Appendix A

SUPPLEMENTAL INFORMATION FOR CHAPTER 2: SEQUENCE-DEPENDENT DYNAMICS OF SYNTHETIC AND ENDOGENOUS RSSS IN V(D)J RECOMBINATION

A.1 Experimental methods

A.1.1 Microscopy components and configuration

All TPM experiments were performed using two Olympus IX71 inverted microscopes with brightfield illumination. Experiments were run in parallel where one microscope imaged a flow cell containing DNA without any RSSs while the other microscope collected data on DNA strands containing the fixed 23RSS sequence and the 12RSS under consideration. Initially, one microscope (Olympus IX73) was outfitted with a 100x objective while another (Olympus IX73) had a 60x objective with a 1.6x magnifier. Both microscopes used Basler A602f-2 cameras. Partway through the study, each microscope was upgraded to larger fields of view for more data-collection by outfitting the hardware with a 60x objective (Olympus) and a 1920-pixel×1200-pixel monochromatic camera with a global shutter (Basler acA1920-155um). The camera is configured in an in-house Matlab image acquisition script to acquire images at a frame-rate of 30 Hz. Each optical set-up is calibrated to relate DNA of lengths ranging from 300 bp to 3000 bp to the root mean squared distance (RMSD) of their tethered beads.

A.1.2 TPM preparation

A schematic of the tethered bead assembly process as discussed in the Materials and Methods of the manuscript is shown in Fig. A.1. All buffers and assembly components are added to the flow cells by gravity flow. After anti-digoxigenin has coated the coverslip surface, flow cell chambers are washed twice with TPM assembly buffer containing 20 mM Tris-HCl (pH 8.0), 130 mM KCl, 2 mM $MgCl_2$, 0.1 mM EDTA, 0.1 mM DTT, 20 µg/mL acetylated bovine serum albumin (BSA), and 3 mg/mL casein. Once washed, DNA tethers are added and diluted in the TPM assembly buffer to a concentration of roughly 2.5 pM. The tethers are allowed to incubate within the cell for 15 minutes, allowing for the digoxigenin-functionalized ends of tethers to attach to

the anti-digoxigenin-coated coverslip. Unbound excess DNA is then removed from the flow cell and custom-ordered streptavidin-coated beads (Bangs Labs) are added to the flow cells, binding the DNA at the biotin ends, and left to incubate for three minutes before flushing excess beads from system. The prepared flow cell chamber is then equilibrated with RAG reaction buffer containing 25 mM Tris-HCl (pH 7.6), 75 mM KCl, 0.05% glyercol, 1 mM DTT, 30 mM potassium acetate, 2.5 mM $MgCl_2$, 5% DMSO and 100 µg/mL acetylated BSA for TPM experiments involving nicking or else the same buffer except with $CaCl_2$ in place of and at the same concentration as $MgCl_2$ for RAG-RSS interactions in the absence of DNA nicking.

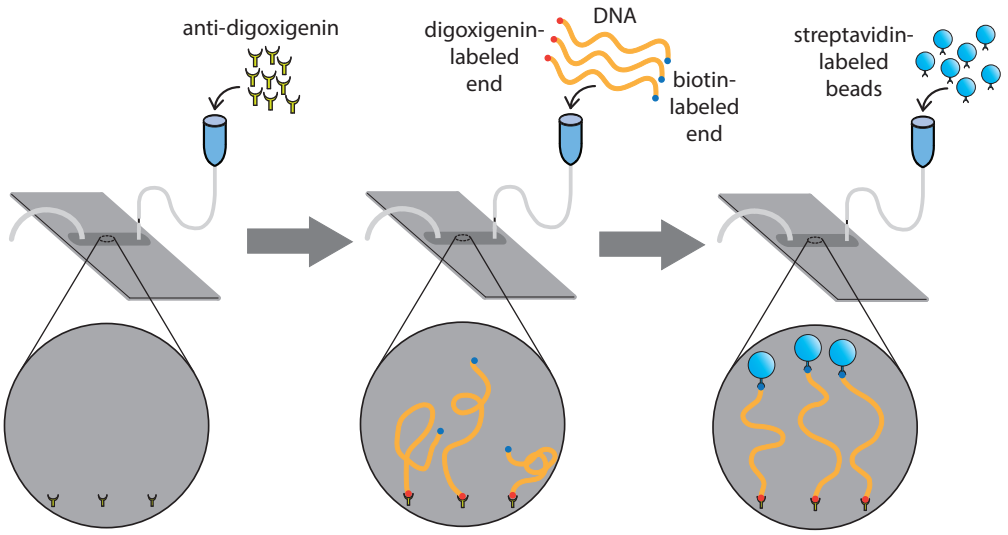


Figure A.1: **Tethered bead preparation process.** Tethered beads are first assembled by adding anti-digoxigenin from Sigma-Aldrich into the flow cell chamber by gravity flow and left to incubate for at least two hours. The fluid is then displaced from the chamber by washing in TPM assembly buffer and introducing DNA tethers containing the desired 12RSS and a constant 23RSS. Unbound DNA tethers are then flushed out and streptavidin-coated beads are introduced to the flow cell. Once the tethered beads have been assembled, chambers are equilibrated with buffer used to study RAG-RSS reaction.

A.1.3 Image processing

Image processing is performed on a field of view in the same manner established by Han *et al.* [1, 2]. After acquiring 60 images over two seconds, beads are identified by setting an intensity threshold before filtering over object sizes. Smaller regions of interest (ROIs) are drawn around each marker identified as

a bead. After initial processing, an additional 120 images over four seconds are acquired and processed by determining intensity-weighted center of masses of beads. The radial root mean squared displacement (RMSD) of the bead position is then determined using the 120 images and compared to the calibration curve based on the expected length of the DNA. Beads are accepted if their RMS values correspond to DNA lengths within 100 bp of their actual lengths for the paired complex assays ($l_{DNA} \approx 2900$ bp). Beads are then further processed to examine their symmetry of motion. After the correlation matrix for the bead position over the 120 frames is obtained, the eigenvalues of the matrix are then obtained, which yield the lengths of the major and minor axes of the range of motion of the bead. If the square root of the ratios of the maximum eigenvalue over the minimum eigenvalue is greater than 1.1, then the asymmetry of the motion is considered to be due to the bead being tethered to multiple DNA strands and is therefore rejected. The remaining beads are kept for data acquisition.

Feedback of the RMSD values of the bead center are obtained during experimentation using a Gaussian filtered by applying an 8-second (240 frame) standard deviation, as done for the post-acquisition processing. To correct for drift in the bead position, often due to the slow unidirectional motion of the microscope stage, the raw data are filtered through a first-order Butterworth filter with a cutoff frequency of 0.05 Hz. All ROI-binned image files can be downloaded from the CaltechDATA research repository under the DOI:10.22002/D1.1288. All code used to analyze these images can be found on the paper website or the paper GitHub repository (DOI:10.5281/zenodo.346571).

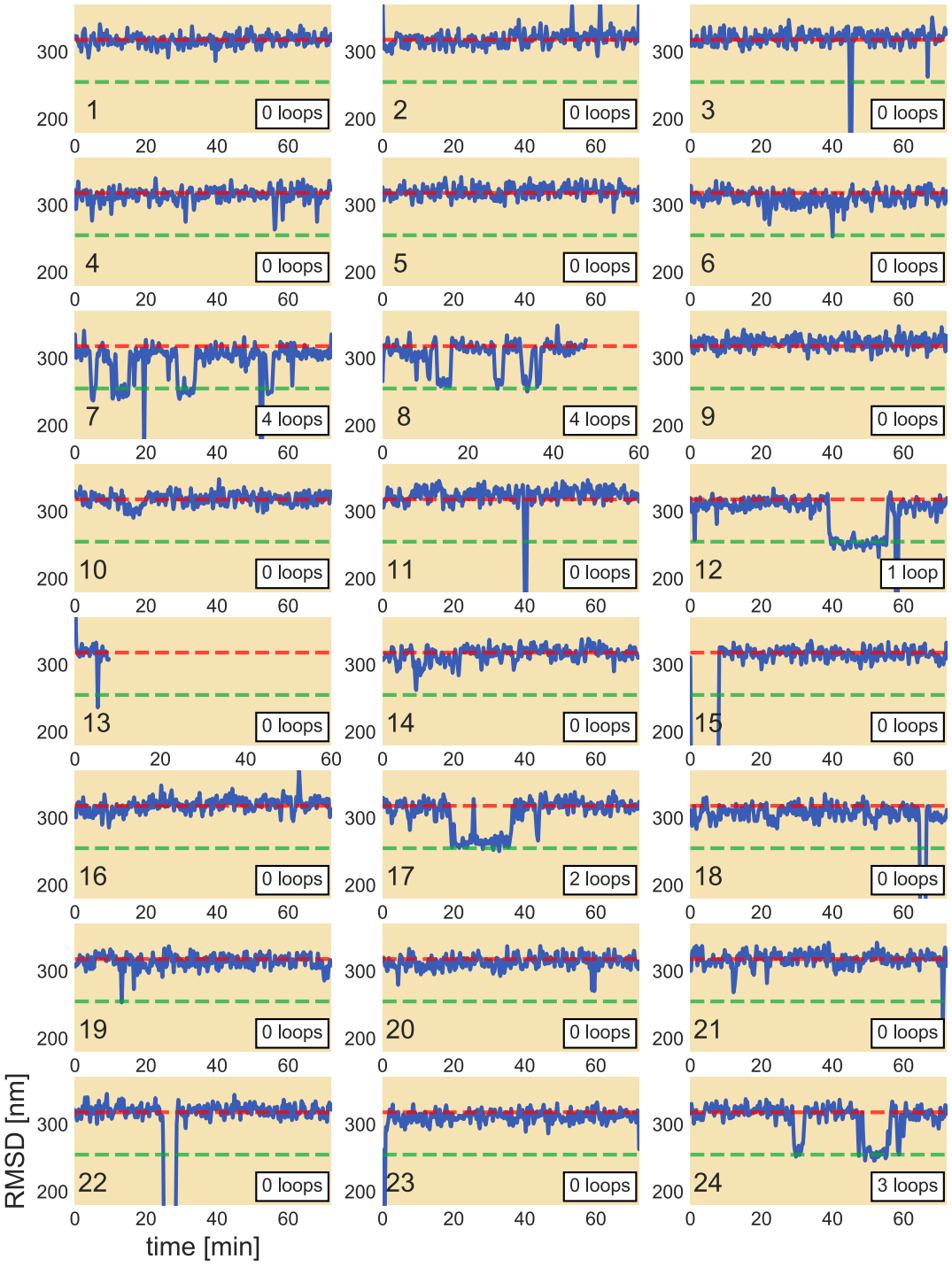
A.2 Data analysis: Extracting all relevant information from bead traces

All of the data reported and used in our results come solely from analyzing the RMSD as a function of time for each individual bead, hereafter called the "bead traces." This source must be further filtered in order to remove beads that passed through the initial image processing steps but still exhibit spurious behaviors, such as sticking to the glass surface or multiple beads falling into the same ROI and confounding the image processing. For each bead, the number of loops formed, the dwell time of each looped state, and whether the loop reverted to the unlooped state or was cleaved by RAG are then extracted and further analyzed through the bootstrapping method for the looping frequency

confidence interval, the Bayesian analysis to obtain our posterior distributions of the cutting probability and the dwell time distributions for our analysis on kinetics of leaving the paired complex state.

A.2.1 Selecting beads for further analysis

Bead selection criteria after preprocessing is applied in the same manner as described elsewhere [1–4]. After correcting for various systematic errors of the experiment, such as slow stage drift, we further smooth the RMSD values of the bead at each instance by applying a Gaussian filter with a -3 dB frequency of 0.0166 Hz corresponding to an 8-second standard deviation. Beads are then manually filtered based upon their RMSD trajectories both before and after introducing RAG and HMGB1 accompanied by 4-second movies of the motion of the bead. Tethers that show multiple attached beads are removed due to a larger variance in the RMSD trajectories for a given state. These beads can also be viewed through a software that shows the raw images at a defined time of the experiment. Furthermore, beads whose traces in the absence of protein lie below the expected RMSD value are considered to be a shorter DNA length than expected or an improperly tethered DNA strand and are also rejected. All other bead trajectories are tracked, as shown with the example set of trajectories from one replicate involving a DNA construct containing the C-to-T deviation at heptamer position 3 of the V4-57-1 (reference) 12RSS in Fig. A.2, until one of four outcomes occurs: 1) RAG cleaves the DNA, causing a sharp increase in RMSD past the tether point and can be observed with the bead diffusing from the ROI (shown for beads 26 and 39). 2) the bead sticks to the glass slide for longer than a few minutes or 3) another bead enters the cropped region enclosing the studied bead due to stage drift that has not been correct, with one of the two outcomes resulting in the truncation of trajectories as in beads 8, 13, 30, or 37. Or, as is also common, 4) the experiment ends, which typically runs for at least one hour of acquisition, without any of these outcomes. In cases where at least one bead reports the looped state at the hour mark without reporting a fate (not shown in this dataset), data acquisition continues until those beads report either unlooping, are untethered, or do not report a fate after roughly 15 minutes of the PC state persisting. The results of one TPM assay, performed with the C-to-T mutation at heptamer position 3 with 39 beads, are shown in Fig. A.2.


RMSD [nm]
time [min]
1 0 loops
2 0 loops
3 0 loops
4 0 loops
5 0 loops
6 0 loops
7 4 loops
8 4 loops
9 0 loops
10 0 loops
11 0 loops
12 1 loop
13 0 loops
14 0 loops
15 0 loops
16 0 loops
17 2 loops
18 0 loops
19 0 loops
20 0 loops
21 0 loops
22 0 loops
23 0 loops
24 3 loops

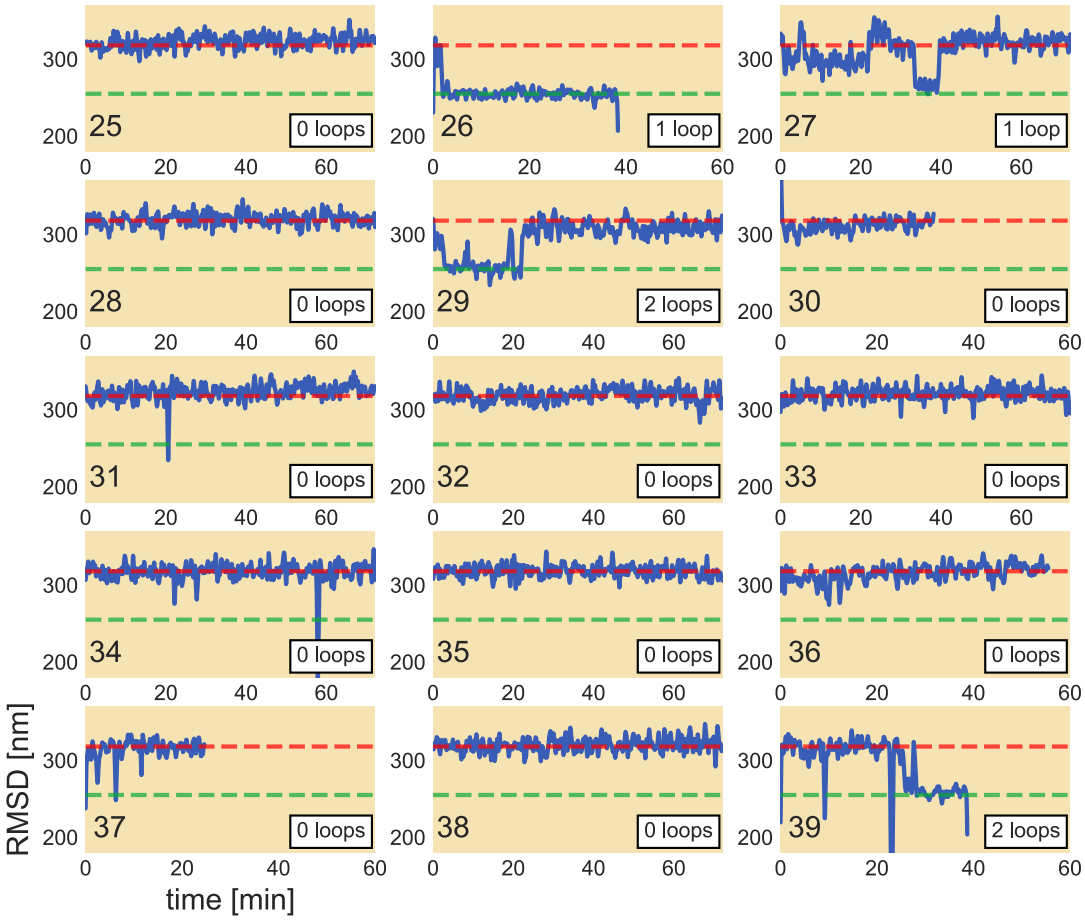


Figure A.2: **Sample bead trajectories for beads that have passed all filters in one replicate.** DNA construct contained the C-to-T alteration at heptamer position 3 of the V4-57-1 (reference) 12RSS. Number in lower left of each bead trajectory denotes bead number. Number of loops denoted in the white box to the lower right of each plot denotes number of loops that the TPM analysis software identifies. Red dashed line shows the empirically-measured root-mean-squared displacement (RMSD) for unlooped DNA length while the green dashed line shows the expected RMSD upon paired complex formation based on the empirically-measured unlooped DNA RMSD. Trajectories where beads reporting a paired complex state stop reporting trajectories, as in beads 26 and 39 are identified as cleaved DNA tethers. Bead trajectories that are truncated before the experiment is terminated but do not show the looped state at the end, such as beads 8, 13, 30, or 37, are not examined past the truncation point because the bead is passively lost from the DNA untethering from the anti-digoxigenin Fab molecule on the coverslip, another bead floating into the field of view and distorting the RMSD analysis, or the bead sticking to the coverslip.

Once the beads have been selected, they are entered into an analysis pipeline that identifies whether a bead is in the unlooped or paired complex state using three thresholding RMSD values at every given instance of data acquisition, as performed in [2]. Looped states are subject to the same 21-second deadfilter as in [2] to be considered as a bona fide paired complex state. In instances where a bead trajectory drops below the minimum RMSD threshold, which is often an indication of temporary adhesion of the bead to the glass slide, or above the maximum RMSD threshold, set due to other temporary aberrations in bead motion, the time that the bead trace spent outside of these bounds are split evenly between the state that the bead was in immediately before and after. With the states of the bead defined at each time point, we can coarse-grain the bead trajectory into the amount of time spent in the paired complex or unlooped states. This allows us not only to determine the lifetime of each paired complex formed but also the number of loops that were formed for a given bead reporter. In addition, all looped states are assigned a binary number based on whether they subsequently lead to unlooping (0) or to the bead untethering (1), the latter of which indicates DNA cleavage by RAG. Data on all beads kept by the TPM data acquisition code, including those that were manually filtered out during post-processing, are available on the CaltechDATA research data repository under the DOI:10.22002/D1.1288.

A.2.2 Bootstrapping looping frequency

As described in Chapter 2, we defined the looping frequency as the total number of observed PC events divided by the total number of beads observed over the experiment. It is tempting to simply repeat this calculation for each experimental replicate, average the results, and report a mean and standard error. However, the number of beads observed can vary greatly from one replicate to another. For example, one replicate may have 20 observed looping events among 100 observed beads, bringing the looping frequency to 0.2. Another replicate of the same RSS may have 0 observed looping events, but among only 10 beads in total, bringing the looping frequency to 0. We would want to apply a penalty to the second measurement as we observed far fewer beads than in the first replicate, however assigning that penalty is also not obvious. To further complicate this calculation, as shown in Fig. A.2, some beads in an experiment will never undergo a looping event while others will show multiple events, making a bead-by-bead calculation of the looping frequency more

challenging.

To address these challenges, we elect to compute and report the looping frequency as the total number of loops observed across all beads and experimental replicates, divided by the number of beads that were studied in total for that particular 12RSS. This metric, being bounded from 0 to ∞, accounts for the fact that for a given 12RSS, looping may occur many times. Furthermore, pooling the beads across replicates results in an appreciably large bead sample size, with the lowest sample size being greater than 80 beads and many RSSs having bead sample sizes in the hundreds.

In order to report a measure of the range of possible looping frequency values that could have been observed for a given RSS, we use the method of bootstrapping on our experimental dataset. In bootstrapping as applied here, we assume that the experimentally-obtained loop count distribution provides the best representation of the population distribution. We can then determine all possible ways we could have obtained the looping frequency by sampling from this empirical distribution. With this bootstrap-generated distribution of possible looping frequency values, we then calculate percentiles to provide confidence intervals on the looping frequency for comparison against the measured looping frequency. To see this in action, suppose our dataset on a particular RSS and salt condition contains N tracked beads across all replicates, with bead i reporting l_i loops. Our measured looping frequency f_{meas} would be $\frac{\sum_i l_i}{N}$. With bootstrapping, we can then determine our confidence interval on the measurement f_{meas} given the bead dataset we obtained with TPM by following the general procedure:

1. Randomly draw N different beads from the dataset of N beads with replacement. This means that the same bead can be drawn multiple times.

2. Sum the total number of loops observed among this collection of N beads and divide by N to get a bootstrap replicate of the looping frequency, $f_{\text{bs},1}$.

3. Repeat this procedure many times. In our case, we obtain 10^6 bootstrap replicates of the looping frequency.

4. For a confidence percentage P, determine the $(50 - \frac{P}{2})^{\text{th}}$ and $(50 + \frac{P}{2})^{\text{th}}$ percentiles from the generated list of 10^6 bootstrapped calculations of the looping frequency.

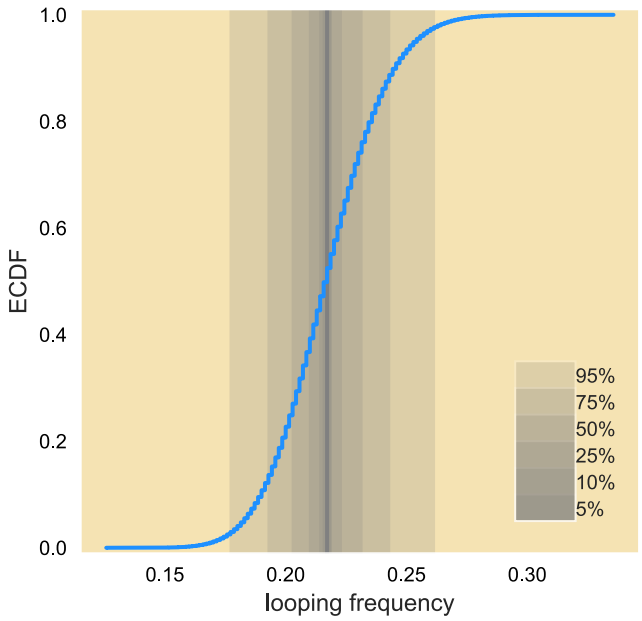


Figure A.3: **Bootstrapped looping frequency and confidence intervals for the V4-57-1 reference sequence.** Empirical CDFs of the bootstrapped looping frequency with 5%, 10%, 25%, 50%, 75% and 95% confidence intervals as represented by the color bar.

As an example, we demonstrate this bootstrap method on the V4-57-1 12RSS, which we also refer to as the reference sequence for our synthetic RSS study. Through TPM, we had tracked 700 beads, each reporting some number of loops l_i. As a result, we draw 700 beads from this dataset with replacement in order to calculate a bootstrap replicate of the looping frequency. We repeat this 10^6 times and obtain the result in Fig. A.3. Although we report the 95% confidence interval in the manuscript, we also offer shades of the 5%, 10%, 25%, 50% and 75% confidence intervals on our website.

A.2.3 Bayesian analysis on probability of cuts

Bayesian analysis on cutting probability is applied in a similar manner to [5]. For a given RSS substrate, to obtain the probability that RAG cuts a paired complex, p_{cut}, we construct a probability density function for p_{cut} conditioned on our data. In this case, our data for each RSS examined is the total number

of loops we observed in TPM, n_{loops}, and the number of loops that were cut, n_{cuts}, so $n_{\text{cuts}} \leq n_{\text{loops}}$. In short, we wish to determine the probability of p_{cut} conditional on n_{loops} and n_{cuts}, or, written concisely, as $P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}})$. Bayes' Theorem tells us that

$$P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}})P(n_{\text{loops}}, n_{\text{cuts}}) = P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}})P(n_{\text{loops}}, p_{\text{cut}}). \quad \text{(A.1)}$$

On the lefthand side Eq. A.1, $P(n_{\text{loops}}, n_{\text{cuts}})$ is the probability of n_{loops} loops and n_{cuts} cut loops, $P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}})$ is the probability that RAG cuts n_{cuts} loops conditional on the n_{loops} total loops examined and the probability that RAG cuts a given loop p_{cut}. $P(n_{\text{loops}}, p_{\text{cut}})$ is the probability of getting n_{loops} total loops and that RAG has a cut probability p_{cut} for the RSS. A rearrangement of the equation shows that

$$P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}}) = \frac{P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}})P(n_{\text{loops}}, p_{\text{cut}})}{P(n_{\text{loops}}, n_{\text{cuts}})}. \quad \text{(A.2)}$$

We can further simplify this equation by noting that the probability of getting n_{loops} loops and a cut probability p_{cut} are independent values. This is evident from the fact that we could have carried out more TPM experiments and in principle p_{cut} should not change from increasing the sample size of loops observed. Thus,

$$P(n_{\text{loops}}, p_{\text{cut}}) = P(n_{\text{loops}})P(p_{\text{cut}}). \quad \text{(A.3)}$$

Furthermore, we can further simplify the probability function in the denominator from noticing that the probability of having n_{loops} total loops and n_{cuts} loops that cut can be broken down into the product of the probability of having n_{cuts} cuts given n_{loops} total loops times the probability of having n_{loops} total loops to begin with, or

$$P(n_{\text{loops}}, n_{\text{cuts}}) = P(n_{\text{cuts}}|n_{\text{loops}})P(n_{\text{loops}}). \quad \text{(A.4)}$$

Inserting equations A.3 and A.4 into equation A.2 gives us

$$\begin{aligned} P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}}) &= \frac{P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}})P(n_{\text{loops}})P(p_{\text{cut}})}{P(n_{\text{cuts}}|n_{\text{loops}})P(n_{\text{loops}})}, \\ &= \frac{P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}})P(p_{\text{cut}})}{P(n_{\text{cuts}}|n_{\text{loops}})}. \end{aligned} \quad \text{(A.5)}$$

We wish to determine the conditional function on the left of Eq. A.5, which we will term our posterior distribution. Here, we construct our posterior distribution from inputting the probabilities on the righthand side of the equation.

We first determine $P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}})$. This conditional probability function is the probability that we observe n_{cuts} loops cut considering we observe n_{loops} loops forming and if the paired complex has a probability of cutting p_{cut}. Here, we would expect that this is similar to flipping a biased coin n_{loops} times and seeing how many instances heads comes up when the true value of the coin coming up heads is p_{cut}. In this case, we expect this conditional probability to be binomially distributed:

$$P(n_{\text{cuts}}|n_{\text{loops}}, p_{\text{cut}}) = \frac{n_{\text{loops}}!}{n_{\text{cuts}}!(n_{\text{loops}} - n_{\text{cuts}})!}(p_{\text{cut}})^{n_{\text{cuts}}}(1-p_{\text{cut}})^{n_{\text{loops}}-n_{\text{cuts}}}. \quad \text{(A.6)}$$

Next, we would like to determine $P(p_{\text{cut}})$. This is our prior distribution and, because this probability function is not conditioned on any data, this distribution function simply comes from our *a priori* knowledge of p_{cut} independent of the data we have in hand. Here, we choose to say that the only knowledge we have of this parameter is that it, like all probabilities, is bounded between zero and one. We assume that p_{cut} can take any value between zero and one equally. Thus,

$$P(p_{\text{cut}}) = \begin{cases} 1 & 0 \leq p_{\text{cut}} \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad \text{(A.7)}$$

Finally, we need to determine the probability that n_{cuts} loops cut given n_{loops} observed loops. This probability is only conditioned on n_{loops} and not p_{cut}, so we can say that n_{cuts} can take on any integer value between 0 and n_{loops}, inclusive. Thus, we have a discrete uniform distribution:

$$P(n_{\text{cuts}}|n_{\text{loops}}) = \frac{1}{n_{\text{loops}} + 1}. \quad \text{(A.8)}$$

By assembling equations A.6, A.7 and A.8 into equation A.5, we get that

$$P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}}) = \frac{(n_{\text{loops}} + 1)!}{n_{\text{cuts}}!(n_{\text{loops}} - n_{\text{cuts}})!}(p_{\text{cut}})^{n_{\text{cuts}}}(1-p_{\text{cut}})^{n_{\text{loops}}-n_{\text{cuts}}}. \quad \text{(A.9)}$$

With the posterior distribution in hand, we compute the most probable value of p_{cut} by determining where the derivative of the posterior distribution with

respect to p_{cut} is 0. For ease of calculation, we will take the logarithm of the posterior distribution and derive with respect to p_{cut}:

$$\begin{aligned}\ln[P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}})] &= \ln\Big[\frac{(n_{\text{loops}}+1)!}{n_{\text{cuts}}!(n_{\text{loops}}-n_{\text{cuts}})!}\Big] + n_{\text{cuts}}\ln(p_{\text{cut}}) \\ &\quad + (n_{\text{loops}}-n_{\text{cuts}})\ln(1-p_{\text{cut}}), \\ \frac{d\ln[P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}})]}{d\,p_{\text{cut}}}\Big|_{p^*_{\text{cut}}} &= \frac{n_{\text{cuts}}}{p^*_{\text{cut}}} - \frac{n_{\text{loops}}-n_{\text{cuts}}}{1-p^*_{\text{cut}}} = 0. \end{aligned} \tag{A.10}$$

Eq. A.10 then tells us that

$$p^*_{\text{cut}} = \frac{n_{\text{cuts}}}{n_{\text{loops}}}. \tag{A.11}$$

To calculate the variance of p_{cut}, we make the assumption that $n_{\text{loops}} \gg 1$ and look to center about the most probable value, p^*_{cut}. With this assumption, we will approximate the posterior distribution as a Gaussian distribution. In order to see this in action, we will define $x \equiv p - p^*_{\text{cut}}$. Then Eq. A.12 becomes

$$P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}}) = \frac{(n_{\text{loops}}+1)!}{n_{\text{cuts}}!(n_{\text{loops}}-n_{\text{cuts}})!}(p^*_{\text{cut}}+x)^{n_{\text{cuts}}}(1-p^*_{\text{cut}}-x)^{n_{\text{loops}}-n_{\text{cuts}}}. \tag{A.12}$$

We also invoke the rule that $ln\, n_{\text{cuts}}! \approx n_{\text{cuts}}\ln n_{\text{cuts}} - n_{\text{cuts}} + \frac{1}{2}\ln[2\pi n_{\text{cuts}}]$. We

can then determine the prefactor of the posterior distribution. Specifically,

$$
\begin{aligned}
\frac{(n_{\text{loops}}+1)!}{n_{\text{cuts}}!(n_{\text{loops}}-n_{\text{cuts}})!} &= \exp\{\ln[(n_{\text{loops}}+1)!] - \ln n_{\text{cuts}}! - \ln[(n_{\text{loops}}-n_{\text{cuts}})!]\}, \\
&\approx \exp\{(n_{\text{loops}}+1)\ln(n_{\text{loops}}+1) - (n_{\text{loops}}+1) + \frac{1}{2}\ln[2\pi\,(n_{\text{loops}}+1)] \\
&\qquad - n_{\text{cuts}}\ln n_{\text{cuts}} + n_{\text{cuts}} - \frac{1}{2}\ln(2\pi n_{\text{cuts}}) \\
&\qquad - (n_{\text{loops}}-n_{\text{cuts}})\ln(n_{\text{loops}}-n_{\text{cuts}}) \\
&\qquad + (n_{\text{loops}}-n_{\text{cuts}}) \\
&\qquad - \frac{1}{2}\ln[2\pi(n_{\text{loops}}-n_{\text{cuts}})]\}, \\
&\approx \exp\Big\{(n_{\text{loops}}+1)\Big[\ln n_{\text{loops}} + \ln\Big(1+\frac{1}{n_{\text{loops}}}\Big)\Big] \\
&\qquad - 1 - n_{\text{cuts}}\ln n_{\text{cuts}} \\
&\qquad - (n_{\text{loops}}-n_{\text{cuts}})\Big[\ln n_{\text{loops}} + \ln\Big(1-\frac{n_{\text{cuts}}}{n_{\text{loops}}}\Big)\Big] \\
&\qquad + \frac{1}{2}\ln\Big[\frac{n_{\text{loops}}+1}{2\pi\, n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}\Big]\Big\}, \\
&\approx \exp\Big\{(n_{\text{loops}}+1)(\frac{1}{n_{\text{loops}}} + \frac{1}{2n_{\text{loops}}^2}) - 1 - n_{\text{cuts}}\ln n_{\text{cuts}} + n_{\text{cuts}}\ln n_{\text{loops}} \\
&\qquad - (n_{\text{loops}}-n_{\text{cuts}})\ln(1-p^*_{\text{cut}}) \\
&\qquad + \frac{1}{2}\ln\Big[\frac{n_{\text{loops}}^3}{2\pi\, n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}\Big]\Big\}, \\
&\approx \frac{1}{\sqrt{2\pi\frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}}\exp\Big\{-n_{\text{cuts}}\ln\Big(p^*_{\text{cut}}\Big) - n_{\text{loops}}(1-p^*_{\text{cut}})\ln(1-p^*_{\text{cut}})\Big\}.
\end{aligned}
\tag{A.13}
$$

Here, we make simplifying assumptions, such as that $n_{\text{loops}}+1 \approx n_{\text{loops}}$ and Taylor expansions for $\frac{1}{n_{\text{loops}}}$.

With the prefactor taken care of, we can rework the entire posterior distribution.

$$
\begin{aligned}
P(p_{\text{cut}}|n_{\text{loops}}, n_{\text{cuts}}) \approx & \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ - n_{\text{cuts}} \ln\Big(p^*_{\text{cut}}\Big) \\
& \qquad - n_{\text{loops}}(1-p^*_{\text{cut}}) \ln(1-p^*_{\text{cut}}) \\
& \qquad + n \ln(p^*_{\text{cut}} + x) \\
& \qquad + (n_{\text{loops}} - n_{\text{cuts}}) \ln(1 - p^*_{\text{cut}} - x) \Big\}, \\
\approx & \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ -n_{\text{cuts}} \ln(p^*_{\text{cut}}) - n_{\text{loops}}(1-p^*_{\text{cut}}) \ln(1-p^*_{\text{cut}}) \\
& \qquad + n_{\text{cuts}} \Big[\ln(p^*_{\text{cut}}) + \ln(1 + \frac{x}{p^*_{\text{cut}}}) \Big] \\
& \qquad + (n_{\text{loops}} - n_{\text{cuts}}) \times \\
& \qquad \Big[\ln(1-p^*_{\text{cut}}) + \ln(1 - \frac{x}{1-p^*_{\text{cut}}}) \Big] \Big\}, \\
\approx & \frac{1}{\sqrt{2\pi \frac{n(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ n_{\text{cuts}} \Big[\ln(1 + \frac{x}{p^*_{\text{cut}}}) \Big] \\
& \qquad + (n_{\text{loops}} - n_{\text{cuts}}) \Big[\ln(1 - \frac{x}{1-p^*_{\text{cut}}}) \Big] \Big\}, \\
\approx & \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ n_{\text{cuts}} \Big[\frac{x}{p^*_{\text{cut}}} - \frac{x^2}{2{p^*_{\text{cut}}}^2} \Big] \\
& \qquad + (n_{\text{loops}} - n_{\text{cuts}}) \Big[- \frac{x}{1-p^*_{\text{cut}}} - \frac{x^2}{2(1-p^2_{\text{cut}})^2} \Big] \Big\}, \\
\approx & \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ n_{\text{loops}}\, x - n_{\text{cuts}} \frac{x^2}{2{p^*_{\text{cut}}}^2} \\
& \qquad - n_{\text{loops}}\, x - (n_{\text{loops}} - n_{\text{cuts}}) \frac{x^2}{2(1-p^*_{\text{cut}})^2} \Big\}, \\
\approx & \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ - n_{\text{cuts}} \frac{x^2}{2{p^*_{\text{cut}}}^2} - (n_{\text{loops}} - n_{\text{cuts}}) \frac{x^2}{2(1-p^*_{\text{cut}})^2} \Big\}, \\
\approx & \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ - n_{\text{loops}} \frac{x^2}{2p^*_{\text{cut}}} - n_{\text{loops}} \frac{x^2}{2(1-p^*_{\text{cut}})} \Big\},
\end{aligned}
$$

$$\approx \frac{1}{\sqrt{2\pi \frac{n(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ -\frac{n_{\text{loops}}\, x^2}{2}\Big(\frac{1}{p^*_{\text{cut}}} + \frac{1}{1-p^*_{\text{cut}}}\Big)\Big\},$$

$$\approx \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ -\frac{n_{\text{loops}}\, x^2}{2}\Big(\frac{1}{p^*_{\text{cut}}(1-p^*_{\text{cut}})}\Big)\Big\},$$

$$\approx \frac{1}{\sqrt{2\pi \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}}} \exp\Big\{ -\frac{(p-p^*_{\text{cut}})^2}{2\Big[\frac{n_{\text{cuts}}\,(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}\Big]}\Big\}. \tag{A.14}$$

Eq. A.14 tells us that, not only is this Gaussian approximation centered at the most probable value of $p_{\text{cut}} = p^*_{\text{cut}}$, as we would expect, but also that the distribution has a variance of $\sigma^2 = \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}$. Thus, we report $p^*_{\text{cut}} = \frac{n_{\text{cuts}}}{n_{\text{loops}}}$ and $\sigma^2 = \frac{n_{\text{cuts}}(n_{\text{loops}}-n_{\text{cuts}})}{n_{\text{loops}}^3}$ in Fig. 2.5C and 2.6C of Chapter 2.

A.2.4 Significance testing of looping frequency, median PC lifetime, and cutting fraction

In Chapter 2, we represent particular point mutations and endogenous sequences demonstrating a statistically significant effect with a colored asterisk ($*$). In this section, we elaborate on our definition of what is deemed statistically significant and outline our computational approach.

Defining the null hypothesis

To assess statistical significance of our measurements, we sought to quantify the probability that the observation could be observed under the null hypothesis. For all quantities computed in this work (i.e. looping frequency, PC dwell time, and cutting probability), the null hypothesis was that the observed value of a quantity was drawn from the same distribution as the observed value for the reference (V4-57-1) 12RSS. For each 12RSS and quantity, we computed the probability that an observation generated under the null hypothesis could be at least as extreme as the actual observed quantity. This probability, often reported as a p-value, can be analytically computed using a variety of well-known statistical tests such as the Students' t-test, the Mann-Whitney U-test, and the unequal variance t-test [6]. However, due to the different definitions of the quantities of interest in this work, we used simulation through random number generation as a means to compute this probability.

For all quantities measured, we wished to compute a p-value for the experimental measurement. To do so, we defined a test statistic as the absolute value of the difference in the quantity of interest between a given sequence and that of the reference V4-57-1 12RSS,

$$\delta^{(\text{observation})} = |m_{\text{mutant}}^{(\text{observation})} - m_{\text{reference}}^{(\text{observation})}|. \tag{A.15}$$

Here, $m^{(\text{observation})}$ represents the observed quantity such as looping frequency, median PC dwell time, or the cutting probability computed from the data.

With this test statistic in hand, we reran the experiment *in silico* as follows:

1. Isolate the raw experimental measurements for a given mutant 12RSS and the reference 12RSS and compute the total number of measurements in each dataset, N_{mutant} and $N_{\text{reference}}$.

2. Combine all measurements from both datasets into a single dataset of size $N_{\text{mutant}} + N_{\text{reference}}$ and randomly shuffle the contents.

3. Take the first N_{mutant} entries of the shuffled vector and compute the quantity of interest, $m_{\text{mutant}}^{(\text{simulation})}$. Using the remaining values of the shuffled vector, compute the reference quantity of interest $m_{\text{reference}}^{(\text{simulation})}$.

4. Given these simulated values, compute the value of the test statistic

$$\delta^{(\text{simulation})} = \left| m_{\text{mutant}}^{(\text{simulation})} - m_{\text{reference}}^{(\text{simulation})} \right|. \tag{A.16}$$

5. Store the value of the test statistic in a vector $\vec{\delta}$ and return to step 2. Repeat this process for a total of $N_{\text{simulations}} = 10^7$ times.

6. From the vector of $N_{\text{simulations}}$ test statistic values, compute the p-value by dividing the total number of test values δ_i in the stored vector $\vec{\delta}$ that are greater than or equal to the empirically observed value $\delta^{(\text{observation})}$ and dividing by the 10^7 simulations, or

$$p\text{-value} = \frac{1}{N_{\text{simulations}}} \sum_{i=1}^{N_{\text{simulations}}} k \text{ where } k = \begin{cases} 1, \text{ if } \delta_i \geq \delta^{(\text{observation})} \text{ for } \delta_i \in \vec{\delta} \\ 0, \text{ otherwise.} \end{cases} \tag{A.17}$$

The computed p-values for each sequence and quantity used in this work can be seen in Fig. A.4. In practice, a measurement is determined to be statistically

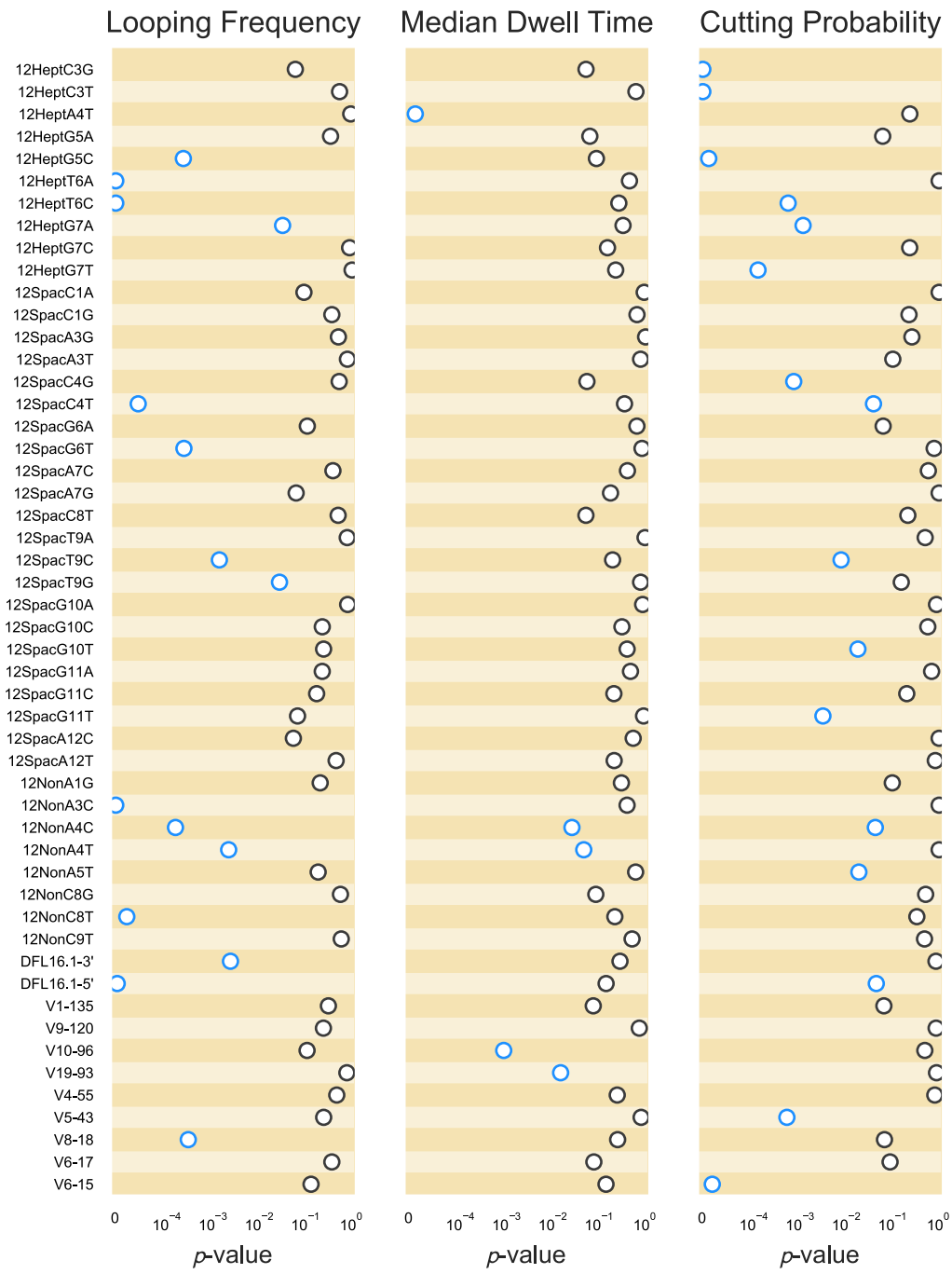


Figure A.4: **Null hypothesis significance testing of looping frequency, median PC lifetime, and cutting fraction for RSSs.** Blue circles denote p-values $p \leq 0.05$.

significant if the p-value is below a given threshold. This threshold is chosen to be the typically chosen 0.05 cutoff value, which indicates that under the null hypothesis, the probability of observing a value at least as extreme as the

experimental measurement is $\leq 5\%$. Measurements with p-value ≤ 0.05 are shown in blue in Fig. A.4.

A.2.5 Relation of looping frequency and cutting probability to bulk *in vitro* cleavage fraction

While we separated different steps of the RAG-RSS reaction through measurements of the looping frequency and cutting probability, we also wanted to know the fraction of DNA tethers that completed the cleavage phase of the reaction. This measurement is applicable to standard bulk *in vitro* cleavage assays where RAG and 12/23RSS-carrying DNA strands are mixed and allowed to carry out the cleavage reaction before running the results on a gel to observe the number of DNA strands cleaved and number that remain intact [7]. Using our TPM data, for each 12RSS construct we calculate the posterior distribution of the fraction of DNA tethers that are cleaved f_{cleaved} based on the total number of tethered beads n_{beads} and the total number of cut tethers n_{cuts},

$$P(f_{\text{cleaved}}|n_{\text{beads}}, n_{\text{cuts}}).$$

By Bayes' theorem,

$$\begin{aligned} P(f_{\text{cleaved}}|n_{\text{beads}}, n_{\text{cuts}})P(n_{\text{beads}}, n_{\text{cuts}}) &= P(n_{\text{cuts}}|n_{\text{beads}}, f_{\text{cleaved}})P(n_{\text{beads}}, f_{\text{cleaved}}), \\ P(f_{\text{cleaved}}|n_{\text{beads}}, n_{\text{cuts}}) &= \frac{P(n_{\text{cuts}}|n_{\text{beads}}, f_{\text{cleaved}})P(n_{\text{beads}}, f_{\text{cleaved}})}{P(n_{\text{beads}}, n_{\text{cuts}})}. \end{aligned} \tag{A.18}$$

Eq. A.18 can be simplified down in the same way as was done with the probability of cutting to yield an equation similar to Eq. A.5.

$$P(f_{\text{cleaved}}|n_{\text{beads}}, n_{\text{cuts}}) = \frac{P(n_{\text{cuts}}|n_{\text{beads}}, f_{\text{cleaved}})P(f_{\text{cleaved}})}{P(n_{\text{cuts}}|n_{\text{beads}})}, \tag{A.19}$$

where p_{cut} is replaced by f_{cleaved} and n_{loops} is replaced by n_{beads}. Each of the functions on the right-hand side of Eq. A.19 shares the same functional form as their counterparts in Eq. A.5: the likelihood function $P(n_{\text{cuts}}|n_{\text{beads}}, f_{\text{cleaved}})$ is also a binomial distribution; we have no prior knowledge of how f_{cleaved} might be distributed, so treat it as uniform from 0 to 1; and n_{cuts} can take on any integer value ranging from 0 (none of the tethers are cleaved) to n_{beads} (all of the tethers are cleaved), so this is also a discrete uniform distribution normalized

by the $n_{\text{beads}}+1$ possible outcomes. Assembling all of these functions into Eq. A.19 yields

$$P(f_{\text{cleaved}}|n_{\text{beads}}, n_{\text{cuts}}) = \frac{(n_{\text{beads}}+1)!}{n_{\text{cuts}}!(n_{\text{beads}}-n_{\text{cuts}})!}(f_{\text{cleaved}})^{n_{\text{cuts}}}(1-f_{\text{cleaved}})^{n_{\text{beads}}-n_{\text{cuts}}}. \tag{A.20}$$

Using the same derivation method as in Section A.2.3, we make a Gaussian approximation to compute the most probable value and standard deviation for f_{cleaved}:

$$f^*_{\text{cleaved}} = \frac{n_{\text{cuts}}}{n_{\text{beads}}}, \tag{A.21}$$

$$\sigma^2_{f_{\text{cleaved}}} = \frac{n_{\text{cuts}}\left(n_{\text{beads}}-n_{\text{cuts}}\right)}{n^3_{\text{beads}}}. \tag{A.22}$$

Fig. A.5 shows, from top to bottom, the looping frequency, cutting probability, and the bead cut fraction. The black dashed line and shaded region for each plot are the point and errorbar equivalent for the V4-57-1 (reference) 12RSS. Specifically, Fig. A.5 shows how the looping frequency and cutting probability can both contribute to limiting the fraction of cleaved DNA tethers. For example, 12NonA3C shows a low looping frequency relative to the reference 12RSS but a similar cutting probability, resulting in a lower bead cut fraction. 12SpacG6T has a higher looping frequency relative to the reference sequence for a comparable cutting probability, yielding a higher bead cut fraction than the reference. We also see that PC cutting probability can limit f_{cleaved}: Even though 12SpacG10T has a similar looping frequency to the reference sequence, the higher cutting probability causes a higher fraction of cleaved tethers. Both changes to heptamer position 3 show that a low cutting probability can abrogate DNA tether cleavage. Both looping frequency and cutting probability as decoupled measurements yield important insights into which processes in the RAG-RSS reaction help or hinder the completion of the cleavage phase in V(D)J recombination.

This observation can also be seen arithmetically. We had defined the looping frequency f_{looped} as the number of loops n_{loops} across all DNA tethers n_{beads}:

$$f_{\text{looped}} = \frac{n_{\text{loops}}}{n_{\text{beads}}}, \tag{A.23}$$

and the most probable value for p_{cut} is the fraction of loops that get cleaved:

$$p^*_{\text{cut}} = \frac{n_{\text{cuts}}}{n_{\text{loops}}}. \tag{A.24}$$

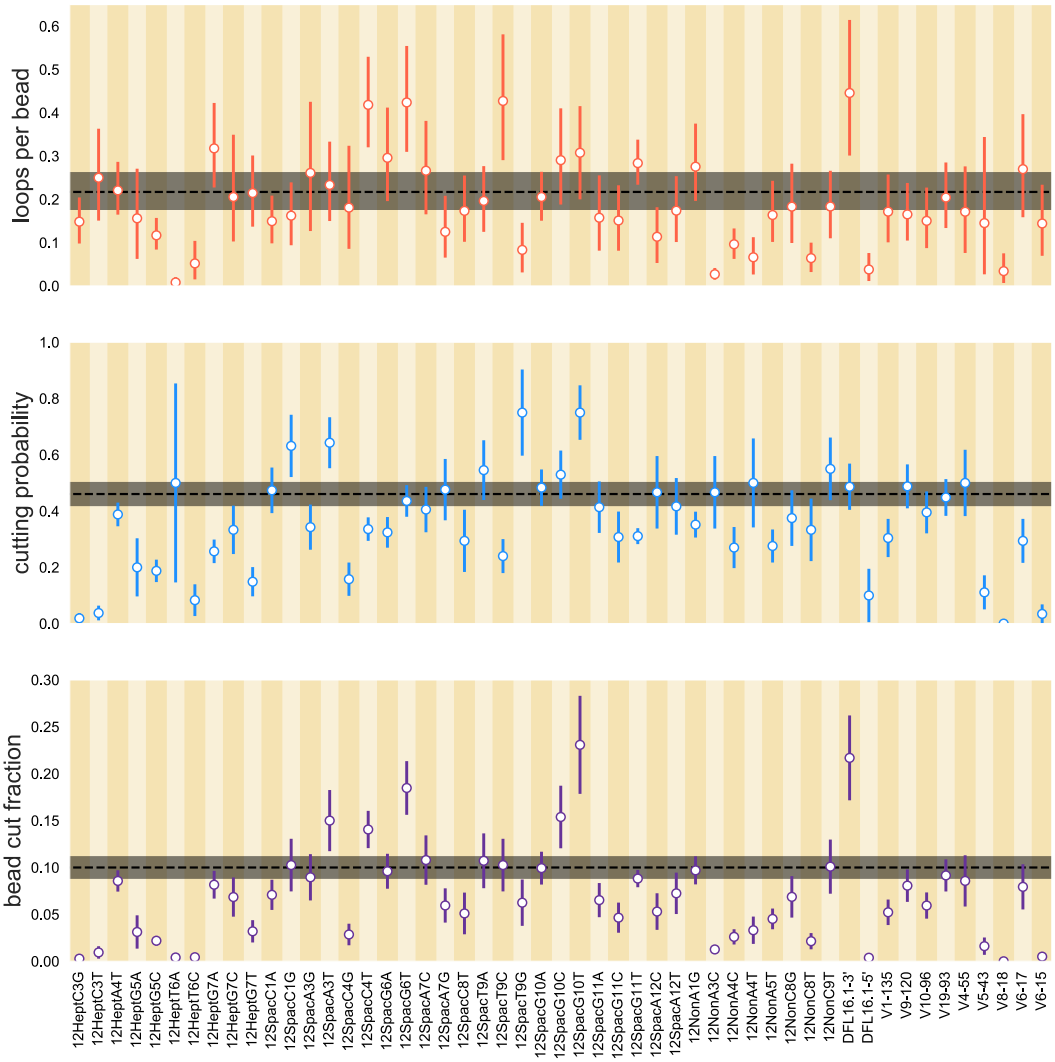


Figure A.5: **Figure stacking of looping frequency (top; red), cutting probability (middle; blue), and bead cut fraction (bottom; purple).** As in the manuscript, looping frequency is shown as the measured value and 95% confidence interval while cutting probability is shown as the most probable p_{cut} and one standard deviation. The bead cut fraction is similarly displayed to the p_{cut} with the most probable f_{cleaved} and one standard deviation. Black dashed line and grey shaded region in each plot corresponds to the measured or most probable value and confidence interval or standard deviation for the V4-57-1 (reference) 12RSS, respectively.

Multiplying both definitions of our metrics allows us to recover Eq. A.21:

$$\begin{aligned} f_{\text{looped}} \times p^*_{\text{cut}} &= \frac{n_{\text{loops}}}{n_{\text{beads}}} \times \frac{n_{\text{cuts}}}{n_{\text{loops}}}, \\ &= \frac{n_{\text{cuts}}}{n_{\text{beads}}}, \\ &= f^*_{\text{cleaved}}. \end{aligned} \tag{A.25}$$

Thus, we recover the relation between the tether cut fraction and the looping frequency and cutting probability, showing that the tether cut fraction will change linearly if one of the metrics is changed due to a change of 12RSS.

A.2.6 Modeling exit from the paired complex as a Poisson process

As discussed in Chapter 2, we attempted to model the kinetics of unlooping and exiting of the paired complex state. In the case of exit, we considered that every paired complex had one of two fates; either the DNA was cleaved and the observed tethered bead was lost or the paired complex dissociated, releasing the bead to its full-length tethered state. We consider these two fates as independent yet competing processes. Under the independence assumption, we can model each process individually as a Poisson process where the time of leaving the paired complex (either through cleavage or unlooping) is exponentially distributed. Mathematically, we can state that the probability of leaving the paired complex at time t_{leave} is defined as

$$P(t_{\text{leave}} \,|\, k_{\text{leave}}) = k_{\text{leave}} e^{-k_{\text{leave}} t_{\text{leave}}}, \tag{A.26}$$

where the leaving rate k_{leave} is defined as the sum of the two independent rates,

$$k_{\text{leave}} = k_{\text{cut}} + k_{\text{unloop}}. \tag{A.27}$$

Therefore, given a collection of paired complex dwell times t_{leave}, we can estimate the most-likely value for k_{leave} providing insight on whether exiting the paired complex can be modeled as a Poisson process.

Rather than reporting a single value, we can determine the probability distribution of the parameter k_{leave}. This distribution, termed the posterior distribution, can be computed by Bayes' theorem as

$$P(k_{\text{leave}} \,|\, t_{\text{leave}}) = \frac{P(t_{\text{leave}} \,|\, k_{\text{leave}}) P(k_{\text{leave}})}{P(t_{\text{leave}})}. \tag{A.28}$$

The posterior distribution $P(k_{\text{leave}} \,|\, t_{\text{leave}})$ defines the probability of a leaving rate *given* a set of measurements t_{leave}. This distribution is dependent on the likelihood of observing the dwell time distribution *given* a leaving rate, represented by $P(t_{\text{leave}} \,|\, k_{\text{leave}})$. All prior information we have about what the leaving rate *could* be is captured by $P(k_{\text{leave}})$ which is entirely independent of the data. The denominator in Eq. A.28 defines the probability distribution of the data marginalized over all values of the leaving rate. For our purposes, this term serves as a normalization constant and will be neglected.

We are now tasked with defining functional forms for the various probabilities enumerated in Eq. A.28. The likelihood already matches the definition in Eq. A.26, so we assign our likelihood as a simple exponential distribution parameterized by the leaving rate. Choosing a functional form for the prior distribution $P(k_{\text{leave}})$ is a much more subjective process. As such, we outline our thinking below.

As written in Eq. A.26, k_{leave} has dimensions of inverse time, meaning that particularly long-lived paired complexes will have $k_{\text{leave}} < 1$ whereas a sequence with unstable paired complexes will have $k_{\text{leave}} > 1$. As we remain ignorant of our data, we consider both of these extremes to be valid values for the leaving rate. However, this parameterization raises technical issues with estimating k_{leave} computationally. We sample the complete posterior using Markov chain Monte Carlo, a computational technique in which the posterior is explored via a biased random walk depending on the gradient of the local probability landscape. With such a widely constrained parameter, effectively sampling very small values of k_{leave} becomes more difficult than larger values. We can avoid this issue by reparameterizing Eq. A.26 in terms of the inverse leaving rate $\tau_{\text{leave}} = \frac{1}{k_{\text{leave}}}$ so that

$$P(t_{\text{leave}} \,|\, \tau_{\text{leave}}) = \frac{1}{\tau_{\text{leave}}} e^{t_{\text{leave}}/\tau_{\text{leave}}}. \tag{A.29}$$

Our parameter of interest now has dimensions of time and can be interpreted as the average life time of a paired complex or, more precisely, the waiting time for the arrival of a Poisson process.

While it is tempting to default to a completely uninformative prior for τ_{leave} to avoid introducing any bias into our parameter estimation, we do have some intuition for what the bounds of the value could be. For example, it is mathematically impossible for the leaving rate to be less than zero. We can also find it unlikely that the leaving rate is *exactly* zero as that would imply irreversible formation of the paired complex. We can therefore say that the value for the leaving rate is positive and can asymptotically approach zero. As we have designed the experiment to actually observe the entry and exit of the paired complex state, we can set a soft upper bound for the leaving rate to be the length of our typical experiment, 60 minutes. With these bounds in place, we can assign some probability distribution between them where it is impossible to equal zero and improbable but not impossible to exceed 60 minutes.

A good choice for such a distribution is an inverse Gamma distribution which has the form

$$P(\tau_{\text{leave}} \,|\, \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \frac{\beta^{\alpha}}{\tau_{\text{leave}}^{(\alpha+1)}} e^{-\beta/\tau_{\text{leave}}}, \tag{A.30}$$

where α and β correspond to the number of arrivals of a Poisson process and their rate of arrival, respectively. Given that only one arrival is necessary to exit a paired complex, we choose α to be approximately equal to 1 and β to be approximately 10. This meets our conditions described previously of asymptotically approaching zero and rarely exceeding 60 minutes.

Combining Eq. A.29 and Eq. A.30 yields the complete posterior distribution. We sampled this distribution for each RSS independently using Markov chain Monte Carlo. Specifically, we used Hamiltonian Markov chain Monte Carlo as is implemented in the Stan probabilistic programming language [8]. The code used to sample this distribution can be accessed on the paper website or GitHub repository.

A.3 Posterior distributions of the endogenous sequences

Fig. A.6 gives the full posterior distributions of the cutting probability for each of the endogenous RSSs. We see clearly that between the two RSSs flanking the DFL16.1 gene segment that RAG is more successful at cleaving the RSS on the 3' side of the gene segment than the RSS on the 5' end. In examining the RSSs adjacent to endogenous Vκ gene segments, we see that the cutting probability is not differentiable across most of the RSSs, but cleavage is dramatically reduced when RAG interacts with the V5-43, V8-18 and V6-15 RSSs. We find that the number of paired complexes formed with the V8-18 12RSS is low to begin with, leading to an uninformative posterior distribution, whereas the V6-15 12RSS may suffer a low cleavage probability due to the T immediately adjacent to the heptamer in the coding flank region, which has been known to broadly reduce recombination efficiency [9–11].

A.4 Ca^{2+}-Mg^{2+} looping frequency comparisons

Although we directly compared the dwell time distributions of three RSS constructs between when the RAG reaction buffer contained Mg^{2+} to allow for nicking and buffer containing Ca^{2+} to prevent nicking, we wanted to know whether the looping frequency would increase when RAG is prohibited from nicking. Our intuition comes from recognizing that without the ability to

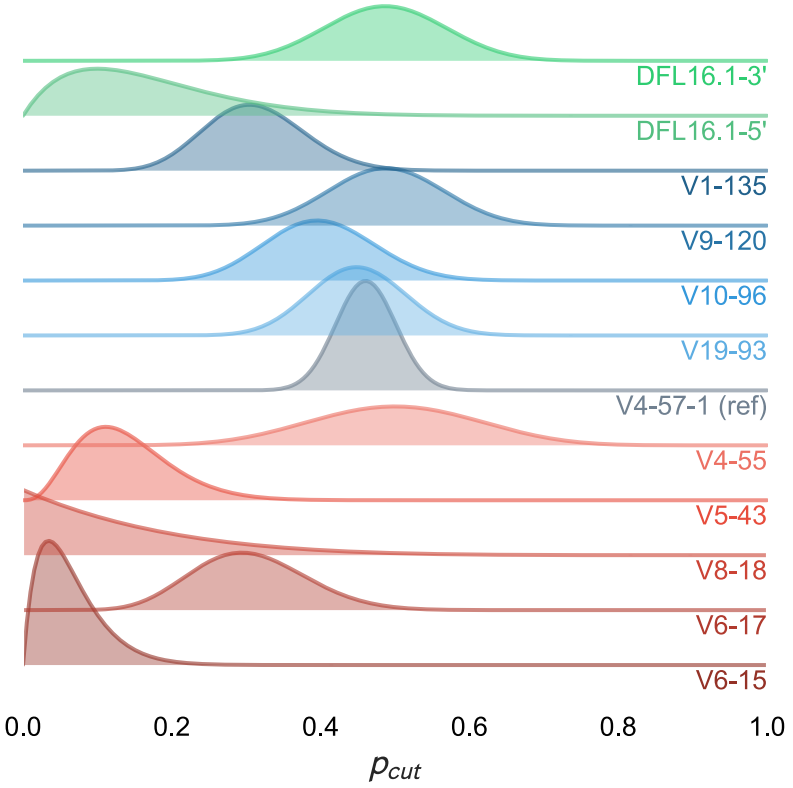


Figure A.6: **Posterior distributions of the cutting probabilities as derived in Section A.2.3 for the endogenous 12RSSs studied.** The top-to-bottom order of the endogenous RSSs is the same as their left-to-right ordering in Fig. 3. Height of the distribution is proportional to the probability of the p_{cut} value.

cleave the DNA, RAG can only release one of the RSSs and leave the paired complex state without cutting the DNA tether. As a result, RAG has an opportunity to form the paired complex with the same DNA tether. We expect that the looping frequency should either increase or remain the same in the Ca^{2+} environment as compared to when Mg^{2+} is used. Fig. A.7 shows that these two outcomes result. Fig. A.7A and A.7C show that RAG forms the paired complex more frequently with the reference sequence and the G-to-T change at the eleventh position of the reference spacer sequence when the reaction occurs in Ca^{2+}. Furthermore, we see that undergoing the reaction with the A-to-T alteration at heptamer position four in Ca^{2+} does not induce much change in the looping frequency as compared to a Mg^{2+} environment (Fig. A.7). Of interest is the fact that the spacer variant, which has a slightly larger measured looping frequency than the reference sequence in Mg^{2+} with over-

lapping 95% confidence intervals, clearly undergoes a more dramatic increase in looping frequency than the reference sequence when the salt is Ca^{2+}. This observation shows that PC formation is more favorable for the spacer variant than the reference sequence. Observed holistically, we find that RAG in the absence of nicking can form loops at least as frequently as when it when it can nick the DNA.

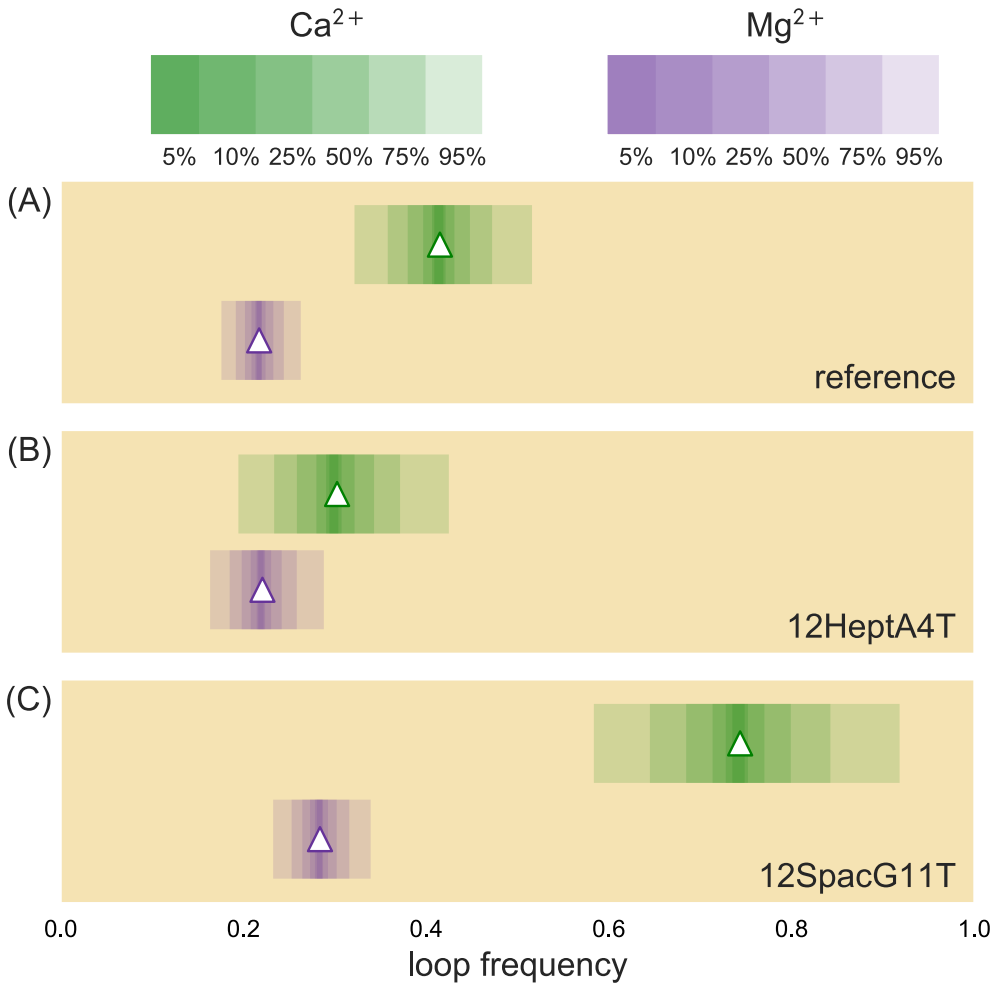


Figure A.7: **Ca^{2+} (green) and Mg^{2+} (purple) looping frequencies** for (A) reference 12RSS, (B) A-to-T change at the fourth position of the heptamer and (C) G-to-T change at the eleventh position of the spacer. Measured looping frequency shown as the triangles. Going from darker shading to lighter shading in rectangle bar indicates increasing of confidence interval percentage of the looping frequency from the bootstrapping method discussed in section A.2.2.

A.5 Coding flank contributions

For our study of the endogenous RSSs, we also modified the coding flanks adjacent to the RSSs during the cloning process to construct the DNA tethers. As shown in table A.1, most of these coding flanks have A and C nucleotides in the two or three base pairs upstream of the heptamer region. However, recent structural work have shown direct contacts between RAG1 residues and the coding flank [12–14]. Furthermore, various bulk assays have demonstrated that coding flank sequence can affect recombination efficiency [9–11]. These bulk assays suggest that coding flanks with A and C nucleotides near the heptamer tend to recombine more efficiently than those that have Ts instead. In attempting to determine whether these A- and C-rich coding flanks have much of an influence on the RAG-RSS dynamics, we looked to two pairs of TPM constructs where within each pair the RSS is identical, but the coding flank sequence is different.

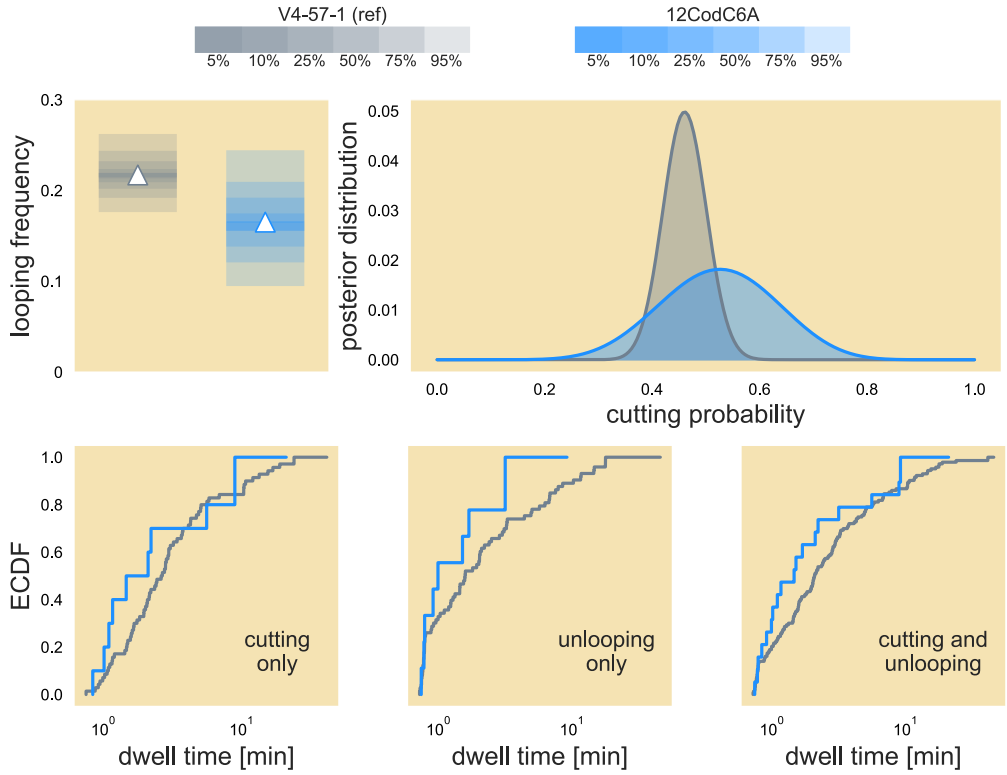


Figure A.8: **V4-57-1 (reference) RSS (grey) and coding flank change (blue) comparison of looping frequency, posterior distribution of the cutting probability and ECDFs of PC lifetimes for PCs that cut, those that unloop, and both.**

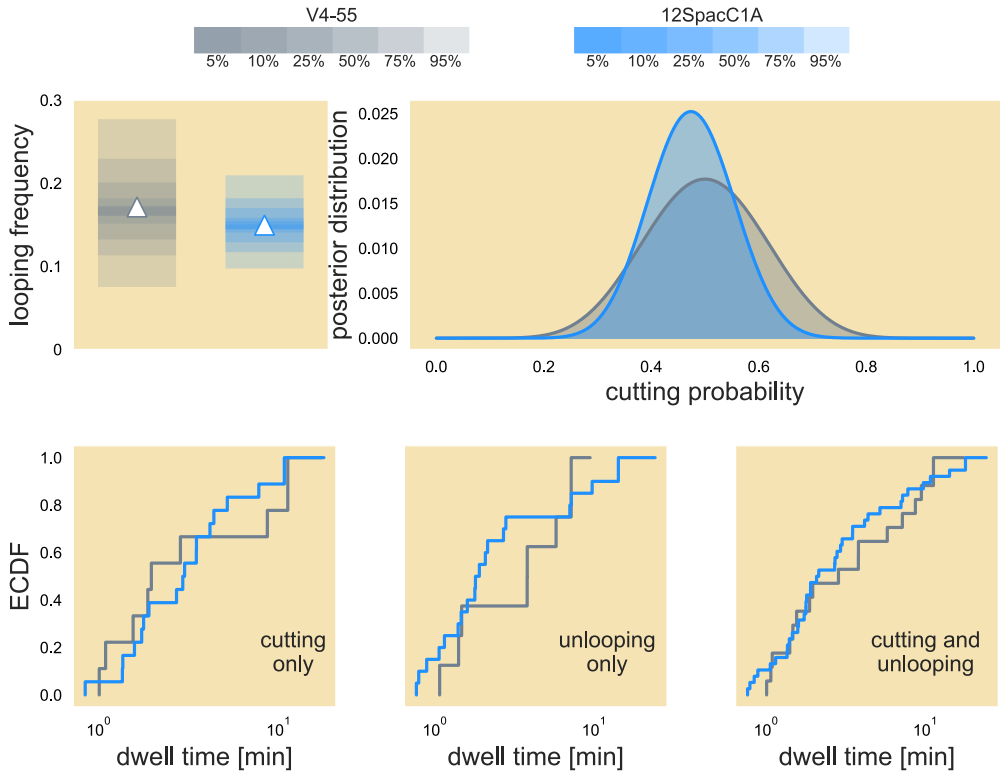


Figure A.9: **V4-55 12RSS (grey) and C-to-A change at spacer position 1 (blue) comparison of looping frequency, posterior distribution of the cutting probability and ECDFs of PC lifetimes for PCs that cut, those that unloop, and both.**

Fig. A.8 shows TPM results on the V4-57-1, or reference, RSS and a single bp change at the nucleotide immediately adjacent to the heptamer, where there is a C-to-A alteration. We find here no distinguishable difference in looping frequency or cleavage probability. Furthermore, we find that the dwell time distributions for PCs that cut, PCs that unloop, and both are identical between the reference and altered coding flank. This finding suggests that at least a single change from C to A near the heptamer does not have a dramatic effect on the RAG-RSS interaction.

We also examined two coding flanks that differ by multiple base pairs. The V4-55 RSS differs from the reference sequence at the first position of the spacer, where the C for the reference is changed to an A for the V4-55 RSS. However, the coding flank sequence differs at five nucleotides. Furthermore, the 6-bp coding flank of V4-55 is composed entirely of Cs and As and removes the Gs and Ts on the reference sequence at the first, third, and fourth positions

of the coding flank (where we index one as six base pairs from the start of the heptamer and six as immediately adjacent). We thus compared the C-to-A change at the spacer position 1 on the reference sequence with the V4-55 coding flank. As Fig. A.9 illustrates that despite the significant difference in sequence between these two constructs at the coding flank, our TPM assay reports little difference that separates these two sequences in looping frequency, dwell time distributions or cutting probability. We thus find that for most of the endogenous RSSs studied, the coding flank has little effect on the overall RAG-RSS interaction. This does not rule out the possibility that Gs or Ts in the first three positions immediately adjacent to the RSS can alter the RAG-RSS dynamics.

A.6 Cloning a different 12RSS in plasmids

To generate the synthetic RSSs used in this work, we used overhang PCR (polymerase chain reaction) and subsequently Gibson assembly (NEB Biolabs) to generate plasmids with the desired change. We selected the endogenous sequence V4-57-1 to serve as the "reference" sequence from which all synthetic RSSs were made. This sequence has been used previously [2] and exhibits a reasonable dwell time distribution, has moderately high looping frequency (compared to the other endogenous sequences), and has close to a 50% cleavage probability, as is shown in this study. This 12RSS sequence is located within the a pZE12 plasmid backbone [15]. The new RSS were inserted into this plasmid via overhang PCR with forward and reverse oligonucleotide primers (IDT) that contain a 15 base-pair overlap with the desired alteration in the middle of the sequence. The primers used in this work are listed in tables A.2 and A.3.

After purification of the PCR fragment and DpnI digestion (NEB Biolabs) of the PCR template, the fragment was circularized using Gibson assembly [16] and transformed into DH5α *Escherichia coli*. Transformants were then cultured and stored for plasmid purification and sequence verification.

A.7 Endogenous RSS sequences

Endogenous 12RSS	Sequence	n_{beads}	n_{loops}	n_{cuts}
DFL16.1-3'	AGCTAC CACAGTGCTATATCCATCAGCAAAAACC	83	37	18
DFL16.1-5'	AATAAA CACAGTAGTAGATCCCTTCACAAAAAGC	263	10	1
V1-135	TCCTCA CACAGTGATTCAGACCCGAACAAAAACT	268	46	14
V9-120	TCCTCC CACAGTGATACAAATCATAACATAAACC	248	41	20
V10-96	TCCTCC CACAATGATATAAGTCATAACATAAACC	286	43	17
V19-93	TCTACC CACAGTGATACAAATCATAACAAAAACC	284	58	26
V4-57-1 (reference)	GTCGAC CACAGTGCTACAGACTGGAACAAAAACC	700	152	70
V4-55	CACCCA CACAGTGATACAGACTGGAACAAAAACC	105	18	9
V5-43	GCCTCA CACAGTGATGCAGACCATAGCAAAAATC	186	27	3
V8-18	TCCCCC CACAGAGCTTCAGCTGCCTACACAAACC	146	5	0
V6-17	TCCTCC CACAGTGCTTCAGCCTCCTACACAAACC	126	34	10
V6-15	TCCTCT CACAGTACTTCAGCCTCCTACATAAACC	201	29	1
Jκ1 23RSS	GGATCC CACAGTGGTAGTACTCCACTGTCTGGCTGTACAAAAACC			

Table A.1: **Table of endogenous 12RSS sequences.** The 6 base pairs before the heptamer, known as the coding flank, is changed in the endogenous RSS studies and is included here. The spacer sequence for each RSS is underlined. Bold blue letters in the heptamer and nonamer regions denote deviations from the consensus sequences, CACAGTG and ACAAAAACC, respectively. The number of beads studied n_{beads}, the number of loops formed among the beads n_{loops}, and the number of cut loops n_{cuts} are given for each RSS. The bottom sequence is of the Jκ1 23RSS applied in all of the DNA constructs used in TPM.

A.8 Synthetic and endogenous 12RSS primer sequences

Tables A.2 and A.3 gives the list of primers used to construct the synthetic and endogenous RSSs. For synthetic RSSs, we apply the nomenclature '12' to denote that the 12RSS is altered, the region of the RSS where the change is made ('Hept' = heptamer, 'Non' = nonamer, 'Spac' = spacer, 'Cod' = coding flank), the original nucleotide, the position number in the region, where indexing starts at 1 and finally the new nucleotide. Therefore, if a change is made to the eighth position of the spacer, where a C is altered to T, the RSS is denoted '12SpacC8T'.

Synthetic 12RSS	**Primer**	$n_{\mathbf{beads}}$	$n_{\mathbf{loops}}$	$n_{\mathbf{cuts}}$
12CodC6A (Fwd)	AACACAGTGCTACAGACTGGAACAAAAACCCTGCAGTC	115	19	10
12CodC6A (Rev)	CTGTAGCACTGTGTTCGACCTGCAGCCCAAGCG			
12HeptC3G (Fwd)	ACCAGAGTGCTACAGACTGGAACAAAAACCCTGCAGTC	357	53	1
12HeptC3G (Rev)	CTGTAGCACTCTGGTCGACCTGCAGCCCAAGCG			
12HeptC3T (Fwd)	ACCATAGTGCTACAGACTGGAACAAAAACCCTGCAGTC	212	53	2
12HeptC3T (Rev)	CTGTAGCACTATGGTCGACCTGCAGCCCAAGCG			
12HeptA4T (Fwd)	ACCACTGTGCTACAGACTGGAACAAAAACCCTGCAGTC	607	134	52
12HeptA4T (Rev)	CTGTAGCACAGTGGTCGACCTGCAGCCCAAGCG			
12HeptG5A (Fwd)	ACCACAATGCTACAGACTGGAACAAAAACCCTGCAGTC	96	15	3
12HeptG5A (Rev)	CTGTAGCATTGTGGTCGACCTGCAGCCCAAGCG			
12HeptG5C (Fwd)	ACCACACTGCTACAGACTGGAACAAAAACCCTGCAGTC	821	96	18
12HeptG5C (Rev)	CTGTAGCAGTGTGGTCGACCTGCAGCCCAAGCG			
12HeptT6A (Fwd)	ACCACAGAGCTACAGACTGGAACAAAAACCCTGCAGTC	246	2	1
12HeptT6A (Rev)	CTGTAGCTCTGTGGTCGACCTGCAGCCCAAGCG			

12HeptT6C (Fwd)	ACCACAGCGCTACAGACTGGAACAAAAACCCTGCAGTC	461	24	2
12HeptT6C (Rev)	CTGTAGCGCTGTGGTCGACCTGCAGCCCAAGCG			
12HeptG7A (Fwd)	ACCACAGTACTACAGACTGGAACAAAAACCCTGCAGTC	343	109	28
12HeptG7A (Rev)	CTGTAGTACTGTGGTCGACCTGCAGCCCAAGCG			
12HeptG7C (Fwd)	ACCACAGTCCTACAGACTGGAACAAAAACCCTGCAGTC	146	30	10
12HeptG7C (Rev)	CTGTAGGACTGTGGTCGACCTGCAGCCCAAGCG			
12HeptG7T (Fwd)	ACCACAGTTCTACAGACTGGAACAAAAACCCTGCAGTC	219	47	7
12HeptG7T (Rev)	CTGTAGAACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacC1A (Fwd)	ACCACAGTGATACAGACTGGAACAAAAACCCTGCAGTC	254	38	18
12SpacC1A (Rev)	CTGTATCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacC1G (Fwd)	ACCACAGTGGTACAGACTGGAACAAAAACCCTGCAGTC	117	19	12
12SpacC1G (Rev)	CTGTACCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacA3G (Fwd)	ACCACAGTGCTGCAGACTGGAACAAAAACCCTGCAGTC	134	35	12
12SpacA3G (Rev)	CTGCAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacA3T (Fwd)	ACCACAGTGCTTCAGACTGGAACAAAAACCCTGCAGTC	120	28	18
12SpacA3T (Rev)	CTGAAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacC4G (Fwd)	ACCACAGTGCTAGAGACTGGAACAAAAACCCTGCAGTC	210	38	6
12SpacC4G (Rev)	CTCTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacC4T (Fwd)	ACCACAGTGCTATAGACTGGAACAAAAACCCTGCAGTC	306	128	43
12SpacC4T (Rev)	CTATAGCACTGTGGTCGACCTGCAGCCCAAGCG			

12SpacG6A (Fwd)	ACCACAGTGCTACAAACTGGAACAAAAACCCTGCAGTC	250	74	24
12SpacG6A (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacG6T (Fwd)	ACCACAGTGCTACATACTGGAACAAAAACCCTGCAGTC	184	78	34
12SpacG6T (Rev)	ATGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacA7C (Fwd)	ACCACAGTGCTACAGCCTGGAACAAAAACCCTGCAGTC	139	37	15
12SpacA7C (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacA7G (Fwd)	ACCACAGTGCTACAGGCTGGAACAAAAACCCTGCAGTC	168	21	10
12SpacA7G (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacC8T (Fwd)	ACCACAGTGCTACAGATTGGAACAAAAACCCTGCAGTC	98	17	5
12SpacC8T (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacT9A (Fwd)	ACCACAGTGCTACAGACAGGAACAAAAACCCTGCAGTC	112	22	12
12SPacT9A (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacT9C (Fwd)	ACCACAGTGCTACAGACCGGAACAAAAACCCTGCAGTC	117	50	12
12SpacT9C (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacT9G (Fwd)	ACCACAGTGCTACAGACGGGAACAAAAACCCTGCAGTC	96	8	6
12SpacT9G (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacG10A (Fwd)	ACCACAGTGCTACAGACTAGAACAAAAACCCTGCAGTC	292	60	29
12SpacG10A (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacG10C (Fwd)	ACCACAGTGCTACAGACTCGAACAAAAACCCTGCAGTC	117	34	18
12SpacG10C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			

12SpacG10T (Fwd)	ACCACAGTGCTACAGACTTGAACAAAAACCCTGCAGTC	65	20	15
12SpacG10T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacG11A (Fwd)	ACCACAGTGCTACAGACTGAAACAAAAACCCTGCAGTC	184	29	12
12SpacG11A (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacG11C (Fwd)	ACCACAGTGCTACAGACTGCAACAAAAACCCTGCAGTC	172	26	8
12SpacG11C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacG11T (Fwd)	ACCACAGTGCTACAGACTGTAACAAAAACCCTGCAGTC	941	267	83
12SpacG11T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacA12C (Fwd)	ACCACAGTGCTACAGACTGGCACAAAAACCCTGCAGTC	132	15	7
12SpacA12C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12SpacA12T (Fwd)	ACCACAGTGCTACAGACTGGTACAAAAACCCTGCAGTC	138	24	10
12SpacA12T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12NonA1G (Fwd)	ACCACAGTGCTACAGACTGGAGCAAAAACCCTGCAGTC	392	108	38
12NonA1G (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12NonA3C (Fwd)	ACCACAGTGCTACAGACTGGAACCAAAACCCTGCAGTC	554	15	7
12NonA3C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12NonA4C (Fwd)	ACCACAGTGCTACAGACTGGAACACAAACCCTGCAGTC	384	37	10
12NonA4C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12NonA4T (Fwd)	ACCACAGTGCTACAGACTGGAACATAAACCCTGCAGTC	151	10	5
12NonA4T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			

12NonA5T (Fwd)	ACCACAGTGCTACAGACTGGAACAATAACCCTGCAGTC	354	58	16
12NonA5T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG			
12NonC8G (Fwd)	GCTACAGACTGGAACAAAAAGCCTGCAGTCAACCTCGA	131	24	9
12NonC8G (Rev)	TTTGTTCCAGTCTGTAGCACTGTGGTCGACCTGCAG			
12NonC8T (Fwd)	GCTACAGACTGGAACAAAAATCCTGCAGTCAACCTCGA	280	18	6
12NonC8T (Rev)	TTTGTTCCAGTCTGTAGCACTGTGGTCGACCTGCAG			
12NonC9T (Fwd)	GCTACAGACTGGAACAAAAACTCTGCAGTCAACCTCGA	109	20	11
12NonC9T (Rev)	TTTGTTCCAGTCTGTAGCACTGTGGTCGACCTGCAG			

Table A.2: **Forward (Fwd) and reverse (Rev) primers of synthetic RSSs.** Underlined sequence denotes the region where change is made. Bold-faced letter denotes the new nucleotide. The number of beads studied n_{beads}, the number of loops formed among those beads n_{loops}, and the number of cut loops n_{cuts} are given with the forward primer sequences.

Endogenous 12RSS	Primer
DFL16.1-3' (Fwd)	AGCTACCACAGTGCTATATCCATCAGCAAAAACCCTGCAGTCGAGTAATGCA
DFL16.1-3' (Rev)	GGTTTTTGCTGATGGATATAGCACTGTGGTATTCGAAGCTTGAGCTCG
DFL16.1-5' (Fwd)	AATAAACACAGTAGTAGATCCCTTCACAAAAAGCCTGCAGTCGAGTAATGCA
DFL16.1-5' (Rev)	GCTTTTTGTGAAGGGATCTACTACTGTGGTATTCGAAGCTTGAGCTCG
V1-135 (Fwd)	TCCTCACACAGTGATTCAGACCCGAACAAAAACTCTGCAGTCAACCTCGAGAAACG
V1-135 (Rev)	AGTTTTTGTTCGGGTCTGAATCACTGTGTGAGGACTGCAGCCCAAGCGTGTAG
V9-120 (Fwd)	TCCTCCCACAGTGATACAAATCATAACATAAACCCTGCAGTCAACCTCGAGAAACG
V9-120 (Rev)	GGTTTATGTTATGATTTGTATCACTGTGGGAGGACTGCAGCCCAAGCGTGTAG
V10-96 (Fwd)	TCCTCCCACAATGATATAAGTCATAACATAAACCCTGCAGTCAACCTCGAGAAACG
V10-96 (Rev)	GGTTTATGTTATGACTTATATCATTGTGGGAGGACTGCAGCCCAAGCGTGTAG
V19-93 (Fwd)	TCTACCCACAGTGATACAAATCATAACAAAAACCCTGCAGTCAACCTCGAGAAACG
V10-93 (Rev)	GGTTTTTGTTATGATTTGTATCACTGTGGGTAGACTGCAGCCCAAGCGTGTAG
V4-55 (Fwd)	CACCCACACAGTGATACAGACTGGAACAAAAACCCTGCAGTCAACCTCGAGAAACG
V4-55 (Rev)	GGTTTTTGTTCCAGTCTGTATCACTGTGTGGGTGCTGCAGCCCAAGCGTGTAG
V5-43 (Fwd)	GCCTCACACAGTGATGCAGACCATAGCAAAAATCCTGCAGTCAACCTCGAGAAACG
V5-43 (Rev)	GATTTTTGCTATGGTCTGCATCACTGTGTGAGGCCTGCAGCCCAAGCGTGTAG
V8-18 (Fwd)	TCCCCCCACAGAGCTTCAGCTGCCTACACAAACCCTGCAGTCAACCTCGAGAAACG
V8-18 (Rev)	GGTTTGTGTAGGCAGCTGAAGCTCTGTGGGGGGACTGCAGCCCAAGCGTGTAG
V6-17 (Fwd)	TCCTCCCACAGTGCTTCAGCCTCCTACACAAACCCTGCAGTCAACCTCGAGAAACG
V6-17 (Rev)	GGTTTGTGTAGGAGGCTGAAGCACTGTGGGAGGACTGCAGCCCAAGCGTGTAG
V6-15 (Fwd)	TCCTCTCACAGTACTTCAGCCTCCTACATAAACCCTGCAGTCAACCTCGAGAAACG
V6-15 (Rev)	GGTTTATGTAGGAGGCTGAAGTACTGTGAGAGGACTGCAGCCCAAGCGTGTAG

Table A.3: **Forward (Fwd) and reverse (Rev) primers for designing TPM constructs with endogenous 12RSSs.** Underlined regions denote the heptamer and nonamer regions.

A.9 Protein purification

A.9.1 Murine core RAG1 and core RAG2 co-purification

Maltose-binding protein(MBP)-tagged murine core RAG1 and core RAG2 are co-transfected into HEK293-6E suspension cells using BioT transfection agent and are expressed in the cells for 48 hours. Cells are centrifuged and collected before resuspending with a lysis buffer consisting of cOmplete Ultra protease inhibitor and Tween-20 detergent before lysis through a cell disruptor. Lysate is centrifuged to remove the cell membrane and the supernatant containing expressed RAG is mixed with amylose resin to bind the MBP region to the resin before loading onto a chromatography gravity column. Amylose-attached RAG is then washed using lysis buffer, wash buffer containing salts before eluting with buffer containing high concentrations of maltose to out-compete the MBP on the resin. RAG-contained eluate is then concentrated and dialyzed in buffer containing 25 mM Tris-HCl (pH 8.0), 150 mM KCl, 2 mM DTT and 10% glycerol before snap-freezing 5-10 µL aliquots and storing at -80°C.

A.9.2 HMGB1 purification

Though not discussed extensively in this paper, the high mobility group box 1 (HMGB1) protein binds nonspecifically to DNA and helps facilitate RAG binding onto the RSS. A plasmid containing a His-tagged HMGB1 gene is transformed into BL21(DE3) cells and grown in liquid cultures until they reach an OD600 of 0.7. Cultures are then induced with isopropyl-β-D-1-thiogalactopyranoside (IPTG) to express HMGB1 for 4 hours at 30°C before cells are collected from the media. Cells are resuspended in binding buffer media containing cOmplete Ultra protease inhibitor, benzonase, Tween-20 and a low imidazole concentration before lysis through the cell disruptor. Lysate is cleared of membrane with an ultracentrifuge and loaded onto a nickel-NTA column to bind HMGB1. Nickel-bound HMGB1 is then washed with more binding buffer before washing with buffer containing low imidazole concentration. Washed HMGB1 are then eluted through the column with elution buffer containing higher concentration imidazole. Degraded HMGB1 is then removed by loading HMGB1 eluate onto SP column and collecting flow-through in 200 µL aliquots with an incrementally increasing salt gradient on the AKTA. Fractions that show highest change in voltage reading on the AKTA are run on a Western blot to confirm that protein of the correct size is collected before collecting. HMGB1 are transferred to a dialysis buffer containing 25 mM

Tris-HCl (pH 8.0), 150 mM KCl, 2 mM DTT and 10% glycerol through a buffer-exchange centrifuge column before snap-freezing 5-10 µL aliquots and freezing at -80°C.

BIBLIOGRAPHY

[1] L. Han et al., *Calibration of tethered particle motion experiments*, 1st ed., vol. 150, New York: Springer-Verlag, 2008, 123.

[2] G. A. Lovely et al., "Single-molecule analysis of RAG-mediated V(D)J DNA cleavage", Proc Natl Acad Sci 112 (2015), E1715.

[3] L. Han et al., "Concentration and length dependence of DNA looping in transcriptional regulation", PLoS ONE 4 (2009), e5621.

[4] S. Johnson, M. Linden, and R. Phillips, "Sequence dependence of transcription factor-mediated DNA looping", Nucleic Acids Res 40 (2012), 7728.

[5] G. Chure et al., "Connecting the dots between mechanosensitive channel abundance, osmotic shock, and survival at single-cell resolution", J Bacteriol 200 (2018), e00460.

[6] G. D. Ruxton, "The Unequal Variance T-Test Is an Underused Alternative to Student's t-Test and the Mann–Whitney U Test", en, Behav Ecol 17 (2006), 688.

[7] A. Agrawal, Q. M. Eastman, and D. G. Schatz, "Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system", Nature 394 (1998), 744.

[8] B. Carpenter et al., "Stan: A probabilistic programming language", J Stat Softw 76 (2017), 1.

[9] R. M. Gerstein and M. R. Lieber, "Coding end sequence can markedly affect the initiation of V(D)J", Genes Dev 7 (1993), 1459.

[10] U. R. Ezekiel et al., "The composition of coding joints formed in V(D)J recombination is strongly affected by the nucleotide sequence of the coding ends and their relationship to the recombination signal sequences", Mol Cell Biol 17 (1997), 4191.

[11] K. Yu and M. R. Lieber, "Mechanistic basis for coding end sequence effects in the initiation of V(D)J recombination", Mol Cell Biol 19 (1999), 8094.

[12] H. Ru et al., "Molecular mechanism of V(D)J recombination from synaptic RAG1-RAG2 complex structures", Cell 163 (2015), 1138.

[13] M.-S. Kim et al., "Cracking the DNA code for V(D)J recombination", Mol Cell 70 (2018), 1.

[14] H. Ru et al., "DNA melting initiates the RAG catalytic pathway", Nat Struct Mol Biol 25 (2018), 732.

[15] R. Lutz and H. Bujard, "Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I_1-I_2 regulatory elements", Nucleic Acids Res 25 (1997), 1203.

[16] D. G. Gibson et al., "Enzymatic assembly of DNA molecules up to several hundred kilobases", Nat Meth 6 (2009), 343.