theodor-Stern-Kai 7, D-60590, Frankfurt am Main, Germany. E-mail: marcel.schulz@em.uni-frankfurt.de

Abstract

RNA·DNA:DNA triple helix (triplex) formation is a form of RNA–DNA interaction which regulates gene expression but is difficult to study experimentally *in vivo*. This makes accurate computational prediction of such interactions highly important in the field of RNA research. Current predictive methods use canonical Hoogsteen base pairing rules, which whilst biophysically valid, may not reflect the plastic nature of cell biology. Here, we present the first optimization approach to learn a probabilistic model describing RNA–DNA interactions directly from motifs derived from triplex sequencing data. We find that there are several stable interaction codes, including Hoogsteen base pairing and novel RNA–DNA base pairings, which agree with *in vitro* measurements. We implemented these findings in *TriplexAligner*, a program that uses the determined interaction codes to predict triplex binding. *TriplexAligner* predicts RNA–DNA interactions identified in all-to-all sequencing data more accurately than all previously published tools in human and mouse and also predicts previously studied triplex interactions with known regulatory functions. We further validated a novel triplex interaction using biophysical experiments. Our work is an important step towards better understanding of triplex formation and allows genome-wide analyses of RNA–DNA interactions.

Keywords: RNA, DNA, Triplex, machine learning, RNA–DNA interaction.

Introduction

Numerous regulatory roles have been ascribed to RNAs [1, 2], which include interactions with both DNA and proteins. The RNA–protein interface includes functions such as transcription factor addressing and recruitment [3], scaffolding of transcription factor machinery [4] and mediation of histone modifications [5]. Various epigenomic consequences have been attributed to RNA–DNA interactions, including the functional role of the *XIST* transcript in the silencing of the X chromosome during dosage compensation [6]. Other examples of RNA–DNA interactions include

Timothy Warwick is a PhD student at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His research pertains to regulatory networks underlying gene expression patterns.
Sandra Seredinski is a PhD student at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. Her research focuses on lncRNAs and their interactions with chromatin.
Nina M. Krause is a PhD student at the Institute for Organic Chemistry and Chemical Biology at Goethe University, Frankfurt. Her research focuses on the biophysics of RNA–DNA interaction.
Jasleen Kaur Bains is a PhD student at the Institute for Organic Chemistry and Chemical Biology at Goethe University, Frankfurt. Her research focuses on the biophysics of RNA conformation.
Lara Althaus is a PhD student at the Institute of Pharmacology and Toxicology, Technical University of Munich. Her research focuses on control of cellular behaviour by lncRNAs.
James A. Oo is a post-doctoral researcher at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His research focuses on lncRNA-mediated control of transcription factors.
Alessandro Bonetti is a post-doctoral researcher at AstraZeneca. His research focuses on reporting of RNA–DNA interactions via next-generation sequencing methods.
Anne Dueck is a group leader at the Institute of Pharmacology and Toxicology, Technical University of Munich. Her research focuses on how lncRNAs define divergent cellular functions.
Stefan Engelhardt is a professor at the Institute of Pharmacology and Toxicology, Technical University of Munich. His research interests cover cardiovascular signal transduction, amongst many other areas.
Harald Schwalbe is a professor at the Institute for Organic Chemistry and Chemical Biology at Goethe University, Frankfurt. His group works on structure and dynamics of biological molecules.
Matthias S. Leisegang is a group leader at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His group works on lncRNAs of relevance to the cardiovascular system.
Marcel H. Schulz is a professor at the Institute of Cardiovascular Regeneration at Goethe University, Frankfurt. His group works on computational epigenomics and systems cardiology.
Ralf P. Brandes is a professor at the Institute for Cardiovascular Physiology at Goethe University, Frankfurt. His group works on the physiology and epigenetics of the cardiovascular system.

Received: June 20, 2022. **Revised:** August 16, 2022. **Accepted:** September 17, 2022

RNA–DNA hybrid G-quadruplex [7] and R-loop [8] formation. R-loops consist of interactions between single-stranded DNA and RNA via Watson-Crick base pairing and have been implicated in chromatin condensation and tumorigenesis [9, 10]. Alongside these, there exists another form of RNA–DNA interaction where DNA structure is maintained and single-stranded RNA binds in the major groove of the double helix, resulting in the formation of an RNA·DNA:DNA triple helix (triplex) [11].

Triplex formation represents an area of epigenetics which, although known of in a biophysical sense for many years [12], remains incompletely understood. There are several reasons for this, chief amongst them being the experimental complexities of studying triplex formation on a genome-wide scale in living cells. Experimental probing of triplex formation in the cellular context has previously relied upon methods capturing the genomic interaction sites of single transcripts [13, 14]. However, regulatory transcripts have been implicated in epigenetic mechanisms across many species, tissues and cell types [15–18]. Herein lies the importance of developing tools which accurately predict points of RNA–DNA interaction as putative sites of triplex formation.

Previously published computational tools have relied upon Hoogsteen base pairing rules [19], which are canonically responsible for triplex formation. Tools implementing Hoogsteen rules include *Triplexator/Triplex Domain Finder* [20, 21] and *LongTarget* [22]. Whilst usage of canonical rules to predict triplex formation provides insight into putative RNA–DNA interactions, benchmarking of *Triplexator* and *LongTarget* using *MEG3* ChOP-seq data [13] revealed substantial room for improvement in this area [23]. These benchmarking data suggest that accurate prediction of triplex formation in a cellular context may require the implementation of as yet unknown base pairing rules which go beyond the currently used Hoogsteen base pairing rules. Related to this, there has been work on prediction of triplex-forming RNA and DNA sequences using previously published triplex interactions [24], although prediction of complete triplex interactions is not possible with this method.

Accompanying computational methods for genome-wide prediction of RNA–DNA interaction have been experimental methods with similar aims. Foremost amongst these, with regard to triplex formation, is genome-wide isolation of triplexes followed by sequencing (triplex-seq) [25]. This method permits the identification of triplex-forming sequences across the genome (triplexDNA-seq) and transcriptome (triplexRNA-seq) but lacks information on the pairing of the sequences with one another. Outside specifically triplex-mediated RNA–DNA interactions, there have been a number of published methods designed to identify all-to-all interactions between transcripts and chromatin [26–29]. However, the methods with most similar nucleotide processing protocols to triplex-seq are RNA And DNA Interacting Complexes Ligated and sequenced (RADICL-seq) [30] and RedC [31]. These methods collectively identify specific interactions between transcripts and regions of the genome through ligation of RNA and DNA via a linker sequence in a proximity-based manner. RADICL-seq and RedC provide rich sources of data on RNA–DNA interaction but also remain relatively novel and experimentally complex. The undertaking of such experiments across a range of steady-state and differential conditions is therefore not feasible at this juncture. Consequently, the most widely applicable use for these data may be as input to machine learning algorithms, which could permit the prediction of RNA–DNA interactions in a condition of interest.

Here, we present a method for the prediction of RNA–DNA interactions based on RNA–DNA binding probabilities learned by expectation-maximization from triplex-forming sequences identified in triplexDNA-seq and triplexRNA-seq. Applying these binding rule sets as substitution matrices in local alignment permitted more accurate recall of RNA–DNA interactions identified from RADICL-seq and RedC when compared with previously published tools. Experimentally validated triplex formation between transcripts and genomic loci could also be recapitulated. A predicted interaction was also subjected to biophysical validation, where triplex formation could be experimentally verified *ex vivo*.

Materials and methods

Triplex sequencing data processing

Publicly accessible triplexDNA-seq data and triplexRNA-seq data [25] (NCBI Short Read Archive accessions SRR7965691, SRR7965692, SRR7965693, SRR7965694, SRR7965701, SRR7965702, SRR7965703) were downloaded and aligned against the *hg38* genome and transcriptome, respectively, using *Bowtie2* (v2.4.4) with default parameters [32]. Peaks were called from alignments using *HOMER findPeaks* (v4.11.1) with default parameters [33]. Sequences underlying identified peaks were then extracted from the *hg38* genome or transcriptome using *bedtools getfasta* (v2.27.1) with the peak coordinates per sample as input [34].

Motif enrichment and processing

Peak sequences were used as input to motif enrichment using *MEME-ChIP* (v5.0.5) [35]. Shuffled peak sequences were supplied as negative sequence input, with the maximum number of enriched *MEME* motifs restricted to 16, with a maximum motif length of 32 nucleotides. Significantly enriched motifs were considered to be those with an $E < 0.05$. Enriched motifs were subjected to motif comparison between samples using *Tomtom* (also part of *MEME Suite*), and matches were considered to be present when $P < 0.05$. *FIMO*, another tool contained within *MEME Suite*, was used to compute the occurrences of enriched triplex motifs across the breadths of triplex DNA and RNA peaks. Triplex motif occurrence was also computed separately for peaks lying in distinct genomic features as defined by the *TxDb.Hsapiens.UCSC.hg38.knownGene* annotation package for *R*, maintained by *Bioconductor* [36–38]. Triplex RNA motif occurrence was computed per transcript biotype, as defined by *EnsDb.Hsapiens.v86* and normalized to transcript length [39, 40]. Enriched triplex DNA and RNA motifs were compared against the *JASPAR 2020* [41] and *ATtRACT* [42] motif databases, respectively, in order to remove any motifs which were identical to transcription factor-binding or RNA-binding protein motifs.

An Expectation-Maximization-based method for RNA–DNA code optimization

Nomenclature

Assume, we are given a DNA motif D of length l, which is a matrix $D^{4\times l}$, over $\Sigma = \{A, C, G, T\}$ denoting the set of characters in the DNA alphabet. Also, we have RNA motif matrix $R^{4\times l}$, for simplicity we assume that the RNA alphabet has been translated to the DNA alphabet by exchanging U→T. For simplicity, all motifs considered here have the same length l. **Note** that a difference in length between an RNA and DNA matrix can easily be accounted for by testing all possible shifts of the smaller matrix against the larger matrix.

We assume that there is a set of DNA motifs $\mathcal{D} = \{D_1, \ldots, D_n\}$ and equivalently a set of RNA motifs $\mathcal{R} = \{R_1, \ldots, R_n\}$. For notational simplicity, we assume that they have the same number of elements n, although in practice this may change.

We assume that there exists a mapping code C^{4x4}, which is a matrix that maps nucleotides from RNA to DNA nucleotides. For example, $C_{A,A}$ denotes the probability to map an *A* RNA nucleotide to an *A* DNA nucleotide. Entries in the row of the matrix sum to one, and thus, it holds that

$$\forall r \in \Sigma, \sum_{d \in \Sigma} C_{r,d} = 1\,. \tag{1}$$

The interest in this formulation is to learn the code *C* that is behind a given set of motifs $\mathcal{R}$ and $\mathcal{D}$.

Code objective value

We define the average column-wise mapping error between a DNA motif *D* and an RNA motif *R*–termed code objective value–given a defined code matrix *C* as

$$objective(R, D, C) = \frac{\sum_{i \in \Sigma, j \in 1,\ldots,l} abs(D_{i,j} - \hat{D}_{i,j})}{l}\,, \tag{2}$$

where $\hat{D}$ is the projected DNA motif after conversion of *R* using *C*

$$\hat{D}_{i,j}, = \sum_{a \in \Sigma} R_{a,j} \cdot C_{a,i}\,, \tag{3}$$

where $i \in \Sigma,\ j \in 1, \ldots, l$.

Obtaining the best code using quadratic programming

Given the two sets of input motifs from DNA $\mathcal{D}$ and RNA $\mathcal{R}$ motifs, we are looking for an optimal code *C* that describes the conversion of an RNA motif to a DNA motif, as would be done when a subsequence of an RNA is aligned to a subsequence in a DNA sequence, a triplex match.

Assume that we had a known pairing *P* of the RNA to DNA motifs, then we would be looking for the code matrix *C* that minimizes the code objective value (Eq. (2)) for the pairing of RNA and DNA matrices

$$\underset{C}{\mathbf{argmin}} = objective(\mathcal{D}, \mathcal{R}, \mathcal{P}, C), C \in \mathcal{C}\,, \tag{4}$$

where $objective(\mathcal{D}, \mathcal{R}, \mathcal{P}, \mathcal{C})$ denotes the sum of code objective values for all defined pairs using Eq. (2) and $\mathcal{C}$ denotes the space of all possible code matrices. The term *error* in Eq. (4) may be interpreted as the error of a learned code given the pairings of motifs it describes. Luckily, we can obtain the code matrix *C* that minimizes the code objective value using quadratic programming efficiently.

An Expectation-Maximization algorithm for finding optimal code sets

While it is straightforward to obtain a code matrix *C* that minimizes the code objective value for a given pairing of RNA and DNA motifs, in practice, the true pairing is not known. Furthermore, it is unknown whether the triplex binding of all RNA–DNA pairs follows the same code and the possibility of several code matrices needs to be considered.

Therefore, we have designed an Expectation-Maximization algorithm for finding a set of code matrices starting from a given set of RNA and DNA motifs for which the correct pairing is unknown.

Conceptually, the algorithm performs the three following steps to find *k* many code matrices

Input: $\mathcal{D}, \mathcal{R}, k$
generate *k* random code matrices $C_1^*, \ldots, C_k^*$
1. $C_1 = C_1^*, \ldots, C_k = C_k^*$
2. obtain the best pairings for elements in $\mathcal{D}$ and $\mathcal{R}$ using one of $C_1, \ldots, C_k$
3. for each: i=1, . . . ,k
C_i^* = minimize code of all paired DNA and RNA motifs that used C_i
repeat at 1. if $(C_1 \neq C_1^*, \ldots, C_k \neq C_k^*)$
Output: final code matrices $C_1^*, \ldots, C_k^*$**, pairing** ***P***

The second step listed above–where the best motif pairings are obtained–refers to obtaining the best pair for each RNA motif in $\mathcal{R}$ with a DNA motif in $\mathcal{D}$ testing all *k* code matrices. The best pair are the indices i, j, h where $error(\mathcal{R}_i, \mathcal{D}_j, \mathcal{C}_h)$ is minimal. In this process, it is allowed that several elements from $\mathcal{R}$ are paired with the same element in $\mathcal{D}$ and *vice versa*.

In summary, the above EM procedure determines the best pairing between RNA and DNA motifs and determines *k* code matrices as a result of the process. Results are output upon convergence of the algorithm, when all code objective values have been minimized and the code matrices are unchanged. Applications using both simulated and real motif pairs in the course of this work have shown that the algorithm converges in a small number of iterations in practice. However, the solution only constitutes a local minimum; therefore, we run the algorithm many times with the same input data, e.g. 11 270 times with the real triplex input motifs.

The code for the steps described above is publicly available at https://github.com/SchulzLab/Codefinder.

Code processing and annotation

Learned code models, which were output from the Expectation-Maximization algorithm, were stratified by their objective values and the total number of motifs incorporated into the model. These metrics were also used to subset the models and identify the most promising candidates for further study (*objective* < 0.75, *total motifs* > 50%). *In vitro* RNA·DNA:DNA triple helix base triplet stability data [43] were used to compare the affinities of learned code models versus a size-matched set of random code models. In short, the normalized dissociation constant of the RNA:DNA interaction as reported in [43] was multiplied by the probabilities of nucleotide interaction contained within the code matrix, and then summed. The relative affinity values for each code model with *total motifs* > 10% were also linearly regressed against the objective values returned from Expectation-Maximization. Following this, high-scoring code models were subjected to hierarchical clustering with Euclidean distances and Ward's method in order to control for redundancy [44]. The resulting dendrogram was then cut to produce eight clusters, and the mean code model for each cluster was computed.

Formulation of TriplexAligner

To be implemented in *TriplexAligner*, probabilistic code model values were converted to log odds scores according to [45]. Subsequent score distributions were computed for each code model with *Biostrings::pairwiseAlignment* [46], using simulated DNA and RNA sequences with matching nucleotide proportions relative to human promoter sequences and known triplex-forming transcript sequences, respectively. Arising scores were fitted using a generalized extreme value distribution [47] with *EnvStats::egevd* [48], using maximum likelihood estimation. The parameter values *K* and λ could then be identified for each code model. Using

these parameters, bit scores (S') and corresponding E values could be calculated from local alignment scores (S) of respective code models according to the following formulas:

$$S' = \frac{\lambda S - ln(K)}{ln2}, \quad (5)$$

$$E = mn\,2^{-S'}. \quad (6)$$

TriplexAligner computes local alignment scores, bit scores and E values between supplied DNA and RNA sequences for each code model using *Biostrings::pairwiseAlignment* and the above formulae, with the log odds code model supplied as the *substitutionMatrix* parameter.

The code for *TriplexAligner* is publicly available at https://github.com/SchulzLab/TriplexAligner, where it is formalized as an *R* package which may be downloaded, installed and used by interested parties.

Computational validation of TriplexAligner using global RNA–DNA interactions

RNA–DNA interactions arising from either RedC (GSE136141) or RADICL-seq (GSE132192) were used to benchmark the performance of *TriplexAligner* compared with the previously published tools *LongTarget* and *Triplexator*. For RedC data, interactions between RNAs and 5 kb genomic bins were used for validation if they were present in two separate replicates. For the RADICL-seq interactions, significant interactions between RNAs and 5 kb genomic bins were identified according to [30]. Interactions between RNAs and genomic bins were expanded to include all possible transcripts of the involved RNA gene, as annotated in *TxDb.Hsapiens.UCSC.hg38.knownGene* [49] or *TxDb.Mmusculus.UCSC.mm10.knownGene* [50] annotation packages for *R* [38]. RADICL-seq interactions were limited to those involving transcripts expressed in accompanying nuclear RNA-seq data, quantified using *Salmon* (v 1.6.0) [51]. For each interaction, involved RNA and DNA sequences were subjected to RNA·DNA:DNA triple helix prediction using *TriplexAligner*, *LongTarget* and *Triplexator*. A corresponding negative dataset was constructed via shuffling of the transcript sequences. *LongTarget* was run with default parameters, and *Triplexator* with *-e 20 -l 5*. Maximum metrics (*TriplexAligner*: $-\log_{10}(E)$; *LongTarget*: MeanStability; *Triplexator*: Score) were identified per gene-bin interaction and used as predictive values in subsequent analyses with *pROC* and *ROCR* [52, 53]. Receiver operating characteristic curves were computed for each method with binomial smoothing and statistically compared by bootstrapping ($n = 2000$).

Electrophoretic mobility shift assay

All hybridization steps were performed in 25 mM HEPES, pH 7.4, 50 mM NaCl and 10 mM MgCl2. DNA oligos, spanning the predicted triplex DNA sequence (20 pmol), were hybridized to DNA duplex in a thermocycler by heating up for 5 min to 95°C followed by a cool down to 24°C with a rate of 1°C/ 30 s. For triplex formation, 10 eq of ssRNA (200 pmol), containing the predicted triplex RNA sequence, was added to the DNA duplex followed by incubation at 60°C for 1 h and a cool down to 24°C with a rate of 1°C/ 30 s. RNase H digestion was performed by adding RNase H to a final concentration of 375 mU/μL to a triplex sample and incubate it for 30 min at 37°C. RNase A was added to a triplex sample with a final concentration of 5 ng/μL and incubated similarly to RNase H. Samples were applied on a native 15% Polyacrylamide gel in a running buffer containing 40 mM Tris-Ac pH 8.3 supplemented with 3 mM magnesium acetate. The gels ran for 6 h at room temperature with 160 V.

CD spectroscopy and melting curve analysis

Circular dichroism spectra were acquired on a Jasco J-810 spectropolarimeter. The measurements were recorded from 210 to 320 nm at 25°C using 1 cm path length quartz cuvette. CD spectra were recorded on 8 μM samples of each DNA duplex, DNA:RNA heteroduplex and DNA:DNA:RNA-triplex (10 equivalents of RNA (80 μM)) in 25 mM HEPES, 50 mM NaCl and 10 mM $MgCl_2$ (pH 7.4). Spectra were acquired with 8 scans and the data were smoothed with Savitzky–Golay filters. Observed ellipticities recorded in millidegree (mdeg) were converted to molar ellipticity $[\Theta] = deg \times cm^2 \times dmol^{-1}$. Melting curves were acquired at constant wavelength using a temperature rate of 1°C/min in a range from 5 to 95°C. All data were evaluated using SigmaPlot 12.5. All melting temperature data were converted to normalised ellipticity and evaluated by the following equation: $f = a/(1 + exp(-(x - x0)/b)) + c/(1 + exp(-(x - x2)/d))$.

Results

Prediction of RNA·DNA:DNA triplex interactions from captured triplex sequences

In order to predict interactions between DNA and RNA mediated by triplex formation (Figure 1A), we developed *TriplexAligner*. *TriplexAligner* is capable of predicting RNA–DNA interactions with high accuracy, surpassing currently available methods. Development of *TriplexAligner* (Figure 1B, Supplementary Figure 1) encompassed multiple stages and included multiple next-generation sequencing datasets, alongside machine learning and biophysical methods. Initially, key sequence elements of triplex-forming RNA and DNA sequences needed to be identified. For this purpose, triplexRNA-seq and triplexDNA-seq data from HeLa cells [25] were analysed and triplex-forming regions were identified by peak calling. RNA and DNA components of the published triplex interaction between the *MEG3* transcript and the gene locus of *COL15A1* [13] could be clearly identified in the dataset (Figure 1C). This satisfied the requirement that the input data sufficiently capture triplex formation taking place in the cellular context.

Identification of short sequences underpinning triplex formation

The next step in development of *TriplexAligner* was the identification of triplex-enriched DNA and RNA regions, along with associated sequences. Peak calling on triplexDNA-seq data identified regions most often in intronic areas of the genome (Figure 2A), although promoter regions were most enriched for peak occurrence relative to the proportion of the genome covered **(Supplementary Figure 2A)**. When calling peaks from triplexRNA-seq data, most peaks were detected in protein-coding transcripts (Figure 2B). Transcripts with retained introns, followed by antisense transcripts, were most enriched for triplexRNA-peaks relative to transcript length **(Supplementary Figure 2B)**.

To identify sequences underpinning triplex-seq peaks, motif enrichment analysis was performed using *MEME-ChIP* [35] on sequences underlying peaks in each sample. Between 22 and 36 significantly enriched motifs were identified per triplexDNA-seq sample (Figure 2C, left). More motifs were identified in the RNA samples, which each returned more than 125 enriched motifs (Figure 2C, right). To investigate the reproducibility of the enriched

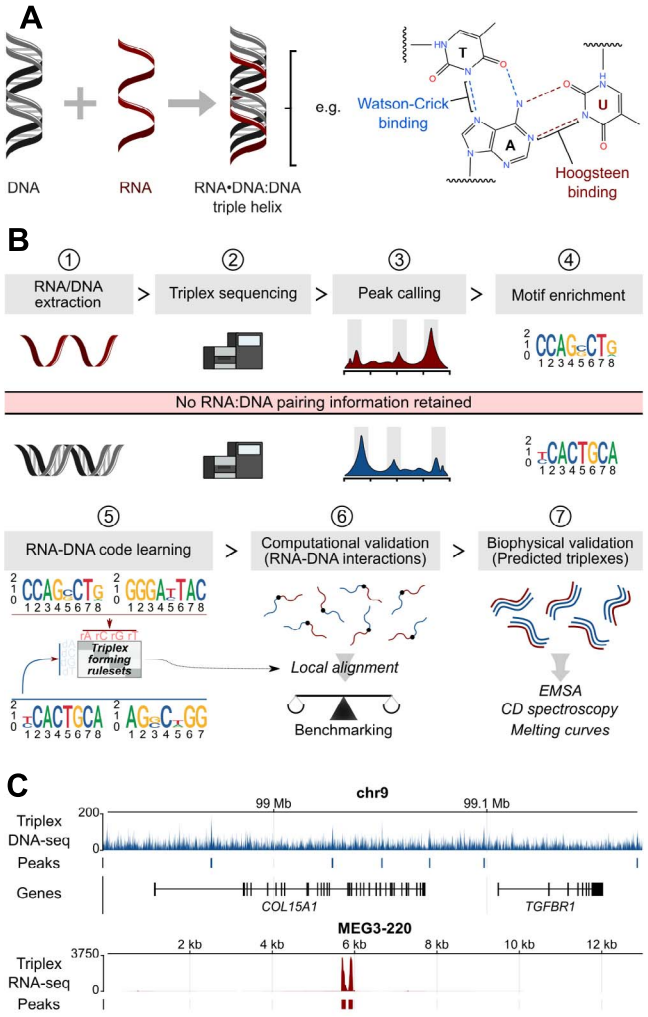

Figure 1. Overview of RNA·DNA:DNA triple helix formation and the development of TriplexAligner from triplex-seq data. **(A)** Schematic of RNA·DNA:DNA triple helix formation and effects on gene expression. **(B)** Overview of the development of TriplexAligner. **(C)** Peak calling on triplexDNA-seq (blue) and triplexRNA-seq data (red). The displayed regions reflect the published RNA·DNA:DNA triple helix interaction between *MEG3* and a DNA site in the locus of *COL15A1*, which results in the regulation of the downstream gene *TGFBR1*.

sequences, motifs were compared between samples using *Tomtom* [35]. Both RNA and DNA samples displayed high degrees of reproducibility, with a minimum of 76% of DNA motifs per sample having similar motifs in another DNA sample (Figure 2D, top). RNA motifs were also reproducible, with the minimum percentage of similar motifs between any two samples being 66% (Figure 2D, bottom). Individual RNA motifs also exhibited more significant enrichment than DNA motifs, observable when examining the five most enriched motifs of each molecule (Figure 2E). The most enriched RNA motif had an *E* value of 8.6×10^{-1382}, in comparison to an *E* value of 4.2×10^{-95} for the most enriched DNA motif.

In order to establish whether enriched triplex motifs reflect putative regulatory functions of triplex formation, motif occurrence in different features and transcripts was compared. Enriched triplex DNA motifs occurred at higher rates in triplex DNA peaks residing in promoters, intergenic regions and introns relative to exonic regions (Figure 2F). Triplex RNA motifs appeared more often in transcripts lacking open-reading frames–specifically lincRNAs, antisense transcripts and transcripts with retained introns–relative to protein coding transcripts (Figure 2G). Taken together, these findings indicate that triplex formation between non-coding RNA and non-coding regions of the genome is best described by reproducible sequence elements. In a positional sense, triplex motifs did not show any specific pattern of occurrence within peak regions **(Supplementary Figure 2C)**.

To select triplex motifs to take forward to the next stage of development, the enriched RNA and DNA motifs were subjected to several stratification steps. Identical motifs between samples were removed from the analysis, and motifs previously implicated in either transcription factor or RNA-binding protein interactions were excluded in an attempt to isolate sequences of most importance for triplex formation. Complementary DNA motifs were also added (Figure 2H), owing to the non-stranded nature of the analysis. The outcome of these steps were two sets of triplex motifs, consisting of 192 DNA motifs and 324 RNA motifs, which were used in the development of *TriplexAligner*.

Expectation-maximization to learn triplex formation rules

To learn the putative nucleotide pairing rules which might govern triplex formation, an Expectation-Maximization (EM) algorithm was used to compute triplex nucleotide pairing probabilities from triplex motifs (Figure 3A). The expectation portion of the algorithm was formed by pairings of RNA and DNA triplex motifs. From these pairs, probabilistic models were computed by quadratic programming, averaged across the pairings and evaluated for their error per motif pair (here termed the code objective value). When the objective value was minimized, motif pairings and probabilistic triplex mapping codes were returned. Initially, simulated motif sets were used to test the algorithm. The algorithm was capable of accurately learning nucleotide pairing probabilities, with an example shown in Figure 3B of the learning of Watson–Crick base pairing rules from 100 simulated motif pairs.

The algorithm was subsequently implemented with enriched triplex motifs as input, across 11 270 separate random initiations (Figure 3C). Results were then probed for several metrics, including the final objective values and proportion of total motifs included in the final motif pairings. Results with objective values less than 0.75 and containing more than 50% of total triplex motifs represented the most promising results and were subset, resulting in 801 putative codes. These subset conditions reflect the intention to find a balance between incorporating as much of the input data as possible, along with discarding low-scoring codes which may reflect aberrant starting points of the algorithm (**Supplementary Figure 3**). High-scoring codes resulting from these steps were annotated with published *in vitro* triplex nucleotide dissociation equilibrium constants [43]. The relative binding strengths were calculated for each code and its reverse complement, with the maximum then being taken as the value for that code. The high-scoring mapping codes presented significantly ($W = 415474$, $P < 2.2 \times 10^{-16}$, Mann–Whitney U test) greater relative binding strengths than an identically sized set of randomly generated codes (Figure 3D). When regressing code objective values from all results containing more than 10% of all motifs versus relative *in vitro* binding strengths, a negative correlation could be seen ($R^2 = 0.394$, $P < 2.2 \times 10^{-16}$, Figure 3E). This correlation suggested that the code objective values reflect experimental data, making them appropriate to stratify the codes by. Given that the algorithm used here minimizes the code objective value, a negative correlation between this value and *in vitro* stability directly supports the approach. Collectively, these data suggest that the probabilistic

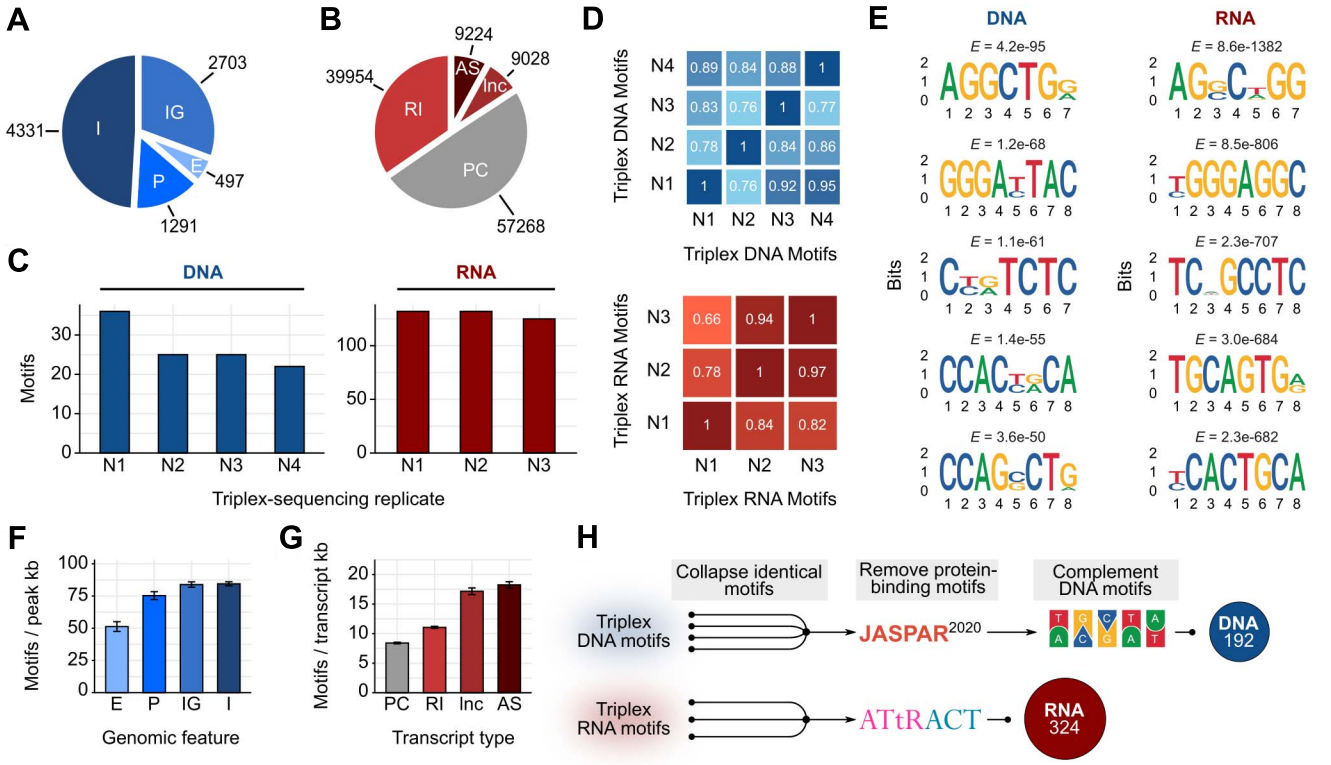

Figure 2. Identification of enriched and reproducible RNA·DNA:DNA triple helix-forming motifs. **(A)** Distribution of triplexDNA-seq peaks across intronic regions (I), intergenic regions (IG), exonic regions (E) and promoter regions (p) as annotated in the hg38 genome build by NCBI. **(B)** Distribution of triplexRNA-seq peaks across antisense transcripts (AS), long non-coding RNAs (lnc), protein-coding transcripts (PC) and transcripts with retained introns (RI). **(C)** Total significantly enriched ($E < 0.01$) triplexDNA and triplexRNA motifs identified per replicate of triplexDNA-seq and triplexRNA-seq. **(D)** Proportions of motifs per replicate with similar ($P < 0.05$, *Tomtom*) motifs in accompanying replicates of triplexDNA-seq (blue) or triplexRNA-seq (red). **(E)** The five most enriched motifs across all replicates of triplexDNA-seq (left) and triplexRNA-seq (right). **(F)** Occurrence of triplexDNA motifs per kilobase of triplexDNA-seq peaks appearing in exonic (E), promoter (P), intergenic (IG) and intronic (I) genomic regions. **(G)** Occurence of triplexRNA motifs per kilobase of protein-coding (PC), retained intron (RI), long non-coding (lnc) and antisense (AS) transcripts. **(H)** Schematic of motif processing steps, including removal of identical motifs, removal of known protein-binding motifs and inclusion of reverse-complement triplexDNA motifs, which resulted in the final sets of triplexRNA (red) and triplexDNA (blue) motifs.

codes learned using an expectation-maximization algorithm from triplex motifs have biophysical relevance in triplex formation.

In order to stratify the codes to be taken forward and implemented in *TriplexAligner*, high-scoring results (*objective value* < 0.75, *total motifs utilised* > 50%) output from the expectation-maximization algorithm were hierarchically clustered and subjected to tree cutting, resulting in eight distinct code clusters (Figure 3F–G). Amongst these were partially redundant pairs of codes (clusters 1 and 2, clusters 3 and 4), resulting from the probabilistic nature of the outputs. Complementary codes (clusters 5 and 7) were also present, reflecting the non-stranded nature of the input sequencing data. Amongst these codes were canonical Hoogsteen base pairing rules (C·G:C, U·A:T, G·G:C, A·A:T) [21] along with previously unreported RNA·DNA:DNA base pairings. The learning of potentially novel base pairings is a key point of the naive, unbiased approach taken herein. To determine the potential worth of these base pairings, an assessment of the returned codes in the prediction of published RNA–DNA interactions was carried out.

In order to perform the validation of the stratified codes returned by the expectation-maximization algorithm, the codes were implemented as scoring matrices in a new software called *TriplexAligner*. *TriplexAligner* is a local alignment program which uses Karlin–Altschul statistics [54] to determine subsequences of triplex formation between RNA and DNA.

TriplexAligner recalls RNA–DNA interactions and known triplexes

Genome-wide validation of *TriplexAligner* was carried out using published RNA–DNA interactions as detected by RADICL-seq [30] and RedC [31]. Significant interactions between transcripts and 5 kb genomic bins were decomposed to RNA and DNA sequences and constituted the positive data set. RADICL-seq interactions were further refined using accompanying nuclear RNA-seq data [30]. Transcript sequences were shuffled in order to generate a negative interaction set whilst maintaining nucleotide frequencies. Triplex formation between sequences was then predicted using *TriplexAligner*, *Triplexator* [20] and *Long Target* [22]. Maximum scores for each method were computed per RNA–DNA interaction (Figure 4A) and used as predictive values. All three tools were able to positively classify RADICL-seq and RedC interactions, with *TriplexAligner* returning the greatest area under the receiver operating characteristic (ROC) curve for both assays (Figure 4B and C, Supplementary Figure 4A). In both cases, the area under the ROC curve was significantly greater for *TriplexAligner* compared with the other tools ($P < 0.05$, bootstrapping $n = 2000$) (Figure 4D).

Upon assessing the performance of each of the individual mapping codes, which constitute *TriplexAligner*, it became clear that code performance varied between both individual codes and assays. Notably, Code 5 displayed the lowest performance on both RADICL-seq and RedC data and was only marginally better

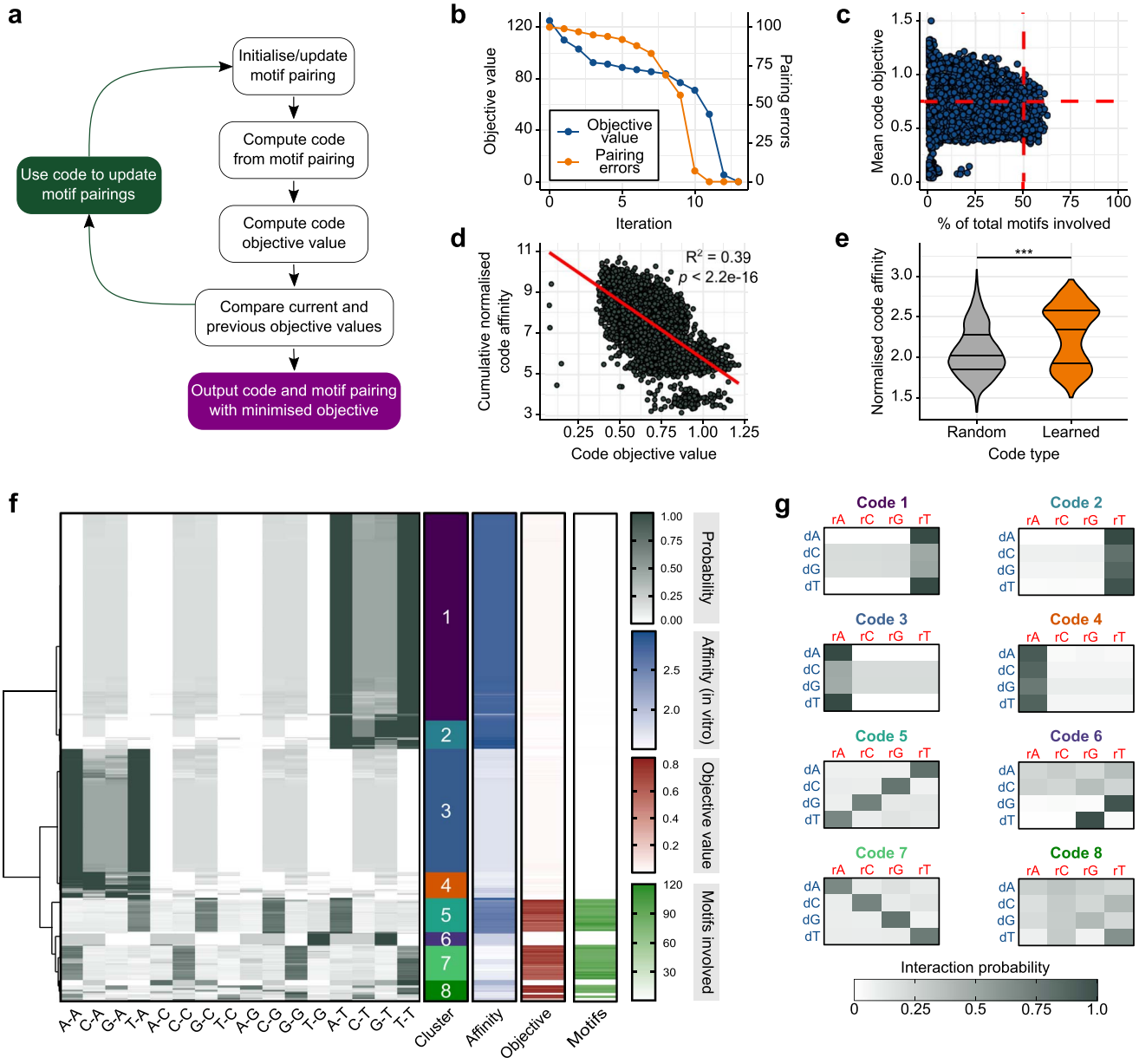

Figure 3. Learning RNA·DNA:DNA triple helix nucleotide pairing rules from motifs using expectation-maximization. **(A)** Schematic of the expectation-maximization algorithm used to learn RNA·DNA:DNA triple helix base pairing probabilities from pairings of enriched triplexRNA and triplexDNA motifs. **(B)** Example use-case of the expectation-maximization algorithm on simulated sets of motifs ($n = 100$) which were paired by Watson–Crick base pairing rules, with corresponding objective values and number of incorrect motif pairs displayed per iteration of the algorithm. **(C)** Output from the expectation-maximization algorithm when run on enriched triplexDNA and triplexRNA motifs identified from triplex-seq, displaying the mean objective values across all code models learned per initiation of the algorithm and the corresponding proportion of motifs included. **(D)** Correlation between code model objective values and *in vitro* RNA·DNA:DNA binding affinities as reported in [43]. Objective values and affinities were subjected to linear regression, with corresponding coefficient of determination (R^2) and *P*-value displayed on the plot. **(E)** Comparison in code model affinities between high-scoring subset (*objective value* < 0.75, *total motifs* > 50%) expectation-maximization results and a size-matched set of randomly generated code models ($P < 0.001$, Wilcoxon signed-rank test). **(F)** High-scoring expectation-maximization results subjected to hierarchical clustering and tree-cutting ($k = 8$), with corresponding clusters, code model affinities, objective values and total motifs assigned displayed. **(G)** Mean probabilistic code models per cluster of expectation-maximization results.

than random code performance (Figure 4E, Supplementary Figure 4B). Different transcripts also showed different preferences for codes. For instance, the two most prevalent lncRNAs in the RedC dataset–*MALAT1* and *NEAT1*–showed completely different code preferences. RedC interactions involving *MALAT1* were most often best-predicted by code 3, compared with interactions of *NEAT1* which were more heterogeneously predicted, with a tendency towards code 7 (Supplementary Figure 4C).

Where the above results showcase the ability of *TriplexAligner* to recall genome-wide RNA–DNA interactions, we also sought to demonstrate that *TriplexAligner* could predict previously published triplex interactions. Triplex interactions between the lncRNA *SARRAH* and a number of cardiac gene promoters (*ITPR2*, *PARP8*, *PDE3A*, *SSBP2* and *GPC6*) have been reported [55]. When using *TriplexAligner* to predict the triplex formation between *SARRAH* and these genomic regions, triplex formation at the

cardiac promoters was predicted with greater $-\log_{10}(E)$ values compared with the control promoter used in the experiment (*GAPDH*) (Figure 4G). Of the promoters considered, the best alignment returned by *TriplexAligner* was between *SARRAH* and *ITPR2* **(Supplementary Figure 4D)**. Other published triplex interactions, between *HOTAIR* and the promoter region of *PCDH7* [56] (Figure 4H, left), as well as the triplex formed between *NEAT1* and the *CYP4F22* promoter [25] (Figure 4H, right) returned $-\log_{10}(E)$ values of 5.84 and 18.0, respectively.

If implemented in a genome-wide manner, *TriplexAligner* could also be used to identify regulatory regions of RNAs which are important to triplex formation. Here, triplex formation between an exemplary RNA (*Neat1*) and promoter sequences of genes differentially expressed after *Neat1* knockout [57] was predicted using *TriplexAligner*. It was evident that a specific region of the *Neat1* transcript was implicated in predicted triplex formation (Figure 4I), indicating a region of putative regulatory importance in the transcript.

TriplexAligner code models are biophysically valid

To assess the biophysical validity of the code models used in *TriplexAligner*, maximal RADICL-seq pair alignments of each code were computed. The alignment with the greatest $-\log_{10}(E)$ value was that implementing code 7. RNA and DNA oligonucleotides representing maximally scoring subsequences of this interaction (Figure 5A) were submitted for analysis by electrophoretic mobility shift assays (EMSA), circular dichroism (CD) spectroscopy and melting curve analysis.

When the double-stranded DNA was incubated with single-stranded RNA and subjected to an EMSA, an RNaseH-resistant band could be observed in the gel separate from the double-stranded DNA alone (Figure 5B). RNaseH resistance indicates that the formed structure was not an R-loop [58] and could therefore be an RNA·DNA:DNA triple helix. When subjected to CD spectroscopy, a distinct negative peak at 230 nm was present when RNA and DNA were mixed, along with a shifted and prominent main peak at 270 nm (Figure 5C). These shifts were not visible when double-stranded DNA alone or mixed single-stranded DNA and single-stranded RNA (heteroduplex) were tested. In melting assays, two melting points could be assigned to the curve obtained from the mixed double-stranded DNA and single-stranded RNA (Figure 5D). In contrast, only single melting points could be assigned to double-stranded DNA alone and heteroduplex inputs.

These results indicate that RNA–DNA interactions positively predicted by *TriplexAligner* have the potential to be biophysically valid triplexes, even when only a small portion of the predicted triplex is tested.

Discussion

Unlike previously published tools for the prediction of triplex formation, *TriplexAligner* uses probabilistic nucleotide pairing models learned from sequencing of triplex-forming DNA and RNA to predict triplexes. Compared with discrete and canonical Hoogsteen base pairing rules, this resulted in the improved recall of all-to-all RNA-DNA interactions. This demonstrates that formation of RNA–DNA interactions is more complex than simple base pairing rules, and therefore, prediction of such interactions requires more malleable models such as those proposed here.

The nature of the input data originating from triplexRNA-seq and triplexDNA-seq [25], and the fact that we wanted to integrate triplex codes into local alignment methods, has led us to our formulation using RNA-to-DNA matrices. We showed that optimization of the code matrices and the corresponding motif pairs can be formulated as a code mixture problem using an expectation-maximization algorithm [59]. The use of this algorithm permitted the estimation of triplex motif pairings, and subsequently RNA–DNA base pairing probabilities whose error (objective) could be minimized. In practice, the algorithm converged in few iterations. The drawbacks of expectation-maximization [60] either had a minimal impact (quick convergence) or could be effectively overcome (many initiations to cover the search space). Thus, the algorithm provided useful information for implementation in *TriplexAligner*. Notably, if more complex formulations of RNA-to-DNA interaction would be considered, the optimization would likely become more challenging and a straightforward integration into existing local alignment approaches may become impossible. Here, we were able to use the established Karlin–Altschul statistic to directly generate *E*-values for the triplex alignments, which is a novel contribution compared with established tools. Nevertheless, more complex formulations of RNA-to-DNA interaction would require alternative machine learning approaches, such as neural network-based approaches.

The naive, unbiased manner in which the expectation-maximization algorithm used herein was formulated meant that the returned code models are a heterogeneous mix of known, canonical Hoogsteen base pairings [21] along with previously unreported RNA–DNA base pairings. Due to the probabilistic nature of the code models, it was initially challenging to determine whether these putative novel base pairings were genuinely interesting, or artefacts. In the validation carried out in the course of this work, it could be shown that these unconventional pairings are also good predictors of RNA–DNA interaction. The question remains, however, whether these base pairings are involved in true biophysical interaction between RNA and DNA molecules, or whether their roles are more complex. For example, seeing as the roles of proteins in triplex formation remain unknown, it could be that these non-canonical base pairings are important for the recruitment of co-factors important for stabilization of the triplex structure.

Interestingly, it could be demonstrated that different transcripts may have different preferences for the code with which their interactions were predicted. The examples given here–*MALAT1* and *NEAT1*–are both lncRNAs which have been previously shown to interact with chromatin [61]. However, when their RNA–DNA interactions are predicted by *TriplexAligner*, they are best-predicted with different codes. These two lncRNAs have been proposed to carry out functions at similar genomic loci–namely actively transcribed genes–but to bind at distinct sites in these regions. Therefore, interactions underpinned by different codes would facilitate this process. Inferring functional roles of different types of triplex formation would constitute an exciting and important future research area.

When investigating the relative performance of each of the codes learned by the expectation-maximization algorithm described herein, we could observe differences in performance of codes between datasets. For example, code 1 was the highest performing code in prediction of RADICL-seq RNA–DNA interactions but was outperformed by codes 4, 2 and 3 in recall of RedC interactions. Given the distinct nature of the methods used to identify these RNA–DNA interactions, it is difficult to establish whether these are true differences in performance or merely reflect the subtle differences between the wet-lab methodologies. For instance, it may be that one of the methods enriches *trans*

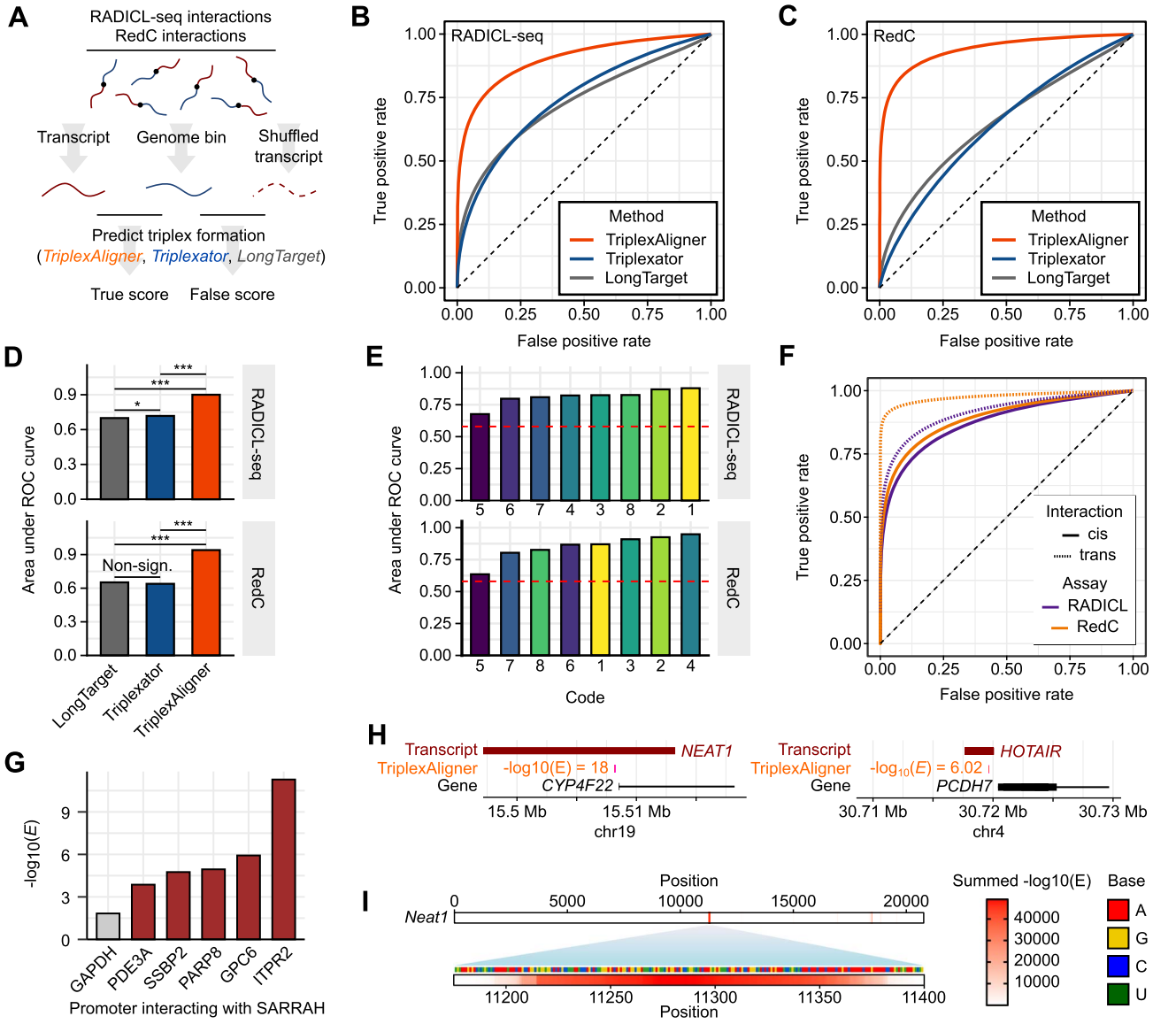

Figure 4. Computational validation of TriplexAligner using RNA–DNA interaction data and published RNA·DNA:DNA triple helix interactions. **(A)** Schematic outlining the computational validation of *TriplexAligner*, using global RNA–DNA interactions identified by either RADICL-seq or RedC and subjecting the corresponding RNA and DNA sequences to prediction of RNA·DNA:DNA triplex formation with *TriplexAligner*, *Triplexator* and *LongTarget*. Negative interaction data were generated by shuffling of RNA sequences. **(B)** ROC curves summarizing performance of *TriplexAligner* (orange), *Triplexator* (blue) and *LongTarget* (grey) in prediction of RADICL-seq RNA–DNA interactions. **(C)** ROC curves summarizing performance of *TriplexAligner* (orange), *Triplexator* (blue) and *LongTarget* (grey) in prediction of RedC RNA–DNA interactions. **(D)** Comparison of area under the ROC curves displayed in B and C (Non-sign. $P > 0.05$, * $P < 0,05$, *** $P < 0.001$, bootstrapping ($n = 2000$)). **(E)** Area under the ROC curves of individual *TriplexAligner* code models for RADICL-seq and RedC RNA–DNA interactions. **(F)***TriplexAligner* ROC curves for *cis* (RNA gene locus and interaction site on the same chromosome, solid line) and *trans* (RNA gene locus and interaction site on different chromosomes, dashed line) RNA–DNA interactions arising from RADICL-seq (purple) and RedC (orange) data. **(G)***TriplexAligner* -log10(E) values for predicted interactions between lncRNA *SARRAH* and published interacting promoters *ITPR2*, *PARP8*, *PDE3A*, *SSBP2* and *GPC6*, in comparison to the negative control promoter of *GAPDH*. **(H)***TriplexAligner* predictions of published RNA·DNA:DNA triplex helix formation between the lncRNAs *NEAT1* and *HOTAIR* and the promoter regions of *CYP4F22* and *PCDH7*, respectively. **(I)** Schematic of the lncRNA *Neat1* showing most commonly predicted sites of RNA·DNA:DNA triple helix formation in the lncRNA against multiple gene promoters dysregulated after *Neat1* knockout.

interactions more so than the other or fails to effectively remove interactions arising from nascent transcription. Another aspect to consider is species differences, given that the RADICL-seq data used here arise from murine cells, and the RedC from human material. Naively, we would not expect this to make a difference. However, so little is known about the conditions which facilitate triplex formation; there could be unknown species-specific co-factors which favour different triplex base pairings.

When compared with *Triplexator*, the most widely used tool for prediction of triplex formation, outputs from *TriplexAligner* differ in a number of aspects. Most notable is that triplexes predicted by *TriplexAligner* tend to be far broader than those predicted by *Triplexator*. There are several potential reasons for this observation, all of which are figurative, given the lack of wet-lab data. From a technical perspective, the implementation of *TriplexAligner* as a local aligner using Karlin–Altschul statistics [54] means that the score metric is highly dependent on the width of the alignment, and thus, broad alignments are more likely to be reported as interacting regions. Triplexes predicted by *Triplexator* are - by default - a minimum length of 20 base pairs. During

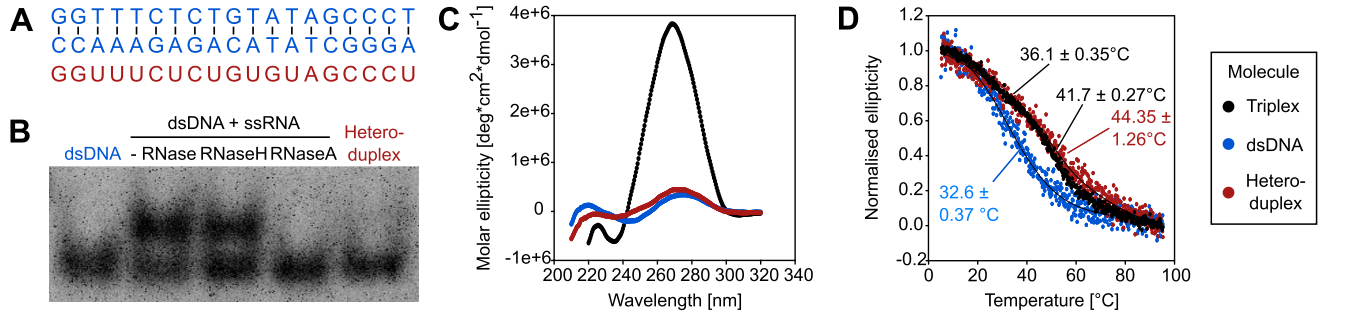

Figure 5. Biophysical validation of interacting DNA and RNA sequences as predicted by TriplexAligner. **(A)** Maximal scoring DNA (blue) and RNA (red) subsequences across RADICL-seq interactions as predicted by *TriplexAligner*, which were synthesized *in vitro* and used in subsequent biophysical experiments investigating RNA·DNA:DNA triple helix formation. **(B)** EMSA using combinations of DNA and RNA (shown in A), as either double-stranded DNA (dsDNA), double-stranded DNA and single-stranded RNA (dsDNA + ssRNA) and single-stranded DNA in combination with single-stranded RNA (heteroduplex). RNA·DNA:DNA triple helix formation was investigated in RNase-free conditions (- RNase), in combination with RNaseH or in combination with RNaseA. **(C)** CD spectroscopy of double-stranded DNA and single-stranded RNA (Triplex, black), double-stranded DNA (dsDNA, blue) and single-stranded DNA with single-stranded RNA (Heteroduplex, red). **(D)** Melting analysis DNA and RNA molecules (described in C), with melting points labelled and annotated.

prediction of RADICL-seq and RedC interactions, the widths of *Triplexator*-predicted interactions did not exceed 30 base pairs. In comparison, the alignments reported by *TriplexAligner* exceeded 100 base pairs on a number of occasions. Due to technical and financial restraints, it is challenging to experimentally determine whether these alignments are reflected in biological systems. However, longer tracts of triplex formation could permit increased specificity of interactions between transcripts and genomic loci, thereby mediating more precise regulatory relationships between RNAs and target genes. Alongside this, longer tracts of interaction could result in increased stability, increasing the robustness of the regulatory mechanism. Alternatively, the long tracts of interaction predicted by *TriplexAligner* may provide the conditions for RNA–DNA interaction to take place along the length of the tract in a dynamic manner. This would entail that the entire length is not interacting at any one time, and instead, sub-tracts of RNA and DNA interact when spatio-temporal conditions are favourable.

Whilst *TriplexAligner* recalls RNA–DNA interactions more accurately than previously published tools, it remains imperfect. There exist a variety of reasons for this. In *TriplexAligner*, the assumption is made that triplex formation takes place between two linear molecules, consequently disregarding the influence of higher order structures. Whilst *TriplexAligner* does not consider chromatin state, it was previously shown that triplexDNA-seq data isare enriched in regions of open chromatin [25]. It is therefore likely that motifs used to develop *TriplexAligner* arose from open chromatin, in spite of the previously reported repressive functions of triplex formation [62–64]. Consequently, triplexDNA-seq and therefore *TriplexAligner* could be biased towards triplex formation with activatory functions. Validating the effects of chromatin conformation on triplex formation would require non-steady-state data, where both differential chromatin states and differential triplexDNA-seq regions could be identified, and these data do not currently exist.

Beyond chromatin conformation, it is also possible that triplex formation is influenced by more complex 3D structures of both RNA and DNA. Sites of predicted triplex formation are correlated with 3D genome structure [65], but it is unclear when triplexes form relative to the establishment of 3D genomic structures. It is also likely that the secondary structures of triplex-forming transcripts affect on the formation of triplexes. Transcripts can fold into complex structures, resulting in regions with divergent accessibility [66]. This would restrict regions of RNA which are free to interact with DNA, alongside forming new interfaces which are irrecoverable from linear molecules. Integration of features beyond linear RNA and DNA sequences therefore represents an important future research topic. Experimental methods to examine high-resolution 3D structures of nucleic acids, such as RNA SHAPE [67], remain complex and challenging. However, progress in this field would provide further insight into the 3D requirements for RNA–DNA interactions to successfully form. Here again, the roles of proteins in the facilitation of RNA–DNA interactions are unclear, but likely highly important. The incorporation of 2D RNA structure prediction, such as tools available from the ViennaRNA package [68], into *TriplexAligner* could be a useful addition to the tool. In addition, the use of RNA aligners which consider secondary and tertiary structures [69–71], could be considered as a downstream analysis option, in order to compare triplex-forming RNA structures and identify important regulatory structures, which may facilitate triplex formation.

TriplexAligner was developed with the aim of providing researchers with a method of predicting RNA–DNA interactions which is grounded in data. By leveraging of triplex-forming sequences captured in next-generation sequencing experiments, *TriplexAligner* reports predictions with a basis in data, and which extend beyond canonical and discrete base pairing rules. As such, *TriplexAligner* is a unique tool with the potential to direct wet-lab research on regulatory RNA networks and thereby further clarify the role of RNA–DNA interactions in epigenetics.

Key Points

- Short, reproducible sequence motifs can be identified from triplexRNA- and triplexDNA-seq data.
- Expectation-Maximization can be used to learn RNA–DNA base pairing rules from enriched motifs.
- Both canonical Hoogsteen base pairing rules and previously unreported base pairing rules were identified by this approach and could be positively correlated with previous *in vitro* work.
- Implementation of the learned RNA–DNA base pairings in a local alignment program permitted the more accurate prediction of RNA–DNA interactions than previously published gold-standard tools.

- An RNA–DNA interaction predicted in the course of this analysis could be shown to be a biophysically valid RNA·DNA:DNA triple helix *in vitro*.

Data availability

Accession numbers for published data used in this study are detailed in *Materials and Methods*.

Code availability

The R package for *TriplexAligner* is available at https://github.com/SchulzLab/TriplexAligner. The code used in the learning of probabilistic RNA-DNA mapping codes is available at https://github.com/SchulzLab/Codefinder.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Authors' contributions

T.W., R.P.B., M.H.S. and M.S.L. designed the study and wrote the manuscript. T.W. and M.H.S. formulated and wrote the algorithms and carried out computational experiments. S.S., N.M.K., J.K.B. and H.S. provided instruments, performed experiments and analysed data. L.A., A.D., S.E., J.A.O. and A.B. provided important data and advice. All authors read and commented on the manuscript.

Funding

Goethe University Frankfurt am Main, the German Centre for Cardiovascular Research (DZHK), the DFG excellence cluster EXS2026 (Cardio-Pulmonary Institute); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 403584255 - TRR 267. Work at BMRZ is supported by the state of Hesse.

References

1. Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* 2015; **16**(2): 71–84.
2. Oo JA, Brandes RP, Leisegang MS. Long non-coding RNAs: novel regulators of cellular physiology and function. *Pflügers Archiv-European Journal of Physiology* 2021;1–14.
3. Takemata N, Ohta K. Role of non-coding RNA transcription around gene regulatory elements in transcription factor recruitment. *RNA Biol* 2017; **14**(1): 1–5.
4. Andric V, Nevers A, Hazra D, et al. A scaffold lncRNA shapes the mitosis to meiosis switch. *Nat Commun* 2021; **12**(1): 1–12.
5. Shimada Y, Mohn F, Bühler M. The RNA-induced transcriptional silencing complex targets chromatin exclusively via interacting with nascent transcripts. *Genes Dev* 2016; **30**(23): 2571–80.
6. Csankovszki G, Nagy A, Jaenisch R. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol* 2001; **153**(4): 773–84.
7. Zhang J-Y, Zheng K-W, Xiao S, et al. Mechanism and manipulation of DNA: RNA hybrid G-quadruplex formation in transcription of G-rich DNA. *J Am Chem Soc* 2014; **136**(4): 1381–90.
8. García-Muse T, Aguilera A. R loops: from physiological to pathological roles. *Cell* 2019; **179**(3): 604–18.
9. Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. *Nat Rev Genet* 2015; **16**(10): 583–97.
10. Crossley MP, Bocek M, Cimprich KA. R-loops as cellular regulators and genomic threats. *Mol Cell* 2019; **73**(3): 398–411.
11. Felsenfeld G, Davies DR, Rich A. Formation of a three-stranded polynucleotide molecule. *J Am Chem Soc* 1957; **79**(8): 2023–4.
12. Escudeé C, Francçois J-C, Sun J-S, et al. Stability of triple helices containing RNA and DNA strands: experimental and molecular modeling studies. *Nucleic Acids Res* 1993; **21**(24): 5547–53.
13. 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**(1): 1–17.
14. Chu C, Kun Q, Zhong FL, et al. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* 2011; **44**(4): 667–78.
15. Vrba L, Futscher BW. Epigenetic silencing of lncRNA MORT in 16 TCGA cancer types. *F1000Research* 2018;7.
16. Jiang M-C, Ni J-J, Cui W-Y, et al. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res* 2019; **9**(7): 1354.
17. Zhou Z, Giles KE, Felsenfeld G. DNA. RNA triple helix formation can function as a cis-acting regulatory mechanism at the human β-globin locus. *Proc Natl Acad Sci* 2019; **116**(13): 6130–9.
18. Garratt H, Ashburn R, Sopić M, et al. Long non-coding RNA regulation of epigenetics in vascular cells. *Non-coding RNA* 2021; **7**(4): 62.
19. Maldonado R, Filarsky M, Grummt I, et al. Purine–and pyrimidine–triple-helix-forming oligonucleotides recognize qualitatively different target sites at the ribosomal DNA locus. *RNA* 2018; **24**(3): 371–80.
20. 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.
21. Kuo C-C, Hänzelmann S, Cetin NS, et al. Detection of RNA–DNA binding sites in long noncoding RNAs. *Nucleic Acids Res* 2019; **47**(6): e32–2.
22. He S, Zhang H, Liu H, et al. LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. *Bioinformatics* 2015; **31**(2): 178–86.
23. Antonov IV, Mazurov E, Borodovsky M, et al. Prediction of lncRNAs and their interactions with nucleic acids: benchmarking bioinformatics tools. *Brief Bioinform* 2019; **20**(2): 551–64.
24. Zhang Y, Long Y, Kwoh CK. Deep learning based dna: Rna triplex forming potential prediction. *BMC bioinformatics* 2020; **21**(1): 1–13.
25. Cetin NS, Kuo C-C, Ribarska T, et al. Isolation and genome-wide characterization of cellular DNA: RNA triplex structures. *Nucleic Acids Res* 2019; **47**(5): 2306–21.
26. Zhou B, Li X, Luo D, et al. GRID-seq for comprehensive analysis of global RNA–chromatin interactions. *Nat Protoc* 2019; **14**(7): 2036–68.
27. Zhong S, Sridhar B, Rivas-Astroza M, et al. Mapping RNA-chromatin interactions. *FASEB J* 2018; **32**:525–2.
28. Wu W, Yan Z, Nguyen TC, et al. Mapping RNA–chromatin interactions by sequencing with iMARGI. *Nat Protoc* 2019; **14**(11): 3243–72.
29. Bell JC, Jukam D, Teran NA, et al. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife* 2018; **7**:e27024.

30. Bonetti A, Agostini F, Suzuki AM, et al. RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. *Nat Commun* 2020; **11**(1): 1–14.
31. Gavrilov AA, Zharikova AA, Galitsyna AA, 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.
32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**(4): 357–9.
33. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010; **38**(4): 576–89.
34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**(6): 841–2.
35. Bailey TL, Johnson J, Grant CE, et al. The MEME suite. *Nucleic Acids Res* 2015; **43**(W1): W39–49.
36. Bioconductor Core Team and Bioconductor Package Maintainer. *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*, 2021, R package version 3.13.0.
37. Morgan M. *BiocManager: Access the Bioconductor Project Package Repository*, 2021, R package version 1.30.16.
38. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.
39. Rainer J. *EnsDb.Hsapiens.v86: Ensembl based annotation package*, 2017, R package version 2.99.0.
40. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res* 2021; **49**(D1): D884–91.
41. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020; **48**(D1): D87–92.
42. Giudice G, Sánchez-Cabo F, Torroja C, et al. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database* 2016; **2016**.
43. Kunkler CN, Hulewicz JP, Hickman SC, et al. 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.
44. Murtagh F, Legendre P. *Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm* arXiv preprint arXiv:1111.6285, 2011.
45. Altschul SF, Wootton JC, Zaslavsky E, et al. The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput Biol* 2010; **6**(7):e1000852.
46. Pagès H, Aboyoun P, Gentleman R, et al. *Biostrings: Efficient manipulation of biological strings*, 2021, R package version 2.60.2.
47. Altschul SF, Bundschuh R, Olsen R, et al. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res* 2001; **29**(2): 351–61.
48. Millard SP. *EnvStats: An R Package for Environmental Statistics*. New York: Springer, 2013.
49. Bioconductor Core Team and Bioconductor Package Maintainer. *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*, 2021, R package version 3.13.0.
50. Bioconductor Core Team and Bioconductor Package Maintainer. *TxDb.Mmusculus.UCSC.mm10.knownGene: Annotation package for TxDb object(s)*, 2019, R package version 3.10.0.
51. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017; **14**(4): 417–9.
52. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; **12**:77.
53. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005; **21**(20): 7881.
54. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci* 1990; **87**(6): 2264–8.
55. Trembinski DJ, Bink DI, Theodorou K, et al. Aging-regulated anti-apoptotic long non-coding RNA sarrah augments recovery from acute myocardial infarction. *Nat Commun* 2020; **11**(1): 1–14.
56. Kalwa M, Hänzelmann S, Otto S, et al. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res* 2016; **44**(22): 10631–43.
57. Mello SS, Sinow C, Raj N, et al. Neat1 is a p53-inducible lincRNA essential for transformation suppression. *Genes Dev* 2017; **31**(11): 1095–108.
58. Amon JD, Koshland D. RNase H enables efficient repair of R-loop induced DNA damage. *Elife* 2016; **5**:e20533.
59. Do CB, Batzoglou S. What is the expectation maximization algorithm? *Nat Biotechnol* 2008; **26**(8): 897–9.
60. Guo Y, Schuurmans D. Convex relaxations of latent variable training. *Advances in Neural Information Processing Systems* 2007; **20**.
61. West JA, Davis CP, Sunwoo H, et al. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol Cell* 2014; **55**(5): 791–802.
62. 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): 474–85.
63. O'Leary VB, Ovsepian SV, Smida J, et al. PARTICLE- The RNA podium for genomic silencers. *J Cell Physiol* 2019; **234**(11): 19464–70.
64. 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): 2264–9.
65. Farabella I, Di Stefano M, Soler-Vila P, et al. Three-dimensional genome organization via triplex-forming RNAs. *Nat Struct Mol Biol* 2021; **28**(11): 945–54.
66. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun* 2021; **12**(1): 1–9.
67. Spitale RC, Crisalli P, Flynn RA, et al. RNA SHAPE analysis in living cells. *Nat Chem Biol* 2013; **9**(1): 18–20.
68. Gruber AR, Lorenz R, Bernhart SH, et al. The vienna RNA websuite. *Nucleic Acids Res* 2008; **36**(suppl_2): W70–4.
69. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 2005; **21**(16): 3352–9.
70. Techa-Angkoon P, Sun Y. glu-RNA: aliGn highLy strUctured ncRNAs using only sequence similarity. In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 2013, 508–17.
71. Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* 2019; **35**(21): 4459–61.