Variable region insert design and cloning strategies

Ethan Holleman

July 30, 2021

Contents

1 Background **2**

2 Insert design **2**
2.1 5' homology arm and KpnI recognition site . . . 3
2.1.1 Anchor region . . . 3
2.1.2 Variable regions . . . 4
Restriction enzyme recognition sites . . . 7
Predicted R-loop probability . . . 7
RNA secondary structure . . . 8
2.1.3 EcoRI site and 3' homology arm . . . 9

3 Assembly of DNA inserts **9**
3.1 T7 initiation series constructs . . . 12
3.1.1 Assembly protocol . . . 13
3.2 T7 termination series constructs . . . 13
3.3 Tac initiation series constructs . . . 16
3.3.1 Tac initiation series primer design . . . 17
3.4 Tac termination series constructs . . . 18
3.4.1 Tac termination series primer design . . . 20
3.5 Validation of insert libraries . . . 21
3.6 Quantification of R-loop formation . . . 21
3.7 Reagents and kits . . . 21
3.8 Additional resources . . . 21

4 Code availability **21**

5 Sequence availability **21**

6 Supplementary materials **22**
6.1 NEB Restriction enzyme activities . . . 22
6.2 Snakemake workflow diagram . . . 22
6.3 Initial R-looper expectation calculations . . . 23

1 Background

R-loops are prevalent, functionally 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 large fractions of the genomes in a cell population.

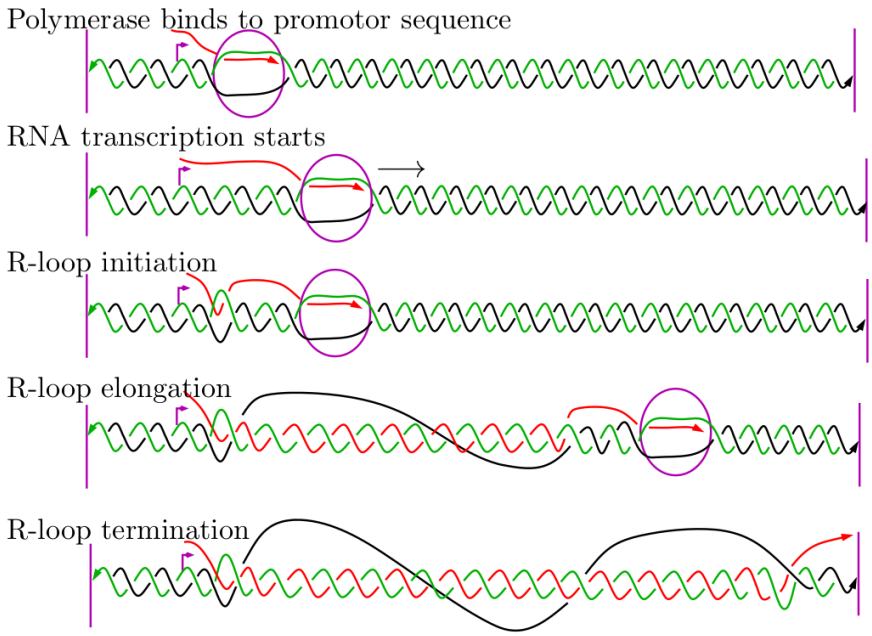


Figure 1: Stages of R-loop formation. The encircled region represents the transcription bubble. The polymerase moves from left to right, starting at the promoter region (purple arrow). The upper strand (in red) in the bubble corresponds to the nascent RNA transcript, generated in the 5' to 3' direction. 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. Generally, their formation is favored by regions with high GC skew (overabundance of G over C nucleotides) and GC content (percentage of G and C nucleotides) as the potential DNA-RNA hybrids formed over these sequences tend to be more energetically favorable with respect to R-loop formation[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 knowledge gap that continues to widen as the recognized scope of these structure's impact on cellular functions continues to grow. Here, I describe a detailed strategy for the systematic investigation of the effects 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 infrastructure regions. The variable region is the sequence of interest with respect to R-loop formation. In contrast, infrastructure sequences are additional nucleotide blocks that allow for the insertion of variable regions into specific plasmid backbones in a modular fashion. The distinct regions that compose each insert are described below.

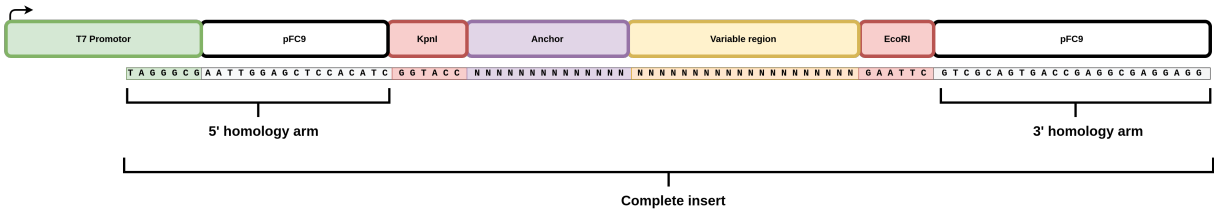


Figure 2: Diagram of a complete insert. Colored boxes represent features insert sequences are derived from while colored sequences represent the actual nucleotide sequence.

2.1 5' homology arm and KpnI recognition site

The first 30 bp of each insert will be composed of the 5' homology arm followed by SacI and KpnI recognition sites. These sequences are included to facilitate Gibson and restriction enzyme cloning into pFC9 and pFC8, respectively.

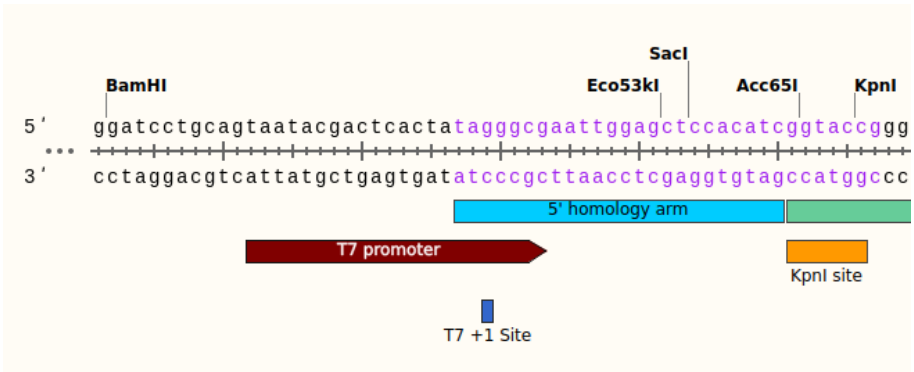


Figure 3: Location in pFC9 from which the 5' homology arm and KpnI site are modeled after. The entire 30 bp sequence is shown in purple with the 5' homology arm highlighted as a feature in blue and the KnpI site in orange.

This strategy will mean recognition sites for Eco53KI, SacI, and Acc65I are 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 region

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 long as the anchor region is present and intact, the number of primers required to amplify the insert region will not be dependent on 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 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 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 3).
- Does not contain single stretches of single nucleotides greater than three nucleotides in length.

name	GC skew	GC content	AT skew	G cluster length
VR-1	0.0	0.4	0.0	NA
VR-2	0.0	0.5	0.0	NA
VR-3	0.0	0.6	0.0	NA
VR-4	0.0	0.7	0.0	NA
VR-5	0.1	0.4	0.0	NA
VR-6	0.1	0.5	0.0	NA
VR-7	0.1	0.6	0.0	NA
VR-8	0.1	0.7	0.0	NA
VR-9	0.2	0.4	0.0	NA
VR-10	0.2	0.5	0.0	NA
VR-11	0.2	0.6	0.0	NA
VR-12	0.2	0.7	0.0	NA
VR-13	0.4	0.4	0.0	NA
VR-14	0.4	0.5	0.0	NA
VR-15	0.4	0.6	0.0	2
VR-16	0.4	0.6	0.0	3
VR-17	0.4	0.6	0.0	4
VR-18	0.4	0.6	0.0	NA
VR-19	0.4	0.6	0.2	2
VR-20	0.4	0.6	0.2	3
VR-21	0.4	0.6	0.2	4
VR-22	0.4	0.6	0.2	NA
VR-23	0.4	0.6	0.4	2
VR-24	0.4	0.6	0.4	3
VR-25	0.4	0.6	0.4	4
VR-26	0.4	0.6	0.4	NA
VR-27	0.4	0.7	0.0	NA
VR-28	0.6	0.4	0.0	NA
VR-29	0.6	0.5	0.0	NA
VR-30	0.6	0.6	0.0	NA
VR-31	0.6	0.7	0.0	NA

Table 1: Properties of all variable regions. NA in the G cluster length column indicates that the guanines are un-clustered. G clusters were distributed so that an equal number of nucleotides was present between each cluster.

2.1.2 Variable regions

Each insert will contain one variable region designed to test the effects of different sequence properties, namely GC / AT skew, GC content, and G / C clustering, on R-loop initiation and termination. A total of 31 distinct variable regions, each 200 bp in length, will ultimately be synthesized. The properties of each region are described in table 1 will ultimately be synthesized.

Each variable region is designed to test sequence effects on R-loop initiation when oriented in the same direction as a proximal promoter and R-loop termination when oppositely oriented to an upstream promoter. The properties of the reverse complements, representing transcription of variable regions in the reverse orientation, are listed in table 2.

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.

name	GC skew	GC content	AT skew	C cluster length
VR-1	0.0	0.6	0.0	NA
VR-2	0.0	0.5	0.0	NA
VR-3	0.0	0.4	0.0	NA
VR-4	0.0	0.3	0.0	NA
VR-5	-0.1	0.6	0.0	NA
VR-6	-0.1	0.5	0.0	NA
VR-7	-0.1	0.4	0.0	NA
VR-8	-0.1	0.3	0.0	NA
VR-9	-0.2	0.6	0.0	NA
VR-10	-0.2	0.5	0.0	NA
VR-11	-0.2	0.4	0.0	NA
VR-12	-0.2	0.3	0.0	NA
VR-13	-0.4	0.6	0.0	NA
VR-14	-0.4	0.5	0.0	NA
VR-15	-0.4	0.4	0.0	2
VR-16	-0.4	0.4	0.0	3
VR-17	-0.4	0.4	0.0	4
VR-18	-0.4	0.4	0.0	NA
VR-19	-0.4	0.4	-0.2	2
VR-20	-0.4	0.4	-0.2	3
VR-21	-0.4	0.4	-0.2	4
VR-22	-0.4	0.4	-0.2	NA
VR-23	-0.4	0.4	-0.4	2
VR-24	-0.4	0.4	-0.4	3
VR-25	-0.4	0.4	-0.4	4
VR-26	-0.4	0.4	-0.4	NA
VR-27	-0.4	0.3	0.0	NA
VR-28	-0.6	0.6	0.0	NA
VR-29	-0.6	0.5	0.0	NA
VR-30	-0.6	0.4	0.0	NA
VR-31	-0.6	0.3	0.0	NA

Table 2: Properties of the reverse complements of all variable regions described in table 1
.

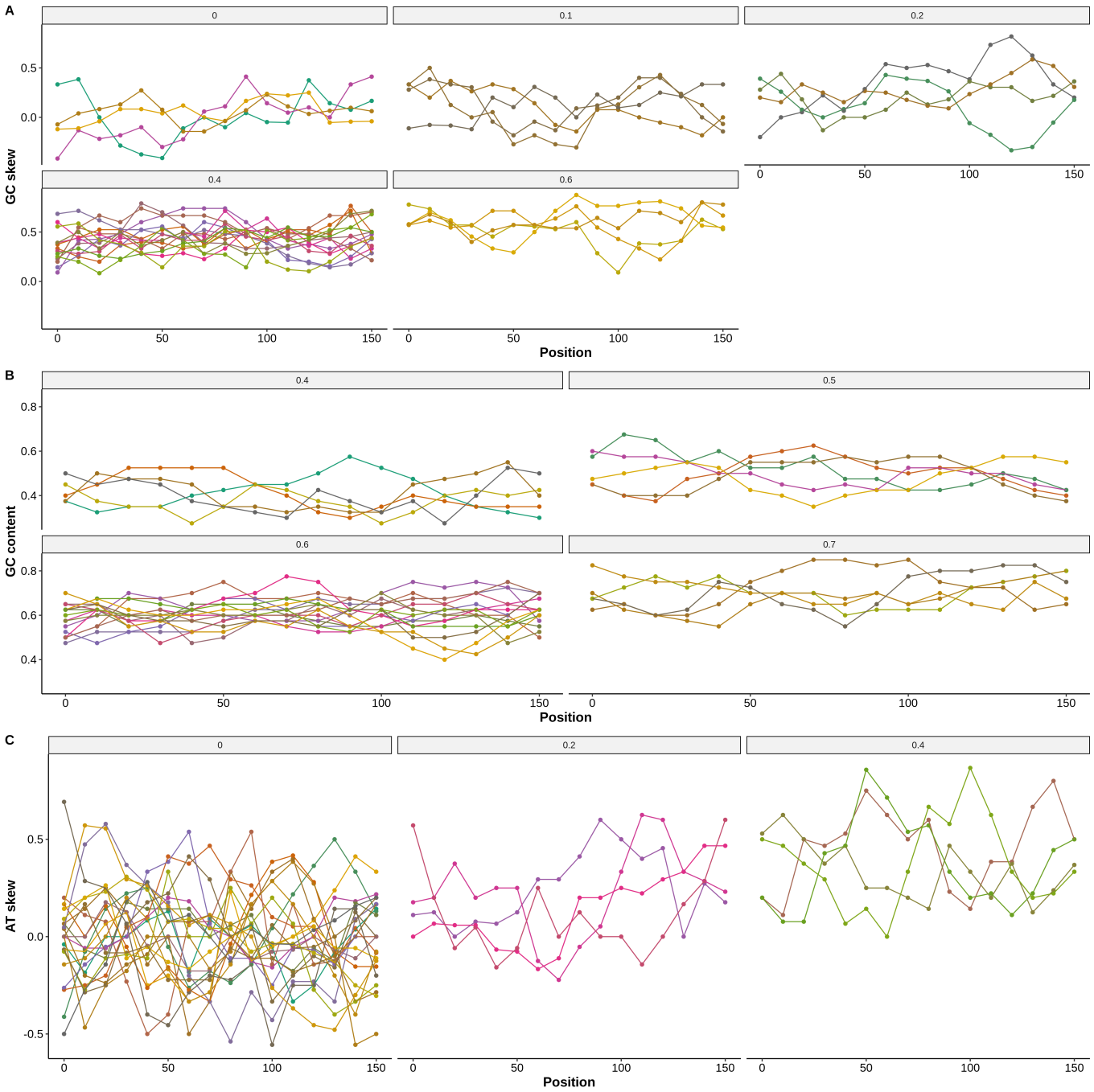

A
0
0.1
0.2
0.4
0.6
GC skew
0.5
0.0
0
50
100
150
Position
B
0.4
0.5
0.6
0.7
GC content
0.8
0.6
0.4
Position
C
0
0.2
0.4
AT skew
0.5
0.0
-0.5
Position

Figure 4

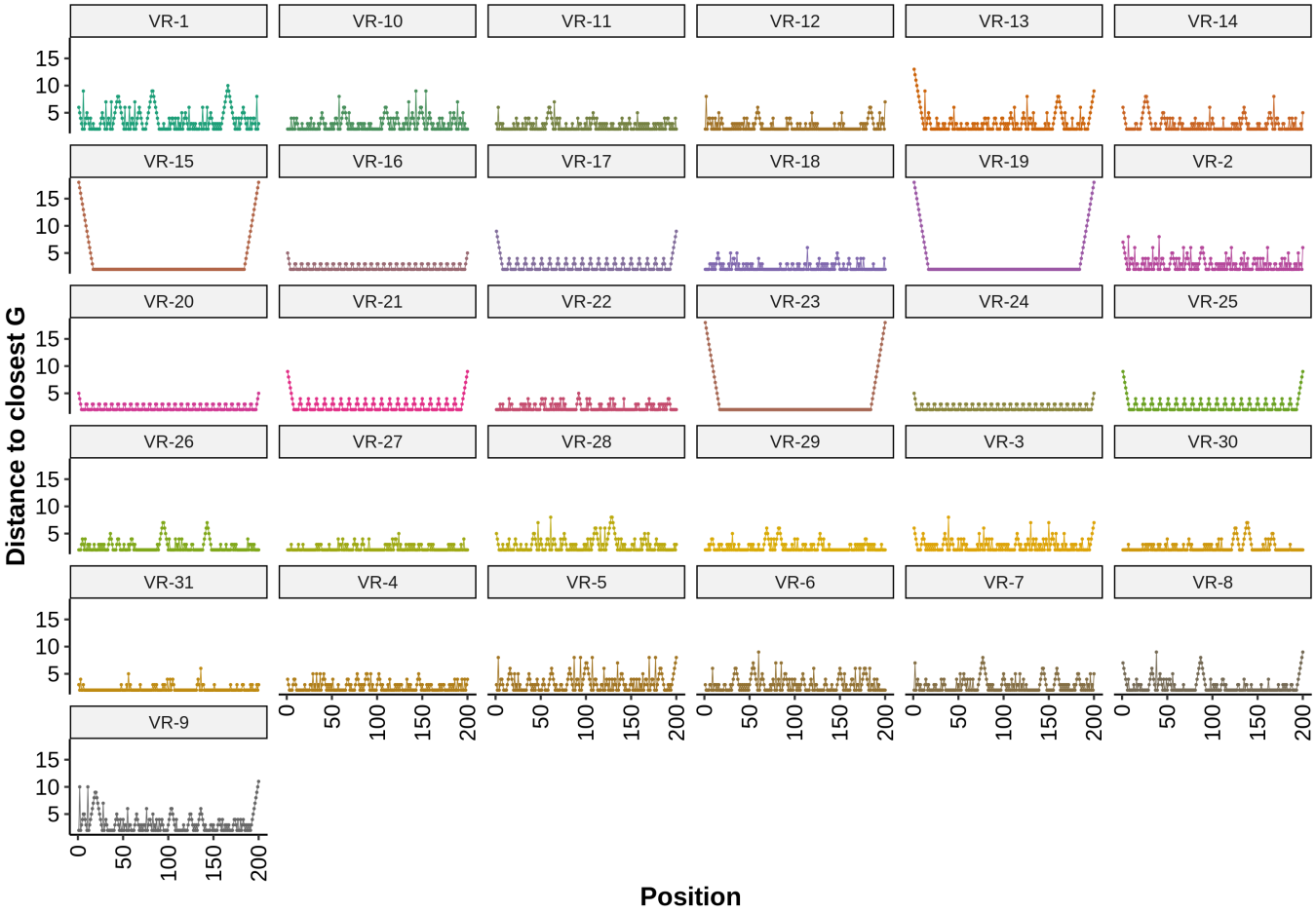


Figure 5

To select one specific sequence from this set, we can sample many sequences and access each one with metrics relevant to the realities of cloning protocols and R-loop formation. Each metric used to access candidate variable regions is described below. All metrics received equal weight in the final candidate selection, and a pool of 100 candidates was generated for each variable region.

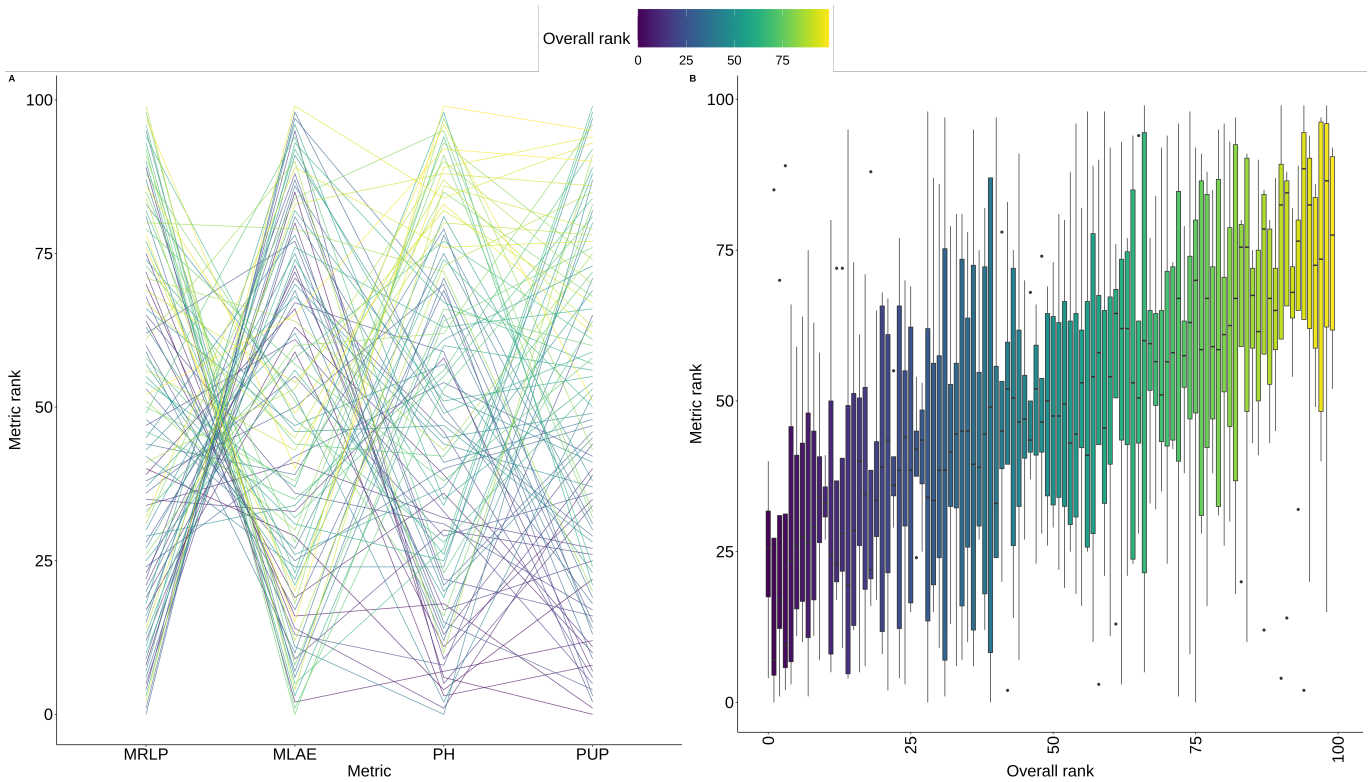


Figure 6: Visualization of sequence selection for an example variable region. Here, 100 candidate variable region sequences were generated and then ranked based on four metrics; mean R-loop probability (MRLP), mean local average energy (MLAE), proportion of nucleotides predicted to participate in RNA hairpins (PH) and proportion of nucleotides predicted to be unpaired (PUP). Sequences were by each 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 It is critical that the inserts are not cut unexpectedly within the variable region by any restriction enzymes used as a part of cloning protocols. Accordingly, before passing on for further downstream analysis, candidate variable regions containing any of the recognition sites listed in table 3 were removed from consideration.

Table 3

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 simulated R-loops overlapping that particular nucleotide. We prefer to select variable regions that are more likely to form R-loops. Accordingly, candidate variable regions with high mean R-loop formation probability and low mean local average energy received better rankings.

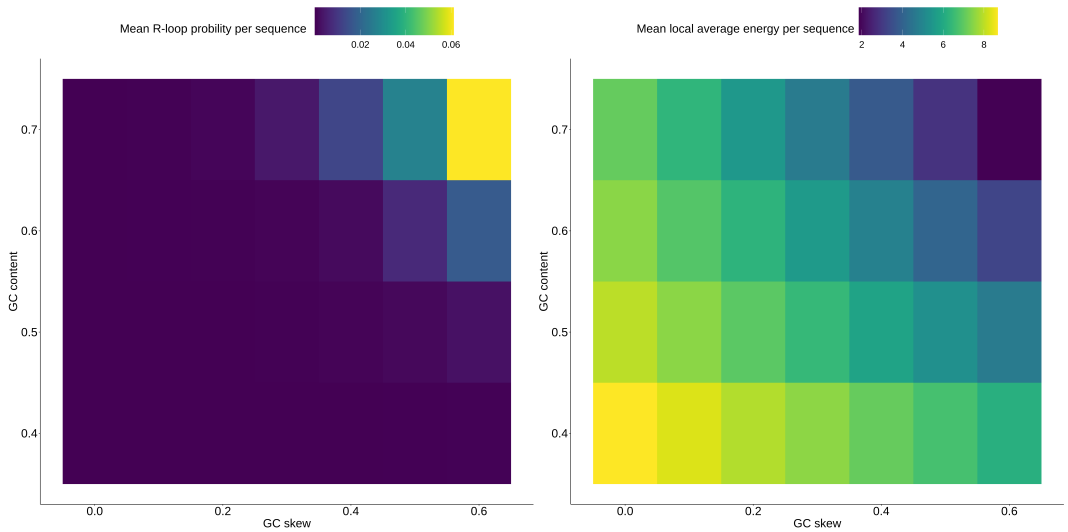


Figure 7: Heatmaps showing mean R-loop probability per sequence (left) and mean local average energy per sequence (right) for 100 sequences of length 200 at various levels of GC skew and content. As GC skew and content increase, the mean R-loop probability also increases while the mean local average energy decreases. This observation 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 7). Comparing 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 generally predicted to affect the likelihood of R-loop formation.

RNA secondary structure Significant amounts of RNA secondary structure, especially large hairpins, are 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 secondary structure prediction program in conjunction with the bpRNA structure annotation program [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 that would remain unpaired (PUP). Candidate variable regions with low PH and high PUP were preferred.

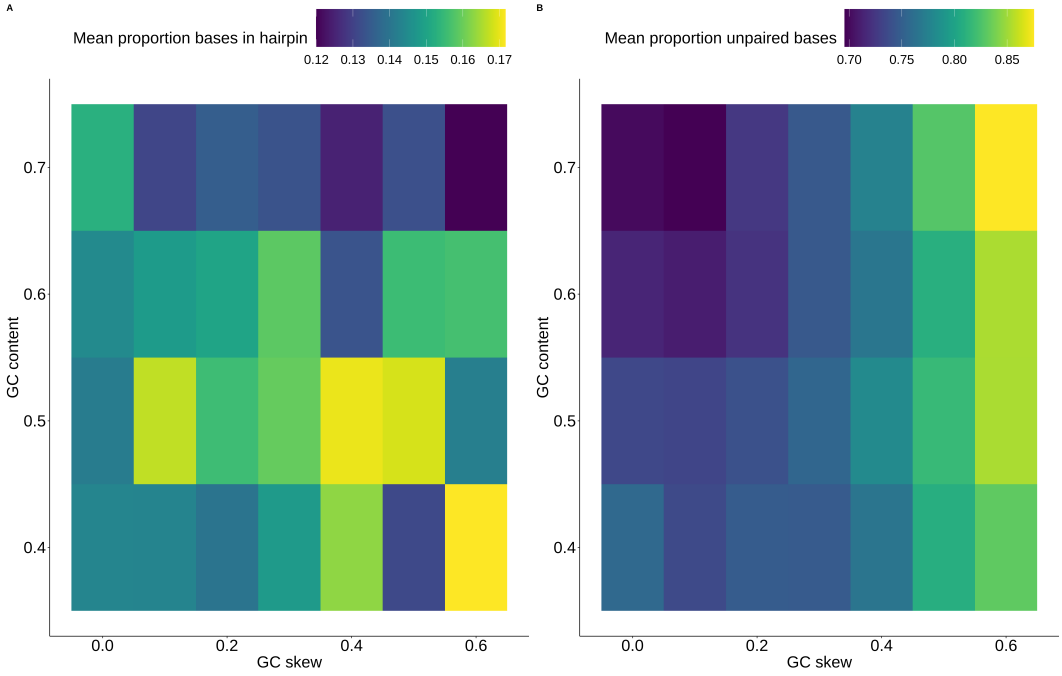


Figure 8: Heatmaps showing PH (left) and PUP (right) for 100 candidate variable regions of length 200 at various levels of both GC skew and content.

2.1.3 EcoRI site and 3' homology arm

The final 30 nucleotides of each insert will be composed of an EcoRI recognition site (6 bp) and the 3' homology arm (24 bp).

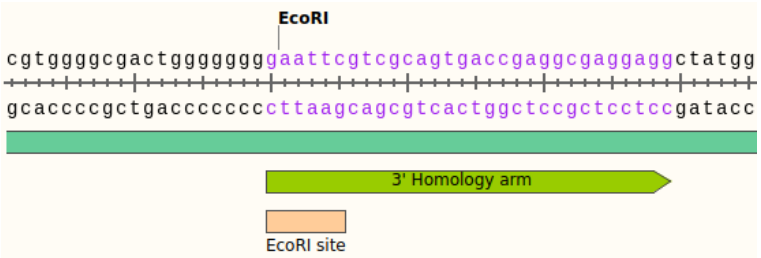


Figure 9: Sequence of the 3' homology arm and EcoRI site. The complete sequence is highlighted in purple.

Both the 5' homology arm + KpnI site and EcoRI site + 3' homology arm was produced using the Jupyter notebook available at this link.

3 Assembly of DNA inserts

Complete insert sequences will be cloned into three different plasmid backbones: pFC9 (fig 11), pFC8 (fig 11), and pFC53tacT_1T_2 (fig 12). pFC9 will be utilized for testing R-loop initiation, pFC8 for R-loop termination and pFC53tacT_1T_2 for multiple and single-round transcription versions of both initiation and termination experiments with Tac polymerase.

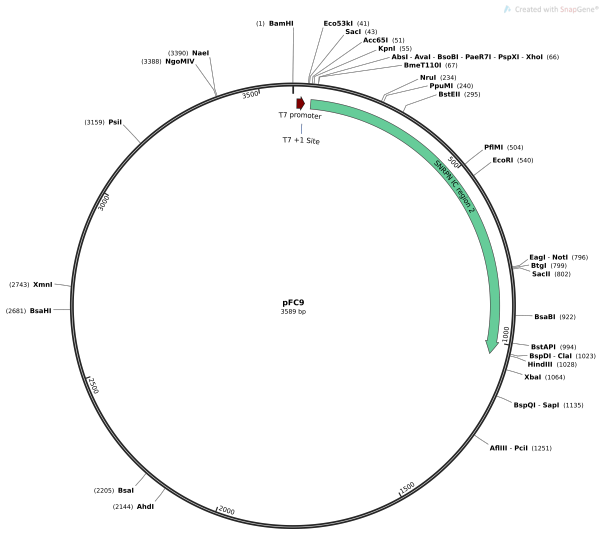


Figure 10: Map of pFC9.

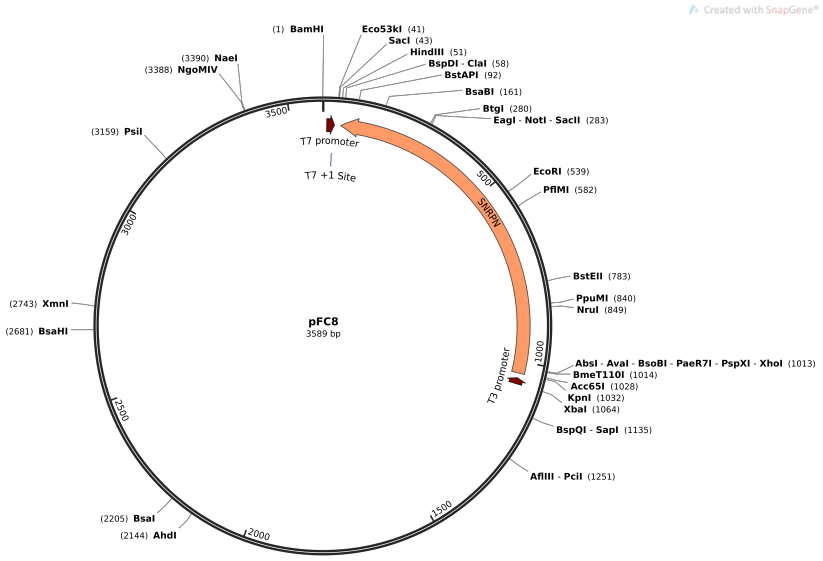


Figure 11: Map of pFC8.

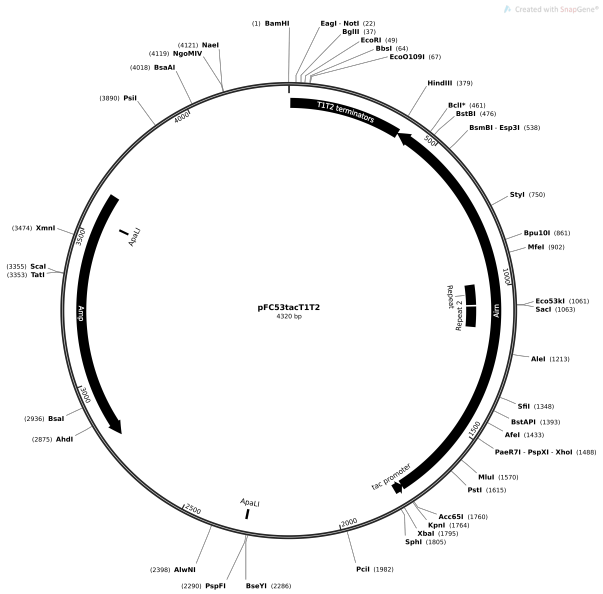


Figure 12: Map of pFC53tacT_1T_2.

3.1 T7 initiation series constructs

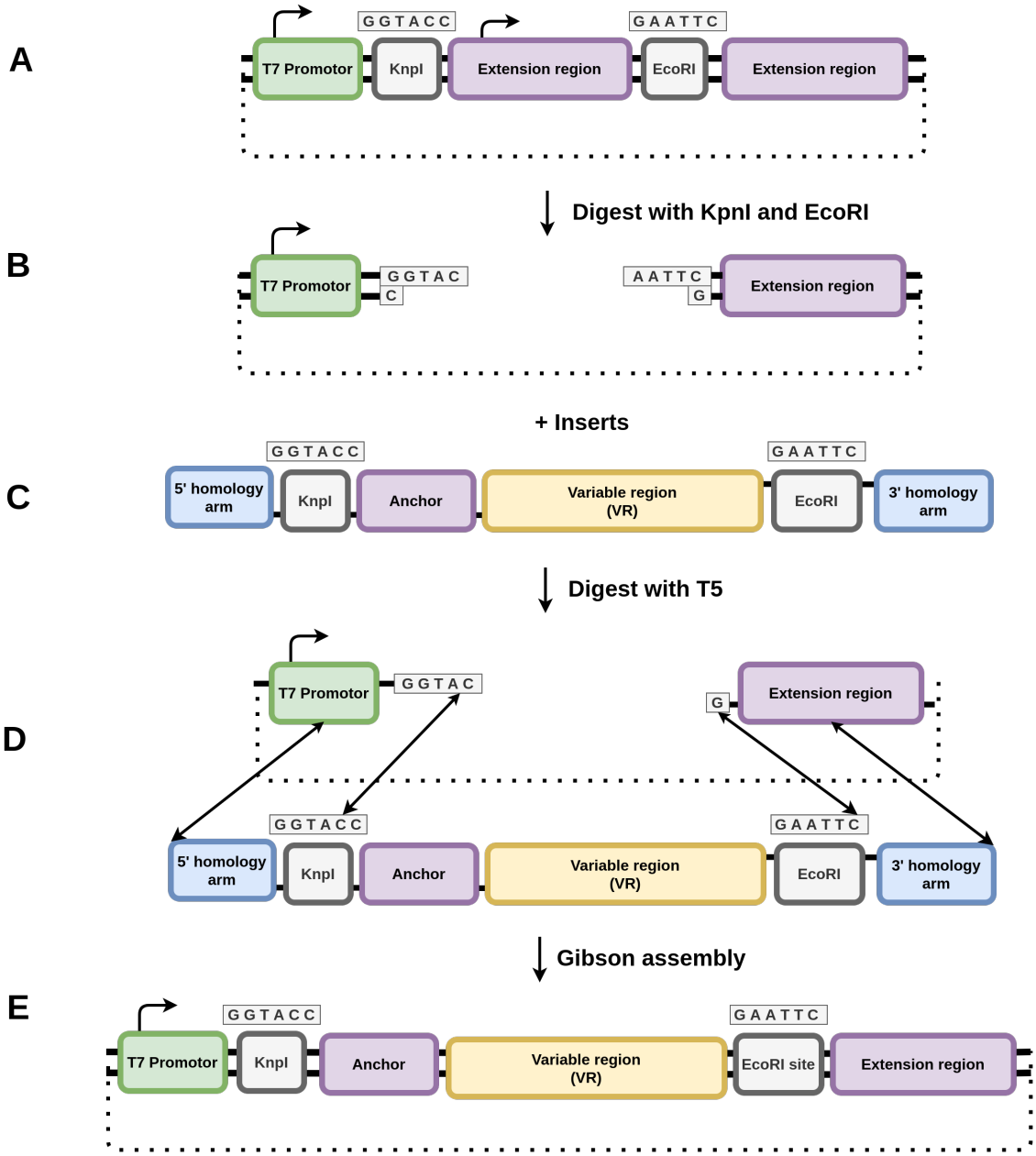


Figure 13: Diagram of pFC9 T7 insertion series cloning strategy.

First, pFC9 will be linearized by digestion with KpnI and EcoRI (fig 13A). The large pFC9 fragment will then be purified via agarose gel and added to a mixture containing all insert sequences in equal molar concentrations (fig 13B). 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 13C). The 5' ends of the insert will also be degraded, exposing the complete KnpI and EcoRI sites. 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. 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 13D). This strategy will result in a library of circular constructs with all inserts located downstream of and oriented forward relative to the pFC9 T7 promoter (fig 13).

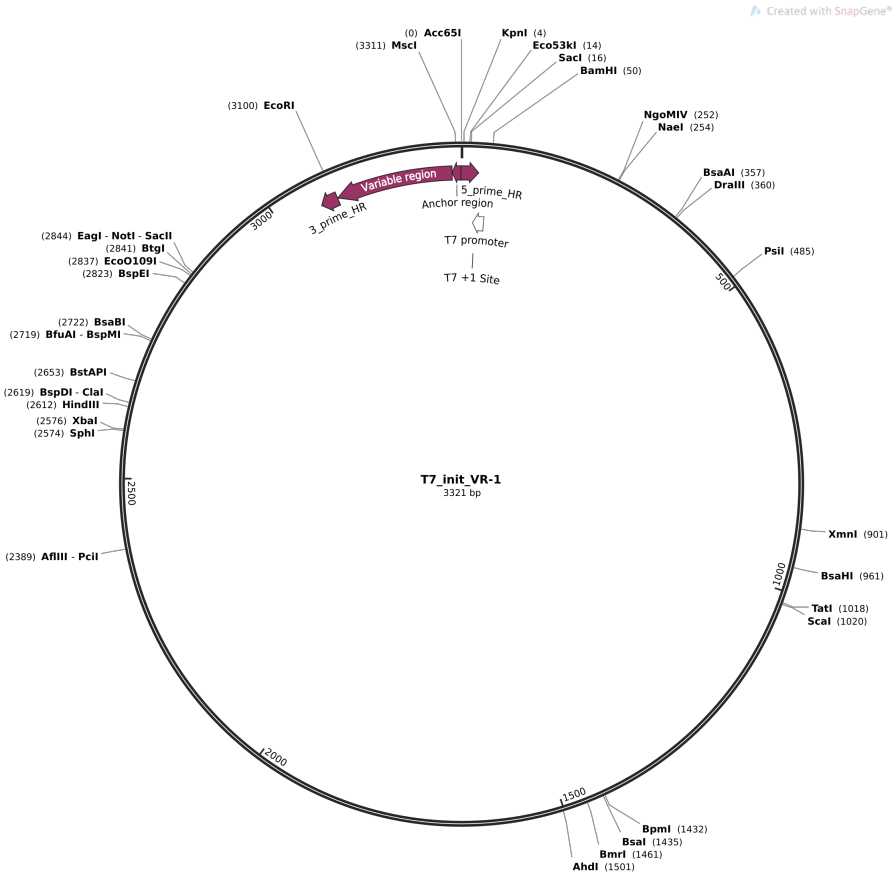


Figure 14: Plasmid map of produced by simulating the T7 initiation series cloning protocol for the insert containing variable region 1. Cloning simulations were ran and confirmed to be successful for all complete inserts sequences.

3.1.1 Assembly protocol

A Jupyter notebook for calculating the volumes of individual inserts, digested pFC9 backbone, and Gibson assembly master mix to complete the Gibson assembly protocol outlined in fig 14 is available at this link. At the time of writing, the concentration of pFC9 fragment and the concentrations of individual inserts are unknown, so placeholder values have been used. All volumes and concentrations of reagents are derived from New England Biolabs Gibson assembly recommendations.

3.2 T7 termination series constructs

After successfully cloning the T7 initiation series, pFC8 will be utilized as the backbone for constructing the T7 termination series library.

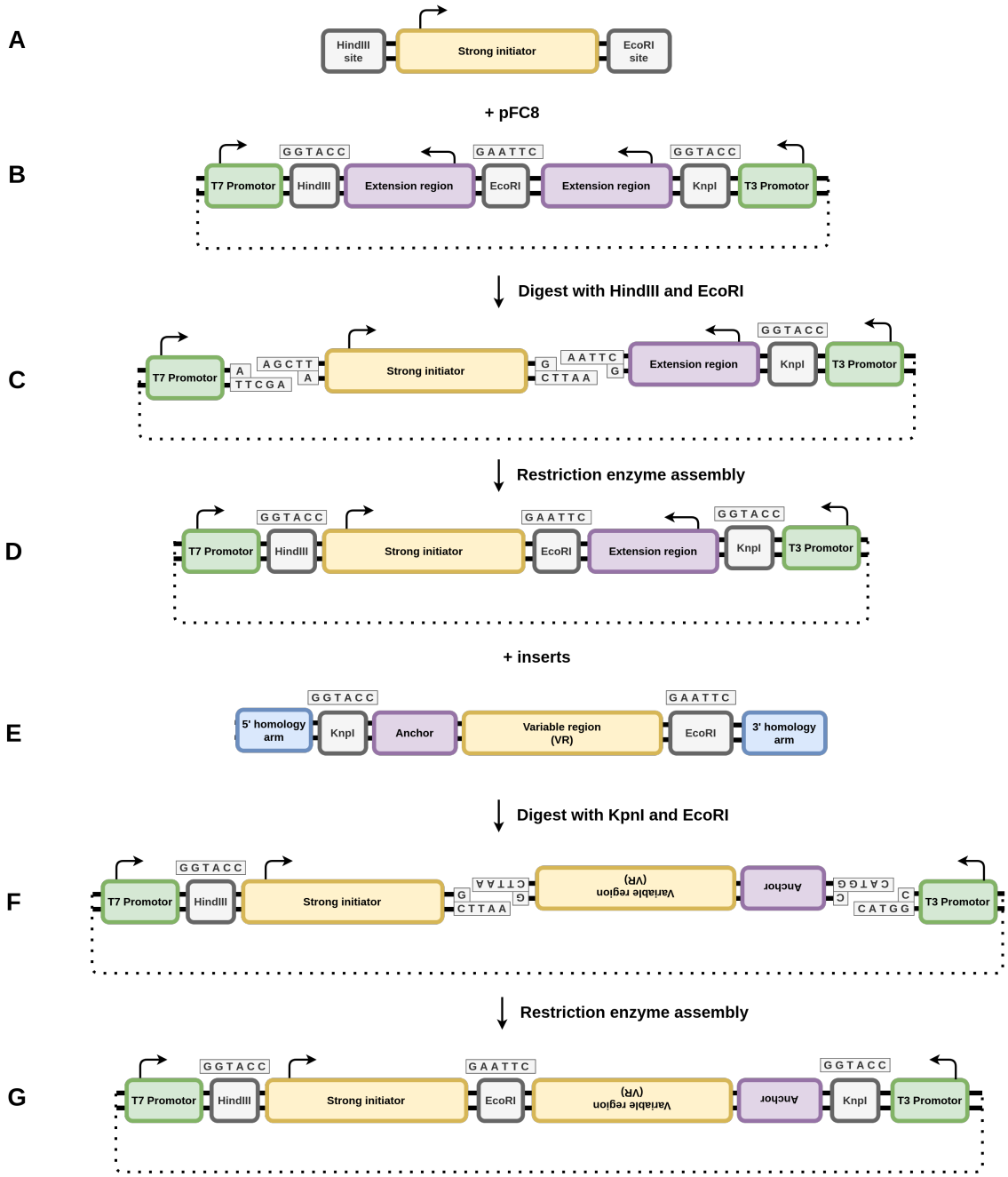


Figure 15: Diagram of pFC8 T7 termination series cloning strategy.

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 (15A). This strong initiator fragment will be added to pFC8 (15B), 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 (15C). Next, all termination inserts will be added to the pFC8-strong-initiator construct in equal molar ratios (15D) and the mixture digested with KpnI and EcoRI (15E). Inserts will then be incorporated into pFC8-strong-initiator constructs via homology between the digested KpnI and EcoRI sites (15F). Since the order of these recognition sites on the pFC8-strong-initiator construct is opposite to that of the insert 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 (15G).

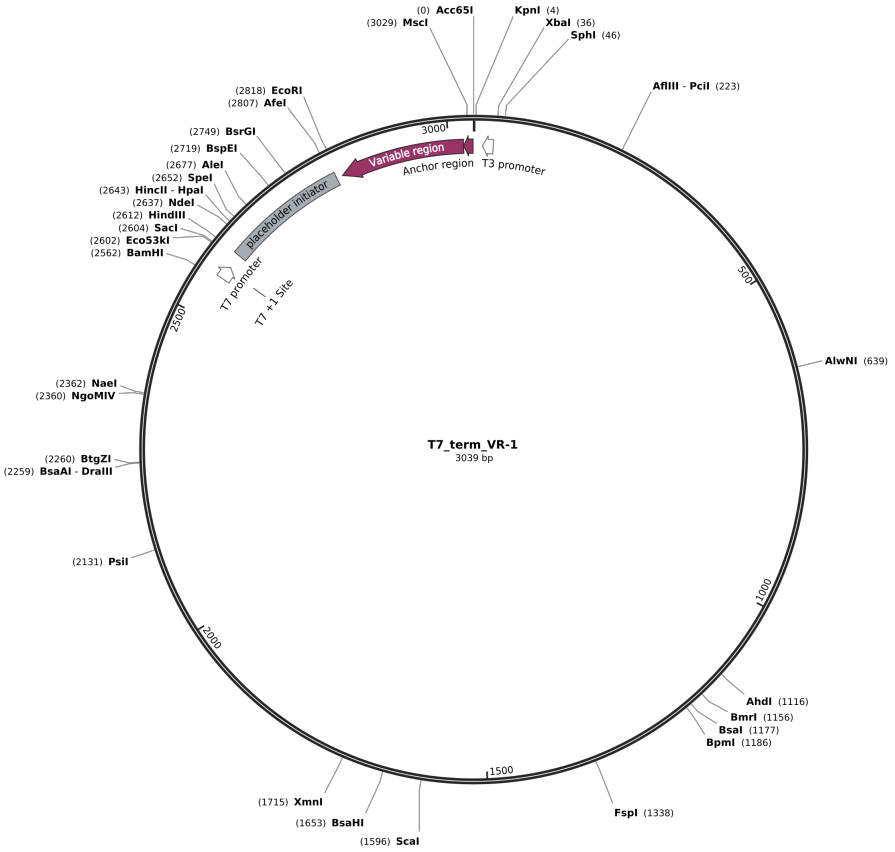


Figure 16: Plasmid map produced by simulating the T7 termination series cloning protocol for the insert containing variable region 1. Cloning simulations were run and confirmed to be successful for all complete inserts sequences. Note that since the strong initiation sequence is undetermined, a simulated but functionally equivalent sequence was used in its place.

3.3 Tac initiation series constructs

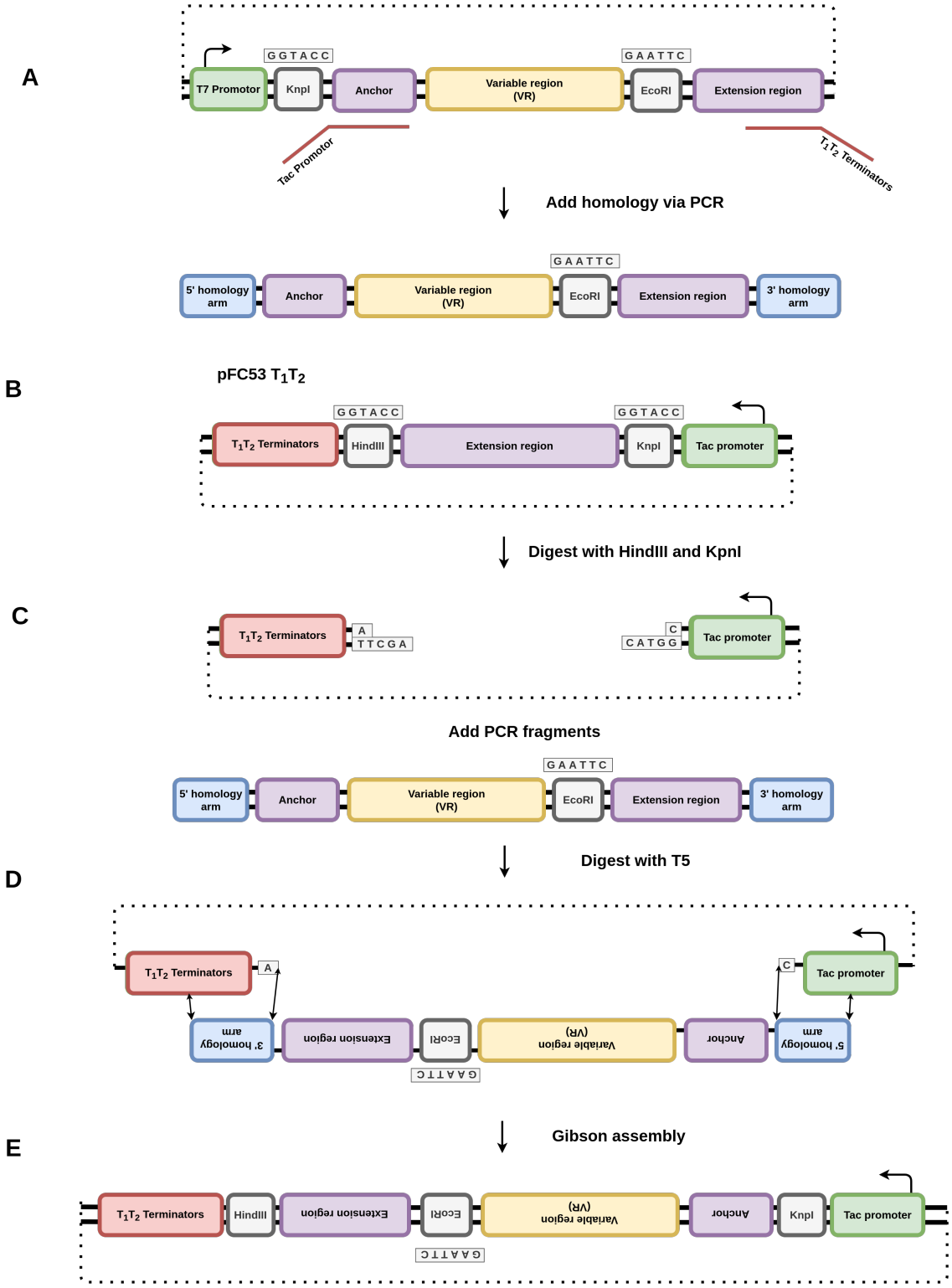


Figure 17: Diagram of pFCT$_1$T$_2$ tac initiation series cloning strategy.

First, primers are used to 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 overhangs with homology to the 5' end of the tac promoter and 3' T1T2 terminator sequences (**??**A). Next, pFCT$_1$T$_2$ is separately digested with HindIII and KpnI, and the large fragment is purified via agarose gel. The PCR products are added to the large fragment and digested with T5 endonuclease (**??**B, C). The final construct is assembled via Gibson assembly, and the HindIII and KpnI sites are regenerated as a result of the insertion.

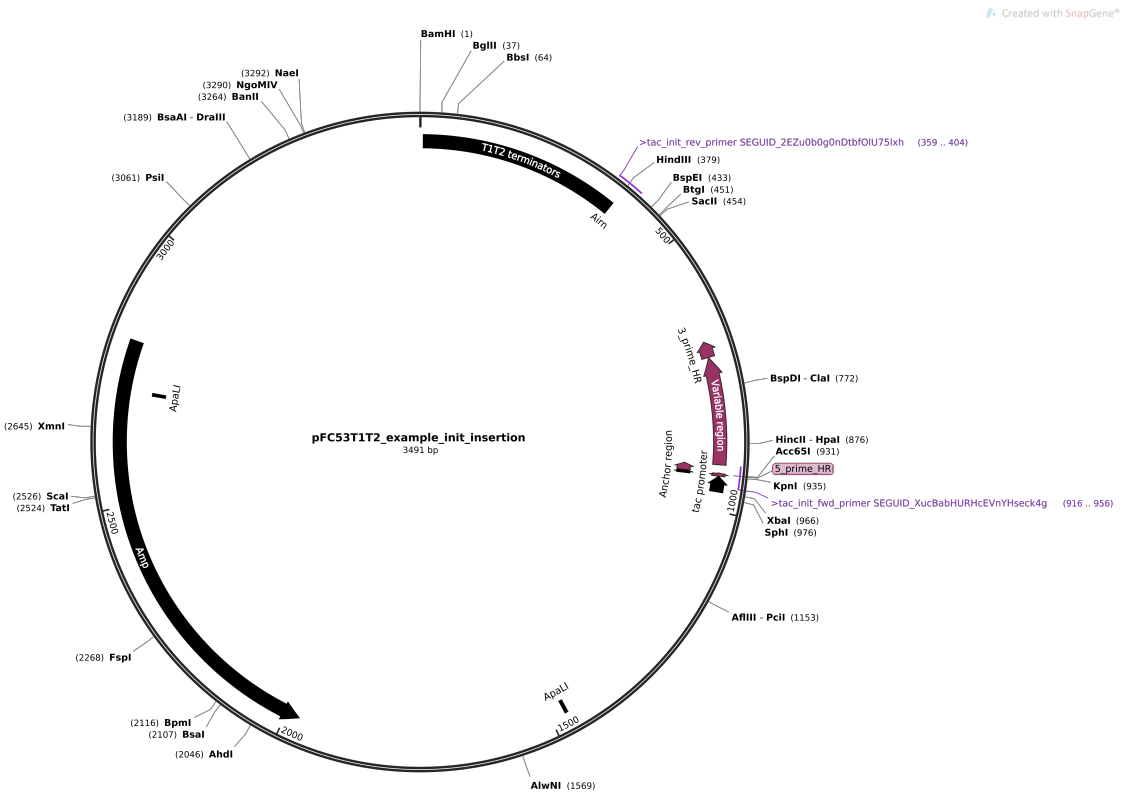
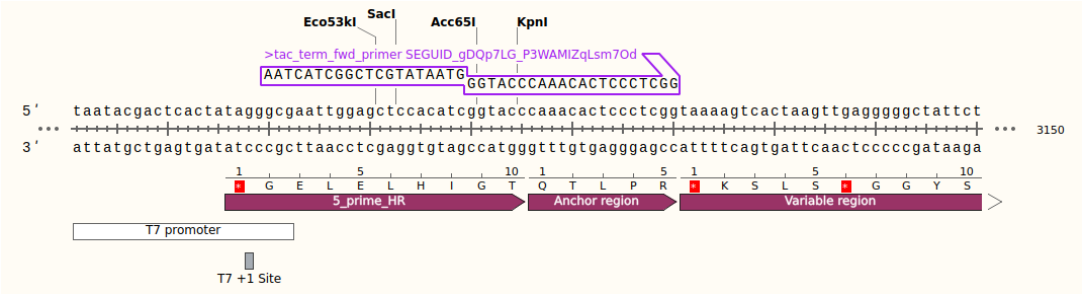
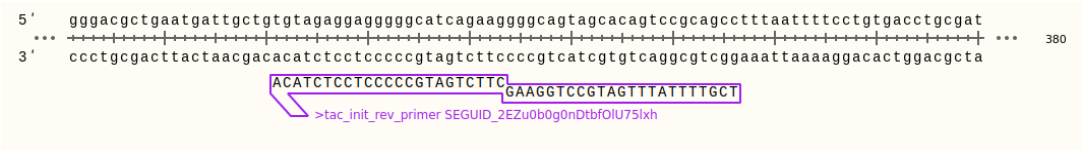


Figure 18: Plasmid map produced by simulating tac initiation series assembly protocols.

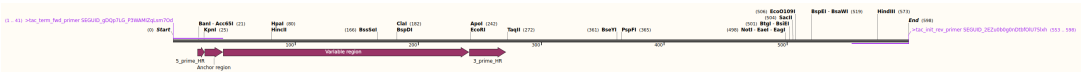
3.3.1 Tac initiation series primer design



(a) Forward primer. Targets the anchor region and adds homology to the downstream end of the Tac promoter of the pFC53T1T2 backbone and regenerates the digested KpnI site.



(b) Reverse primer. Targets an arbitrary region 500 nucleotides downstream of the end of the insert in order to include an "extension region" downstream of the variable region. Adds homology to the upstream end of the T1T2 terminator sequence of the pFC53T1T2 backbone and regenerates the digested HindIII site.



(c) Amplicon produced by the forward and reverse primers.

Figure 19: PCR primers and resulting amplicon used to produce tac initiation series constructs.

```
|95°C|95°C                |    |tmf:73.4
|____|_____          72°C|72°C|tmr:78.5
|5min|30s  \ 64.4°C _____|____|45s/kb
|    |      \______/ 0:59|5min|GC 55%
|    |       30s         |    |1318bp
```

Figure 20: PCR program for Tac polymerase for tac initiation series primers.

3.4 Tac termination series constructs

The assembly protocol for the tac termination series is identical to the initiation series, the difference being the insert is sourced from T7 termination series constructs and accordingly a different set of primers is utilized to amplify the insert region and add homology for Gibson assembly into pFCT$_1$T$_2$.

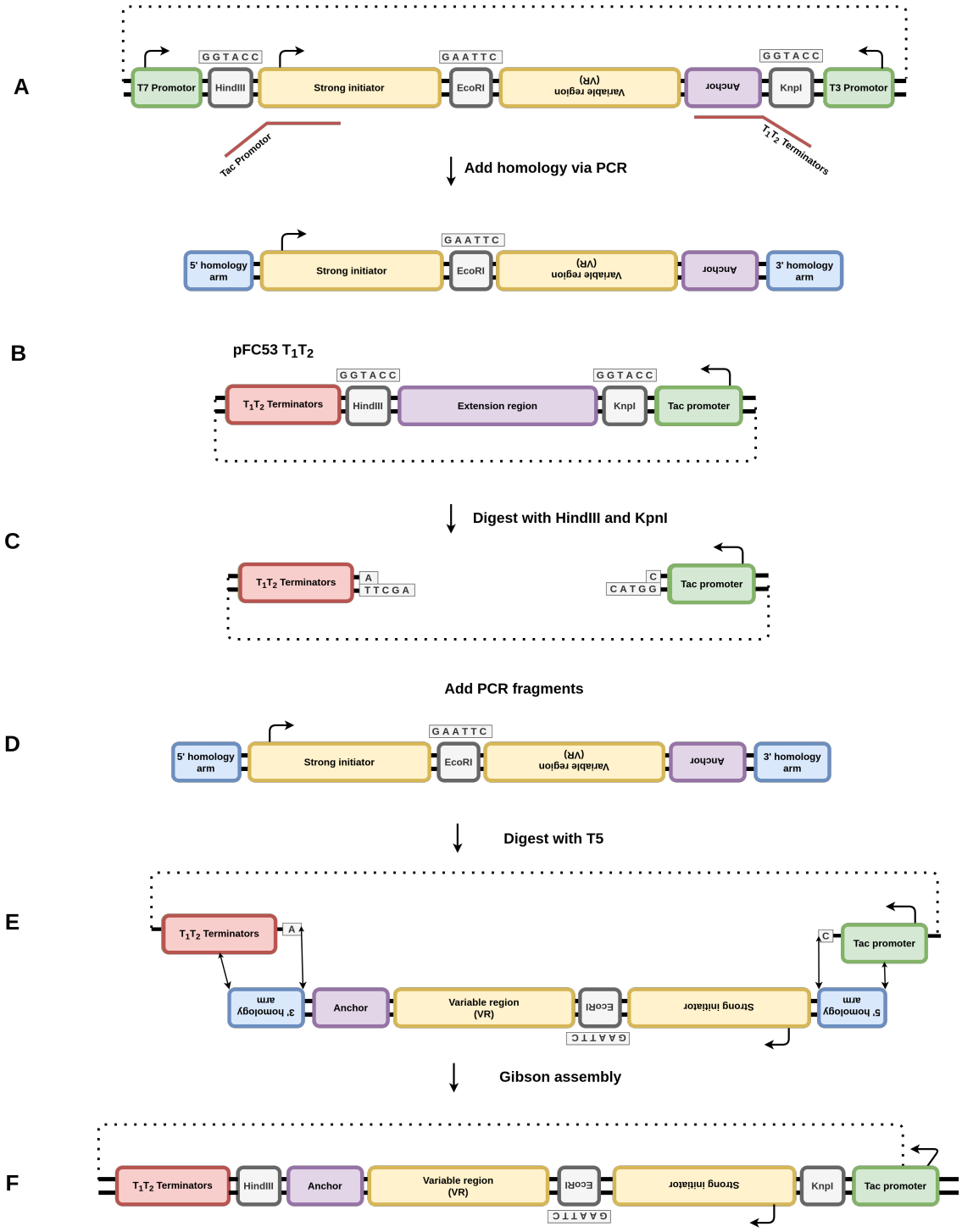


Figure 21: Diagram of pFCT$_1$T$_2$ termination series cloning strategy.

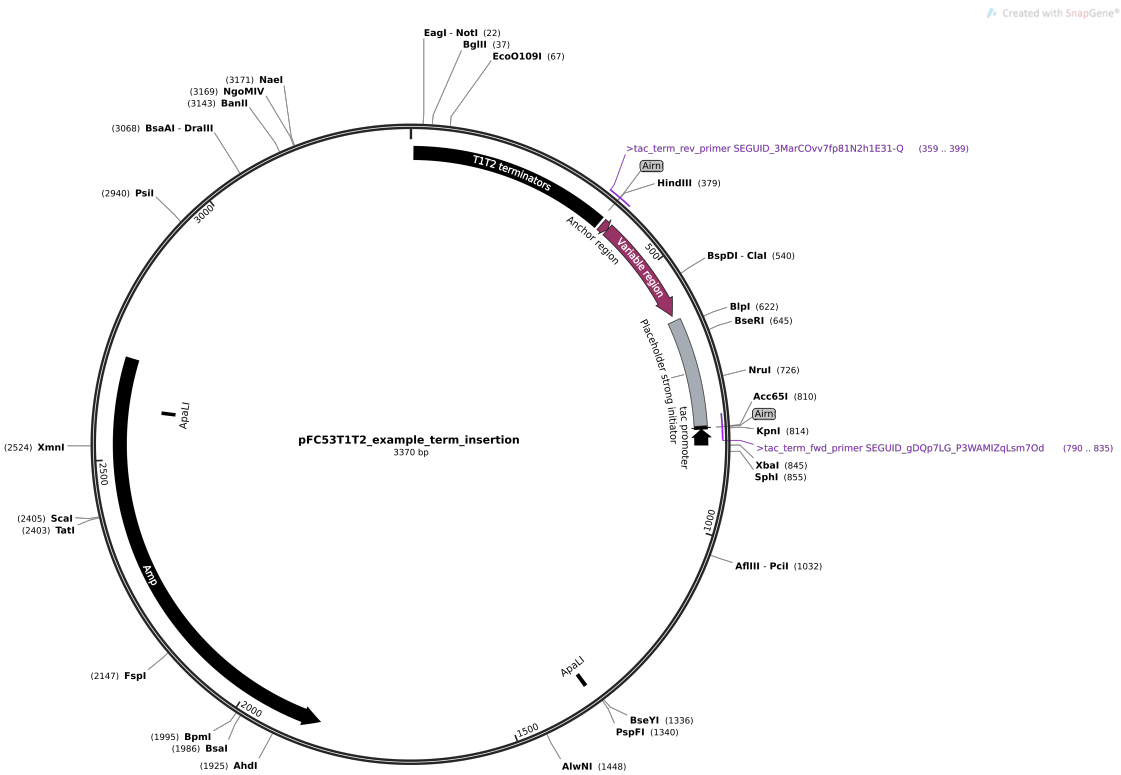
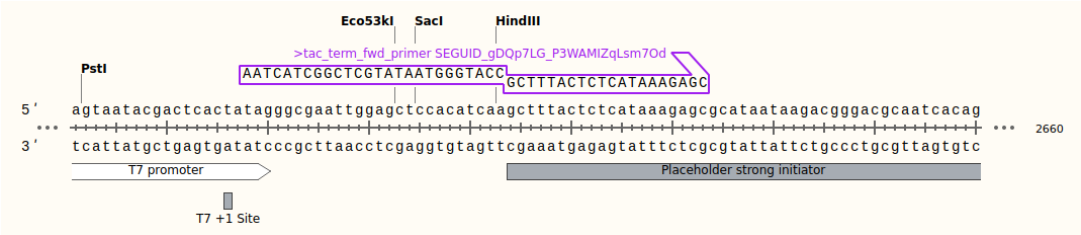
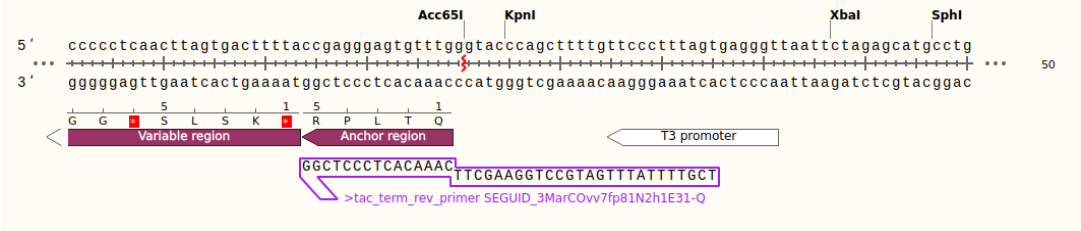


Figure 22: Plasmid map produced by simulating tac termination series assembly protocols.

3.4.1 Tac termination series primer design



(a) Forward primer. Targets the 5' end of the placeholder strong initiator and adds homology to the downstream end of adds homology to the downstream end of the Tac promoter of the pFC53T1T2 backbone and regenerates the KpnI site.



(b) Reverse primer. Targets the anchor region and adds homology to the to the upstream end of the T1T2 terminator sequence of the pFC53T1T2 backbone and regenerates the HindIII site.

(c) Amplicon produced by the forward and reverse primers.

Figure 23: PCR primers and resulting amplicon used to produce tac termination series constructs.

A PCR program is not included here as a placeholder initiation region was used to demonstrate how tac termination series primers would be designed. Therefore, until an actual strong initiator region is determined and verified,

these primers should *not* be ordered. Code to generate both initiation and termination series primers and PCR programs is available as a Jupyter notebook at this link. This notebook was later converted to a functionally equivalent version that is run as part of the insert calculation workflow.

3.5 Validation of insert libraries

Since all inserts of a given series will be assembled into their respective backbone in the same pot, it will be necessary to confirm that each insert-backbone assembly is present in relatively equal concentrations in the final library. Since the variable region of each insert-backbone assembly will be unique, qPCR utilizing primers that target a subset or all variable regions can be used to measure relative insert-backbone concentrations in the library. Potential primers specific to each variable region are generated using Primer3 as part of the insert calculation workflow.

3.6 Quantification of R-loop formation

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

3.7 Reagents and kits

- NEB Gibson assembly master mix
- NEB High-Fidelity DNA Polymerase
- QIAquick gel extraction kit.
- DpnI, optionally use for digesting plasmid DNA after amplication of inserts

3.8 Additional resources

- NEB Gibson assembly protocols
- Janet B. Matsen: Open Wet Ware Gibson assembly

4 Code availability

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

5 Sequence availability

All sequence data is available as a part of the insert region calculation pipeline and can be found at this link.

6 Supplementary materials

6.1 NEB Restriction enzyme activities

Enzyme	1.1	2.1	3.1	CutSmart
EcoRI	25	100	50	50
EcoRI-HF	10	100	<10	100
KpnI	100	75	<10	50
KpnI-HF	100	25	10	100
HindIII	25	100	50	50
HindIII-HF	10	100	100	100

6.2 Snakemake workflow diagram

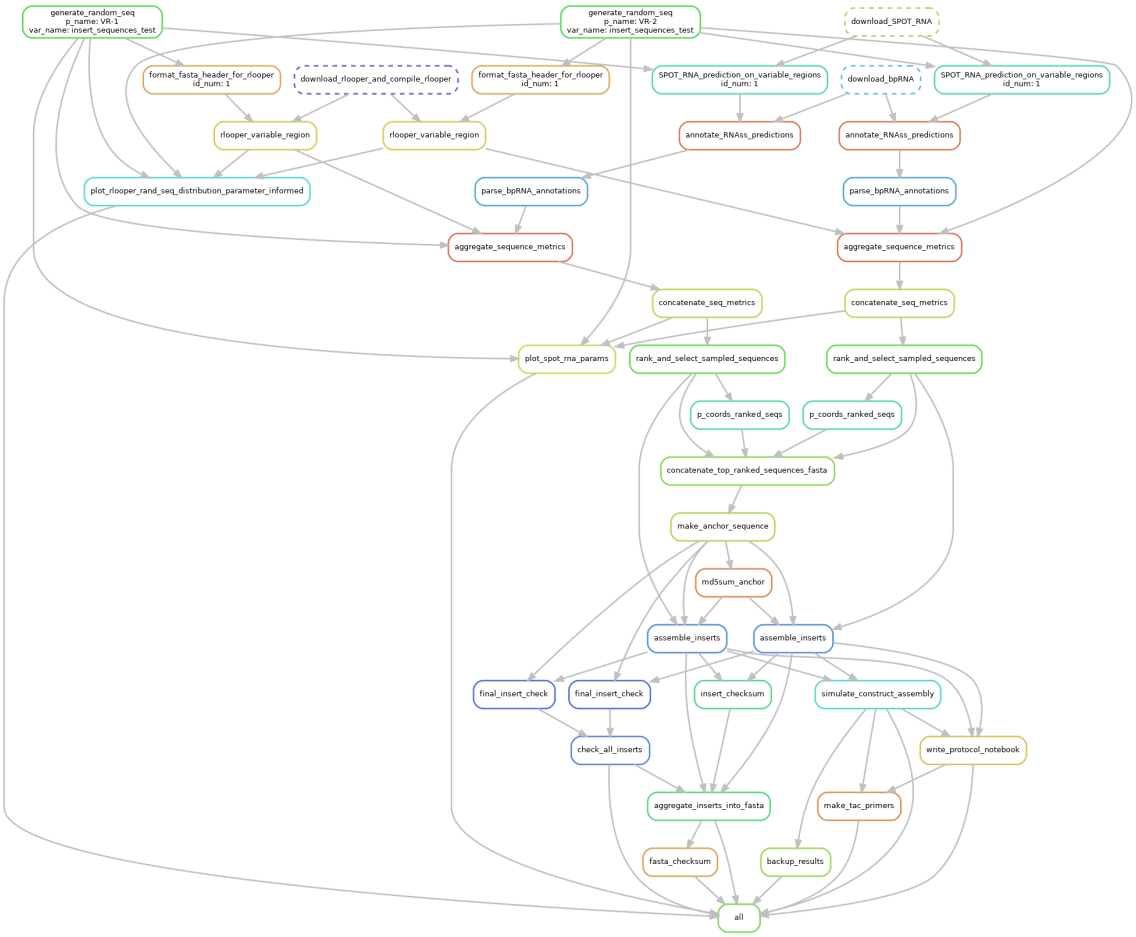


Figure 24: Visualization of insert calculation workflow.

6.3 Initial R-looper expectation calculations

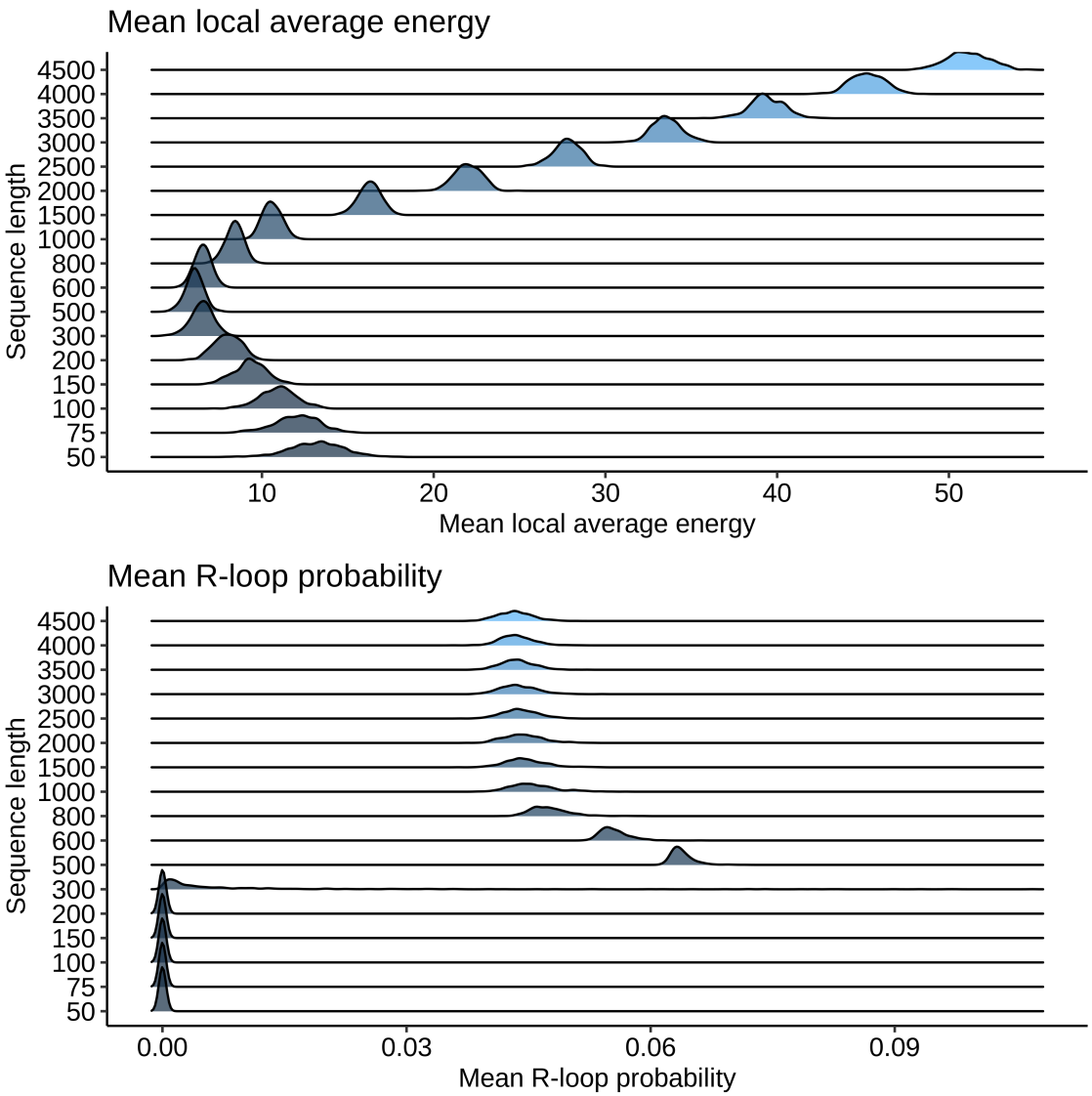


Figure 25: Initial rlooper expectation plots based on completely random sequences of increasing length.

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.