Variable region design and cloning

Ethan Holleman

July 4, 2021

1 Background

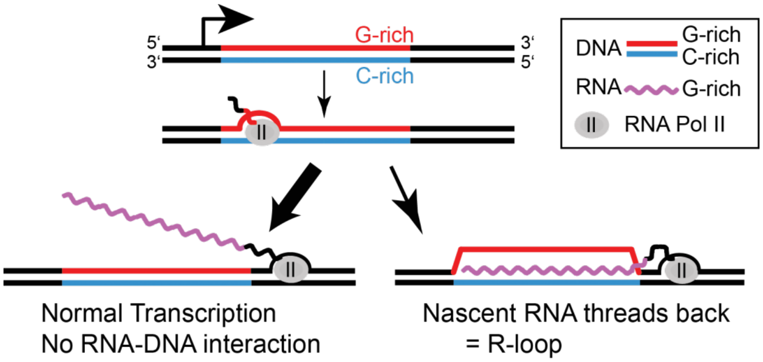

Figure 1

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.

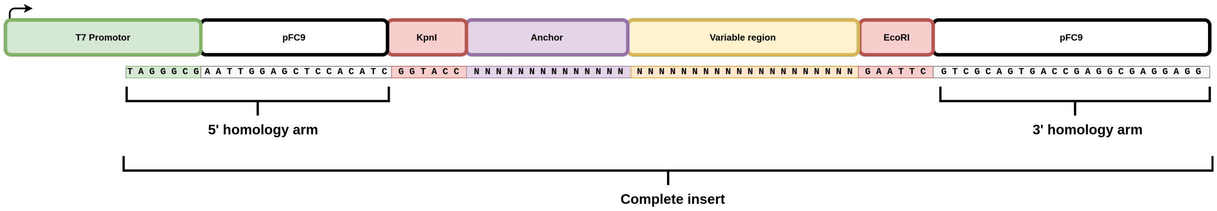

Figure 2: Diagram of a complete insert. Colored boxes represent features insert sequences are derived from while colored sequences represent the actual nucleotide sequence.

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.

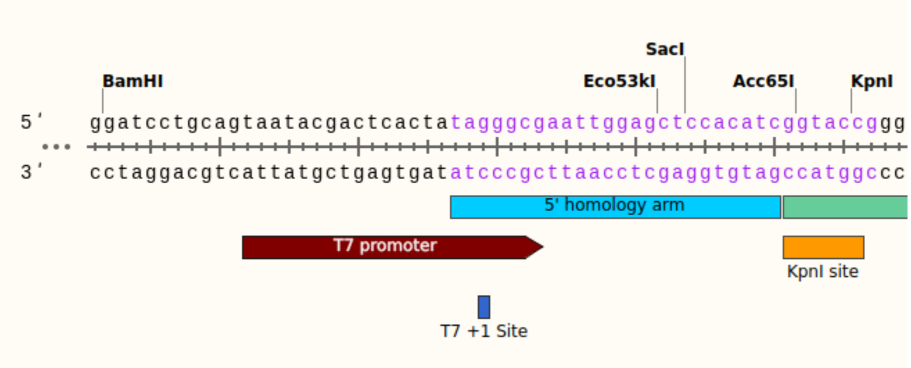
SacI
BamHI
Eco53kI
Acc65I
KpnI
5'
ggatcctgcagtaatacgactcactatagggcgaattggagctccacatcggtaccggg
...
3'
cctaggacgtcattatgctgagtgatatcccgcttaacctcgaggtgtagccatggccc
5' homology arm
T7 promoter
KpnI site
T7 +1 Site

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.

2.2 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 **??** and **??**.

Since this region is intended to be used as part of a PCR primer done to make sure it was unique across all plasmid backbones and insert sequences and did not contain any restriction enzyme recognition sites used in any of the cloning experiments.

2.3 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 reliablely 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.

2.3.1 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

2.3.2 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.

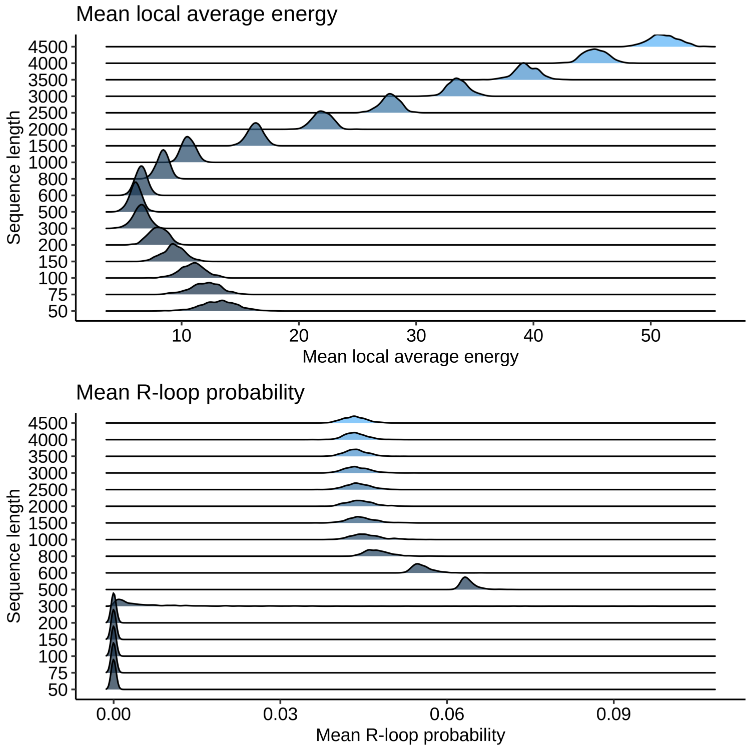

Figure 4: Map of pFC9 showing major features and all restriction enzyme recognition sites.

2.3.3 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.

2.4 EcoRI site

2.5 3' homology arm

3 Assembly of DNA inserts

3.1 T7 initiation series constructs

The first series of constructs will be utilize pFC9 as the plasmid backbone.

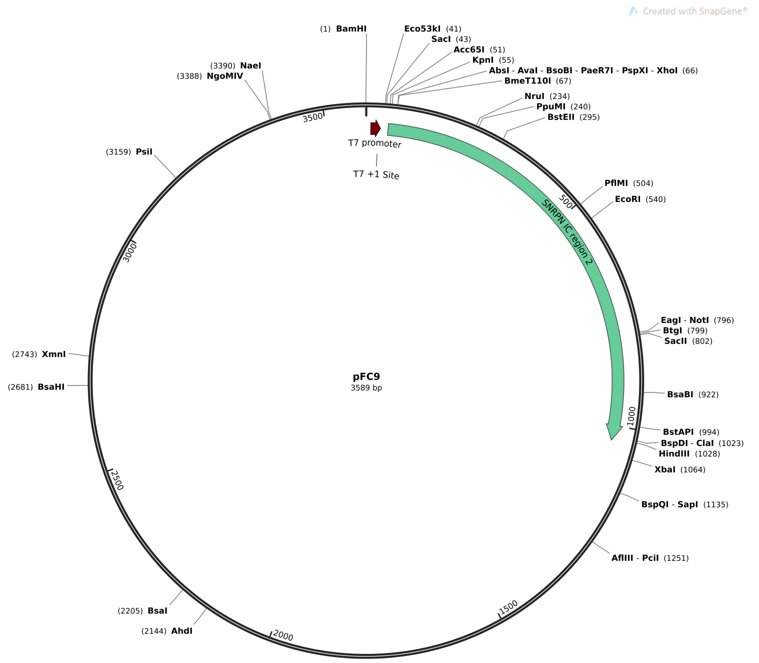

Figure 5: Map of pFC9 showing major features and all restriction enzyme recognition sites.

First, pFC9 will be cut using Eco53KI, producing blunt ends just downstream of the T7 promotor (fig 4). The complete ensemble of inserts will then be added in equal concentrations. Using the NEB Gibson assembly kit and protocol, the 5' and 3' homology arms will anneal to the T7 and extension region downstream of the Eco53KI cut site of pFC9 respectively producing a library of circular pFC9 plasmids with each insertion sequence in theoretically equivalent concentrations.

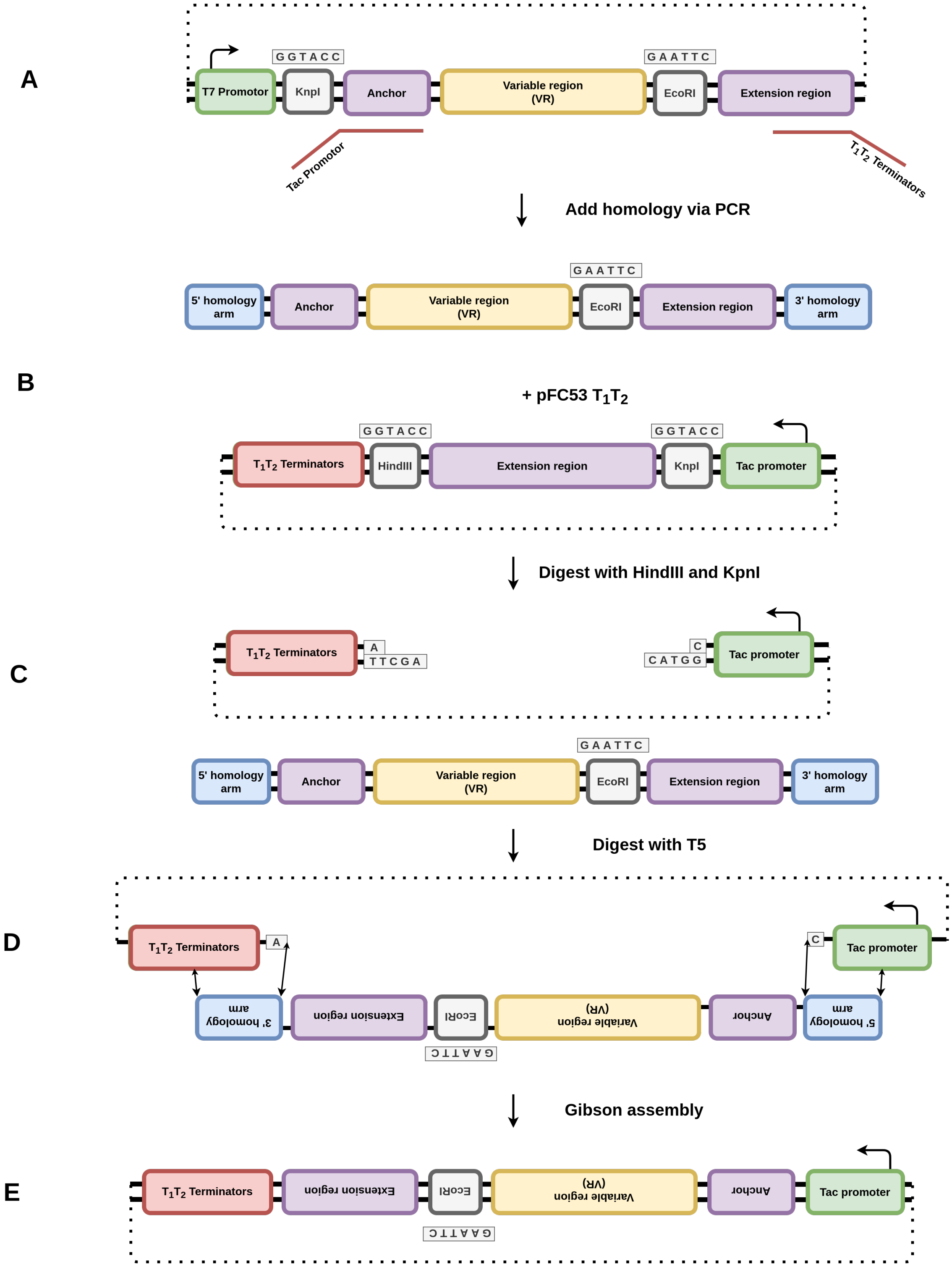

Figure 6: Diagram of pFC9 insertion series cloning strategy.

Primers will be designed for a subset of initiation regions and the relative concentrations of each insert will be measured by qPCR. Having confirmed that inserts are present at relatively equal concentrations the library will be transcribed *in vitro* and prepared for single molecule R-loop footprinting using uniquely bar-coded PCR primers to facilitate the computational removal of PCR duplicates.

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. First, the strongest and most consistent R-loop initiator identified from the T7 initiation series will be cloned into pFC9 without the presence of any other inserts using the methods described in section 3.1.

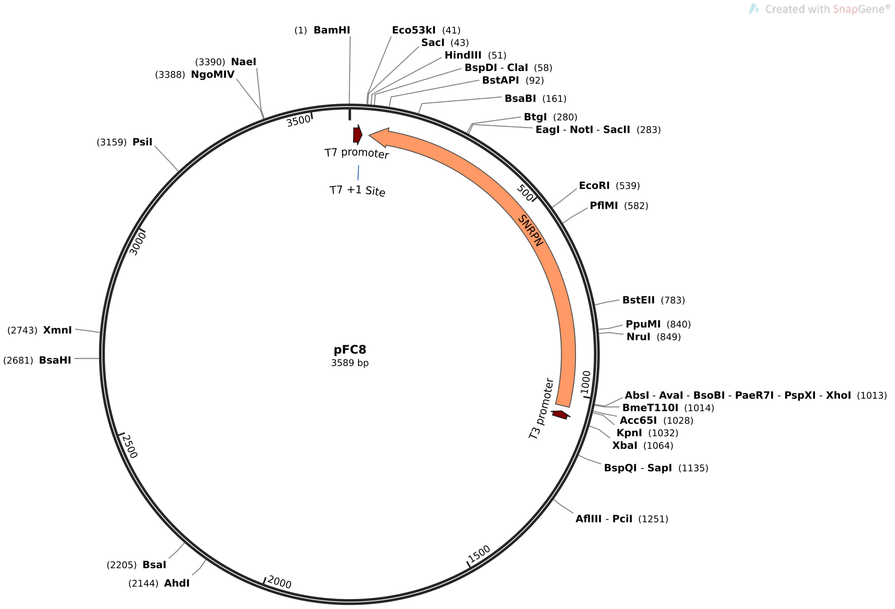

Figure 7: Map of pFC8.

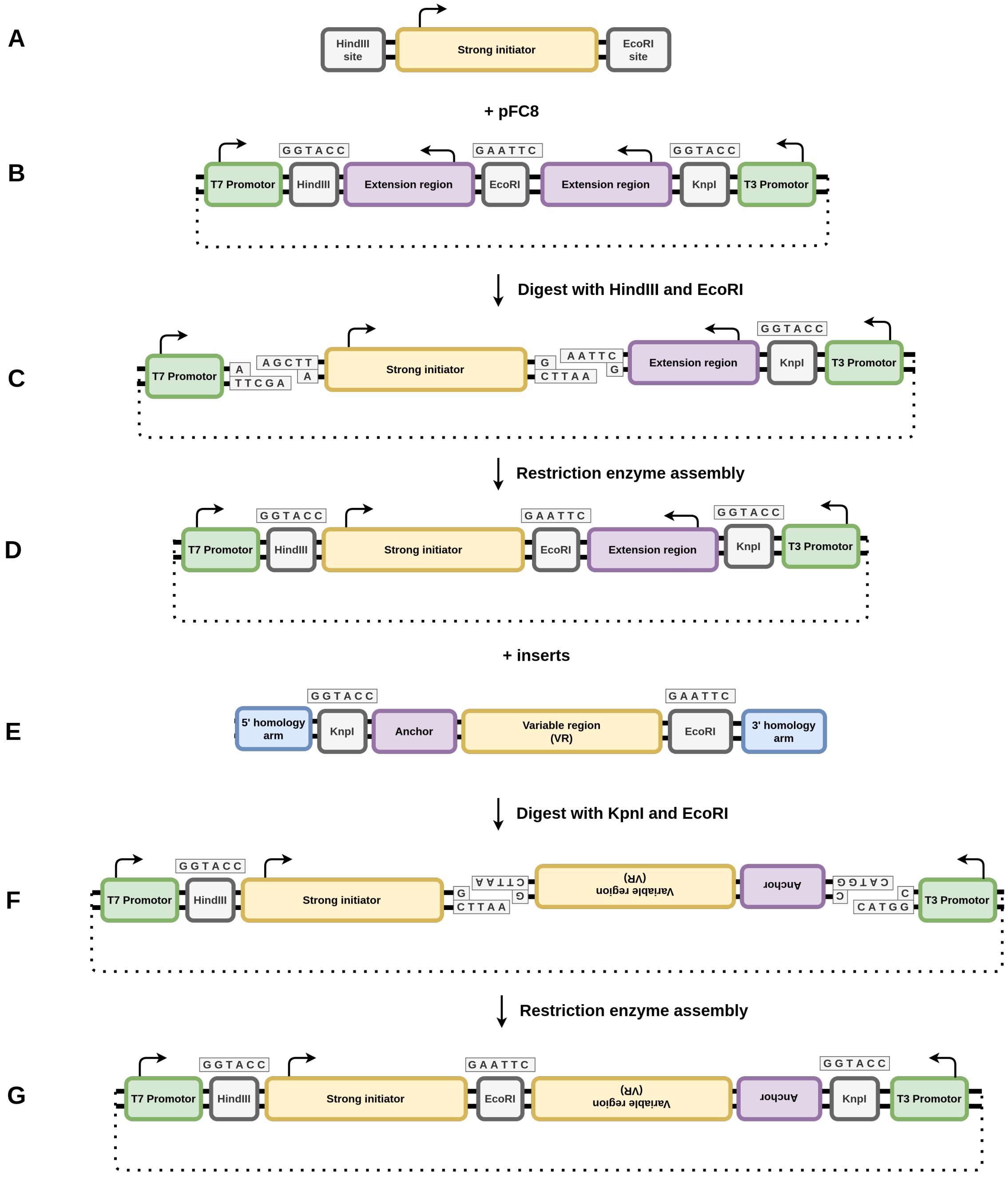

Figure 8: Diagram of pFC9 insertion series cloning strategy.

3.3 Tac initiation series constructs

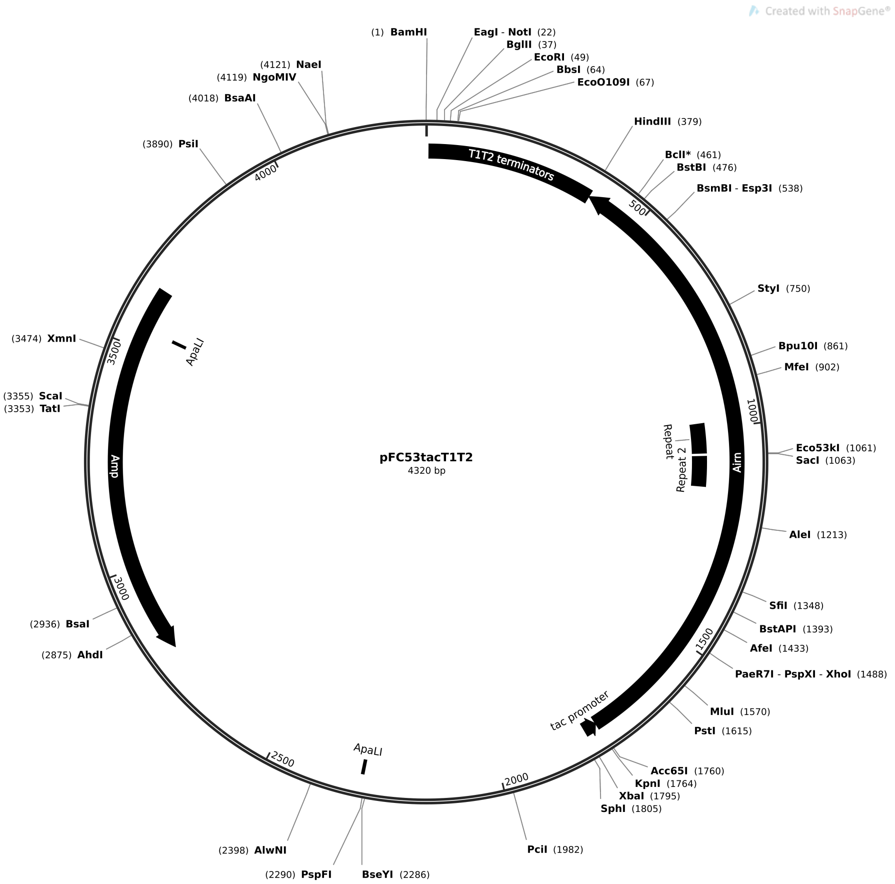

Figure 9: Map of pFC53tacT1T2.

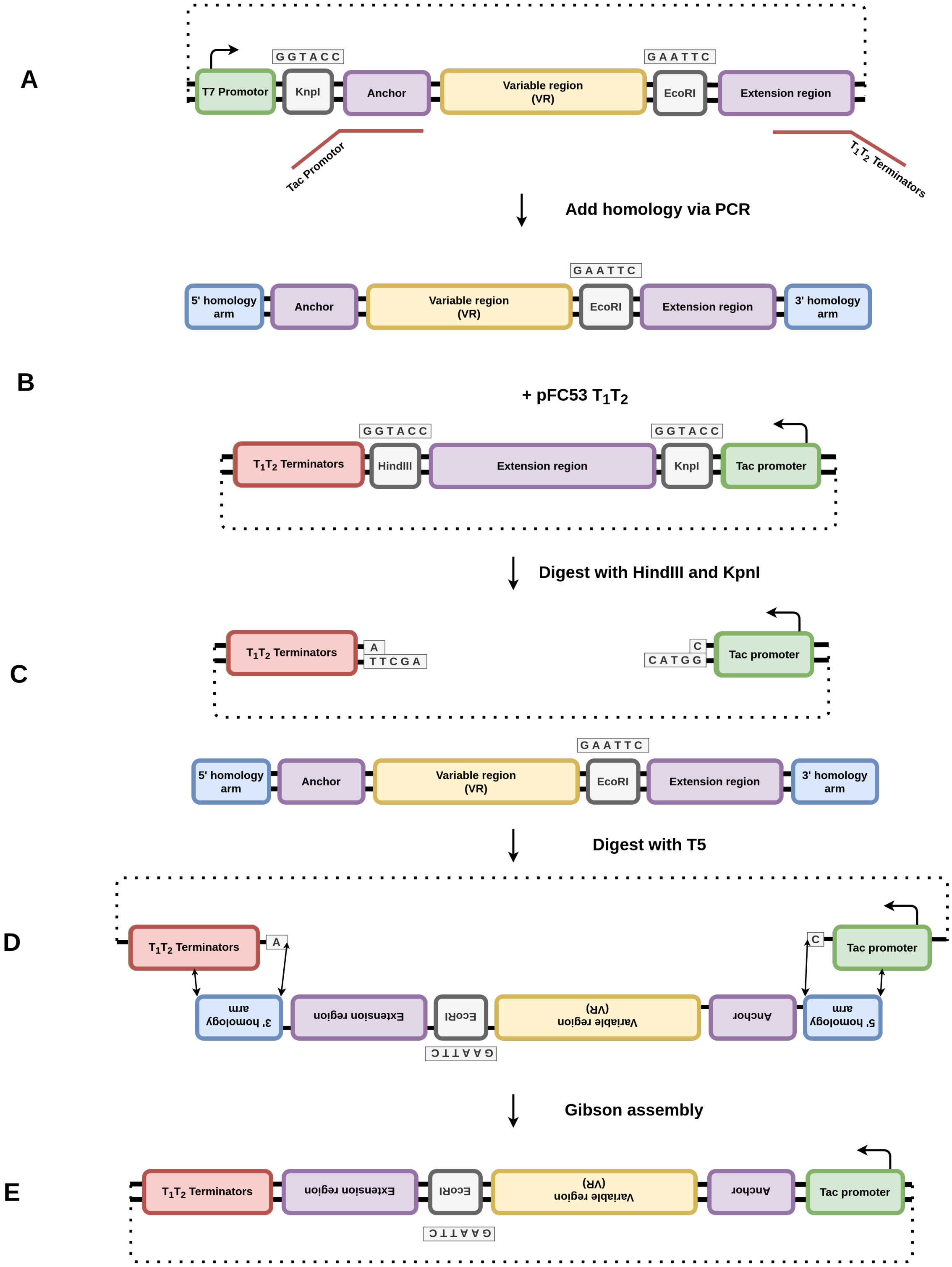

Figure 10: Diagram of pFC9 insertion series cloning strategy.

3.4 Tac termination series constructs

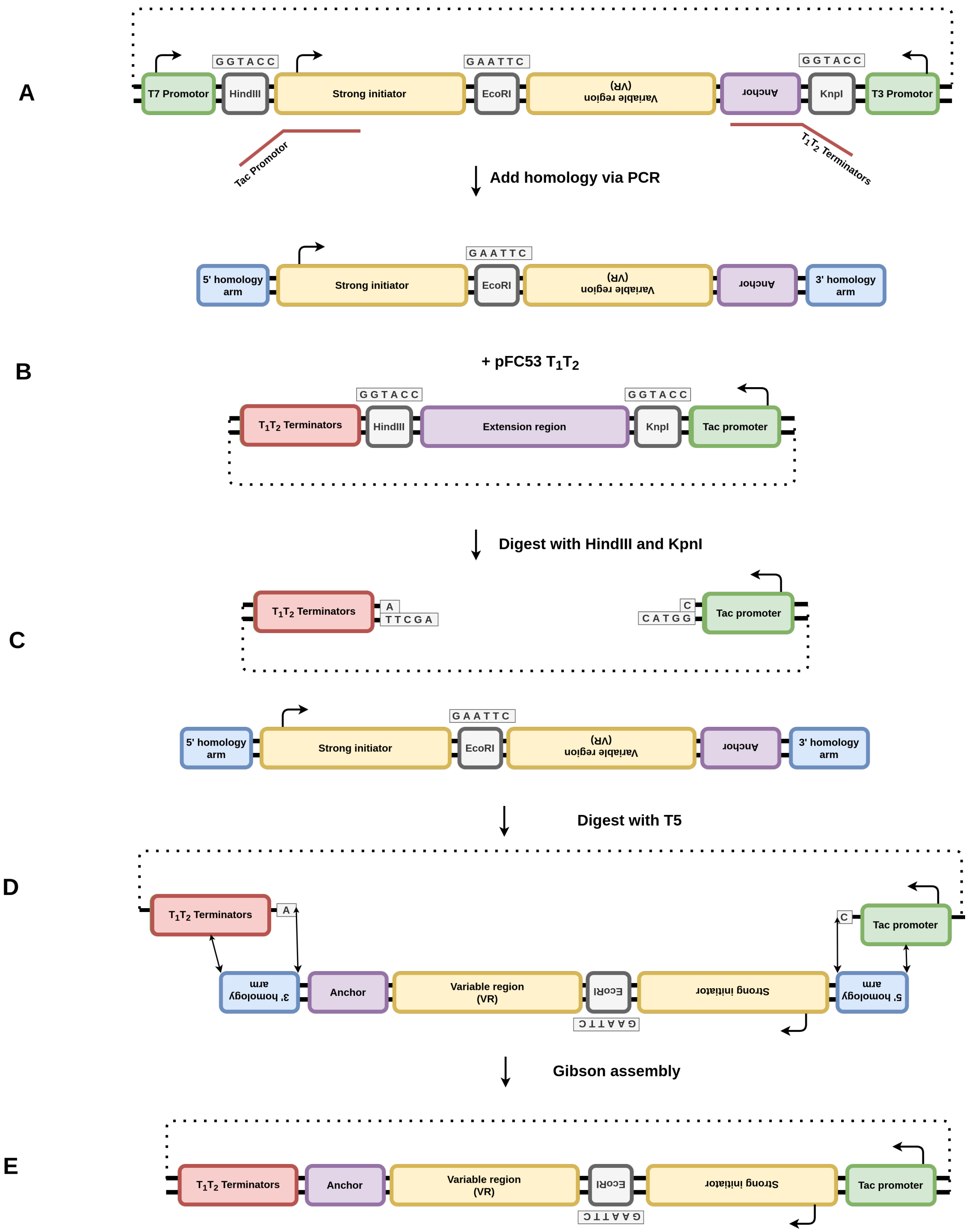

Figure 11: Diagram of pFC9 insertion series cloning strategy.

References