Variable region design and cloning

Ethan Holleman

July 7, 2021

Abstract

Abstract text here

Contents

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].

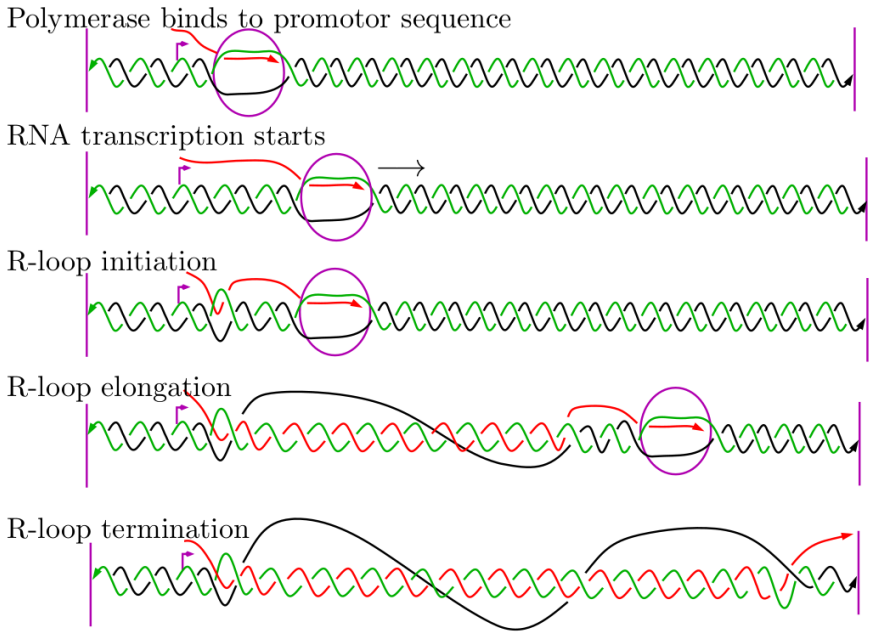

Figure 1: Stages of R-loop formation. The encircled region represents the transcription bubble. The polymerase moves from left to right, starting at the promoter region (purple arrow). The upper strand (in red) in the bubble corresponds to the nascent RNA transcript, generated in the 5' to 3' direction.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

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 sequences on R-loop formation *in vitro*. Inserts are composed of two types of DNA components, the variable region and flanking infrastructure regions. The variable region contains the sequence we are interested in observing R-loop dynamics over. Infrastructure sequences are additional nucleotide blocks that allow for the insertion of variable regions into specific plasmid backbones in a modular fashion.

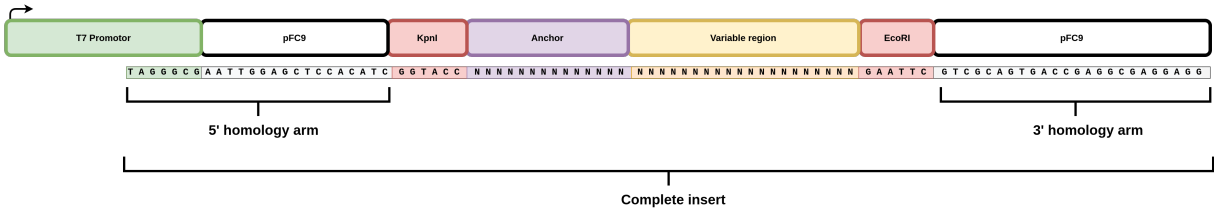

Figure 2: Diagram of a complete insert. Colored boxes represent features insert sequences are derived from while colored sequences represent the actual nucleotide sequence.

From right to left (5' to 3') each insert will contain a 5' homology arm with complementary to the last 7 nucleotides of the T7 promoter, 17 nucleotides complementary to the pFC9 plasmid immediately downstream of the T7 promoter for a total of 24 bp. This will be followed by a KpnI recognition site, and an "anchor" region will will be composed of a constant sequence which can be utilized for targeting by PCR primers. The following region will contain the variable region which defines the identity of each complete insert. Finally a EcoRI recognition

site and 24 taken from the region downstream of pFC9's EcoRI recognition site will be included. This design will allow for insertion of variable regions in either forward or reverse orientations using combinations of Gibson and restriction enzyme cloning, as well as allowing for later extraction via PCR or restriction enzymes. Specific methodolgies are discussed in greater detail in section **??**. Each insert will be completely synthesized as double strand DNA from an outside company and therefore require no additional assembly.

2.1 Insert components

2.1.1 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 pFC9 respectively. These 30 bp are taken directly from nucleotides 27 - 56 of pFC9.

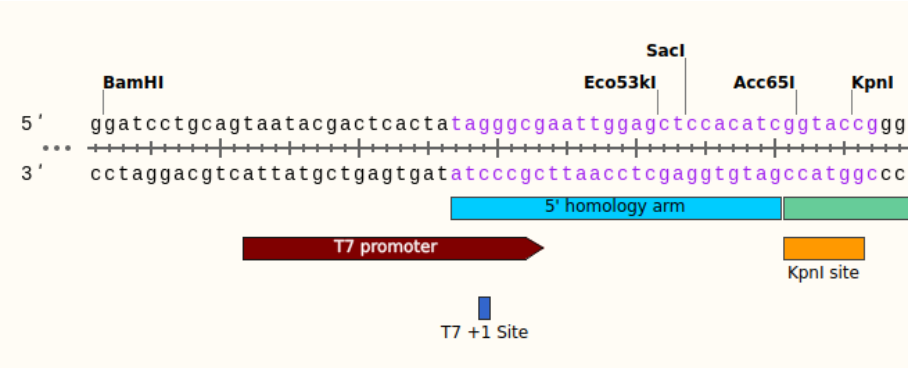

Figure 3: Location in pFC9 from which the 5' homology arm and KpnI site are modeled after. The entire 30 bp sequence is shown in purple with the 5' homology arm highlighted as a feature in blue and the KnpI site in orange.

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

2.1.2 Anchor

The anchor region serves as a constant sequence that will always be present in any final assembled construct adjacent to the 5' end of the variable region. Having this region within the insertion sequence means for a given plasmid backbone, at most 1 pair of primers will be required to amplify any sequence downstream of the 3' end of the anchor. This is useful when working with libraries containing all insertion sequences as each unique insertion will not require its own primer pair to amplify. 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 is selected from candidate random sequences that satisfy the requirements below.

- Does not contain any common sub-strings of minimun 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℃.

2.1.3 Variable regions

Each insert will contain 1 variable region which will be designed to test the effects of different sequence properties on R-loop dynamics, namely GC / AT skew, GC content and G / C clustering. Each variable region will be 200 bp in length. Twenty nine different variable regions with the properties described in table 1 will be generated using

the variable region design workflow. Each variable region, as part of the larger insertion sequence, will be cloned downstream of a promoter so it is transcribed in the forward direction.

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

A subset of the variable regions will also be cloned further downstream of the same promoter species and oriented so the reverse complement of the sequence is transcribed in order to access these regions capacity for R-loop termination. The properties of the transcripts of these sequences are listed in table 2.

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

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 initiated R-loops in order to study their downstream termination.

Table 3

Total synthesized inserts	Total contructs
29	47

While the global sequence properties for each variable region are well defined, properties such as GC-skew, content or clustering do not determine what nucleotide should occur at position n 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 one specific sequence from this set we can sample a large number of sequences and access each one with metrics relevant to the realities of the cloning protocols and R-loop formation.

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 therefore 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 potential variable regions containing any of these recognition sites were thrown out.

Table 4

Enzyme	Recognition sequence
KnpI	GGTACC
EcoRI	GAATTC
HindIII	AAGCTT

Predicted R-loop probability The Chedin lab has previously devloped and the work of Dr. Robert has devloped R-looper here we are using it to access where in distrabution sequences tend to fall using predictions for average probility of R-loop formation across the length of a sequence and the mean local average energy.

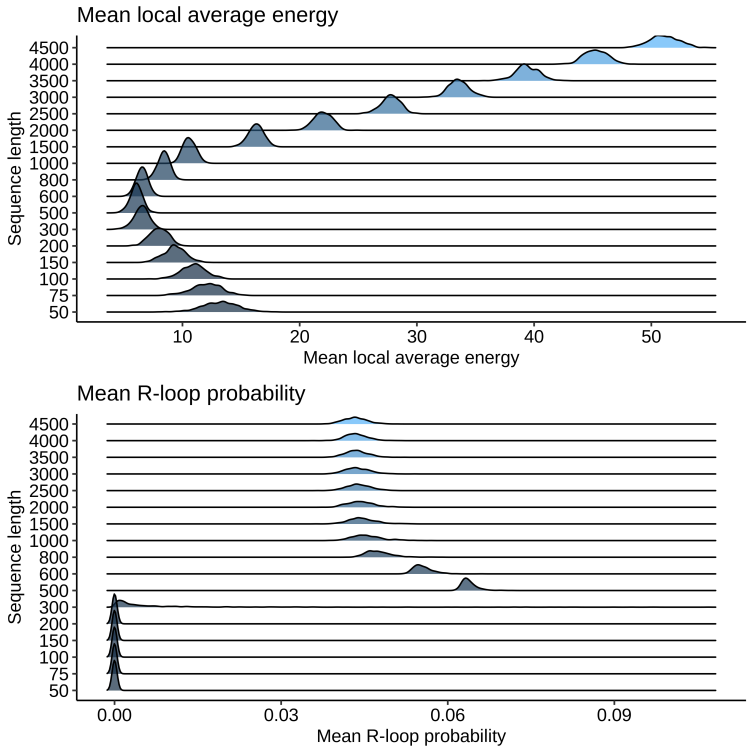

Figure 4: Map of pFC9 showing major features and all restriction enzyme recognition sites.

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.

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).

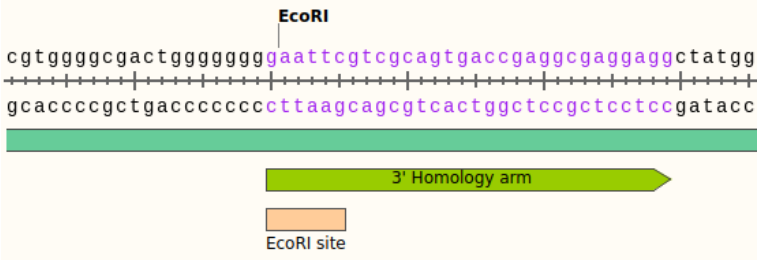

Figure 5

3 Assembly of DNA inserts

Complete insert sequences will be cloned into three different plasmid backbones: pFC9 (fig 7), pFC8 (fig 7), and pFC53tacT_1T_2 (fig 8). 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.

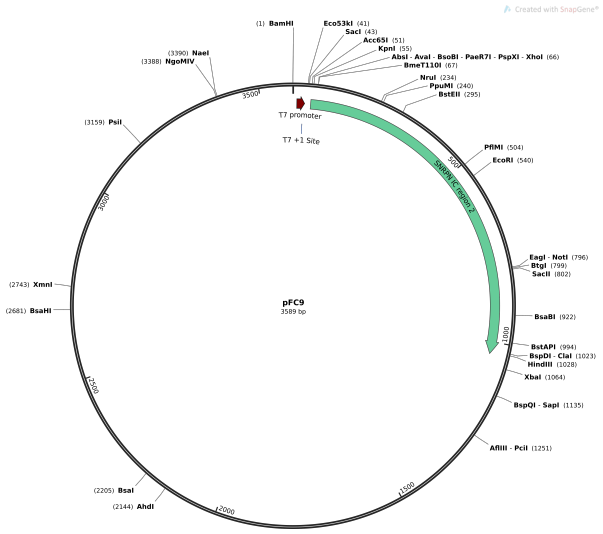

Figure 6: Map of pFC9.

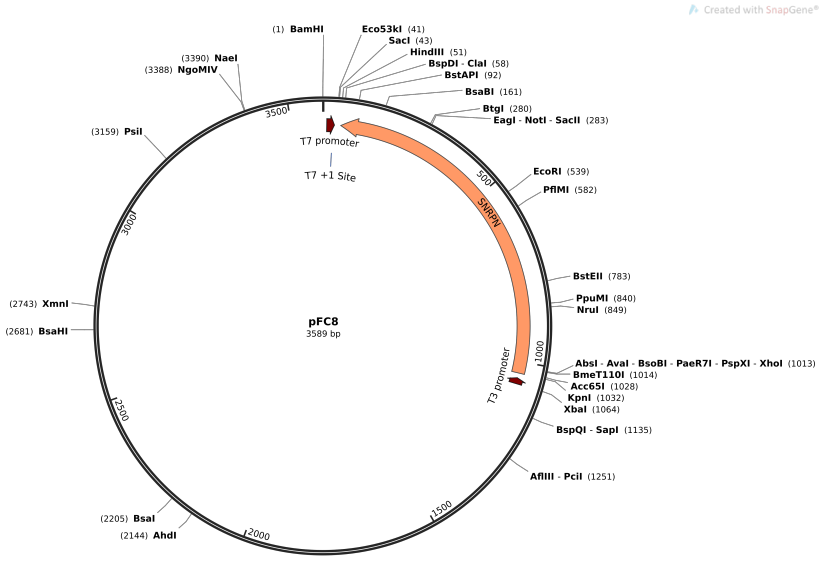

Figure 7: Map of pFC8.

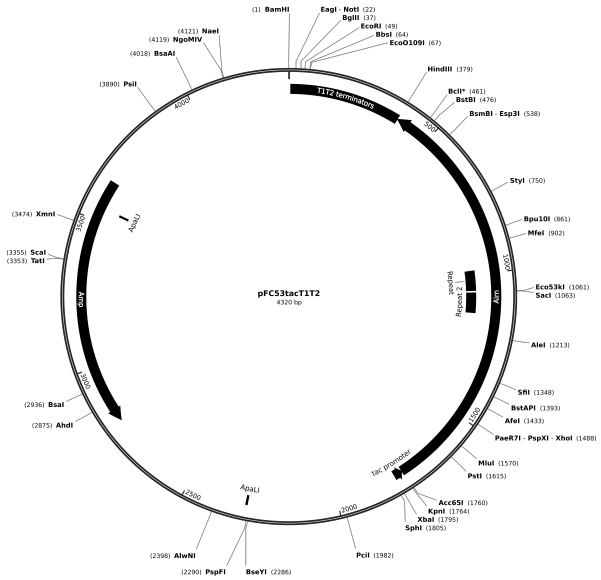

Figure 8: Map of pFC53tacT_1T_2.

3.1 T7 initiation series constructs

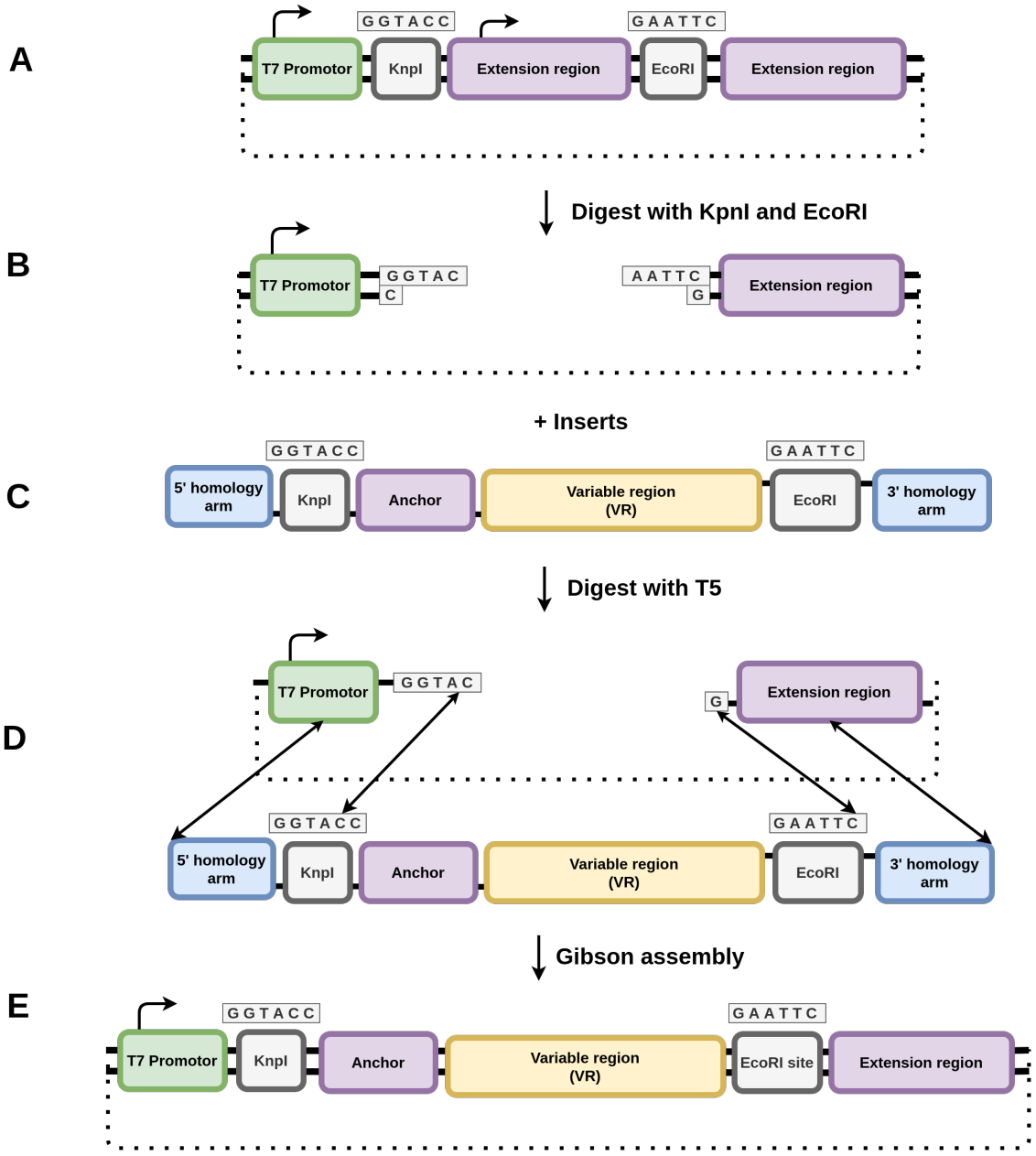

Figure 9: Diagram of pFC9 insertion series cloning strategy.

First, pFC9 will be linearized by digestion with KpnI and EcoRI (fig 9A). The large pFC9 fragment will then be purified and added to a mixture containing all insert sequences in equal concentration (fig 9B). 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 9C). 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 9D). 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 9).

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.

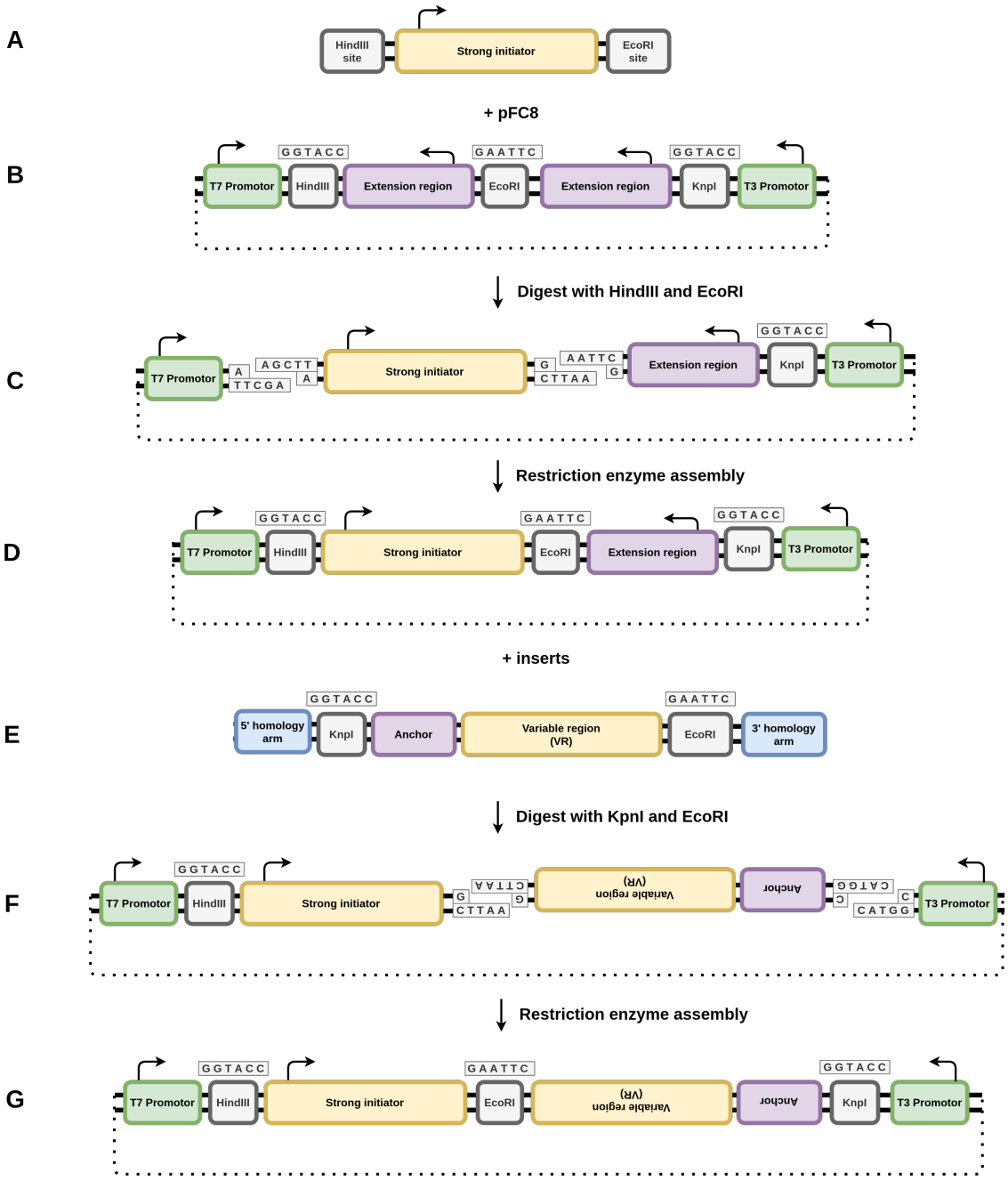

Figure 10: Diagram of pFC9 insertion series cloning strategy.

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 (10A). This strong initiator fragment will be added to pFC8 (10B) 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 (10C). Next all termination inserts will be added to the pFC8-strong-initiator construct in equal concentrations (10D) and the mixture digested with KpnI and EcoRI (10E). Inserts will then be incorporated into pFC8-strong-initiator constructs via homology between the digested KpnI and EcoRI sites (10F). 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 (10G).

3.3 Tac initiation series constructs

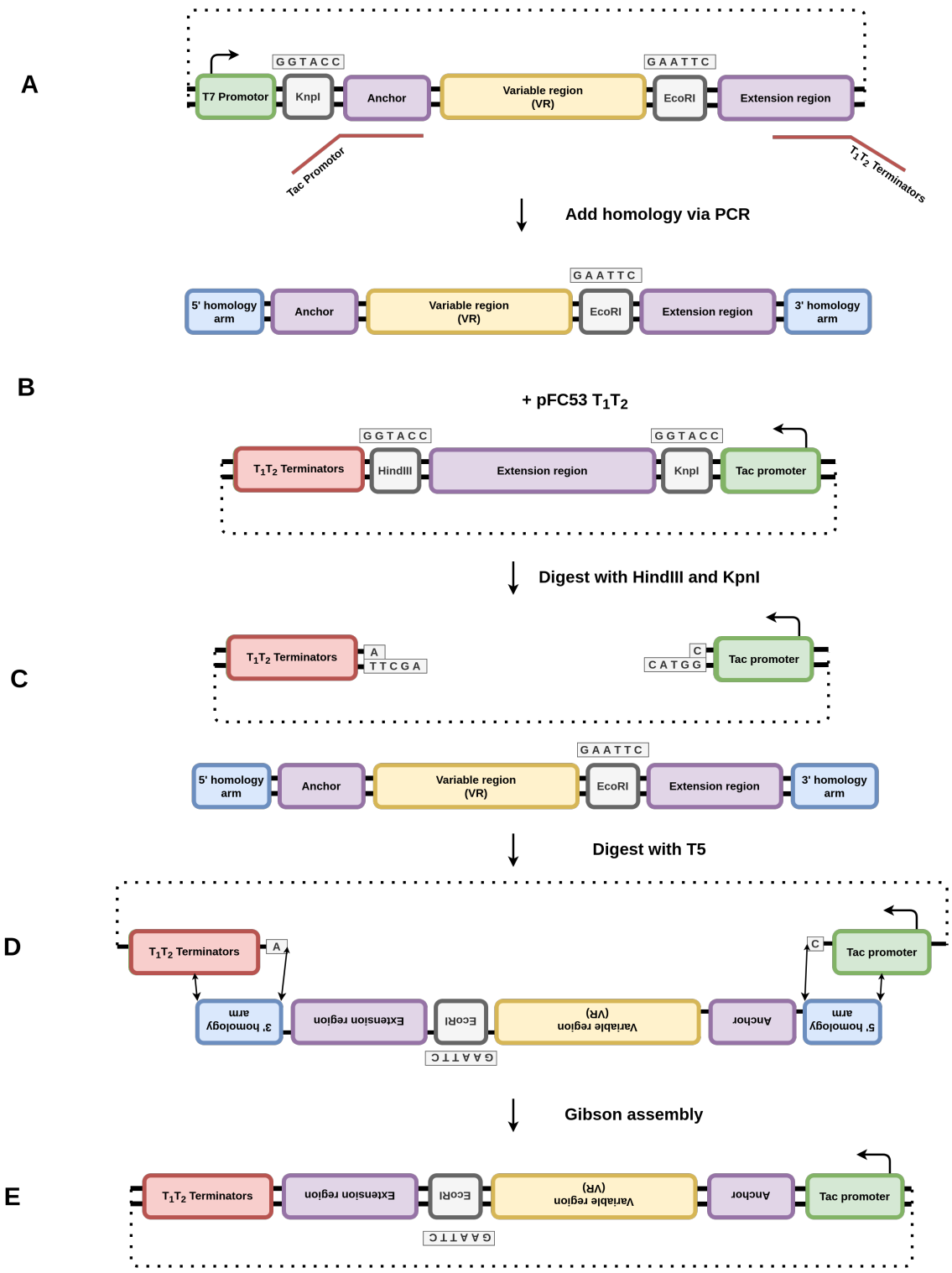

Figure 11: Diagram of pFC9 insertion series cloning strategy.

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

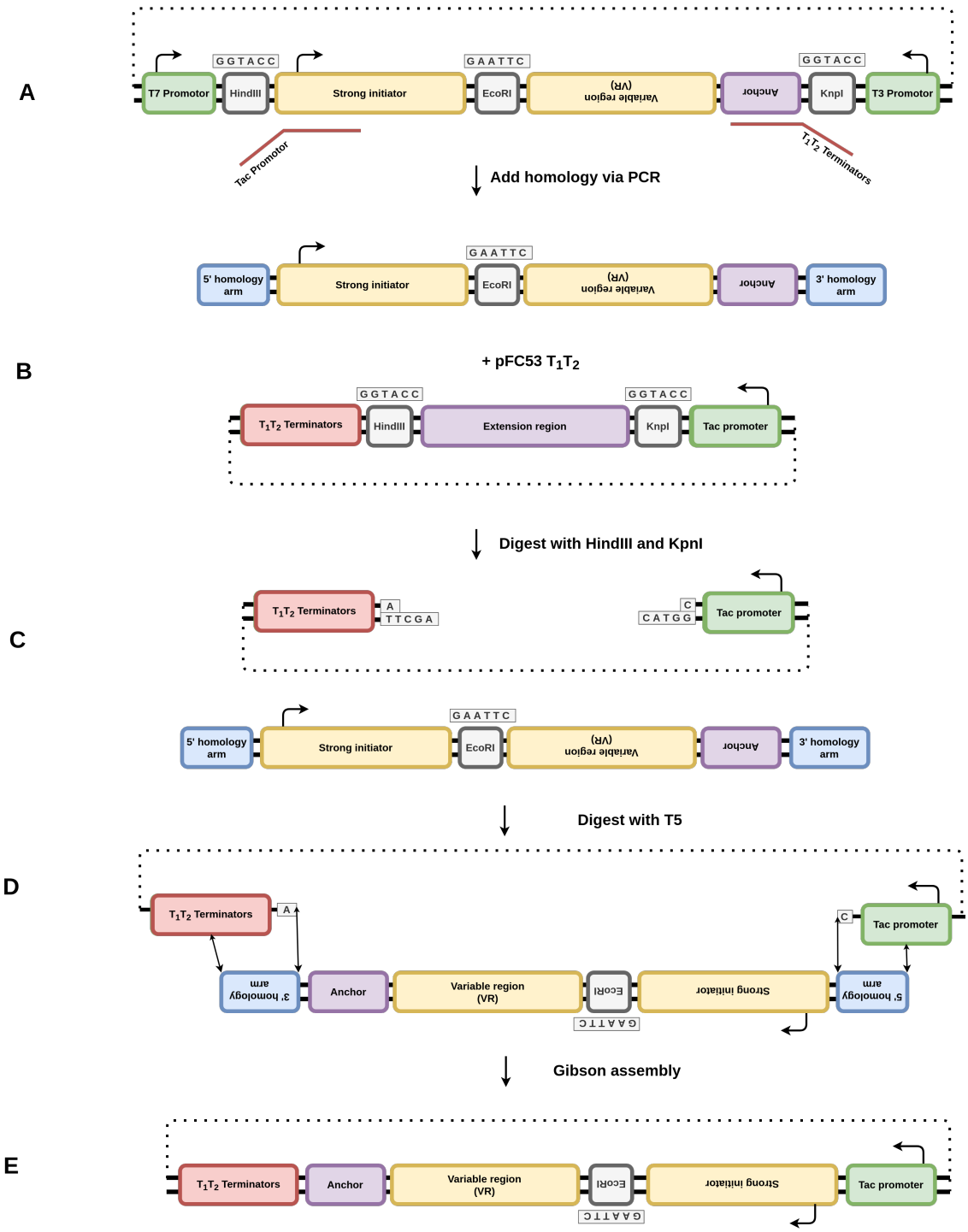

Figure 12: Diagram of pFC9 insertion series cloning strategy.

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 Code availability

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

5 Supplementary materials

5.1 Variable region snakemake pipeline diagram

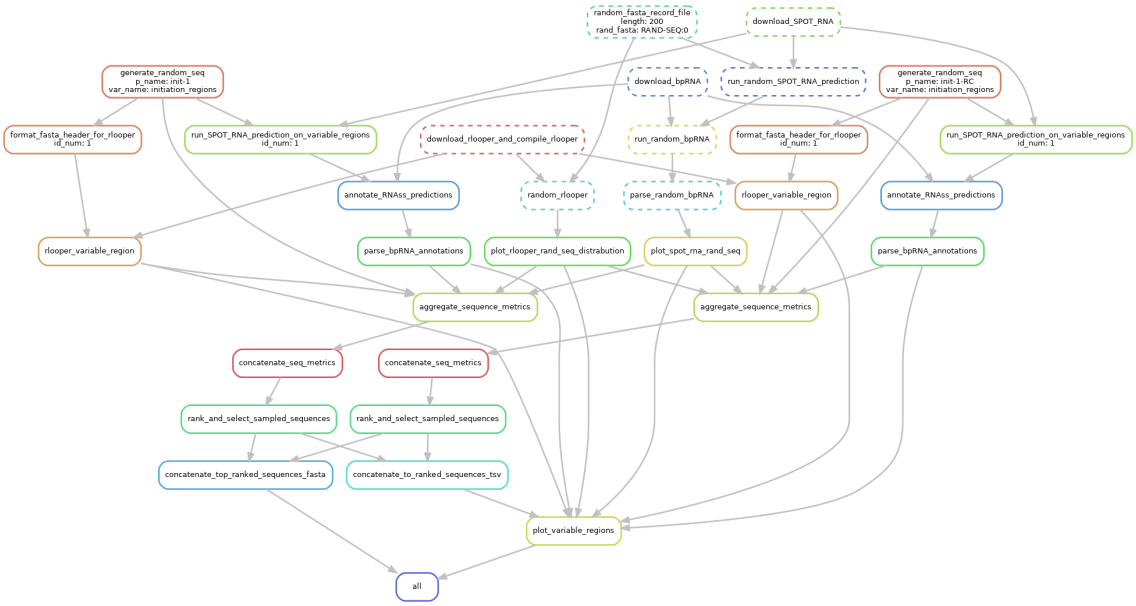

Figure 13: Place holder dag image.

5.2 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.

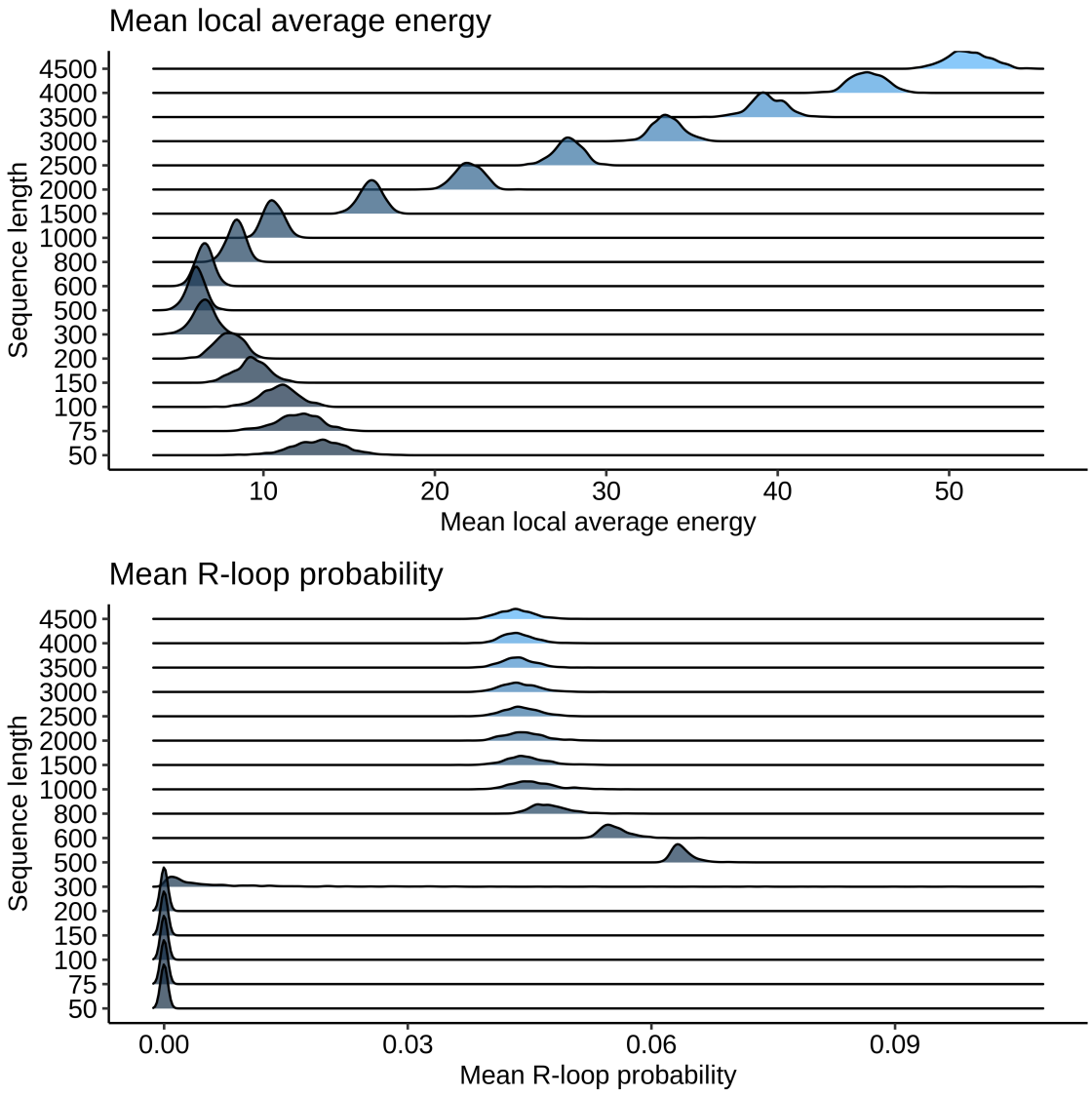

Figure 14: Initial rlooper expectation plots based on completely random sequences.

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.

5.3 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.

5.4 Variable regions in fasta format

```
>VR_init-1-5_initiation_region_1_GCskew:0.1_GCcontent:0.4_ATskew:0_ATcontent:0.6_Clustered:False
TTAACTATATATGTATCCTTCCACACCCTTATCTAAATTCTCTTAGATTTTCAATCCTATAGTGTAGCTGTGGCAGATG
CAACTTTACAGGCGCGAGTTGCTACTACACCCAGACTAGTAATAGTCCAGTTTAGACAAGGGAAGAGGTAAATCGTCTA
CTAAAACTAGGACTCTGTATAAGTTACCAACGGTGTTACCGC
>VR_init-1-5_initiation_region_1_GCskew:0.1_GCcontent:0.4_ATskew:0_ATcontent:0.6_Clustered:False
```

TTAACTATATATGTATCCTTCCACACCCTTATCTAAATTCTCTTAGATTTTCAATCCTATAGTGTAGCTGTGGCAGATG
CAACTTTACAGGCGCGAGTTGCTACTACACCCAGACTAGTAATAGTCCAGTTTAGACAAGGGAAGAGGTAAATCGTCTA
CTAAAACTAGGACTCTGTATAAGTTACCAACGGTGTTACCGC
>VR_init-1-5_initiation_region_1_GCskew:0.1_GCcontent:0.4_ATskew:0_ATcontent:0.6_Clustered:False
TTAACTATATATGTATCCTTCCACACCCTTATCTAAATTCTCTTAGATTTTCAATCCTATAGTGTAGCTGTGGCAGATG
CAACTTTACAGGCGCGAGTTGCTACTACACCCAGACTAGTAATAGTCCAGTTTAGACAAGGGAAGAGGTAAATCGTCTA
CTAAAACTAGGACTCTGTATAAGTTACCAACGGTGTTACCGC

5.5 Complete inserts in fasta format

5.6 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.