Prediction of RNA:DNA:DNA triple helix formation using next-generation sequencing data

Timothy Warwick  
Institute for Cardiovascular Physiology, Goethe University  
https://orcid.org/0000-0002-6804-2931

Sandra Seredinski  
Institute for Cardiovascular Physiology, Goethe University

Nina Krause  
Institute for Organic Chemistry and Chemical Biology, Centre for Biomolecular Magnetic Resonance (BMRZ), Goethe University

Jasleen Kaur Bains  
Institute for Organic Chemistry and Chemical Biology, Centre for Biomolecular Magnetic Resonance (BMRZ), Goethe University

Lara Althaus  
Institute of Pharmacology and Toxicology, Technical University of Munich

James Oo  
Institute for Cardiovascular Physiology, Goethe University

Alessandro Bonetti  
Translational Genomics, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca

Anne Dueck  
Institute of Pharmacology and Toxicology, Technical University of Munich  
https://orcid.org/0000-0002-7956-6327

Stefan Engelhardt  
Institute of Pharmacology and Toxicology, Technical University of Munich  
https://orcid.org/0000-0001-5378-8661

Harald Schwalbe  
Institute for Organic Chemistry and Chemical Biology, Centre for Biomolecular Magnetic Resonance (BMRZ), Goethe University  
https://orcid.org/0000-0001-5693-7909

Matthias Leisegang  
Institute for Cardiovascular Physiology, Goethe University

Marcel Schulz  
Vascular Research Centre, Goethe-University, Frankfurt am Main

Ralf Brandes (✉ brandes@vrc.uni-frankfurt.de)  
Institute for Cardiovascular Physiology, Goethe University  
https://orcid.org/0000-0002-8035-0048
Prediction of RNA:DNA:DNA triple helix formation using next-generation sequencing data

Timothy Warwick\textsuperscript{1,4}, Sandra Seredinski\textsuperscript{1,4}, Nina M. Krause\textsuperscript{3}, Jasleen Kaur Bains\textsuperscript{3}, Lara Althaus\textsuperscript{5,6}, James A. Oo\textsuperscript{1,4}, Alessandro Bonetti\textsuperscript{7}, Anne Dueck\textsuperscript{5,6}, Stefan Engelhardt\textsuperscript{5,6}, Harald Schwalbe\textsuperscript{3}, Matthias S. Leisegang\textsuperscript{1,4}, Marcel H. Schulz\textsuperscript{2,4*} and Ralf P. Brandes\textsuperscript{1,4*}

\textsuperscript{1*}Institute for Cardiovascular Physiology, Goethe University, Frankfurt am Main, Germany.
\textsuperscript{2*}Institute for Cardiovascular Regeneration, Goethe University, Frankfurt am Main, Germany.
\textsuperscript{3}Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance (BMRZ), Goethe University, Frankfurt am Main, Germany.
\textsuperscript{4}German Center for Cardiovascular Research (DZHK), Partner site Rhein-Main, Germany.
\textsuperscript{5}Institute of Pharmacology and Toxicology, Technical University of Munich, Munich, Germany.
\textsuperscript{6}German Center for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Munich, Germany.
\textsuperscript{7}Translational Genomics, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden.

*Corresponding author(s). E-mail(s):
marcel.schulz@em.uni-frankfurt.de; brandes@vrc.uni-frankfurt.de;
Contributing authors: warwick@vrc.uni-frankfurt.de;
seredinski@vrc.uni-frankfurt.de; krause@nmr.uni-frankfurt.de;
bains@nmr.uni-frankfurt.de; lara.althaus@tum.de;
oo@vrc.uni-frankfurt.de; alessandro.bonetti@astrazeneca.com;
Warwick et al.: TriplexAligner

anne.dueck@tum.de; stefan.engelhardt@tum.de;
schwalbe@nmr.uni-frankfurt.de; leisegang@vrc.uni-frankfurt.de;

Abstract

RNA:DNA:DNA triple helix (triplex) formation regulates gene expression, but is difficult to study experimentally in vivo. This makes accurate computational prediction of triplex formation highly important in the field of RNA research. Current predictive methods have used canonical Hoogsteen base pairing rules, which whilst biophysically valid, may not reflect the plastic nature of cell biology. Herein, we present TriplexAligner, a local alignment tool implementing probabilistic scoring matrices learned from triplex-forming sequences captured in published triplexRNA-seq and triplexDNA-seq experiments. Short, conserved sequence elements were found to be enriched at points of triplex formation. Probabilistic mapping codes between RNA and DNA sequences were learned by Expectation-Maximisation, and used as scoring matrices for local alignment. TriplexAligner predicts RNA-DNA interactions identified in all-to-all sequencing data more accurately than previously published tools, and also predicts previously studied triplex interactions with known regulatory functions.

Keywords: RNA, DNA, triplex, Machine learning

Numerous regulatory roles have been ascribed to RNAs [25, 40], which include interactions with both DNA and proteins. The RNA-protein interface includes functions such as transcription factor addressing and recruitment [54], scaffolding of transcription factor machinery [4], and mediation of histone modifications [52]. 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 [14]. Other examples of RNA-DNA interactions include RNA-DNA hybrid G-quadruplex [59] and R-loop [19] formation. R-loops consist of interactions between single-stranded DNA and RNA via Watson-Crick base pairing, which have been implicated in chromatin condensation and tumorigenesis [48, 13]. 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) [17].

Triplex formation represents an area of epigenetics which, although known of in a biophysical sense for many years [15], 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 \cite{36,12}. However, regulatory transcripts have been implicated in epigenetic mechanisms across many species, tissues, and cell types \cite{57,27,62,20}. 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 \cite{33}, which are canonically responsible for triplex formation. Tools implementing Hoogsteen rules include \textit{Triplexator/Triplex Domain Finder} \cite{11,31} and \textit{LongTarget} \cite{23}. Whilst usage of canonical rules to predict triplex formation provides insight into putative RNA-DNA interactions, benchmarking of \textit{Triplexator} and \textit{LongTarget} using \textit{MEG3 ChOP-sequencing} data \cite{36} revealed substantial room for improvement in this area \cite{5}. Further, these data suggest that triplex formation in a cellular context might extend beyond canonical base-pairing rules.

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) \cite{51}. 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 \cite{61,60,58,7}. However, the methods with most similar nucleotide processing protocols to triplex-seq are RNA And DNA Interacting Complexes Ligated and sequenced (RADICL-seq) \cite{10} and RedC \cite{21}. 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-maximisation on 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 to previously published tools, and the prediction of triplex formation between experimentally validated triplex-forming transcripts and genomic loci. Predicted interacting sequences were also subjected to biophysical validation, where triplex formation could be experimentally verified \textit{ex vivo}.

1 Results

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

In order to predict interactions between DNA and RNA mediated by triplex formation (Fig. 1a) we developed \textit{TriplexAligner}. \textit{TriplexAligner} is capable of predicting RNA-DNA interactions with high accuracy, surpassing currently available methods. Development of \textit{TriplexAligner} (Fig. 1b) 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 [51] 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 \textit{COL15A1} [36] could be clearly identified in the dataset (Fig. 1c). This satisfied the requirement that the input data sufficiently capture triplex formation taking place in the cellular context.
1.2 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 (Fig. 2a), although promoter regions were most enriched for peak occurrence relative to the proportion of the genome covered (Supplementary Fig. 1a). When calling peaks from triplexRNA-seq data, most peaks were detected in protein-coding transcripts (Fig. 2b). Transcripts with retained introns, followed by antisense transcripts, were most enriched for triplexRNA-peaks relative to transcript length (Supplementary Fig. 1b).

To identify sequences underpinning triplex-sequencing peaks, motif enrichment analysis was performed using MEME-ChIP [6] on sequences underlying peaks in each sample. Between 22 and 36 significantly enriched motifs were identified per triplexDNA-seq sample (Fig. 2c, left). More motifs were identified in the RNA samples, which each returned more than 125 enriched motifs (Fig. 2c, right). To investigate the reproducibility of the enriched sequences, motifs were compared between samples using Tomtom [6]. 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 (Fig. 2d, top). RNA motifs were also reproducible, with the minimum percentage of similar motifs between any two samples being 66% (Fig. 2d, bottom). Individual RNA motifs also exhibited more significant enrichment than DNA motifs, observable when examining the five most enriched motifs of each molecule (Fig. 2e). The most enriched RNA motif had an $E$ value of $8.6 \times 10^{-1382}$, in comparison to an $E$ value of $4.2 \times 10^{-95}$ 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 (Fig. 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 (Fig. 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 Fig. 1c).

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, 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 (Fig. 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.

1.3 Expectation-maximisation 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 identified in the previously described steps (Fig. 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 minimised, 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 Fig. 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 (Fig. 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. Subset codes were annotated with published in vitro triplex nucleotide dissociation equilibrium constants. 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 subset 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 (Fig. 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 (\(R^2 = 0.394, p < 2.2 \times 10^{-16},\)
Fig. 3e). These data - together with published in vitro work - suggest that
the probabilistic codes learned using an expectation-maximisation algorithm
from triplex motifs have biophysical relevance in triplex formation.

In order to stratify the codes to be taken forward and implemented
in TriplexAligner, subset results (objective value < 0.75, total motifs
utilised > 50%) output from the expectation-maximisation algorithm were
hierarchically clustered and subjected to tree cutting, resulting in eight distinct
code clusters (Fig. 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. All
codes were subsequently implemented as scoring matrices in TriplexAligner, a
local alignment program which uses Karlin-Altschul statistics [29] to determine
points of triplex formation between RNA and DNA.

1.4 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 [10] and RedC [21].
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 [10]. 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 [11] and Long Target [23]. Maximum scores for each method
were computed per RNA-DNA interaction (Fig. 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 (Fig. 4b-c, Supplementary Fig. 2a). In both cases, the area under the ROC curve was significantly greater for TriplexAligner compared to the other tools ($p < 0.05$, bootstrapping, $n = 2000$) (Fig. 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 than random code performance (Fig. 4e, Supplementary Fig. 2b-c).

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 [56]. 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 to the control promoter used in the experiment ($GAPDH$) (Fig. 4g). Of the promoters considered, the best alignment returned by TriplexAligner was between SARRAH and $ITPR2$ (Supplementary Fig. 2d). Other published triplex interactions, between $HOTAIR$ and the promoter region of $PCDH7$ [28] (Fig. 4h, left), as well as the triplex formed between $NEAT1$ and the $CYP4F22$ promoter [51] (Fig. 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 differential after $Neat1$ knockout [34] was predicted using TriplexAligner. It was evident that a specific region of the transcript was implicated in predicted triplex formation (Fig. 4i), indicating a region of putative regulatory importance in the transcript.

### 1.5 TriplexAligner code models are biophysically valid

To assess the biophysical validity of the code models used in TriplexAligner, maximal 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 (Fig. 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 (Fig. 5b). RNaseH resistance indicates that the formed structure was not an R-loop [3], 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 (Fig. 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 (Fig. 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.

2 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 to 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 requires more malleable models such as those proposed here.

When compared to 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 [29] means that the score metric is highly dependent on the width of the alignment, 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 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.

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 is enriched in regions of open chromatin [51]. 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 [41, 39, 50]. 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 formation, it is also possible that triplex formation is influenced by more complex three-dimensional structures of both RNA and DNA. Sites of predicted triplex formation are correlated with three-dimensional genome structure [16], but it is unclear when triplexes form relative to the establishment of three-dimensional 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 [49]. 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.

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.

3 Methods

3.1 Triplex sequencing data processing

Publicly accessible triplexDNA-seq data and triplexRNA-seq data [51] (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 [24]. 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 [44].

3.2 Motif enrichment and processing

Peak sequences were used as input to motif enrichment using MEME-ChIP (v5.0.5) [6]. 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 $E$-value $< 0.05$. 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 [55, 37, 45]. Triplex RNA motif occurrence was computed per transcript biotype, as defined by EnsDb.Hsapiens.v86, and normalised to transcript length [46, 26]. Enriched triplex DNA and RNA motifs were compared against the JASPAR 2020 [18] and ATtRACT [22] motif databases, respectively, in order to remove any motifs which were identical to transcription factor-binding or RNA-binding protein motifs.

3.3 An Expectation-Maximisation based method for RNA-DNA code optimization

3.3.1 Nomenclature

Assume we are given a DNA motif $D$ of length $l$, which is a matrix $D^{4 \times l}$, over the DNA alphabet $\Sigma = \{A, C, G, T\}$. 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 $U \rightarrow T$. 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 $D = \{D_1, \ldots, D_n\}$ and equivalently a set of RNA motifs $R = \{R_1, \ldots, R_n\}$. 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 $C^{4 \times 4}$, 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, thus it holds that:

$$\forall r \in \Sigma, \sum_{d \in \Sigma} C_{r,d} = 1. \quad (1)$$

The interest in this formulation is to learn the code $C$ that is behind a given set of motifs $R$ and $D$. 
3.3.2 Code error

We define the average column-wise mapping error between a DNA motif $D$ and an RNA motif $R$ given a defined code matrix $C$ as:

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

(2)

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

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

(3)

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

3.3.3 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 error (Eq. (2)) for the pairing of RNA and DNA matrices:

$$\arg\min_{C} = error(\mathcal{D}, \mathcal{R}, P, C), C \in \mathcal{C},$$

(4)

where $error(\mathcal{D}, \mathcal{R}, P, C)$ denotes the sum of code errors for all defined pairs using Eq. (2) and $\mathcal{C}$ denotes the space of all possible code matrices. We refer to the error in Eq. (4) as the objective value of the formulation. Luckily, we can obtain the code matrix $C$ that minimizes the code error using quadratic programming efficiently.

3.3.4 An Expectation-Maximisation algorithm for finding optimal code sets

While it is straightforward to obtain a code matrix $C$ that minimizes the code error for a given pairing of RNA and DNA motifs, in practice the true pairing is not known. Further, 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-Maximisation 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 following steps to find \( k \) many
code matrices:

**Input:** \( D, R, k \)

1. \( C_1 = C_1^*, \ldots, C_k = C_k^* \)
2. obtain the best pairings for elements in \( D \) and \( R \) using one of \( C_1, \ldots, C_k \)
3. for each: \( i=1, \ldots, k \)
   
   \( C_i^* = \text{minimize code of all paired DNA and RNA motifs that used } C_i \)

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 \( P \)

Step 2, where the best pairings are obtained, refers to obtaining the
best pair for each RNA motif in \( R \) with a DNA motif in \( D \) testing all \( k \)
code matrices. The best pair are the indices \( i, j, h \) where \( \text{error}(R_i, D_j, C_h) \) is
minimal. In this process, it is allowed that several elements from \( R \) are paired
with the same element in \( D \) 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. Applications 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 several times with the same input
data.

### 3.4 Code processing and annotation

Learned code models, which were output from the Expectation-Maximisation
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 [30] was used to compare the affinities of learned
code models versus a size-matched set of random code models. In short, the
normalised dissociation constant of the RNA:DNA interaction as reported in [30] 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-Maximisation. Following this, subset code models were subjected to hierarchical clustering with Euclidean distances and Ward’s method [38]. The resulting dendrogram was then cut to produce eight clusters, and the mean code model for each cluster was computed.

3.5 Formulation of TriplexAligner

To be implemented in TriplexAligner, probabilistic code model values were converted to log odds scores according to [2]. Subsequent score distributions were computed for each code model with Biostrings::pairwiseAlignment [42], 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 to generalised extreme value distributions [1] with EnvStats::egevd [35], using maximum likelihood estimation. The parameter values $K$ and $\lambda$ could then be identified for each code model. Using these parameters, bit scores and corresponding $E$ values could be calculated from local alignment scores of respective code models according to the following formulae:

\[
S' = \frac{\lambda S - \ln(K)}{\ln 2}
\]

\[
E = mn 2^{-S'}
\]

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.

3.6 Computational validation of TriplexAligner using global RNA-DNA interactions

RNA-DNA interactions arising from either RedC (GSE136141) or RADICL-sequencing (GSE132192) were used to benchmark the performance
of \textit{TriplexAligner} compared to the previously published tools \textit{LongTarget} and \textit{Triplexator}. For RedC data, interactions between RNAs and 5kb genomic bins which were present in two separate replicates were used for validation. For the RADICL-sequencing interactions, significant interactions between RNAs and 5kb genomic bins were identified according to [10]. Interactions between RNAs and genomic bins were expanded to include all possible transcripts of the involved RNA gene, as annotated in the \textit{TxDb.Hsapiens.UCSC.hg38.knownGene} [9] or \textit{TxDb.Mmusculus.UCSC.mm10.knownGene} [8] annotation packages for R [45]. RADICL-sequencing interactions were limited to those involving transcripts expressed in accompanying nuclear RNA-sequencing data, quantified using \textit{Salmon} (v 1.6.0) [43]. For each interaction, involved RNA and DNA sequences were subjected to RNA:DNA:DNA triple helix prediction using \textit{TriplexAligner}, \textit{LongTarget} and \textit{Triplexator}. A corresponding negative dataset was constructed via shuffling of the transcript sequences. \textit{LongTarget} was run with default parameters, and \textit{Triplexator} with \texttt{-e 20 -l 5}. Maximum metrics (\textit{TriplexAligner}: $-\log_{10}(E)$; \textit{LongTarget}: MeanStability; \textit{Triplexator}: Score) were identified per gene-bin interaction and used as predictive values in subsequent analyses with \textit{pROC} and \textit{ROCR} [47, 53]. Receiver operating characteristic curves were computed for each method with binomial smoothing, and statistically compared by bootstrapping (n = 2000).

### 3.7 Electrophoretic mobility shift assay

All hybridization steps were performed in 25 mM HEPES, pH 7.4, 50 mM NaCl and 10 mM MgCl$_2$. 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 hours at room temperature with 160 V.

3.8 CD spectroscopy and melting curve analysis

Circular dichroism spectra were acquired on a Jasco J-810 spectropolarimeter. The measurements were recorded from 210 nm 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, 10 mM MgCl₂ (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 [Θ] = deg × cm² × dmol⁻¹.

Melting curves were acquired at constant wavelength using a temperature rate of 1 °C/min in a range from 5°C 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:

\[ f = a/(1 + exp(-(x - x0)/b)) + c/(1 + exp(-(x - x2)/d)). \]
Supplementary information.

- **Supplementary Figure 1**: Further analysis of triplex-sequencing peaks and motifs.
- **Supplementary Figure 2**: Further detail of the computational validation of TriplexAligner.

Acknowledgments. This work was supported by the Goethe University Frankfurt am Main, the German Centre for Cardiovascular Research (DZHK), the DFG excellence cluster EXS2026 (Cardio-Pulmonary Institute), and the DFG Transregio project TRR267 (TP A06, TP B06, TP Z03). Work at BMRZ is supported by the state of Hesse.

Declarations

**Funding.** Detailed in Acknowledgements.

**Conflict of interest.** The authors declare the following competing interests: Alessandro Bonetti is an employee and shareholder of AstraZeneca.

**Ethics approval.** Not applicable.

**Consent to participate.** Not applicable

**Consent for publication.** Not applicable.

**Availability of data and materials.** Accession numbers for published data used in this study are detailed in 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.

**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.
References


REFERENCES


[21] Alexey A Gavrilov, Anastasiya A Zharikova, Aleksandra A Galitsyna,


REFERENCES


[52] Yukiko Shimada, Fabio Mohn, and Marc Bühler. The RNA-induced
transcriptional silencing complex targets chromatin exclusively via
interacting with nascent transcripts. *Genes & Development*,


[54] Naomichi Takemata and Kunihiro Ohta. Role of non-coding RNA
transcription around gene regulatory elements in transcription factor

[55] Bioconductor Core Team and Bioconductor Package Maintainer.
*TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb

[56] D Julia Trembinski, Diewertje I Bink, Kosta Theodorou, Janina Sommer,
Ariane Fischer, Anke van Bergen, Chao-Chung Kuo, Ivan G Costa,
Christoph Schürmann, Matthias S Leisegang, et al. Aging-regulated
anti-apoptotic long non-coding RNA sarrah augments recovery from acute

[57] Lukas Vrba and Bernard W Futscher. Epigenetic silencing of lncRNA

[58] Weixin Wu, Zhangming Yan, Tri C Nguyen, Zhen Bouman Chen, Shu
Chien, and Sheng Zhong. Mapping RNA–chromatin interactions by

Tan. Mechanism and manipulation of DNA: RNA hybrid G-quadruplex
formation in transcription of G-rich DNA. *Journal of the American

[60] Sheng Zhong, Bharat Sridhar, Marcelo Rivas-Astroza, Tri C Nguyen,
Weizhong Chen, Zhangming Yan, Xiaoyi Cao, and Lucie Hebert. Mapping

[61] Bing Zhou, Xiao Li, Daji Luo, Do-Hwan Lim, Yu Zhou, and Xiang-Dong
Fu. GRID-seq for comprehensive analysis of global RNA–chromatin

helix formation can function as a cis-acting regulatory mechanism at the
human β-globin locus. *Proceedings of the National Academy of Sciences*,
Fig. 1 | Overview of RNA:DNA:DNA triple helix formation and the development of TriplexAligner from triplex-sequencing 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-sequencing (blue) and triplexRNA-sequencing 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.
Fig. 2 | Identification of enriched and reproducible RNA:DNA:DNA triple helix-forming motifs. a, Distribution of triplexDNA-sequencing 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-sequencing peaks across antisense transcripts (AS), long non-coding RNAs (Inc), 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-sequencing and triplexRNA-sequencing. d, Proportions of motifs per replicate with similar (p < 0.05, Tomtom) motifs in accompanying replicates of triplexDNA-sequencing (blue) or triplexRNA-sequencing (red). e, The five most enriched motifs across all replicates of triplexDNA-sequencing (left) and triplexRNA-sequencing (right). f, Occurrence of triplexDNA motifs per kilobase of triplexDNA-sequencing peaks appearing in exonic (E), promoter (P), intergenic (IG) and intronic (I) genomic regions. g, Occurrence of triplexRNA motifs per kilobase of protein-coding (PC), retained intron (RI), long non-coding (Inc) 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.
Fig. 3 | Learning RNA:DNA:DNA triple helix nucleotide pairing rules from motifs using expectation-maximisation. 

a, Schematic of the expectation-maximisation 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-maximisation 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-maximisation algorithm when run on enriched triplexDNA and triplexRNA motifs identified from triplex-sequencing, 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 [29]. Objective values and affinities were subjected to linear regression, with corresponding coefficient of determination (R²) and p value displayed on the plot. 

e, Comparison in code model affinities between subset (objective value < 0.75, total motifs > 50%) expectation-maximisation results and a size-matched set of randomly generated code models (p < 0.001, Wilcoxon signed-rank test). 

f, Subset expectation-maximisation 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-maximisation results.
Fig. 4 | Computational validation of TriplexAligner using RNA-DNA interaction data and published RNA:DNA:DNA triplex helix interactions. 

a, Schematic outlining the computational validation of TriplexAligner, using global RNA-DNA interactions identified by either RADICL-sequencing 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 was generated by shuffling of RNA sequences. 
b, Receiver operating characteristic (ROC) curves summarising performance of TriplexAligner (orange), Triplexator (blue) and LongTarget (grey) in prediction of RADICL-sequencing RNA-DNA interactions. 
c, ROC curves summarising 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 (* 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-sequencing (purple) and RedC (orange) data. 
g, TriplexAligner –log_{10}(E) values for predicted interactions between IncRNA 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 IncRNAs NEAT1 and HOTAIR and the promoter regions of CYP4F22 and PCDH7, respectively. 
i, Schematic of the IncRNA Neat1 showing most commonly predicted sites of RNA:DNA:DNA triple helix formation by TriplexAligner across multiple gene promoters dysregulated after Neat1 knockout.
Fig. 5 | Biophysical validation of interacting DNA and RNA sequences as predicted by TriplexAligner. a, Maximal scoring DNA (blue) and RNA (red) subsequences across RADICL-sequencing interactions as predicted by TriplexAligner, which were synthesised in vitro and used in subsequent biophysical experiments investigating RNA:DNA:DNA triple helix formation. b, Electrophoretic mobility shift assay (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, Circular dichroism 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.
Supplementary Fig. 1 | a, Enrichment of triplexDNA peaks in intergenic (IG), intronic (I), exonic (E) and promoter (P) regions of the genome, relative to the total coverage of the respective features throughout the whole genome. b, Enrichment of triplexRNA peaks in IncRNAs (Inc), protein-coding (PC), antisense (AS) and retained intron (RI) transcript biotypes, relative to the coverage of those biotypes across the whole transcriptome. c, Positional occurrence of triplexDNA motifs in triplexDNA peaks (blue) and triplexRNA motifs in triplexRNA peaks (red).
Supplementary Fig. 2 | a, TriplexAligner ROC curves for individual RADICL-sequencing samples. b, TriplexAligner ROC curves per code, across all RADICL-sequencing samples. c, TriplexAligner ROC curves per code, across RedC interactions in human fibroblasts. d, Individual alignments of lncRNA SARRAH against cardiac and control (GAPDH) promoter regions, with $-\log_{10}(E)$ values returned by TriplexAligner displayed in orange.