Variable region design and cloning

Ethan Holleman

July 12, 2021

Abstract

Abstract text here

Contents

1 Background **2**

2 Insert design **2**
2.1 5' homology arm and KpnI recognition site . . . 3
2.1.1 Anchor . . . 3
2.1.2 Variable regions . . . 4
Restriction enzyme recognition sites . . . 6
Predicted R-loop probability . . . 6
RNA secondary structure . . . 7
2.1.3 EcoRI site and 3' homology arm . . . 8

3 Assembly of DNA inserts **8**
3.1 T7 initiation series constructs . . . 10
3.2 T7 termination series constructs . . . 11
3.3 Tac initiation series constructs . . . 13
3.3.1 Tac initiation series primer design . . . 14
3.4 Tac termination series constructs . . . 14
3.4.1 Tac termination series primer design . . . 14
3.5 Validation of insert libraries . . . 14
3.6 Quantification of R-loop formation . . . 15

4 Sequence availability **15**

5 Code availability **15**

6 Supplementary materials **15**
6.1 Variable region snakemake pipeline diagram . . . 15
6.2 Initial R-looper expectation calculations . . . 15
6.3 Homology arm calculations . . . 16
6.4 Variable regions in fasta format . . . 16
6.5 Complete inserts in fasta format . . . 20
6.6 RNA secondary structure expectation plots . . . 20

1 Background

R-loops are prevalent, functional important, non-B DNA structures that form co-transcriptionally when the nascent RNA strand hybridizes back to the DNA template forming a DNA-RNA hybrid [1]. R-loops are formed ubiquitously across the genomes of all organisms often occupying a large fractions of the genomes in a cell population.

Figure 1: Stages of R-loop formation. The encircled region represents the transcription bubble. The polymerase moves from left to right, starting at the promoter region (purple arrow). The upper strand (in red) in the bubble corresponds to the nascent RNA transcript, generated in the 5' to 3' direction. The R-loop initiates in frame 3, elongates in frame 4, and terminates in frame 5. Vazquez, Chedin and Natasa, 2020.

R-loops have been shown to cover approximately 5% of the human genome, and generally, their formation is favored by regions with high GC-skew and GC content as the potential DNA-RNA hybrids formed over these sequences tend to be more energetically favorable [2]. Beyond these broad associations there has been little investigation into the specific sequence characteristics that influence R-loop initiation and termination, representing a significant gap in knowledge that continues to become increasingly pressing as the recognized scope of these structure's impact of cellular functions continues to grow. Here, I describe a detailed strategy for the systematic investigation of the impact of DNA sequence on R-loop initiation and termination which seeks to implement the methods described by Vazquez, Chedin and Natasa in NSF proposal number 2054347.

2 Insert design

The overall goal of this series of experiments is to transcribe carefully controlled DNA sequences, referred to as inserts, in order to systematically observe the effects of these controlled sequences on R-loop initiation and termination *in vitro*. Inserts are composed of two types of DNA components; the variable and flanking infrastructure regions. The variable region is the actual region of interest with respect to R-loop formation while infrastructure sequences are additional nucleotide blocks that allow for the insertion of variable regions into specific plasmid backbones (see section **??**) in a modular fashion.

Figure 2: Diagram of a complete insert. Colored boxes represent features insert sequences are derived from while colored sequences represent the actual nucleotide sequence.

2.1 5' homology arm and KpnI recognition site

The first 30 bp of each insert will be composed of the 5' homology arm and a KpnI recognition site. These sequences are included to facilitate Gibson and restriction enzyme cloning into pFC9 and pFC8 respectively. These 30 bp are taken directly from nucleotides 27 - 56 of pFC9.

Figure 3: Location in pFC9 from which the 5' homology arm and KpnI site are modeled after. The entire 30 bp sequence is shown in purple with the 5' homology arm highlighted as a feature in blue and the KnpI site in orange.

This strategy will mean recognition sites for Eco53KI, SacI and Acc65I will be included in the 5' homology arms. However, none of these enzymes will be utilized in any cloning protocols and so only the KpnI site is noted in diagrams.

2.1.1 Anchor

The anchor region serves as a constant 15 nucleotide sequence that is always present in any final assembled construct adjacent to the 5' end of the variable region. This region provides significant utility when attempting to clone inserts already present in a plasmid into an alternative backbone as the number of primers required to amplify the insert region will not be dependent on the the number of variable regions but the number of unique backbones.This utility is more explicitly shown in sections 3.3 and 3.4.

Since this region is intended to be used as part of a PCR primer it was selected from candidate random sequences that satisfied the requirements below.

- Does not contain any common sub-strings of minimum length equal to 75% of the anchor's own length with any variable region or plasmid backbone.
- Does not contain the recognition sites for any restriction enzyme used in any relevant cloning procedure (see table 4).
- Has a melting temperature estimated by using nearest neighbor thermodynamics of at least 50℃.

The final anchor region sequence is available for download at this link.

Table 1: Properties of all syntheized inserts. The G clustering number refers to the number of G nucleotides in a given cluster with 0 being no clustering.

GC Skew	AT Skew	GC Content	G Clustering	Reverse Complement	Insert Number
0.2	0.0	0.4	0	1	0
0.1	0.0	0.3	0	0	1
0.6	0.0	0.6	0	0	2
0.4	0.0	0.3	0	1	3
0.2	0.0	0.3	0	1	4
0.0	0.0	0.3	0	1	5
0.0	0.0	0.6	0	0	6
0.6	0.0	0.5	0	0	7
0.1	0.0	0.6	0	0	8
0.4	0.0	0.6	0	0	9
0.2	0.0	0.6	0	0	10
0.0	0.0	0.5	0	1	11
0.6	0.0	0.4	0	0	12
0.0	0.0	0.4	0	1	13
0.1	0.0	0.5	0	0	14
0.6	0.0	0.3	0	0	15
0.4	0.0	0.5	0	1	16
0.2	0.0	0.5	0	1	17
0.1	0.0	0.4	0	0	18
0.4	0.0	0.4	0	1	19
0.4	0.0	0.6	2	1	20
0.4	0.2	0.6	2	1	21
0.4	0.4	0.6	2	1	22
0.4	0.0	0.6	3	1	23
0.4	0.2	0.6	3	1	24
0.4	0.4	0.6	3	1	25
0.4	0.0	0.6	4	1	26
0.4	0.2	0.6	4	1	27
0.4	0.4	0.6	4	1	28

2.1.2 Variable regions

Each insert will contain 1 variable region designed to test the effects of different sequence, namely GC / AT skew, GC content and G / C clustering, on both R-loop initiation and termination. Each variable region will be 200 bp in length and twenty nine different variable regions with the properties described in table 1 will ultimately be synthesized.

Each variable region is designed so that it may be used to test sequence effects on R-loop initiation when transcribed in the forward direction and placed adjacent to a promoter or R-loop termination when transcribed in reverse and placed more distantly to a promoter. Variable regions that will be tested as initiators and terminators are marked with a 1 in the reverse complement column of 1 and the properties of their reverse complements are described in table 2.

The cloning experiments required to produce these termination constructs will be undertaken after initiation experiments are completed. This will allow for the identification of a strong initiation sequence that can be used to reliability form R-loops.

While the global sequence properties for each variable region are well defined, properties such as GC-skew, content or nucleotide clustering do not determine which particular nucleotide should occur at a specific position in a given sequence. In this way, the parameters that define and separate each variable region can be thought of as bounding the set of all possible nucleotide sequences of length 200. In order to select one specific sequence

Table 2

GC Skew	AT Skew	GC Content	C Clustering	Reverse Complement of Insert
-0.2	-0.0	0.4	0	0
-0.4	-0.0	0.3	0	3
-0.2	-0.0	0.3	0	4
-0.0	-0.0	0.3	0	5
-0.0	-0.0	0.5	0	11
-0.0	-0.0	0.4	0	13
-0.4	-0.0	0.5	0	16
-0.2	-0.0	0.5	0	17
-0.4	-0.0	0.4	0	19
-0.4	-0.0	0.6	2	20
-0.4	-0.2	0.6	2	21
-0.4	-0.4	0.6	2	22
-0.4	-0.0	0.6	3	23
-0.4	-0.2	0.6	3	24
-0.4	-0.4	0.6	3	25
-0.4	-0.0	0.6	4	26
-0.4	-0.2	0.6	4	27
-0.4	-0.4	0.6	4	28

Table 3

Total synthesized inserts	Total contructs
29	47

from this set we can sample a large number of sequences and access each one with metrics relevant to the realities of the various cloning protocols and R-loop formation.

Figure 4: Visualization of sequence selection for an example variable region. Here, 100 candidate variable region sequences where generated and then ranked based on four metrics; mean R-loop probability (MRLP), mean local average energy (MLAE), proportion of nucleotides predicted to participate in RNA hairpins (PH) and proportion of nucleotides predicted to be unpaired (PUP). Sequences were by each individual metric (A) and an overall rank was calculated by selecting the candidate with the lowest average rank (B). The candidate with the best (lowest) overall rank was then selected as the final sequence.

Restriction enzyme recognition sites Over the course of all planned cloning experiments, all the restriction enzymes in table **??** will be utilized in some capacity. It is critical that the inserts are not cut unexpectedly within the variable region by any of these enzymes. Accordingly, before passing on for further downstream analysis candidate variable regions containing any of the recognition sites listed in table 4 were removed from consideration.

Table 4

Enzyme	Recognition sequence
KnpI	GGTACC
EcoRI	GAATTC
HindIII	AAGCTT

Predicted R-loop probability Previous Chedin lab member Dr. Robert Stolz developed rlooper, a physics based equilibrium model for predicting the probability of R-loop formation over a given DNA sequence [2]. Rlooper output includes a predicted probability of R-loop formation at each base in the input sequence, and the calculated average free energy (local average energy) from all R-loops overlapping that particular nucleotide. Since, for this analysis, we prefer to select variable regions that are more likely to form R-loops these metrics are also used to compare and rank candidate variable regions with highest ranking sequences having low mean local average energy and high average R-loop probability compared to all sequences generated with the same parameters.

Figure 5: Heatmaps showing mean R-loop probability (left) and mean local average energy per sequence (right) for 100 sequences of length 200 at various levels of both GC skew and content. As GC skew and GC content increase, the mean R-loop probability also increase while the mean local average energy tends to decrease which aligns with the basic expectation that increased GC skew and content favor R-loop formation.

Both local average energy and somewhat consequentially, mean R-loop probability are sensitive to both the GC skew and content of the input sequence (fig 5). Comparing between sequences generated with the same parameters and therefore GC skew and content ensures that we are mainly evaluating how the arrangement of a particular set of nucleotides is predicted in generally effect the likelihood of R-loop formation.

RNA secondary structure Significant amounts of RNA secondary structure, especially large hairpins, can be expected to reduce the likelihood of R-loop formation by causing competition for binding to the nascent RNA strand between itself and the DNA template. To avoid selecting sequences that are predicted to form an anomalous degree of RNA secondary structure the SPOT-RNA in conjunction with bpRNA structure annotation [3, 4] were used to calculate the estimated proportion of ribonucleotides in a given transcribed variable region that would participate in hairpin structures (PH) and the estimated proportion of that would remain unpaired (PUP). Candidate variable regions with low PH and high PUP were preferred.

Figure 6: Sequence of the 3' homology arm and EcoRI site. The complete sequence is highlighted in purple.

Need to update the above with same datapoints as R-looper

2.1.3 EcoRI site and 3' homology arm

The final 20 nucleotides of each insert will be composed of a EcoRI recognition site (6 bp) and the 3' homology arm (24 bp).

Figure 7: Sequence of the 3' homology arm and EcoRI site. The complete sequence is highlighted in purple.

Both the 5' homology arm + KpnI site and EcoRI site + 3' homology arm were produced using the jupyter notebook available at this link and the final sequences are available in genbank format at this link.

3 Assembly of DNA inserts

Complete insert sequences will be cloned into three different plasmid backbones: pFC9 (fig 9), pFC8 (fig 9), and pFC53tacT_1T_2 (fig 10). pFC9 will be utilized for testing R-loop initiation, pFC8 for R-loop termination and pFC53tacT_1T_2 for multiple round and single-round transcription versions of both initiation and termination experiments with Tac polymerase.

Figure 8: Map of pFC9.

Figure 9: Map of pFC8.

Figure 10: Map of pFC53tacT_1T_2.

3.1 T7 initiation series constructs

Figure 11: Diagram of pFC9 insertion series cloning strategy.

First, pFC9 will be linearized by digestion with KpnI and EcoRI (fig 11A). The large pFC9 fragment will then be purified and added to a mixture containing all insert sequences in equal concentration (fig 11B). Next, T5 exonuclease is added to digest the 5' ends of all DNA in the mixture. This will leave the 3' overhang of the digested KpnI site intact but degrade the 5' overhang of EcoRI (fig 11C). The 5' ends of the insert will also be degraded exposing exposing the complete KnpI and EcoRI sites present in each insert. Next, during Gibson assembly the 5' homology arm and KnpI site will anneal to the pFC9 large fragment, overhanging the digested KnpI site by 1 nucleotide; a C. Similarly, the 3' homology arm and intact EcoRI site of the inserts will anneal to the pFC9 large fragment, with only the last nucleotide (C) of the insert's intact EcoRI site annealing to the 3' G present at the digested EcoRI site of the pFC9 large fragment (fig 11D). This will result in a library of circular constructs with all inserts located downstream of and oriented forward relative to, the pFC9 T7 promoter (fig 11).

Figure 12: Plasmid map of produced by simulating the T7 initiation series cloning protocol for the insert containing variable region 1. Cloning simulations were ran and confirmed to be successful for all complete inserts sequences.

3.2 T7 termination series constructs

After the successful sequencing of the T7 initiation series, pFC8 will be utilized as the backbone for construction of the termination series library.

Figure 13: Diagram of pFC9 insertion series cloning strategy.

First, the strongest and most consistent R-loop initiator identified from the T7 initiation series will serve as the substrate for an additional synthetic double stranded DNA fragment containing the strong initiator sequence flanked by HindIII and EcoRI recognition sequences (13A). This strong initiator fragment will be added to pFC8 (13B) and then the mixture digested with HindIII and EcoRI. The strong initiator will then anneal to the large pFC8 fragment via homology between the digested HindIII and EcoRI recognition sites (13C). Next all termination inserts will be added to the pFC8-strong-initiator construct in equal concentrations (13D) and the mixture digested with KpnI and EcoRI (13E). Inserts will then be incorporated into pFC8-strong-initiator constructs via homology between the digested KpnI and EcoRI sites (13F). Since the order of these recognition sites on the pFC8-strong-initiator construct is opposite to that of the insert with respect to the T7 promoter the inserts will be present in the final construct in the reverse orientation and place the variable region downstream of the strong initiator (13G).

Figure 14: Plasmid map of produced by simulating the T7 termination series cloning protocol for the insert containing variable region 1. Cloning simulations were ran and confirmed to be successful for all complete inserts sequences. Note that since the strong initiation sequence that was used is to be determined a simulated but functionally equivalent sequence was used in its place and is not shown in the map image.

3.3 Tac initiation series constructs

Figure 15: Diagram of pFC9 insertion series cloning strategy.

First primers two pairs of primers are used to add amplify the anchor region, variable region, EcoRI recognition site and extension region from the T7 initiation construct library (3.3). These primers will also contain 15 bp overhangs with homology to the 5' end of the tac promoter and 3' T1T2 terminator sequences. Next, the PCR products are added to pFC53T1T2

3.3.1 Tac initiation series primer design

3.4 Tac termination series constructs

Figure 16: Diagram of pFC9 insertion series cloning strategy.

3.4.1 Tac termination series primer design

3.5 Validation of insert libraries

Use qPCR with primers for subset of variable regions with divergent characteristics should be present in equal quantities or just use primers for everything but then need to design unique primers or at least one primer that ends in the variable region so is specific to each insert.

3.6 Quantification of R-loop formation

Short description of SMRF-seq protocol using barcoded PCR primers for amplification and anticipated data analysis.

4 Sequence availability

Final versions of all complete insert sequences in fasta and genbank format are available at this link. For convenience, a copy of the fasta formatted file is also included in this document in section 6.4.

5 Code availability

All code used in insert creation and analysis is available here. The source code for this document is available at the GitHub repository at this link.

6 Supplementary materials

6.1 Variable region snakemake pipeline diagram

Figure 17: Place holder dag image.

6.2 Initial R-looper expectation calculations

Originally, expectations for R-looper results were derived by measuring average per base pair probability and of R-loop formation average per base pair local average energy calculated by rlooper for a large number of random sequences of a given length.

Figure 18: Initial rlooper expectation plots based on completely random sequences.

The top figure shows the distribution of average R-loop energy for a given base pair, averaged over all R-loops that contain that base pair for random sequences of a given length.

Bottom plot

Thanks to conversations with Craig and Robert about this also decided using parameterized sequences would be more informative resulting in fig 1.

6.3 Homology arm calculations

Jupyter notebook that includes all code used to generate homology arms and insert restriction sites is available at this link. The genbank formatted files of both homology arms can be found in the same repository in the resources directory.

6.4 Variable regions in fasta format

Headers for each complete insert begin with the SEGUID string (A SEquence Globally Unique IDentifier) [5] for the sequence followed by a space, the word insert, an underscore and finally the insert name.

```
>uI2oMODk-22mA4uMFkAWb8MNQEQ insert_init-1
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCATTAACTATATATGTA
```

TCCTTGCACACCCTTATCTAAATTGTCTTAGATTTTCAATCCTATAGTGTAGGTGTGGCA
GATGCAACTTTACAGGCGCGAGTTGGTACTACACGCAGACTAGTAATAGTGCAGTTTAGA
GAAGGGAAGAGGTAAATCGTCTAGTAAAACTAGGACTCTGTATAAGTTACCAACGGTGTT
ACCGCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>xxGE8Gg_FuXoRcMCzcbS8H6P8wk insert_init-2
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCACCTTCGCCGAATGGG
AAGCGGGGCAGGGGCCGGAGTCTAACCGAAGAGGGACAAGGGTCGACGCCGCTGGACGCG
GGGTAGGTTGGTTGTGGTCAGAGGGGCGGCGTGCGGGGGGCGGTGGCAACGTGGTATCGT
CGTCATGGACGTGACGCGCTGGGGGGCGATACACGGGGGGGGGGTGGCTTCAGAGTGCGG
TGGCCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>QgaasmOlhlPX5b3t-z1qbJu3S9o insert_init-3
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAATTATGGGGCTGGCT
ACGGAGGTGTAATTTCTAGGGAGGGGGAGCATTAGACGGTGGCAGGTTAGTGAGGGGGGG
GGACAGTCGAGCAGAGGCCGCGTATATGGGCGGGATGGTGGATGGACAAGGGAAGGTGCC
ACTCGAGGGTGGGTGTAAGGGGTGGTAGTTAGGAGGACTGGATTTGGCTGCGGTAGGGTG
GCACGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>HZfblNGgM64FV1hPYV8cTLZ5uOM insert_init-4
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCATGTGGTGCTGCGGCT
TGGGGAGGCGTCTTGTCGAGTGTCGATCGACAGGCGGTGGCCTGGGGGATCCGACTTCCG
GCTCCTTGGGGCCGCGCGAAGGGGGACATTCGGAGACACAGGGCCCGCTACCCCGCAGCG
AGCGGGGCGAAGCTGACCACGCAGCTCCCCCGTGCGACACGATGGCCTCCAGCGAGATCC
GGACCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>M3VLqJl5oQIRO7wZeEKKr_hUr4O insert_init-5
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAACGCGGGATTGCCGG
ACATCGGAGAGCCCAACGAGCGGAATTTCCGTGACGAGGTCAGAGCTGATGGAACTCCCT
CAGTTACTCAGGTATGAACGTACTCAAAGGCCCTGTGTCGTGGGTGCACCGCTCCCTGGC
AGGGCTATTCCGCCGATACGCAGTCATCAGGGGACCGCGTGTTGCTACAACTGCCCCGTC
TCGTCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>Cyy1TuR8girckLFRFzzJQtffY78 insert_init-6
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCACACATTAGAGTACGT
AGAGGCACTGCATCCTGGGCGGTTTGAATTAGGTTAACTCGAGGGTAACTTGTACTTAAC
ATAGAGGAGAGGCATTCCACGCTTGGTCTCTAAAACTGTTTAGTATTGGGTAAACCCGAA
ATCTAATTTCAACGAGCTCTATAGGAAATAATTAATGGAAGATAGTGGTGTTATCGTTAC
ATTTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>nnPuzS7AblFb1Bw2eAE_7h8SePY insert_init-7
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAATTATGGGGCTGGCT
ACGGAGCTGTAATTTCTAGCCAGGGGGAGCATTAGACGGTGGCAGGTTACTGAGGGGGCG
CGACAGTCCAGCAGAGGCCGCGTATATGGGCGGGATGGTGGATGGACAAGGGAACGTGCC
ACTCGAGGCTGGGTGTAAGCGGTGGTACTTAGGAGGACTGGATTTGGCTGCGGTAGGCTG
GCACGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>48li-aMEdMWrqlvMwBVGdXkIoic insert_init-8
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAAGCTAGAGTGTAGAT
GATCTAAGAATAGTATGGAGGGATTTAAGTGAGAGTTCGTAGAGGGAACTGCAGTGGCAG
GTGGGATGGAAATGTGAGCGTAGGTGTCTGCTTGGTAATAGGCTGGATCTGCAGGGTCGG
TGTCGGTGGGACTATGGGAGGGGTAATAGGCGCAAGCTGGGGTGATGCATAAGCGTTAAT
AGTTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>k1DC5eUm6s6LHQ6IBWL3fh21sOU insert_init-9
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAACGCGGGATTGCCGG
ACATCGGAGAGCCCAACGAGCGGAATTTCCGTGACGAGGTCAGAGCTGATGGAACTCCCT
CAGTTACTCAGGTATGAACGTACTCAAAGGCCCTGTGTCGTGGGTGCACCGGTCCCTGGC
AGGGCTATTCCGCCGATAGGCAGTCATGAGGGGACCGCGTGTTGCTACAACTGCCGGGTC
TCGTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>PtfE-xWOCu1CcZAF_GmNkRvARG8 insert_init-10

TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAACCTAGAGTCTACAT
CATCTAACAATACTATGGAGCGATTTAACTGAGAGTTCGTAGACCGAACTGCACTGGCAC
GTGGGATGCAAATGTGACCGTAGGTGTCTGCTTGCTAATAGGCTGGATCTGCAGGGTCGC
TGTCGCTCCCACTATCGGACGGGTAATAGGCGCAAGCTCGCCTCATCCATAACCCTTAAT
AGTTCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>ZfSsyr5JKJMmzfxVoihBQwhX_NE insert_init-11
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCATGTGGTGCTGCGGCT
TGGGGAGGGGTCTTGTGGAGTGTCGATCGAGAGGCGGTGGCCTGGGGGATCCGACTTCCG
GCTCCTTGGGGCCGCGCGAAGGGGGACATTCGGAGACACAGGGCCCGCTACCCCGGAGCG
AGCGGGGCGAAGGTGACCACGCAGCTCCGCCGTGCGACACGATGGCCTGCAGCGAGATCC
GGACCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>ABOve2-vLjsIguhmvORXjxcgBfI insert_init-12
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAATTATGGGGCTGGCT
ACGGATCTTTAATTTCTAGGCATGGGTATCATTAGACGTAGTCAGGTAACTGAGGGTGGG
CGACAGACGAGCAGAGGCCTCGAATAAGGGCTGGATGTAGGATGGACAAGTGAAGTTTCC
ACTCGAGGGTGGGAGTAAGGGGTGTTACTAATGATTACTGGAAATGGCTGCTGTATGCTG
GCACGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>OajKeeoLFH_-_WR3FlsYUghJ_wo insert_init-13
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAACGCGGGATTGCCGG
ACATCGGAGAGCGCAACGAGCGGAATTTCCGTGACGAGGTCAGAGCTGATGGAAGTCCGT
CAGTTAGTCAGGTATGAACGTACTCAAAGGCCCTGTGTGGTGGGTGCACCGGTCCCTGGC
AGGGCTATTCCGCCGATAGGCAGTCATGAGGGGACCGCGTGTTGCTACAAGTGCCGGGTC
TCGTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>5vd9OwN61JjxUtv5yJRgi5W4rkQ insert_init-14
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCACTCATTACAGTACGT
TCACGCACTGCAGCCTCGGCGGTTGCAATTTGGTTAACTCGAGGGGAACTGCTACTTAAC
ATAGTGGAGACGCTTGCCACGCTGGCTCGCGATATCGGTGTAGTAGTGGCGAAACCCGAT
AGCTAATGTCAACGACCGCGTTACGTAATAAGGAATCGAAGATAGGGGTCTGATCGTTAC
ATTTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>vOi3nAJBZKQbAAixEIhLho5HRGO insert_init-15
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCATTAACTATATATGTA
TCCTTGCACACCCTTATCTAAATTGTCTTAGATTTTCAATCCTATAGTGTAGGTGTGGCA
GATGCAACTTTACAGGCGCGAGTTGCTACTACACCCAGACTAGTAATAGTGCAGTTTAGA
CAAGGGAAGAGGTAAATCGTCTACTAAAACTAGGACTCTGTATAAGTTACCAACGGTGTT
ACCGCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>F-umROlWXPNmCnU9huosc24irXk insert_init-16
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAAAATGGAGAGGTTAG
TTGGAATTTGGCGTAGGTGAAGGCAAAAATAGAAATCGGTATAATGTTTCAAGGATGTTA
ACAAGGTGTGAAAGAGGTCGGTGACAGAAGTCTTTAAATAGTTATGTCTCAAATTACTCA
GTAGTTTGAGGAGGTATTGTAGGACGTTATTCCGATAGGGTTTATGTGCGTGATGGGGTG
GTGAGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>nfcEqNf7lCB4XRpZ2RAu7O-sjyO insert_init-17
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCACGTTCGCGGAATGGG
AAGCGGGGGAGGGGGGGGAGTCTAAGCGAAGAGGGACAAGGGTCGACGCCGGTGGACGCG
GGGTAGGTTGGTTGTGGTGAGAGGGGCGGGGTGCGGGGGGGGGTGGCAAGGTGGTATCGT
CGTGATGGAGGTGACGCGCTGGGGGGCGATACACGGGGGGGGGTGGCTTCAGAGTGGGG
TGGCCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>xMIFrhuHyxOkfOAkHZv3Gt7avOM insert_init-18
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAACTCGGGCGCGGAAG
GCAGCCACCGGCCCGAGGCCGTGTCCCCCGGGGCTTGGGCGTACAACTGGACTTCCCGGG
CGGAGGTCCTTCTCGACCACATAGACCGCGGCGGGAGGTGGGACGCCTTCCGTTTTAGCG
TCGTGCGCCGTAGGGCGAGGAGCTCCCGTTCAGACATCCAGGCTCGCAGCACATGCCCCC
GACCGGAATTCGTCGCAGTGACCGAGGCGAGGAGG

>_bPA1dDaKHyM7wwFymFjj8iaOHA insert_init-19
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCACACATTAGAGTACGT
AGAGGCAGTGCATCCTGGGCGGTTTGAATTAGGTTAACTCGAGGGTAAGTTGTACTTAAC
ATAGAGGAGAGGCATTCGAGGCTTGGTCTGTAAAACTGTTTAGTATTGGGTAAAGCGGAA
ATCTAATTTCAACGAGCTCTATAGGAAATAATTAATGGAAGATAGTGGTGTTATCGTTAG
ATTTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>z-VWjJs4HOVPRiFTKkB2CMhOAeQ insert_init-20
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAATTATGGGGCTGGCT
ACGGATCTTTAATTTCTAGCCATCGGTATCATTAGACGTAGTCAGCTAACTGAGGGTGCG
CGACAGACCAGCAGAGGCCTCGAATAAGGGCTGCATGTAGGATGGACAAGTGAACTTTCC
ACTCGAGGCTGCGAGTAAGCGGTGTTACTAATGATTACTGGAAATGGCTGCTGTATGCTG
GCACGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>a1rQbSGsl2x7SCwWDrGsQcxG2Jo insert_init-51
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCACACATTAGAGTACGT
AGAGGCACTGCATCCTGGGCGGTTTGAATTAGGTTAACTCGAGGGTAACTTGTACTTAAC
ATAGAGGAGAGGCATTCCAGGCTTGGTCTGTAAAACTGTTTAGTATTGGGTAAAGCGGAA
ATCTAATTTCAACGAGCTCTATAGGAAATAATTAATGGAAGATAGTGGTGTTATCGTTAC
ATTTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>i5xBsEuOX_D3UAsMD5_D5RBKrRk insert_init-52
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAAAAGGGAGAGGTTTG
GGGGAATTTGGCCTACCTGAAGGCTATAATAGAAATCGGTAGTATGTTGCAAGGTTCTTT
ACAACGTGTGAAAGTGGTCCGGGACACATCTCGTTAAATTCTGATGTCGCAAAGGACGCA
GTAGGTGCACGAGCTATTGTAGGACCTGATGCCGTTACGGTTGATCTCCGTGAGCGGCGG
GGGACGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>Bl5NH1DmQHE_PhlW8RrNLnBKT24 insert_init-53
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAACGCGGGATTGCCGG
ACATGGGAGAGCGCAACGAGCGGAATTTCCGTGACGAGGTCAGAGCTGATGGAAGTCCGT
CAGTTAGTGAGGTATGAAGGTACTCAAAGGGCCTGTGTGGTGGGTGCACCGGTCCGTGGC
AGGGCTATTCCGCCGATAGGCAGTCATGAGGGGAGCGCGTGTTGCTACAAGTGCCGGGTC
TCGTGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>BceLUjEU9O8GsQVx9xW1Kxn6SUI insert_init-54
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAGCGCTGAACGGGGGA
GGGTCGGGGCGGGCGGCGTGAGAGCTCAGGCTCCGGGGTGTACGGGAGGTGCAGGGGGGC
GCATCCATGGGTAGGCGGGAGGCAAAGTATCGAAGGGCCCACAGATGGGTCGTTGTCCGT
GGAGTTCGCTGACCTCTGCACCCGCGGCTGAGCCGGGGGTCCCCGCACGGAGTTGGATGG
GTCCAGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>kSnZfaYyMmtaXNeKUrJYDrVc6is insert_init-55
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCATTAACTATATATGTA
TGGTTGCAGACCGTTATGTAAATTGTCTTAGATTTTCAATCGTATAGTGTAGGTGTGGCA
GATGGAACTTTAGAGGCGCGAGTTGGTACTACAGGGAGACTAGTAATAGTGGAGTTTAGA
GAAGGGAAGAGGTAAATGGTCTAGTAAAACTAGGACTGTGTATAAGTTACCAAGGGTGTT
ACGGGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>izktT5yxYm4tK4Ise9OtCjMWvoA insert_init-56
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAAGCTAGAGTGTAGAT
GATCTAAGAATAGTATGGAGGGATTTAACTGAGAGTTCGTAGAGGGAACTGCAGTGGCAC
GTGGGATGGAAATGTGAGCGTAGGTGTCTGCTTGGTAATAGGCTGGATCTGCAGGGTCGC
TGTCGGTGGGACTATCGGAGGGGTAATAGGCGCAAGCTGGGGTGATGCATAAGCGTTAAT
AGTTCGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>D_X_8FNcqgcekMkN6HwwXxxeSWo insert_init-57
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAAGTGGATCATGGTGT
TGGGTATTGTTCTTGAGGAGTGAGGATCGAGAGGGGGTGGGCTGGGGGAAGCGACAACCG
GCTGGAAGTGTCGGGGGGAAGGGGGAGATTCTGAGAGAGAGTGGGCTCTACCGCGGAGCG
AGGGTGGCTAAGGTGACCACTGATGTGCGCCGTGGTACACTATGTCGTGGATGGAGATCC

```
GGACGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
>LrNW4QmsWQQFWxdEJX_k9mBQBCo insert_init-58
TAGGGCGAATTGGAGCTCCACATCGGTACCCACGTTTGGCCACCAATGTGGGGGCGGGCT
TCGGAGCGGTTAGGGCGAGCCAGGGGGAGCAGGTGACGGTGGCAGGGTTCGGAGGGGGCG
CGTCAGTCCAGCAGTGGCCGCGTTGATGGGCGGGATGGTGGATGGACATGGGAAGGGGCC
TCTCGAGGCGGGGTGTAAGCGGGGGGACGTAGGAGGACTGGATTGGGCGGCGGTAGGCTG
GCACGGAATTCGTCGCAGTGACCGAGGCGAGGAGG
```

6.5 Complete inserts in fasta format

6.6 RNA secondary structure expectation plots

References

[1] Frédéric Chédin. Nascent connections: R-loops and chromatin patterning. *Trends in genetics : TIG*, 32(12):828–838, December 2016.

[2] Robert Stolz, Shaheen Sulthana, Stella R. Hartono, Maika Malig, Craig J. Benham, and Frederic Chedin. Interplay between DNA sequence and negative superhelicity drives r-loop structures. *Proceedings of the National Academy of Sciences*, 116(13):6260–6269, March 2019.

[3] Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1), November 2019.

[4] Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research*, 46(11):5381–5394, May 2018.

[5] György Babnigg and Carol S. Giometti. A database of unique protein sequence identifiers for proteome studies. *PROTEOMICS*, 6(16):4514–4522, August 2006.