Sequence-Dependent Dynamics of Synthetic and Endogenous RSSs in V(D)J Recombination

Soichi Hirokawa[a], Griffin Chure[b], Nathan M. Belliveau[b,c], Geoffrey A. Lovely[d], Michael Anaya[b], David G. Schatz[e], David Baltimore[b], and Rob Phillips[b,f,1]

[a]Department of Applied Physics, California Institute of Technology, Pasadena, CA, USA; [b]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA; [c]Present Address: Howard Hughes Medical Institute and Department of Biology, University of Washington, Seattle, WA, USA; [d]National Institute on Aging, National Institutes of Health, Baltimore, MD, USA; [e]Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA; [f]Department of Physics, California Institute of Technology, Pasadena, CA, USA

This manuscript was compiled on October 3, 2019

Developing lymphocytes in the immune system of jawed vertebrates assemble antigen-receptor genes by undergoing large-scale reorganization of spatially separated V, D, and J gene segments through a process known as V(D)J recombination. The RAG protein initiates this process by binding and cutting recombination signal sequences (RSSs) composed of conserved heptamer and nonamer sequences flanking less well-conserved 12- or 23-bp spacers. Little quantitative information is known about the contributions of individual RSS positions over the course of the RAG-RSS interaction. We employ a single-molecule method known as tethered particle motion to quantify the formation, stability, and cleavage of the RAG-12RSS-23RSS paired complex (PC) for numerous synthetic and endogenous 12RSSs. We thoroughly investigate the sequence space around a RSS by making 40 different single-bp changes and characterizing the reaction dynamics. We reveal that single-bp changes affect RAG function based on their position: loss of cleavage function (first three positions of the heptamer); reduced propensity for forming the PC (the nonamer and last four bp of the heptamer); or variable effects on PC formation (spacer). We find that the rare usage of some endogenous gene segments can be mapped directly to their adjacent 12RSSs to which RAG binds weakly. The 12RSS, however, cannot explain the high-frequency usage of other gene segments. Finally, we find that RSS nicking, while not required for PC formation, substantially stabilizes the PC. Our findings provide detailed insights into the contribution of individual RSS positions to steps of the RAG-RSS reaction that previously have been difficult to assess quantitatively.

V(D)J recombination | recombination-activating gene | recombination signal sequence | single-molecule biophysics

Jawed vertebrates call upon developing lymphocytes to undergo a genomic cut-and-paste process known as V(D)J recombination, where disparate gene segments that do not individually code for a protein are systematically combined to assemble a complete, antigen receptor-encoding gene (1). V(D)J recombination supports the production of a vast repertoire of antibodies and T-cell receptors that protect the host organism from a broad array of pathogens. However, gene segment combinations are not made in equal proportions; some gene segment combinations are produced more frequently than others (2–5). Although V(D)J recombination requires careful orchestration of many enzymatic and regulatory processes to ensure functional antigen receptor genes whose products do not harm the host, we strip away these factors and focus on the initial stages of V(D)J recombination. Specifically, we investigate how the dynamics between the enzyme that carries out the cutting process and its corresponding DNA-binding sites adjacent to the gene segments influence the initial stages of recombination for an array of synthetic and endogenous binding site sequences.

The process of V(D)J recombination (schematized in Fig. 1) is initiated with the interaction between the recombination-activating gene (RAG) protein complex and two short sequences of DNA neighboring the gene segments, one that is 28 bp and another that is 39 bp in length. These recombination signal sequences (RSSs) are composed of a well-conserved heptamer region immediately adjacent to the gene segment, a more variable 12- (for the 12RSS) or 23-bp (for the 23RSS) spacer sequence and a well-conserved nonamer region. For gene rearrangement to begin, RAG must bind to both the 12- and the 23RSS to form the paired complex (PC) state (Fig. 1B). Throughout the binding interaction between RAG and either RSS, RAG has an opportunity to nick the DNA (Fig. 1B zoom in) (6). RAG must nick both RSSs before it cleaves the DNA adjacent to the heptamers to expose the gene segments and to create DNA hairpin ends (Fig. 1C). DNA repair proteins complete the reaction by joining the gene segments to each other and the RSSs to one another (Fig. 1D).

RSS sequence-conservation studies across many organisms have shown a vast diversity of 12- and 23RSS sequences, mainly found through heterogeneity in the spacer region (7). Bulk assays reveal that changing an RSS sequence can significantly influence the RAG-RSS interaction and ultimately the success

Summary

V(D)J recombination is a genomic cut-and-paste process for generating diverse antigen-receptor repertoires. The RAG enzyme brings separate gene segments together by binding the neighboring sequences called RSSs, forming a paired complex (PC) before cutting the DNA. There are limited quantitative studies of the sequence-dependent dynamics of the crucial intermediate steps of PC formation and cleavage. Here, we quantify individual RAG-DNA dynamics for various RSSs. While RSSs of frequently-used segments do not comparatively enhance PC formation or cleavage, the rare use of some segments can be explained by their neighboring RSSs crippling PC formation and/or cleavage. Furthermore, PC lifetimes reveal DNA-nicking is not required for forming the PC, but PCs with nicks are more stable.

S.H., G.A.L., D.G.S., D.B., R.P. designed research. S.H. performed research. S.H., N.M.B., G.A.L., M.A., D.B., R.P. provided new reagents and analytical tools. S.H., G.C., N.M.B., G.A.L., D.G.S., D.B., R.P. analyzed data. S.H., G.C., N.M.B., G.A.L., D.G.S., D.B., R.P. wrote paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: phillips@pboc.caltech.edu

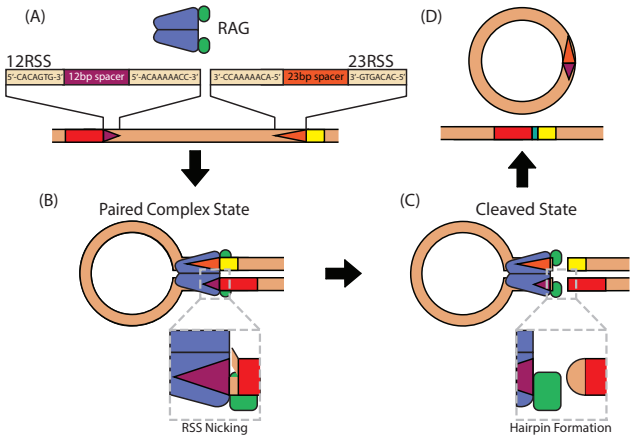

Fig. 1. Schematic focusing on the initial steps of V(D)J recombination. (A) The RAG protein complex binds to the 12- and 23RSSs (purple and orange triangles, respectively) neighboring gene segments (shown as red and yellow boxes on the DNA), (B) forming the paired complex (PC). At any point when it is bound to an RSS, RAG can introduce a nick in the DNA between the heptamer and gene segment (shown with the magnified 12RSS) and must do so to both sites before (C) it cleaves the DNA to expose the gene segments. As indicated by the magnified gene segment end, the exposed DNA strands of the gene segment are connected to form a DNA hairpin. (D) Additional proteins join these segments together. In this work, the stages subsequent to DNA cleavage are not monitored.

rate of completing recombination (8–12). Recent structural results provide evidence that RAG binding is sensitive to base-specific contacts and the local flexibility or rigidity of the 12- and 23RSS (13–15). Despite this extensive characterization on the interaction, little is known about how a given RSS sequence affects each step of the RAG-RSS reaction. In this work, we provide one of the most comprehensive studies of how RSS sequences govern the initial steps of V(D)J recombination and provide a quantitative measure of their effects on the formation frequency, lifetime, and cleavage probability of the PC.

We employ a single-molecule technique known as tethered particle motion (TPM) in which an engineered strand of DNA containing a 12RSS and 23RSS is attached to a glass coverslip at one end and to a polystyrene bead at the other (Fig. 2A). Using brightfield microscopy, we collect the root mean squared displacement (RMSD) of the bead over time to identify the state of the RAG-RSS interaction. As illustrated in Fig. 2B, when RAG forms the PC with the RSSs, the DNA tether is shortened, constraining the motion of the bead which is manifest in a reduction of the RMSD. When RAG cleaves the PC, the bead is released and diffuses away from the tether site (Fig. 2C). TPM has been applied to track the dynamic behavior of various protein-DNA systems, including RAG and RSS (16–21). It is with the temporal resolution provided by TPM that we can track the full progression of individual RAG-RSS interactions from PC formation to cleavage.

We were interested in using TPM to determine the extent to which endogenous RSSs dictate the usage frequency of their neighboring gene segments and, for those RSS positions that do seem to influence gene segment usage, identify the steps in the RAG-RSS reaction when the RSSs help or hurt their gene segment. We first examine single bp changes to a designated reference RSS, thereby establishing a mechanistic understanding of the contribution of individual nucleotide positions to RAG-RSS dynamics. With the synthetic RSSs providing context, we study a set of endogenous RSSs, each of whose sequences can be directly related to the reference sequence and a subset of the characterized synthetic RSSs. This selection of RSSs was also chosen from repertoires where the usage frequencies of their gene segments are known. Finally, we report on why our attempts to uncover deeper quantitative insights on the kinetics of the system from the PC lifetimes were met with strong disagreement between our intuited model and the TPM data, and what that consequently says about our understanding of the molecular details of the RAG-RSS reaction. As this study resulted in a wealth of data on a large number of RSS sequences, we have developed an interactive online resource for visualizing the dataset in its entirety (22).

Results

Synthetic RSSs. We chose a 12RSS flanking the immunoglobulin κ variable (*Ig*κV) gene segment, V4-57-1, as the reference sequence due to its use in a previous TPM study on RAG-RSS interactions (20) for our reference sequence. This sequence has also been used in structural studies of RAG-RSS complexes (13, 15), allowing us to compare our results with known information on the RAG-RSS structure. To explore how RAG-RSS interactions are affected by single bp changes, we examined 40 synthetic RSSs consisting of single bp changes across 21 positions of the V4-57-1 12RSS, with a particular focus on altering the spacer which is the least well-understood element in the RSS. We also studied changes made to positions 3-7 of the heptamer and various positions of the nonamer. The first three positions of the heptamer are perfectly conserved (7), likely to support DNA distortions needed for nicking and for base-specific interactions with the cleavage domain on RAG1 after nicking (13–15), while heptamer positions 4-7 also mediate base-specific interactions with RAG (13). The nonamer is bound by a nonamer-specific binding domain on

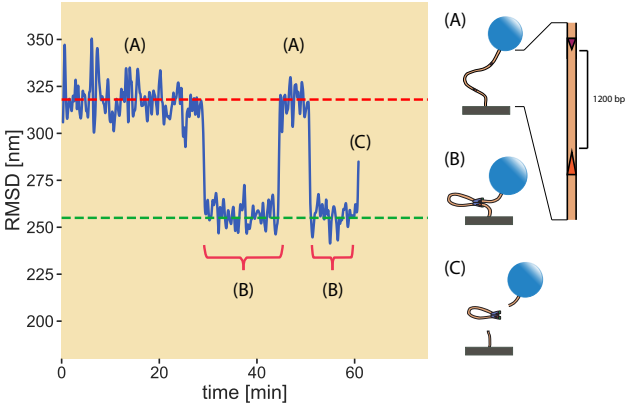

Fig. 2. Sample data output of TPM. By tracking the root mean square displacement (RMSD) of the tethered bead position undergoing restrained Brownian motion, we discern when the DNA tether is (A) in the unlooped state, (B) in the PC, or looped, state and (C) cleaved. Red brace shows the measured PC lifetime. The dashed horizontal lines distinguish the unlooped (red) and looped (green) states of the DNA, and are drawn before examining the bead trajectories and based on the length of the DNA tether and the distance between the RSSs along the strand, the extent to which HMGB1, a protein that binds nonspecifically to DNA and helps facilitate RAG binding, kinks the DNA and a set of calibration experiments relating the range of motion of the bead to the length of its tether. As depicted in the zoom-in on the DNA in (A), the distance between the 12RSS and 23RSS is fixed at 1200 bp.

RAG (13, 23). Throughout our synthetic and endogenous RSS study, we used the same concentrations of the purified forms of the two proteins that make up RAG (RAG1 and RAG2) and the high mobility group box 1 (HMGB1) protein, which binds nonspecifically to DNA and helps facilitate RAG binding to the RSSs (12). We also fixed the distance between the two binding sites to be 1200 bp, thereby constraining our study to the influence of binding site sequence on RAG-RSS dynamics alone. In addition, all of these endogenous 12RSSs are partnered with a well-characterized 23RSS (13, 15, 20) adjacent to the frequently-used gene segment from the mouse Vκ locus on chromosome 6 (5). The sequence of this RSS is provided in Table S1 of the SI text.

We pooled the relevant data across experimental replicates to characterize synthetic RSSs by three empirical properties, namely the frequency of entering the PC (looping frequency), the quartiles of the PC lifetime (dwell time) distribution, and the probability of exiting the PC through DNA cleavage (cutting probability). We define the looping frequency as the ratio of distinct PCs observed to the total number of beads monitored over the course of the experiment. Because a single DNA tether can loop and unloop multiple times over the course of the experiment, the looping frequency can in principle range from 0 to ∞. The dwell times were obtained from measuring the lifetimes of each PC state, irrespective of whether the PC led to a cleavage event or simply reverted to an unlooped state. To compute the cutting probability, we considered the fate of each PC as a Bernoulli trial with cleavage probability p_{cut}. The cutting probability reported here is the ratio of observed cleavage events to total number of observed PCs. A more detailed discussion of these calculations and the corresponding error estimates are provided in Materials & Methods and SI Text. We provide a detailed record of the data for the synthetic RSSs on the website. This webpage includes heatmaps to qualitatively illustrate how the synthetic RSSs differ in the three defined metrics. By clicking on a particular cell in any of the heatmaps, the interactive displays the measured looping frequency of the synthetic RSS with the corresponding bp change with several confidence intervals. In addition, the webpage shows for the RSS empirical cumulative distribution functions (ECDFs) of PC lifetimes in three groups: PCs that are cleaved, PCs that are unlooped, and both together. Finally, this webpage includes the complete posterior probability distribution of the cleavage probability for each synthetic RSS.

Fig. 3 illustrates the significant effect that a single bp change to an RSS can have on the formation (A), stability (B), and cleavage (C) of the PC, reaffirming that RSS sequence plays a role in regulating the initial steps of recombination. Of interest is the observed difference in phenomena between changes made to the third position and those made to the last four bases of the heptamer region. Bulk assays showing that deviating from the consensus C at heptamer position 3 essentially eliminates recombination (8, 10), yet we found that changes to G or T did not inhibit PC formation (Fig. 3B). In fact, these alterations showed similar looping frequencies and PC lifetimes (Fig. 3B) as found for the reference sequence. However, both synthetic RSSs almost completely suppress cleavage (Fig. 3C). We provide the full probability distribution for the estimate of the cutting probability (this posterior distribution is fully defined in the Materials & Methods and SI text) for these two RSSs in Fig. 3D. Nearly all of the distribution is concentrated below 10%, showing that cutting the PC is exceedingly rare. Thus, having a C at the third position of the heptamer is critically important for the cleavage step but is not necessary for RAG to form the PC with the RSS. Although deviations from the consensus nucleotide at the heptamer position 3 do not prevent RAG from forming the PC, they do impede DNA cleavage.

We find that PC formation is reduced compared to the reference sequence when the last four bases of the heptamer are altered, particularly at the fifth and sixth positions. Of more than 400 DNA tethers with the 12RSS containing a T-to-A change at heptamer position 6, we observed the PC only once, which subsequently led to cleavage. This result is consistent with recent findings that the consensus TG dinucleotide at the last two positions of the heptamer supports a kink in the DNA and may be critical for RAG binding (14). We notice that some changes increase the median time spent in the PC such as with the heptamer position 4 (Fig. 3B). This RSS also had one of the widest dwell time distributions of all of the synthetic RSSs studied. Furthermore, we find that the cutting probability decreased when we altered any one of the last four positions of the heptamer, but to a lesser extent than for changes made to the heptamer 3. The single bp change that had the greatest effect, located at heptamer position 6 (T to C) showed 2 out of 28 PCs led to cleavage.

Although we observed only modest differences in the median dwell times when we altered the reference sequence in the spacer region, some alterations substantially affected the looping frequency and cutting probability. The C-to-T change at spacer position 4 doubled the frequency of observing the PC while a T-to-G change at the ninth position reduced PC formation nearly as much as changes made at heptamer position 6. These two changes made to the spacer reflect the observed extremes of spacer sequence effects on the looping frequency. While many of the changes in the spacer region do not alter the cutting probability, we can still find spacer-altered RSSs that improve or inhibit cleavage in this region. Fig. 3D shows that changing the fourth position from C to G reduces the cleavage probability, while altering the tenth position of the spacer from G to T increases the cleavage probability as well as the frequency of PC formation (Fig. 3A). RAG1 makes contacts along the entire length of the 12RSS spacer (14), helping to explain our finding that changes to the spacer can substantially alter the probability of PC formation and cutting, thereby playing more of a role than simply separating the heptamer and nonamer sequences.

Similar to spacer changes, most nonamer changes alter the PC dwell time by $\approx$1 minute relative to the reference sequence. However, unlike spacer-modified RSSs, most synthetic-nonamer RSSs reduced the frequency of PC formation. Disruptions to the poly-A sequence in the center of the nonamer cause a substantial reduction in looping frequency, most notably the near complete inhibition of PC formation with the A-to-C change at nonamer position 3. This detrimental effect of deviating from the poly-A tract agrees with previous work demonstrating numerous protein-DNA interactions in this region and with the proposal that the rigidity produced from the string of A nucleotides is a critical feature for RAG1 to bind the nonamer (14, 23). Furthermore, this reduction in looping frequency extends to changes made at the eighth and ninth positions. The sequence deviations in the nonamer

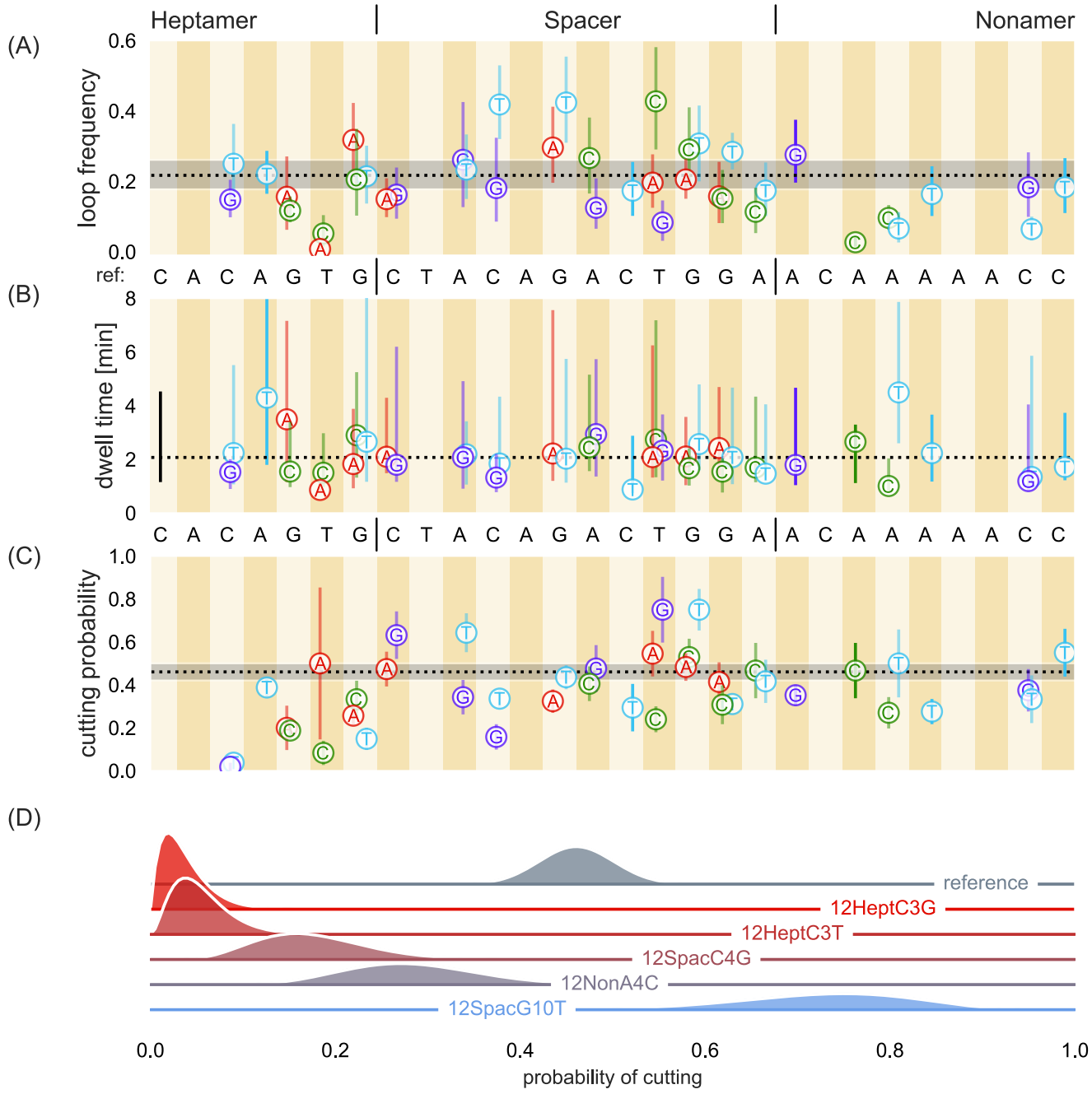

Fig. 3. TPM data for single bp changes introduced at various positions of the reference 12RSS. (A) Loop frequency with 95% confidence interval, (B) dwell time with median represented as a point and lines extending to the first and third quartiles of the distribution and (C) cutting probability with standard deviations. The dotted black line in (A) is set at the reference loop frequency, 0.23, with shaded area denoting the extent of the 95% confidence interval for the reference. The dotted black line in (B) denotes the reference 12RSS median dwell time, 1.8 minutes, with the black bar at the left denoting the first and third quartiles of the distribution. Dotted black line in (C) is the most probable cutting probability for the reference sequence, roughly 0.4, with the grey shaded region setting one standard deviation. The reference sequence is provided along the x-axis for ease of determining the position where the change was made and the original nucleotide. The introduced nucleotide is provided in the figure with the letter and color-coded (red for A, green for C, light blue for T and purple for G). Heptamer, spacer, and nonamer regions are also separated by vertical lines in the sequences. (D) Ridgeline plot of posterior distributions of the cutting probability, given the number of loops observed and loops that cut (see SI) for a subset of the synthetic RSSs (labeled and colored along the zero-line of the respective ridgeline plot). Height of the distribution is proportional to the probability of a given cutting probability.

region, however, do not significantly affect cleavage once the PC has formed, as evidenced by the overlap in the posterior distributions of the reference sequence and its nonamer variant with the greatest reduction in the cleavage probability (position 4, A to C), in Fig. 3D. Overall, synthetic-nonamer RSSs have negative effects on PC formation with minimal effects on subsequent DNA cleavage, consistent with extensive biochemical and structural evidence that the primary function of the nonamer is in facilitating RAG-DNA binding (23).

Endogenous RSSs. To build on our study of single bp effects on RAG-RSS dynamics, we selected a set of endogenous RSSs based on existing gene usage frequency data and whether their sequences were superpositions of the reference RSS and subsets of our studied synthetic RSSs. Specifically, we selected gene segments from the mouse Vκ locus on chromosome 6 from data collected by Aoki-Ota *et al.*, including a variety of frequently-used gene segments (V1-135, V9-120, V10-96, V19-93, V6-15 and V6-17), two moderately-used gene segments (V4-55 and V5-43) and two rarely-used (V4-57-1 and V8-18) gene segments (5). The V4-57-1 12RSS is identical to the reference 12RSS from our synthetic RSS study. In addition, we examined DFL16.1, the most frequently used D gene segment from the murine immunoglobulin heavy chain (*Igh*) locus on chromosome 12 (4, 24). Unlike the Vκ gene segments, which only need to combine with one gene segment, D gene segments must combine with two other gene segments to encode a complete protein. As a result, DFL16.1 is flanked on both its 5' and 3' sides by distinct 12RSS sequences, denoted DFL16.1-5' and DFL16.1-3', respectively, both of which are examined in this study. The sequences of all endogenous RSSs studied here are provided in the SI Text. We apply TPM on these sequences to determine whether their involvement in the RAG-RSS reaction could both provide insight into the usage frequency of their flanking gene segments and be predicted based on the activity profile of the synthetic RSSs.

To develop a better sense for how RAG interacts with these RSSs in their endogenous context, the 6 bp coding flank sequence adjacent to the heptamer of all but the V4-57-1 RSS was chosen to be the natural flank provided by the endogenous gene segment. RAG interacts with the coding flank during DNA binding and PC formation (13–15) and coding flank sequence can influence recombination efficiency, particularly the two bp immediately adjacent to the heptamer (25–27). Two T nucleotides and to some extent even a single T immediately 5' of the heptamer inhibit the nicking step of cleavage and thus reduce recombination efficiency (25–27). We did not extensively analyze the contribution of coding flank sequence in this study, and only V6-15 RSS among the studied RSSs would be predicted to be detrimental due to the T flanking the heptamer; all other coding flanks have combinations of A and C as the two terminal coding flank bases. We kept the same coding flank for the V4-57-1 RSS as in a previous study (20) to facilitate closer comparison of the results of the synthetic RSSs. We do not expect much difference between the endogenous coding flank sequence (5'-CACTCA, where the two nucleotides closest to the heptamer are underlined) and the coding flank used here (5'-GTCGAC) because the two terminal coding flank bases are similar to those of all but the V6-15 RSS and for reasons discussed in the Discussion and SI Text. The coding flank sequences for all studied endogenous RSSs are included in the SI text. We present the results of the RAG-endogenous RSS interaction in Fig. 4 and provide an interactive tool for exploring these data on the paper website. This webpage includes an interactive feature where the looping frequency, ECDFs of looping lifetimes, and probability distribution of the cleavage probability of any two endogenous RSS can be directly compared.

The variable nature of all three metrics [looping frequency (Fig. 4A), dwell time (B), and cutting probability (C)] across RSSs highlights how, similar to the synthetic RSSs, endogenous

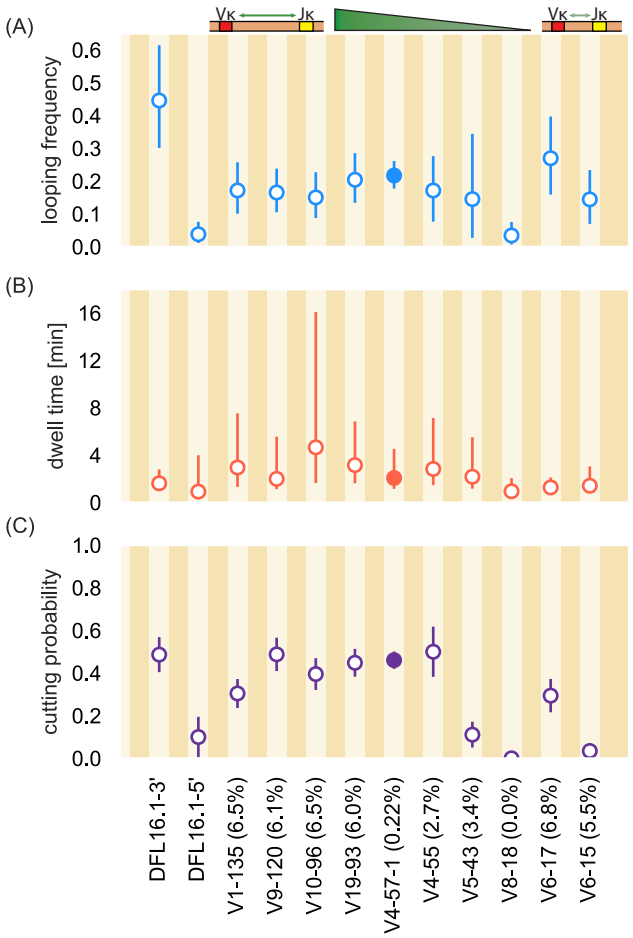

Fig. 4. Observed dynamics between RAG and endogenous RSS sequences. (A) Frequency of PC formation (looping frequency) with 95% confidence interval. (B) Median PC lifetime with the lower error bar extending to the first quartile and the upper error bar extending to the third quartile. (C) Probability of DNA cleavage (cutting probability) of RAG with error bars showing one standard deviation. For discussion of the errors in Fig. 4A and 4C, see the SI text. DFL16.1-3' and DFL16.1-5' flank the same gene segment but in different orientations on the Igh chromosome. As shown in the graphic above Fig. 4A, Vκ gene segments listed are ordered by their position along the chromosome, with linear distance from the Jκ gene segments decreasing from left to right. Numbers in parentheses next to Vκ gene segment denote percentage of usage in repertoire (5). The V4-57-1 12RSS has a filled in circle to denote that it is the reference sequence for examining the effects of single base changes.

sequences influence formation, stability, and cleavage of the PC differently. Of particular interest is the behavior of DFL16.1-3' which shows the highest propensity for PC formation but some of the shortest PC lifetimes. Despite this short median dwell time, the probability of the PC successfully proceeding to DNA cleavage is approximately 0.5. Notably, the frequency of PC formation and the probability of cleavage are both greatly reduced for DFL16.1-5' as compared to DFL16.1-3', although their median PC dwell time and the width of the dwell time distributions are approximately equal. Reduced function of DFL16.1-5' relative to DFL16.1-3' is consistent with prior studies (24, 28, 29) and is addressed further in the Discussion.

The endogenous RSSs of the Vκ gene segments show varying degrees of PC formation and cleavage probabilities. Many of the endogenous RSSs studied here, including those of gene

segments used frequently *in vivo* (V1-135, V9-120, V10-96, V19-93, V6-17 and V6-15), demonstrate looping frequencies between 15 and 30 events per 100 beads. Gene segments V4-57-1 and V4-55 are used with low and modest frequency, respectively, yet in our experiments, they enter the PC with comparable frequency (approximately 20 to 30 loops per 100 beads). In general, we find these two sequences to behave almost identically in our experimental system, illustrating that other biological phenomena, such as higher-order DNA structure, govern the segment usage *in vivo* (4, 30). The endogenous V8-18 12RSS exhibits infrequent PC formation and cleavage and short median PC lifetimes, much like the DFL16.1-5' 12RSS. Using the V8-18 12RSS, only 5 looping events were detected from 146 DNA tethers and cleavage was never observed. Despite the similarities in reaction parameters for the V8-18 and DFL16.1-5' RSSs, DFL16.1 is the most frequently used D gene segment in the repertoire (4) while V8-18 is never used (5). One possible explanation for this difference is that the DFL16.1-5' 12RSS does not participate in recombination until after its gene segment has undergone D-to-J recombination and has moved into the RAG-rich environment of the "recombination center." This relocation is thought to facilitate RAG binding to the 5' RSS of the committed D gene segment (31, 32). In contrast, V8-18, like other Vκ gene segments, must be captured into the PC by RAG that has previously bound to a Jκ RSS in the recombination center.

Fig. 4B demonstrates that, with the exception of the V10-96 RSS, PC lifetimes are similarly distributed across the endogenous RSSs examined in this work. Most RSSs have median dwell times between 1 to 3 minutes with the V8-18 12RSS displaying the shortest-lived median dwell time of roughly 40-50 seconds. While most endogenous RSSs here have a similar range between the first and third quartiles (see interactive figure on the paper website), the V10-96 12RSS distribution is noticeably wider, with the first quartile of the distribution being a longer lifetime than the median lifetime for most endogenous RSS distributions and the third quartile of this RSS extending out to over 19 minutes. These observations suggest a similar stability of the PC for all but the V10-96 RSS once RAG manages to bind simultaneously to both 12- and 23RSSs.

Fig 4C indicates that six endogenous RSS sequences from V1-135 to V4-55 have comparable cutting probabilities ranging from 0.4 to 0.5. Considering that the less-frequently used V4-57-1 and V4-55 gene segments have 12RSSs that show similar cutting probabilities and looping frequencies to the 12RSSs of more frequently-selected gene segments, other factors appear to prevent their selection. The low probability of cutting with the V6-15 12RSS is particularly noteworthy, with the low cutting probability of about 0.05 indicating that RAG tends to easily break the looped state rather than commit to cleavage. However, this low cutting probability might be attributed to the T in the coding flank immediately adjacent to the heptamer. Other features of the system must dictate the high-frequency usage of V6-15 *in vivo* (5).

Kinetic Modeling of the PC Lifetime Distribution. Figs. 3B and 4B show that the vast majority of median looping lifetimes ranged between 1 to 3 minutes with rare exceptions, suggesting similar dwell time distributions for many of the RSS variants. However, many of these synthetic and endogenous RSSs have different probabilities of DNA cleavage, suggesting that at the very least the rate of cutting changes. As TPM has been used to extract kinetic parameters for various other protein-DNA systems (17, 18, 33, 34), we used the distributions of the PC lifetimes to estimate the rates of unlooping and cutting for each RSS and to attempt to discern a deeper connection between RSS sequence and fate of the PC. We developed a simple model in which a PC state can have two possible fates: either simple unlooping of the DNA tether or cleavage of the DNA by RAG. We characterized each of these outcomes as independent yet competing processes with rates k_{unloop} and k_{cut} for unlooping and DNA cleavage, respectively. If the waiting time distribution t_{unloop} or t_{cut} for each process could be measured independently where only one of the two outcomes was permitted to occur, one would expect the probability densities of these waiting times given the appropriate rate to be single exponential distributions of the form

$$P(t_{\text{unloop}} \,|\, k_{\text{unloop}}) = k_{\text{unloop}} e^{-k_{\text{unloop}}\, t_{\text{unloop}}} \qquad [1]$$

for the unlooping process and

$$P(t_{\text{cut}} \,|\, k_{\text{cut}}) = k_{\text{cut}} e^{-k_{\text{cut}}\, t_{\text{cut}}} \qquad [2]$$

for DNA cleavage. However, as these two Poisson processes are competing, we cannot estimate k_{cut} solely from the waiting time distribution of paired complex states that led to DNA cleavage nor k_{unloop} using the states which simply unlooped. As each individual cutting or unlooping event is assumed to be independent of all other cutting and unlooping events, the distribution of the dwell time t before the PC either unloops or undergoes cleavage can be modeled as an exponential distribution parameterized by the sum of the two rates,

$$P(t \,|\, k_{\text{leave}}) = k_{\text{leave}}\, e^{-k_{\text{leave}}\, t}, \qquad [3]$$

where $k_{\text{leave}} = k_{\text{unloop}} + k_{\text{cut}}$.

Given the collection of waiting time distributions measured for each RSS, we estimated the values of k_{leave} which best describe the data. We find that the observed dwell times are not exponentially distributed for any 12RSS sequence analyzed, either endogenous or synthetic. Examples of these waiting time distributions along with an exponential distribution parameterized by the 95% credible region for k_{leave} can be seen for twelve of the RSS variants in Fig. 5. In general, the observed dwell times are underdispersed relative to a simple exponential distribution with an overabundance of short-lived PCs. We also find that the observed dwell time distributions are heavily tailed with exceptionally long dwell times occurring more frequently than expected for an exponential distribution.

The ubiquity of this disagreement between our simple model and the observed data across all of the examined RSSs indicates that leaving the PC state either by reverting to the unlooped state or committing to the cleaved state is not a one-step process, suggesting that at least one of the two fates for the PC state on its own is not single-exponentially distributed as assumed in our null model of the dynamics.

One hypothesis for the disagreement between the model given in Eq. 3 and the data is that other processes, such as nicking of the DNA by RAG, create effects in the tethered bead trajectories that are too subtle to be detected in the TPM assays. Nicking creates a more stable RAG-single RSS complex (though this effect on PC stability had not been previously quantified) (13, 35) and can occur at any time after RAG binds to the RSS (6), making it exceedingly difficult to

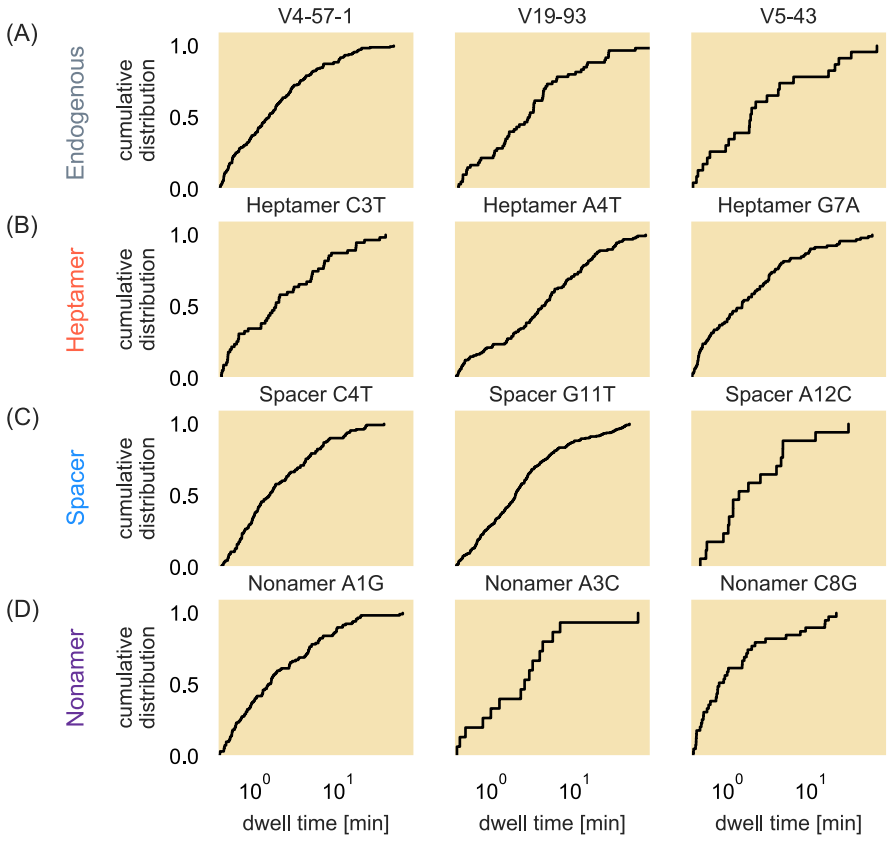

Fig. 5. Non-exponential waiting time distributions for endogenous and synthetic 12RSSs. The empirical cumulative distribution of the measured PC lifetimes (black lines) are shown for representative endogenous sequences (A) as well as for the synthetic RSSs with single point alterations made in the heptamer (B), spacer (C), or nonamer (D) regions. The shaded area corresponds to the 95% credible region of a true exponential distribution parameterized in Eq. 3 given a posterior distribution for k_{leave}, the rate of the arrival of either an unlooping or cleavage event.

determine whether a given PC has one, both or neither of the RSSs nicked. As a result, we are unable to model the combined kinetics of unlooping and cleavage without also identifying when RAG nicks the RSSs to which it is bound.

Substitution of Ca^{2+} in place of Mg^{2+} in the reaction buffer allows RAG to bind the RSSs but blocks both nicking and cleavage (36), leaving unlooping as the only possible fate of a PC. To determine if unlooping could be modeled as a simple Poisson process, we measured the PC dwell time distribution for a subset of the RSSs in a reaction buffer containing Ca^{2+}.

While we observe no cleavage of PCs in the Ca^{2+}-based buffer, the dwell times of PC events are still not in agreement with an exponential distribution (left panels of Fig. 6A-C), indicating that the process of unlooping itself is not a Poisson process and that there are other looped states which our experimental system cannot detect. We also note that for each of the RSS variants the observed PC lifetimes are short lived compared to those in the Mg^{2+}-based buffer, as can be seen in the bottom plots of Fig. 6. Because Ca^{2+} does not significantly alter DNA flexibility compared to Mg^{2+} (37), our data strongly argue that nicking itself results in a more stable PC. This is notable in light of recent structural evidence showing that nicking and the associated "flipping out" of two bases at the RSS-gene segment junction away from their complementary bases creates a more stable RAG-RSS binding conformation (13). With the more stable conformation from nicking one or both RSSs, the PC state can last longer than if RAG could not nick either RSS, which is reflected in the longer dwell time distributions when using Mg^{2+}.

Discussion

Through the temporal resolution provided by TPM, we have discerned how RAG forms and cleaves the PC for a series of synthetic and endogenous RSSs. We find that the RSSs of frequently-used gene segments typically do not support more efficient PC formation or cleavage than those neighboring gene segments of more modest usage. This observation is consistent with recent findings that RSS strength, as assessed by the RSS information content (RIC) algorithm (11, 38–40), is only one of multiple parameters needed to be able to predict gene segment usage frequency (30, 41). Furthermore, we found from analyzing single bp variations of the V4-57-1 RSS that the efficiencies of PC formation and cleavage are sensitive to single bp changes depending upon the conservation level at the respective position. We see that altering the perfectly-conserved third position of the heptamer almost completely blocked cleavage by RAG, but did not significantly alter PC formation frequency or dwell time distribution. In contrast, most deviations from the consensus nucleotide at the last four positions of the heptamer or in the nonamer decreased the frequency of PC formation. Finally, even though few positions of the spacer have a consensus nucleotide (7), formation and cleavage of the PC can still be strongly affected by a single bp change in the spacer. In fact, sequence-context effects might help explain why some of these synthetic RSSs in less conserved positions of the spacer have such a strong influence on PC formation and cleavage on their own.

We asked to what extent we could account for the behavior of an endogenous RSS based on its constituent nucleotides as revealed by our synthetic RSS study. The comparative

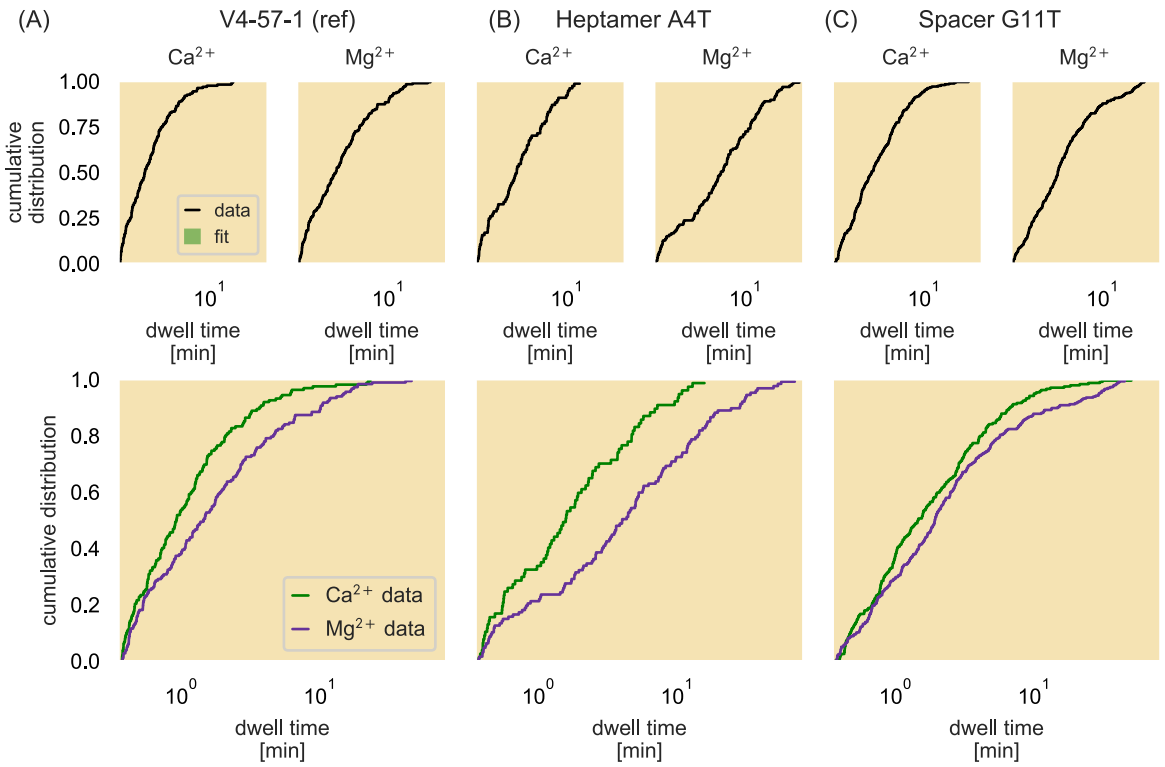

Fig. 6. Empirical cumulative distributions of PC lifetimes with different divalent cations. The empirical cumulative dwell time distributions are plotted in black over the 95% credible region of the fit to an exponential distribution (top row) for the reference sequence (A), a base pair change in the heptamer region of the 12RSS (B), and a base pair change in the spacer region (C). The bottom plots show direct comparisons of the empirical cumulative dwell time distributions collected in either Ca^{2+}- (green) or Mg^{2+}- (purple) supplemented reaction buffer for each RSS.

interactive tool on the paper website allows one to select an endogenous RSS to reveal not only its data on PC formation, PC lifetime distributions, and cleavage probability distributions, but also data for each nucleotide difference between it and the reference 12RSS through the relevant synthetic RSSs. Although we were unable to construct a quantitative model that could directly relate endogenous RSS behavior to the effects measured for each individual sequence deviation, these results provide several insights into the relation between RSS function and its constituent nucleotides. In particular, the data reveal a subset of RSS positions, including some in the spacer, that appear to strongly influence RAG-RSS interactions.

The synthetic RSS with the G-to-T change at the spacer position 10 strongly increases the cleavage probability and also enhances PC formation (Fig. 3A, C, D). These improvements might be due to the 5'-TG-3' dinucleotide created by this change at spacer positions 10 and 11. Such a pyrimidine-purine (YR) pairing is inherently deformable (42) and a substantial 60° bend in the 12RSS is seen at this location in the spacer in RAG-RSS complexes (14). Hence, as noted previously (14), a YR combination at the 3' end of the spacer in the 12RSS is favorable for DNA binding, consistent with our data. The DFL16.1-5' RSS contains a T at spacer position 10 (Table S1), as well as several other nucleotides in the spacer that each individually increase PC formation (see the paper website), but this RSS hampers PC formation (Fig. 4A). Because spacer position 11 is also a T in the DFL16.1-5' RSS, the T at position 10 does not create a YR pair and instead, the last seven bp of the spacer are all pyrimidines. A spacer with such a sequence might be particularly poor at supporting the DNA distortions needed for RAG-12RSS binding. This example of the importance of sequence context in determining how a particular bp will influence RSS function supports a concept borne out of the development of the RIC algorithm (11, 38, 40).

The contributions that coding flanks make to RAG-RSS dynamics (13) are important considerations to quantitatively model the RAG-DNA interactions, as each endogenous RSS neighbors a different coding flank. We attributed the low cleavage probability of the V6-15 RSS to the T immediately adjacent to the RSS in the coding flank, which has been shown to be detrimental to recombination efficiency (25–27). Because the other endogenous RSSs studied are rich in C and A nucleotides in the two bp adjacent to the heptamer, we compared data for two pairs of DNA constructs that differed only in coding flank sequence. One comparison involves the substrate containing the coding flank sequence used on the V4-57-1 RSS (5'-GTCGAC) and a substrate with a C-to-A change adjacent to the heptamer (5'-GTCGAA). The other pair is the V4-55 endogenous RSS substrate and the synthetic RSS substrate containing a C-to-A alteration at spacer position 1, where, fortuitously, the RSSs are identical and the coding flanks differ by five base pairs (5'-CACCCA for V4-55 and 5'-GTCGAC for the synthetic RSS). In both cases, the looping frequencies, PC lifetime distributions, and cutting probability distributions are

similar for the respective pairs, arguing that these coding flank differences contribute little to the overall RAG-RSS reaction (see SI Text). Hence, coding flank differences present in all of the endogenous RSS substrates analyzed here, with the exception of the V6-15 RSS, are unlikely to have a strong influence on RAG-RSS dynamics. However, a more extensive examination of coding flank, particularly for G- and T-rich sequences, in a dynamic experimental method such as TPM will help to shed light on the extent to which these RSS-adjacent sequences influence the various steps of V(D)J recombination.

The V5-43 12RSS has a low level of PC formation, likely because of its C-to-T change at nonamer position 8, while its poor cutting probability can be attributed to a collection of sequence changes that reduce cleavage probability. The low frequency of PC formation with the V9-120 and V6-15 RSSs is likely driven primarily by the A-to-T change at nonamer position 4, with additional negative contributions coming from altering the reference spacer. And the DFL16.1-3' RSS, which supported the highest frequency of PC formation across all RSSs studied, differ from the reference RSS at the fourth and sixth positions of the spacer that each in their own synthetic RSSs strongly stimulated PC formation. These findings support the important conclusion that spacer sequence can influence RSS synapsis by RAG.

We find that the DFL16.1-5' RSS is much less competent or PC formation and cleavage than the DFL16.1-3' RSS. Weaker activity of the 5' RSS compared to the 3' RSS is consistent with the results of recombination assays performed using plasmid substrates in cells (28, 29) and for chromosomal recombination when DFL16.1 was placed approximately 700 bp from its *Igh* J gene segment partner, J_H1 (24). However, when assayed in their natural location over 50 kb from the J_H gene segments, the two RSSs support roughly equal levels of recombination as long as they are in the same orientation relative to the J_H 23RSSs (24). The existing data argue that the DFL16.1-5' RSS is intrinsically less active for recombination than the DFL16.1-3' RSS, but this difference can be minimized over large chromosomal distances when chromatin "scanning" by RAG is the dominant mechanism for bringing RSSs together to form the PC (24, 43).

Our study of both synthetic and endogenous RSSs explains the low usage of the V8-18 gene segment in the *Igκ* repertoire and further highlights the strong impact that can be exerted from a single nucleotide change to an RSS. The V8-18 RSS contributes to inefficient PC formation and further interrogating each sequence mismatch between the V8-18 and reference RSSs revealed that its T-to-A alteration at heptamer position 6 is sufficient to virtually abrogate PC formation. This result provides a mechanistic explanation for why the VκA2b gene segment is underutilized in the antibody repertoire of Navajos, which in turn has been proposed to account for the high susceptibility of Navajos and several genetically-related groups of Native Americans to *Haemophilus influenza* type b infection (44). The VκA2b RSS differs in sequence from the more common and efficiently recombined VκA2a RSS by a single T-to-A change at the heptamer position 6 (44–46). We conclude that the inefficient recombination caused by this alteration is due to a defect in PC formation and suggest that any gene segment whose RSS contains an A at the sixth position of the heptamer will recombine poorly. Consistent with this, A is almost never observed at the sixth position of the heptamer in either the 12- or 23RSS (7).

Our attempts to obtain quantitative insight into the kinetics of RAG-RSS dynamics led to two interesting findings on the nature of the interaction. Upon first applying our fitting procedure to determine the rates of unbinding and cleavage, we learned that at least one of these two processes did not behave as a simple Poisson process. Thinking that our inability to detect nicking was the culprit, we examined the rate of unlooping in the absence of nicking by using Ca^{2+} instead of Mg^{2+} in our reactions. Here, our finding that the PC lifetimes were not exponential for any of the studied RSSs further thwarted our efforts to obtain a pure measurement of the rate of unlooping. These Ca^{2+} results suggest that the PC state may have multiple conformations like the *lac* repressor (47) in that the two RAG1/2 dimers may have multiple states, or that binding to the heptamer and to the nonamer on each RSS are actually separate sequential processes. One possible source of distinct conformations is the dramatic 180° rotation of the DNA that must occur prior to nicking. Rotated and unrotated configurations of un-nicked RSSs have been identified in recent structural studies (14, 15), but would be indistinguishable in the TPM assay. Despite these challenges to obtaining a quantitative description, our data demonstrate that nicking of an RSS is not a prerequisite for RAG to form the PC state, consistent with previous gel shift analyses performed either in Ca^{2+} or with RAG mutants lacking catalytic activity (48, 49). In addition, our findings demonstrate that PCs with nicked RSSs are more stable than those where RSSs cannot be nicked, extending previous findings made with RAG bound to single RSSs (35). To our knowledge, this study is the first attempt to obtain kinetic rates of unlooping from and cutting of the PC and reveals that there are still key details in the reaction that are left unaccounted for.

The work presented here leaves open several questions about the RAG-RSS dynamics. Although our TPM assay detects PC formation and cleavage, it does not detect nicking, preventing us from determining how the RSSs studied influence the rate of nicking or when nicking occurs relative to PC formation. Even without nicking, we see that the unlooping dynamics behave differently from a simple Poisson process. This result suggests a need for an experimental method such as single-molecule FRET (50) that can detect such subtle conformational changes that occur between RAG and the RSS. Finally, we have left the 23RSS unchanged in this study, but it is possible that the trends that we see for our synthetic or endogenous 12RSSs may change with a different partner RSS and shed more light on the "beyond 12/23 rule" (11, 51, 52). Ultimately, these finer details in the RAG-RSS interaction can provide a more complete kinetic description of the initial phases of V(D)J recombination. While we changed the 12RSS sequence in this work, the TPM assay in principle allows us to titrate other parameters, such as the distance between RSSs, or introduce more biochemical players to better contextualize our work in the bigger picture of recombination *in vivo*.

Materials and Methods

Protein purification. The two RAG components, core RAG1 and core RAG2 (RAG1/2), are purified together as outlined in (20). Maltose binding protein-tagged murine core RAG1/core RAG2 were co-expressed by transfection in HEK293-6E suspension cells in a 9:11 w/w ratio for 48 hours before purifying using amylose resin.

HMGB1 is purified as outlined in (20). His-tagged HMGB1 was expressed in isopropyl-β-D-1-thiogalactopyranoside-induced BL21 cells for 4 hours at 30°C before purification. For more details, see the SI Text.

Flow cell assembly. TPM flow cells were assembled by drilling four holes along each length of a glass slide before cleaning the slides and cover slips. The slides and cover slips were treated with an epoxidizing solution for at least an hour and a half. Upon completion of the treatment, flow cells are assembled by cutting four channels into double-sided tape to connect the drilled holes at opposite ends of the glass slide before adhering to the cover slip on one side and the glass slide on the other. Short connective tubes are inserted into each of the holes to serve as inputs and outputs for fluids and sealed using 5-minute epoxidizing solution. The constructed flow cells are baked for fifteen minutes on the hot plate.

Tethered bead assembly. Tethered beads are constructed by incubating anti-digoxigenin in the flow cell channels for two hours to allow for sticking to the glass surfaces. After washing away excess anti-digoxigenin in a buffer solution containing Tris-HCl, KCl, $MgCl_2$, DTT, EDTA, acetylated BSA and casein, engineered strands of 2900 bp-long DNA containing a 12RSS and a 23RSS located 1200 bp apart and tagged with digoxigenin on one end and biotin at the other end are injected into the flow cells to attach the digoxigenin end of the DNA to the anti-digoxigenin-scattered surfaces. After excess DNA is washed out, 490 nm streptavidin-coated polystyrene beads are added to the channels and incubated for no more than 3 minutes to bind the biotin-labeled end of the DNA. Excess beads are washed away and the TPM assembly buffer is replaced with a RAG reaction buffer containing Tris-HCl, KCl, glyercol, DTT, potassium acetate, $MgCl_2$, DMSO and acetylated BSA. For Ca^{2+} studies, $CaCl_2$ is used in place of $MgCl_2$ in the RAG reaction buffer and in the same concentration. See SI Text for visual demonstration of TPM preparation.

TPM experiment. TPM experiments involve the simultaneous acquisition of bead trajectories from two different channels on separate microscopes. One of the channels contains tethered DNA with a 12RSS and a 23RSS oriented toward each other (nonamer regions on both RSSs closest to each other). Properly tethered beads are filtered using various methods to ensure proper spacing from neighboring beads and that individual beads are tethered by a single strand of DNA. The trajectories of the selected beads are then examined in the absence of RAG and HMGB1 for ten minutes before flowing in 9.6 nM murine core RAG1/core RAG2 and 80nM full-length HMGB1 and acquiring bead trajectories for at least one hour. Additional information on bead selection criteria and identification of PCs are provided in the SI Text.

Statistical inference. We used both Bayesian and Frequentist methods in this work to calculate parametric and nonparametric quantities, respectively. The PC formation frequencies were assigned confidence intervals via bootstrapping. Briefly, the observed beads and their reported PC formation counts were sampled with replacement to generate a simulated data set of the same length as the number of observations. The looping frequency was then calculated as the total loops formed among the generated dataset divided by the number of beads and the distribution was resampled again. This procedure was performed 10^6 times and we report various percentiles of these bootstrap replicates, as shown both in the main text and on the paper website.

To compute the cleavage probability and PC leaving rate k_{leave}, we used a Bayesian definition of probability and constructed a posterior distribution for each as is explicitly laid out in the SI Text. The displayed posterior distributions for the cleavage probability were generated by numerically evaluating the posterior distribution over a range of cleavage probabilities bounded from 0 to 1. The reported values for the cleavage probability and uncertainty were computed analytically and is derived in the SI text.

To estimate k_{leave} we again constructed a posterior distribution. Here, we chose an exponential form for the likelihood and assumed an inverse Gamma distribution as a prior on the leaving rate. This posterior was then sampled using Markov chain Monte Carlo as is implemented in the Stan probabilistic programming language. A more detailed derivation of the posterior distribution is provided in the SI Text. All models and code for this inference are available on the paper website.

Data and code availability. All data and code are publicly available. Raw image files can be obtained upon request. Preprocessed image data can be downloaded from CaltechDATA research data repository under the DOI:10.22002/D1.1288. Processed data files, Matlab, and Python code used in this work can be downloaded either from the paper website or on the dedicated GitHub repository (DOI:10.5281/zenodo.346571).

ACKNOWLEDGMENTS. We thank members of the David G. Schatz, David Baltimore, and Rob Phillips labs for useful discussions and Caltech's Protein Expression Center for supplying resources and equipment for protein purification. We also thank Helen Beilinson, Justin Bois, Zev Bryant, Heun Jin Lee, Stephanie Johnson, Eddy Rubin, Charlie Starr, Yuhang Zhang, and Haojie Zhuang for discussions. This work was supported by R01GM085286 and 1R35 GM118043 Maximizing Investigators' Research Award (MIRA) (to R.P.). S.H. was also supported by the Caltech Center for Environmental Microbial Interactions (CEMI) (R.P.), the Foundational Questions Institute (FQXI) (R.P.), and the Sackler Foundation (D.B.).

1. Tonegawa S (1983) Somatic generation of antibody diversity. *Nature* 302:575–581.
2. Nadel B, Tang A, Guia E, Lugo G, Feeney AJ (1998) Sequence of the spacer in the recombination signal sequence affects V(D)J rearrangement frequency and correlates with nonrandom Vκ usage in vivo. *J. Exp. Med* 187(9):1495–1503.
3. Weinstein JA, Jiang N, White RAI, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324(5928):804–807.
4. Choi NM, et al. (2013) Deep sequencing of the murine Igh repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *Journal of Immunology* 191(5):2393–2402.
5. Aoki-Ota M, Torkamani A, Ota T, Schork N, Nemazee D (2012) Skewed primary Igκ repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. *Journal of Immunology* 188(5):2305–2315.
6. Schatz DG, Swanson PC (2011) V(D)J recombination: mechanisms of initiation. *Annu Rev Genet* 45(1):167–202.
7. Ramsden DA, Baetz K, Wu GE (1994) Conservation of sequence in recombination signal sequence spacers. *Nucleic Acids Res* 22(10):1785–1796.
8. Hesse JE, Lieber MR, Mizuuchi K, Gellert M (1989) V(D)J recombination: a functional definition of the joining signals. *Genes Dev* 3:1053–1061.
9. Ramsden DA, Wu GE (1991) Mouse κ light-chain recombination signal sequences mediate recombination more frequently than do those of λ light chain. *PNAS* 88(23):10721–10725.
10. Akamatsu Y, et al. (1994) Essential residues in V(D)J recombination signals. *Journal of Immunology* 153:4520–4529.
11. Lee AI, et al. (2003) A functional analysis of the spacer of V(D)J recombination signal sequences. *PLoS Biol* 1(1):56–69.
12. Swanson PC (2004) The bounty of RAGs: recombination signal complexes and reaction outcomes. *Immunol Rev* 200(1):90–114.
13. Ru H, et al. (2015) Molecular mechanism of V(D)J recombination from synaptic RAG1-RAG2 complex structures. *Cell* 163(5):1138–1152.
14. Kim MS, et al. (2018) Cracking the DNA code for V(D)J recombination. *Molecular Cell* 70(2):1–13.
15. Ru H, et al. (2018) DNA melting initiates the RAG catalytic pathway. *Nature Structural & Molecular Biology* 25(8):732–742.
16. Schafer DA, Gelles J, Sheetz MP, Landick R (1991) Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature* 352(6334):444–448.
17. Han L, et al. (2009) Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS ONE* 4(5):e5621–17.
18. Johnson S, Linden M, Phillips R (2012) Sequence dependence of transcription factor-mediated DNA looping. *Nucleic Acids Res* 40(16):7728–7738.
19. Plénat T, Tardin C, Rousseau P, Laurence S (2012) High-throughput single-molecule analysis of DNA-protein interactions by tethered particle motion. *Nucleic Acids Res* 40(12).
20. Lovely GA, Brewster RC, Schatz DG, Baltimore D, Phillips R (2015) Single-molecule analysis of RAG-mediated V(D)J DNA cleavage. *PNAS* 112(14):E1715–23.
21. Brunet A, et al. (2017) How does temperature impact the conformation of single DNA molecules below melting temperature? *Nucleic Acids Res* 46(4):2074–2081.
22. Hirokawa S, et al. (2019) Sequence-dependent dynamics of endogenous and synthetic RSSs in V(D)J recombination (https://www.rpgroup.caltech.edu/vdj_recombination/). Deposited September 27, 2019.
23. Yin FF, et al. (2009) Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. *Nat Struct Mol Biol* 16(5):499–508.
24. Zhang Y, et al. (2019) The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. *Nature* 573(7775):600–604.
25. Gerstein RM, Lieber MR (1993) Coding end sequence can markedly affect the initiation of V(D)J. *Genes Dev* 7(7B):1459–1469.

26. Ezekiel UR, Tianhe S, Bozek G, Storb U (1997) 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(7):4191–4197.
27. Yu K, Lieber MR (1999) Mechanistic basis for coding end sequence effects in the initiation of V(D)J recombination. *Mol Cell Biol* 19(12):8094–8102.
28. Gauss GH, Lieber MR (1992) The basis for the mechanistic bias for deletional over inversional V(D)J recombination. *Genes Dev* 6(8):1553–1561.
29. Pan PY, Lieber MR, Teale JM (1997) The role of recombination signal sequences in the preferential joining by deletion in DH-JH recombination and in the ordered rearrangement of the IgH locus. *Int Immunol* 9(4):515–522.
30. Gopalakrishnan S, et al. (2013) Unifying model for molecular determinants of the preselection Vβ repertoire. *PNAS* 110(34):E3206–15.
31. Subrahmanyam R, et al. (2012) Localized epigenetic changes induced by DH recombination restricts recombinase to DJH junctions. *Nat Immunol* 13(12):1205–1212.
32. Schatz DG, Ji Y (2011) Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* 11(4):251–263.
33. Wong OK, Guthold M, Erie DA, Gelles J (2008) Interconvertible lac repressor-DNA loops revealed by single-molecule experiments. *PLoS Biol* 6(9):e232.
34. Kumar S, et al. (2014) Enhanced tethered-particle motion analysis reveals viscous effects. *Biophys J* 106(2):399–409.
35. Grawunder U, Lieber MR (1997) A complex of RAG-1 and RAG-2 proteins persists on DNA after single-strand cleavage at V(D)J recombination signal sequences. *Nucleic Acids Res* 25(7):1375–1382.
36. Hiom K, Gellert M (1997) A stable RAG1-RAG2-DNA complex that is active in V(D)J cleavage. *Cell* 88(1):65–72.
37. Guilbaud S, Salomé L, Destainville N, Manghi M, Tardin C (2019) Dependence of DNA persistence length on ionic strength and ion type. *Phys Rev Lett* 122(2):028102.
38. Cowell LG, Davila M, Kepler TB, Kelsoe G (2002) Identification and utilization of arbitrary correlations in models of recombination signal sequences. *Genome Biol* 3(12):1–20.
39. Cowell LG, Davila M, Yang K, Kepler TB, Kelsoe G (2003) Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling. *J Exp Med* 197(2):207–220.
40. Cowell LG, Davila M, Ramsden D, Kelsoe G (2004) Computational tools for understanding sequence variability in recombination signals. *Immunol Rev* 200(1):57–69.
41. Bolland DJ, et al. (2016) Two mutually exclusive local chromatin states drive efficient V(D)J recombination. *Cell Reports* 15(11):2475–2487.
42. Lankaš F, Šponer J, Langowski J, Cheatham TEI (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys J* 85(5):2872–2883.
43. Jain S, Ba Z, Zhang Y, Dai HQ, Alt FW (2018) Ctcf-binding elements mediate accessibility of RAG substrates during chromatin scanning. *Cell* 174(1):102–116.e14.
44. Feeney AJ, Goebel P, Espinoza CR (2004) Many levels of control of V gene rearrangement frequency. *Immunol Rev* 200(1):44–56.
45. Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G (1996) A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to Haemophilus influenzae type b disease. *J. Clin. Invest* 97(10):2277–2282.
46. Nadel B, et al. (1998) Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus influenzae type b disease. *Journal of Immunology* 161(11):6068–6073.
47. Johnson S, van de Meent JW, Phillips R, Wiggins CH, Linden M (2014) Multiple LacI-mediated loops revealed by Bayesian statistics and tethered particle motion. *Nucleic Acids Res* 42(16):10265–10277.
48. Hiom K, Gellert M (1998) Assembly of a 12/23 paired signal complex: a critical control point in V(D)J recombination. *Molecular Cell* 1(7):1011–1019.
49. Fugmann SD, Villey IJ, Ptaszek LM, Schatz DG (2000) Identification of two catalytic residues in RAG1 that define a single active site within the RAG1/RAG2 protein complex. *Molecular Cell* 5(1):97–107.
50. Zagelbaum J, et al. (2016) Real-time analysis of RAG complex activity in V(D)J recombination. *PNAS* 113(42):11853–11858.
51. Drejer-Teel AH, Fugmann SD, Schatz DG (2007) The beyond 12/23 restriction is imposed at the nicking and pairing steps of DNA cleavage during V(D)J recombination. *Mol Cell Biol* 27(18):6288–6299.
52. Banerjee JK, Schatz DG (2014) Synapsis alters RAG-mediated nicking at Tcrb recombination signal sequences: implications for the "beyond 12/23" rule. *Mol. Cell. Biol* 34(14):2566–2580.

Supplementary Information for "Probing the Sequence-Dependent Dynamics of Synthetic and Endogenous RSSs in V(D)J Recombination"

Soichi Hirokawa[a], Griffin Chure[b], Nathan M. Belliveau[b,c], Geoffrey A. Lovely[d], Michael Anaya[b], David G. Schatz[e], David Baltimore[b], and Rob Phillips[b,f,1]

[a]Department of Applied Physics, California Institute of Technology, Pasadena, CA, USA
[b]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA
[c]Present Address: Howard Hughes Medical Institute and Department of Biology, University of Washington, Seattle, WA, USA
[d]National Institute on Aging, National Institutes of Health, Baltimore, MD, USA
[e]Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA
[f]Department of Physics, California Institute of Technology, Pasadena, CA, USA

October 3, 2019

Contents

S1 Experimental Methods

S1.1 Microscopy Components and Configuration

All TPM experiments were performed using two Olympus IX71 inverted microscopes with bright-field 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 studied 12RSS. Each microscope is outfitted 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 of their tethered beads.

S1.2 TPM Preparation

A schematic of the tethered bead assembly process as discussed in the Materials & Methods of the manuscript is shown in Fig. S1. 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 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.

S1.3 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 beads range of motion. 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.

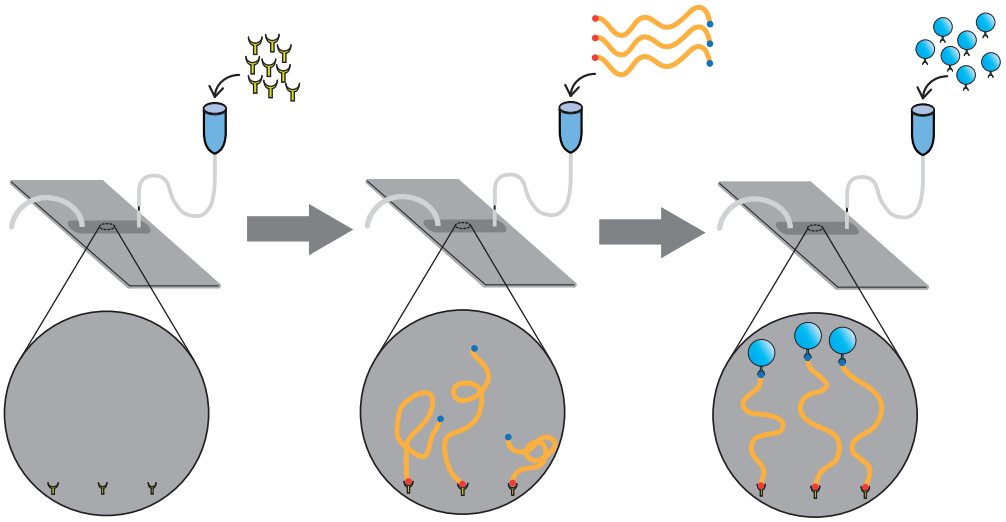

Figure S1: Tethered bead preparation process. Tethered beads are first assembled by adding anti-digoxigenin from Sigma-Aldrich into the flow cell chamber by gravity flow and left to incubate for at least two hours. The fluid is then displaced from the chamber by washing in TPM assembly buffer and introducing 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.

RMSD values of the bead center are obtained using a sliding window of 120 images acquired over four-second intervals. 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:XXXX. All code used to analyze these images can be found on the paper website or the paper GitHub repository (DOI: XXXX).

S2 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. Information on the valid beads 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.

S2.1 Selecting Beads for Further Analysis

Bead selection criteria after preprocessing is applied in the same manner as [1, 2, 3, 4]. After correcting for various systematic errors of the experiment, such as slow stage drift, beads are

manually filtered based upon their RMSD trajectories both before and after introducing RAG and HMGB1. Tethers that have multiple beads attached 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 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; 2) the bead sticks to the glass slide for longer than a few minutes; 3) another bead enters the cropped region enclosing the studied bead due to stage drift that has not been correct or 4) the experiment ends, which typically runs for at least one hour of acquisition.

Once the beads have been selected, they are entered into an analysis 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]. 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 led 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:XXX.

S2.2 Bootstrapping Looping Frequency

While measuring the PC dwell time or the probability of PC cleavage is a straight-forward measurement, it is less clear how the propensity to enter the looped state should be calcuated. As described in the main text, 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 loop 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, 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 elect to apply bootstrapping. In bootstrapping as applied here, we treat the beads studied as the best representation of the population distribution of loop counts, as we do not have an idealized system where we could study infinitely many beads and track the number of paired complexes formed for each DNA tether. With this assumption that the experimentally-obtained loop count distribution provides the best representation of the population distribution, we can 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.

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. S2. 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.

S2.3 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, and the number of loops that were cut, n, so $n \leq N$. In short, we wish to determine the probability of p_{cut} conditional on N and n, or, written concisely, as $P(p_{cut}|N, n)$. Bayes' Theorem tells us that

$$P(p_{cut}|N, n)P(N, n) = P(n|N, p_{cut})P(N, p_{cut}). \tag{S1}$$

On the lefthand side Eq. S1, $P(N, n)$ is the probability of N loops and n cut loops, $P(n|N, p_{cut})$ is the probability that RAG cuts n loops conditional on the N total loops examined and the

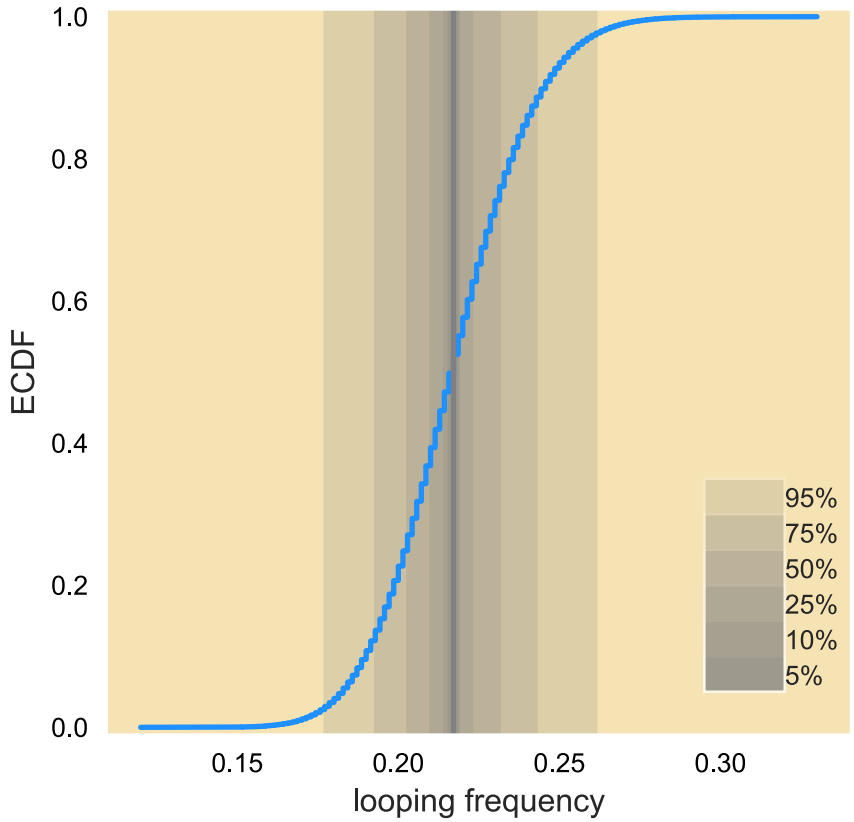

Figure S2: Bootstrapped looping frequency and confidence intervals for the V4-57-1 reference sequence. Empirical CDFs of the bootstrapped looping frequency with 5%, 10%, 25%, 50%, 75% and 95% confidence intervals as represented by the color bar.

probability that RAG cuts a given loop p_{cut}. $P(N, p_{cut})$ is the probability of getting N total loops and that RAG has a cut probability p_{cut} for the RSS. A rearrangement of the equation shows that

$$P(p_{cut}|N, n) = \frac{P(n|N, p_{cut})P(N, p_{cut})}{P(N, n)}. \tag{S2}$$

We can further simplify this equation by noting that the probability of getting N 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,

$$P(N, p_{cut}) = P(N)P(p_{cut}). \tag{S3}$$

Furthermore, we can further simplify the probability function in the denominator from noticing that the probability of having N total loops and n cut loops can be pieced apart as the probability of having n cut loops given N total loops times the probability of having N total loops to begin with, or

$$P(N, n) = P(n|N)P(N). \tag{S4}$$

Inserting equations S3 and S4 into equation S2 gives us

$$P(p_{cut}|N, n) = \frac{P(n|N, p_{cut})P(N)P(p_{cut})}{P(n|N)P(N)},$$

$$= \frac{P(n|N, p_{cut})P(p_{cut})}{P(n|N)}. \tag{S5}$$

We wish to determine the conditional function on the left of Eq. S5, 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 $P(n|N, p_{cut})$. This conditional probability function is the probability that we observe n loops cut considering we observe N 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 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:

$$P(n|N, p_{cut}) = \frac{N!}{n!(N-n)!}(p_{cut})^n(1-p_{cut})^{N-n}. \tag{S6}$$

Next, we would like to determine $P(p_{cut})$. 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,

$$P(p_{cut}) = \begin{cases} 1 & 0 \leq p_{cut} \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{S7}$$

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

$$P(n|N) = \frac{1}{N+1}. \tag{S8}$$

By assembling equations S6, S7 and S8 into equation S5, we get that

$$P(p_{cut}|N, n) = \frac{(N+1)!}{n!(N-n)!}(p_{cut})^n(1-p_{cut})^{N-n}. \tag{S9}$$

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}:

$$\begin{aligned} ln[P(p_{cut}|N, n)] &= ln\Big[\frac{(N+1)!}{n!(N-n)!}\Big] + n\, ln(p_{cut}) + (N-n)\, ln(1-p_{cut}), \\ \frac{d\, ln[P(p_{cut}|N, n)]}{d\, p_{cut}}\Big|_{p^*_{cut}} &= \frac{n}{p^*_{cut}} - \frac{N-n}{1-p^*_{cut}} = 0. \end{aligned} \tag{S10}$$

Eq. S10 then tells us that

$$p^*_{cut} = \frac{n}{N}. \tag{S11}$$

To calculate the variance of p_{cut}, we make the assumption that $N \gg 1$ 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 $x \equiv p - p^*_{cut}$. Then Eq. S12 becomes

$$P(p_{cut}|N, n) = \frac{(N+1)!}{n!(N-n)!}(p^*_{cut} + x)^n(1-p^*_{cut} - x)^{N-n}. \tag{S12}$$

We also invoke the rule that $ln\, n! \approx n\, ln\, n - n + \frac{1}{2}\, ln[2\pi n]$. We can then determine the prefactor of the posterior distribution. Specifically,

$$\begin{aligned}
\frac{(N+1)!}{n!(N-n)!} &= exp\{ln[(N+1)!] - ln\, n! - ln[(N-n)!]\}, \\
&\approx exp\{(N+1)ln(N+1) - (N+1) + \frac{1}{2}ln[2\pi\,(N+1)] - n\, ln\, n + n - \frac{1}{2}ln(2\pi n) \\
&\qquad - (N-n)ln(N-n) + (N-n) - \frac{1}{2}ln[2\pi(N-n)]\}, \\
&\approx exp\Big\{(N+1)\Big[ln\, N + ln\left(1+\frac{1}{N}\right)\Big] - 1 - n\, ln\, n - (N-n)\Big[ln\, N + ln\Big(1-\frac{n}{N}\Big)\Big] \\
&\qquad + \frac{1}{2}ln\Big[\frac{N+1}{2\pi\, n(N-n)}\Big]\Big\}, \\
&\approx exp\Big\{(N+1)(\frac{1}{N} + \frac{1}{2N^2}) - 1 - n\, ln\, n + n\, ln\, N - (N-n)ln(1-p^*_{cut}) \\
&\qquad + \frac{1}{2}ln\Big[\frac{N^3}{2\pi\, n(N-n)}\Big]\Big\}, \\
&\approx \frac{1}{\sqrt{2\pi\frac{n(N-n)}{N^3}}} exp\Big\{ - n\, ln\Big(p^*_{cut}\Big) - N(1-p^*_{cut})ln(1-p^*_{cut})\Big\}. \qquad \text{(S13)}
\end{aligned}$$

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

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

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

$$\approx \frac{1}{\sqrt{2\pi\frac{n(N-n)}{N^3}}} exp\Big\{ -\frac{N\,x^2}{2}\Big(\frac{1}{p^*_{cut}(1-p^*_{cut})}\Big)\Big\},$$
$$\approx \frac{1}{\sqrt{2\pi\frac{n(N-n)}{N^3}}} exp\Big\{ -\frac{(p-p^*_{cut})^2}{2\Big[\frac{n\,(N-n)}{N^3}\Big]}\Big\}. \tag{S14}$$

Eq. S14 tells us that, not only is this Gaussian approximation centered at the most probable value of $p_{cut} = p^*_{cut}$, as we would expect, but also that the distribution has a variance of $\sigma^2 = \frac{n(N-n)}{N^3}$. Thus, we report $p^*_{cut} = \frac{n}{N}$ and $\sigma^2 = \frac{n(N-n)}{N^3}$ in Fig. 3C and 4C of the main text.

S2.4 Modeling Exit from the Paired Complex As A Poisson Process

As discussed in the main text, 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

$$P(t_{\text{leave}} \,|\, k_{\text{leave}}) = k_{\text{leave}} e^{-k_{\text{leave}} t_{\text{leave}}}, \tag{S15}$$

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

$$k_{\text{leave}} = k_{\text{cut}} + k_{\text{unloop}}. \tag{S16}$$

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

$$P(k_{\text{leave}} \,|\, t_{\text{leave}}) = \frac{P(t_{\text{leave}} \,|\, k_{\text{leave}}) P(k_{\text{leave}})}{P(t_{\text{leave}})}. \tag{S17}$$

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. S17 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. S17. The likelihood already matches the definition in Eq. S15, 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. S15, 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. S15 in terms of the inverse leaving rate $\tau_{\text{leave}} = \frac{1}{k_{\text{leave}}}$ so that

$$P(t_{\text{leave}} \,|\, \tau_{\text{leave}}) = \frac{1}{\tau_{\text{leave}}} e^{t_{\text{leave}}/\tau_{\text{leave}}}. \tag{S18}$$

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

$$P(\tau_{\text{leave}} \,|\, \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \frac{\beta^{\alpha}}{\tau_{\text{leave}}^{(\alpha+1)}} e^{-\beta/\tau_{\text{leave}}}, \tag{S19}$$

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. S18 and Eq. S19 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 [6]. The code used to sample this distribution can be accessed on the paper website or GitHub repository.

S3 Posterior Distributions of the Endogenous Sequences

Fig. S3 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 [7, 8, 9].

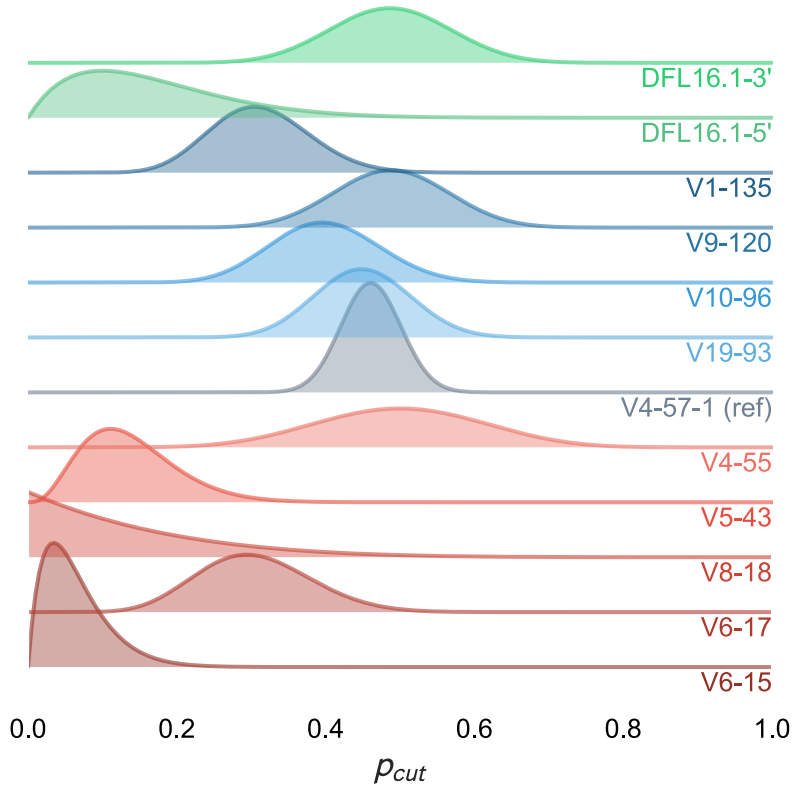

Figure S3: Posterior distributions of the cutting probabilities as derived in SI Section S2.3 for the endogenous 12RSSs studied. The top-to-bottom order of the endogenous RSSs is the same as their left-to-right ordering in Fig. 3. Height of the distribution is proportional to the probability of the p_{cut} value.

S4 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 S1, 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 [10, 11, 12]. Furthermore, various bulk assays have demonstrated that coding flank sequence can affect recombination efficiency [7, 8, 9]. 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.

Fig. S4 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.

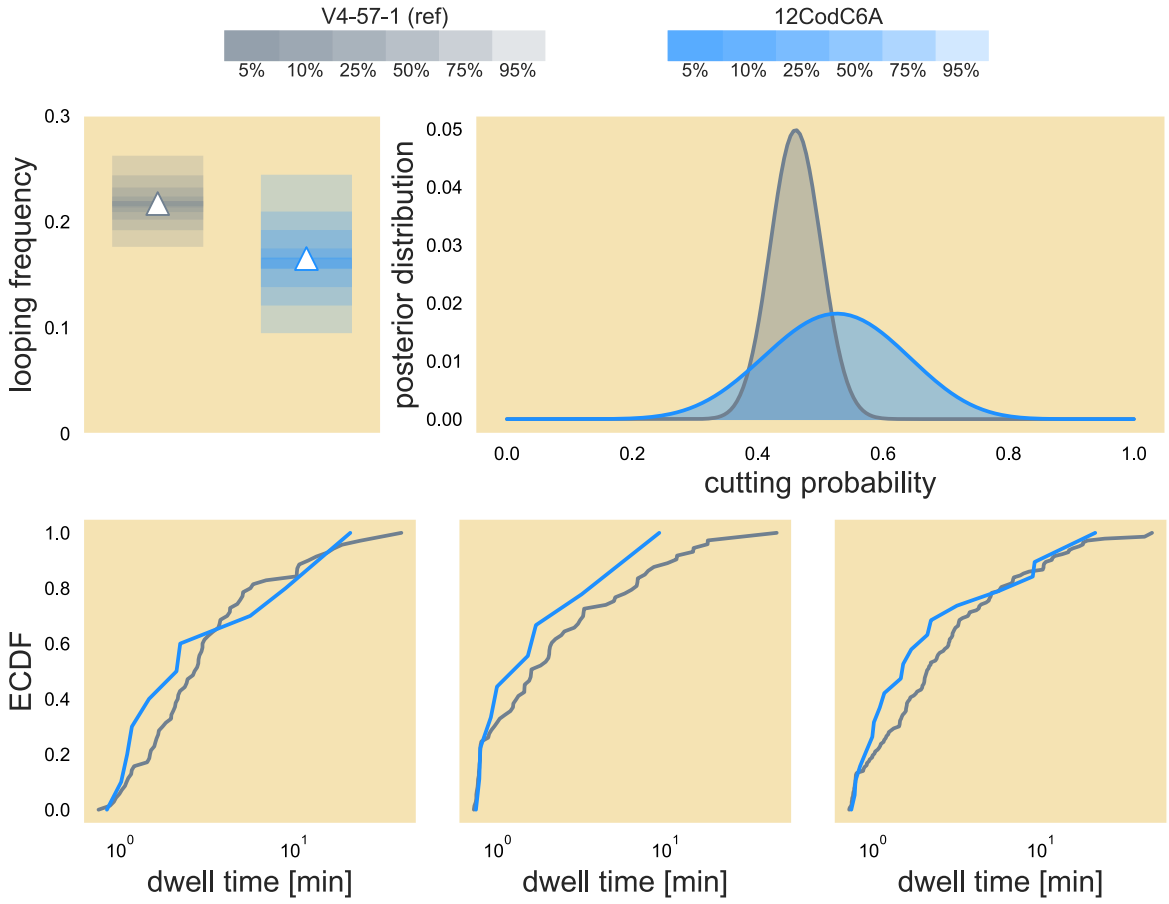

Figure S4: V4-57-1 (reference) RSS (grey) and coding flank change (blue) comparison of looping frequency, posterior distribution of the cutting probability and ECDFs of PC lifetimes for PCs that cut, those that unloop, and both.

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. S5 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.

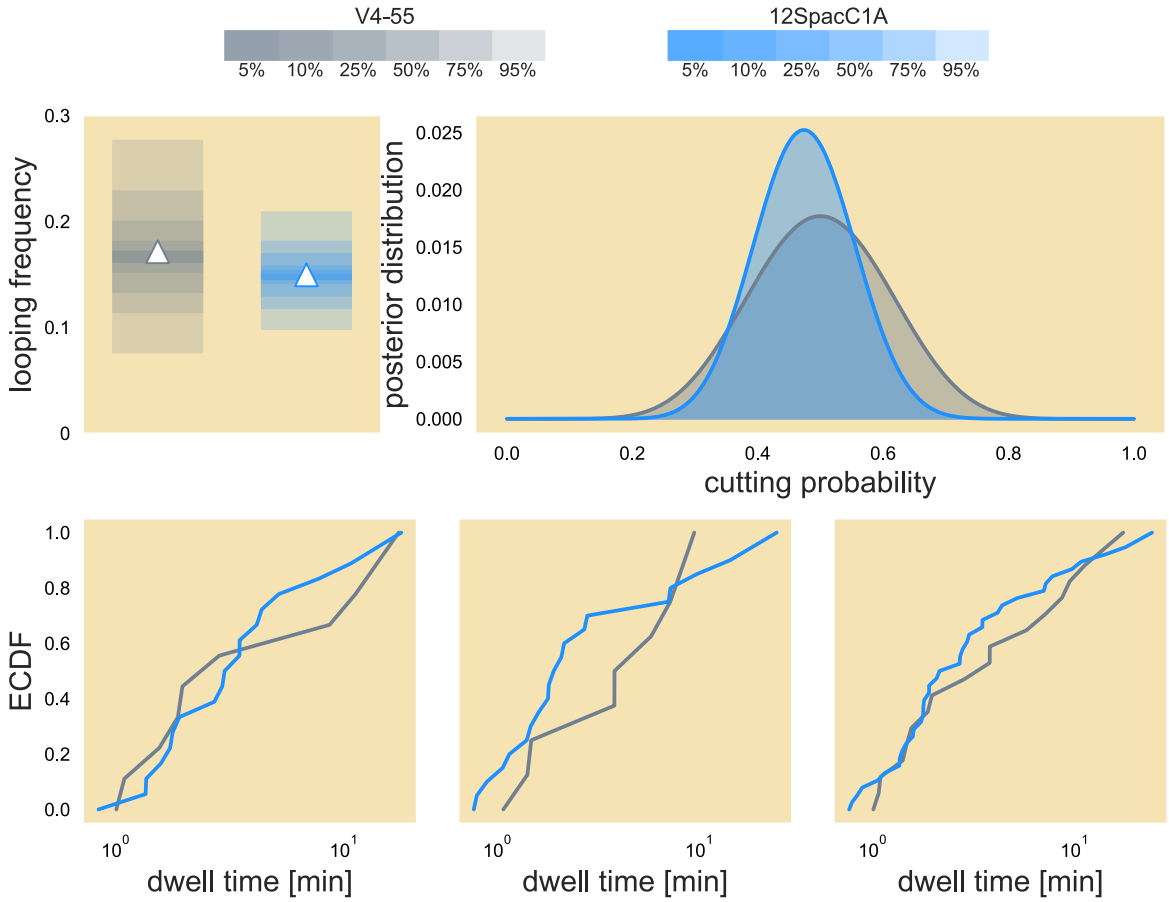

Figure S5: V4-55 12RSS (grey) and C-to-A change at spacer position 1 (blue) comparison of looping frequency, posterior distribution of the cutting probability and ECDFs of PC lifetimes for PCs that cut, those that unloop, and both

S5 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 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. S6 shows that these two outcomes result. Fig. S6A and S6C 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. S6). 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 overlapping 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.

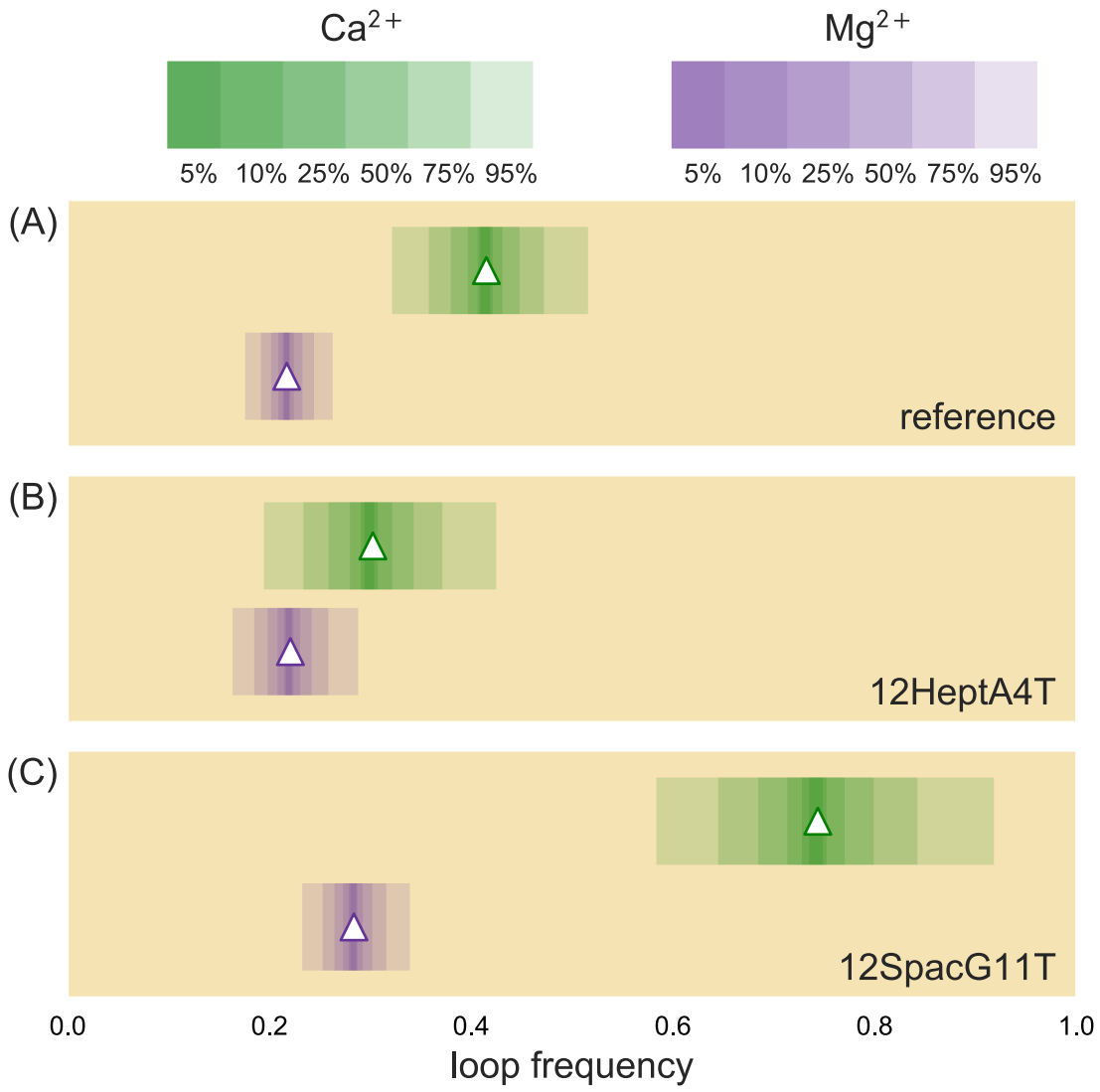

Figure S6: Ca^{2+} (green) and Mg^{2+} (purple) looping frequencies for (A) reference 12RSS, (B) A-to-T change at the fourth position of the heptamer and (C) G-to-T change at the eleventh position of the spacer. Measured looping frequency shown as the triangles. Going from darker shading to lighter shading in rectangle bar indicates increasing of confidence interval percentage of the looping frequency from the bootstrapping method discussed in section S2.2.

S6 Endogenous RSS Sequences

Endogenous 12RSS	Sequence
DFL16.1-3'	AGCTAC CACAGTGCTATATCCATCAGCAAAAACC
DFL16.1-5'	AATAAA CACAGTAGTAGATCCCTTCACAAAAAGC
V1-135	TCCTCA CACAGTGATTCAGACCCGAACAAAAACT
V9-120	TCCTCC CACAGTGATACAAATCATAACATAAACC
V10-96	TCCTCC CACAATGATATAAGTCATAACATAAACC
V19-93	TCTACC CACAGTGATACAAATCATAACAAAAACC
V4-57-1 (reference)	GTCGAC CACAGTGCTACAGACTGGAACAAAAACC
V4-55	CACCCA CACAGTGATACAGACTGGAACAAAAACC
V5-43	GCCTCA CACAGTGATGCAGACCATAGCAAAAATC
V8-18	TCCCCC CACAGAGCTTCAGCTGCCTACACAAACC
V6-17	TCCTCC CACAGTGCTTCAGCCTCCTACACAAACC
V6-15	TCCTCT CACAGTACTTCAGCCTCCTACATAAACC
Jκ1 23RSS	GGATCC CACAGTGGTAGTACTCCACTGTCTGGCTGTACAAAAACC

Table S1: **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 bottom sequence is of the Jκ1 23RSS applied in all of the DNA constructs used in TPM.

S7 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 [13]. 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 S2 and S3.

After purification of the PCR fragment and DpnI digestion (NEB Biolabs) of the PCR template, the fragment was circularized using Gibson assembly [14] and transformed into DH5α *Escherichia coli*. Transformants were then cultured and stored for plasmid purification and sequence verification.

S8 Synthetic 12RSS Primers

Tables S2 and S3 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**
12CodC6A (Fwd)	AACACAGTGCTACAGACTGGAACAAAAACCCTGCAGTC
12CodC6A (Rev)	CTGTAGCACTGTGTTCGACCTGCAGCCCAAGCG
12HeptC3G (Fwd)	ACCAGAGTGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptC3G (Rev)	CTGTAGCACTCTGGTCGACCTGCAGCCCAAGCG
12HeptC3T (Fwd)	ACCATAGTGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptC3T (Rev)	CTGTAGCACTATGGTCGACCTGCAGCCCAAGCG
12HeptA4T (Fwd)	ACCACTGTGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptA4T (Rev)	CTGTAGCACAGTGGTCGACCTGCAGCCCAAGCG
12HeptG5A (Fwd)	ACCACAATGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptG5A (Rev)	CTGTAGCATTGTGGTCGACCTGCAGCCCAAGCG
12HeptG5C (Fwd)	ACCACACTGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptG5C (Rev)	CTGTAGCAGTGTGGTCGACCTGCAGCCCAAGCG
12HeptT6A (Fwd)	ACCACAGAGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptT6A (Rev)	CTGTAGCTCTGTGGTCGACCTGCAGCCCAAGCG
12HeptT6C (Fwd)	ACCACAGCGCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptT6C (Rev)	CTGTAGCGCTGTGGTCGACCTGCAGCCCAAGCG
12HeptG7A (Fwd)	ACCACAGTACTACAGACTGGAACAAAAACCCTGCAGTC
12HeptG7A (Rev)	CTGTAGTACTGTGGTCGACCTGCAGCCCAAGCG
12HeptG7C (Fwd)	ACCACAGTCCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptG7C (Rev)	CTGTAGGACTGTGGTCGACCTGCAGCCCAAGCG
12HeptG7T (Fwd)	ACCACAGTTCTACAGACTGGAACAAAAACCCTGCAGTC
12HeptG7T (Rev)	CTGTAGAACTGTGGTCGACCTGCAGCCCAAGCG
12SpacC1A (Fwd)	ACCACAGTGATACAGACTGGAACAAAAACCCTGCAGTC
12SpacC1A (Rev)	CTGTATCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacC1G (Fwd)	ACCACAGTGGTACAGACTGGAACAAAAACCCTGCAGTC
12SpacC1G (Rev)	CTGTACCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacA3G (Fwd)	ACCACAGTGCTGCAGACTGGAACAAAAACCCTGCAGTC
12SpacA3G (Rev)	CTGCAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacA3T (Fwd)	ACCACAGTGCTTCAGACTGGAACAAAAACCCTGCAGTC
12SpacA3T (Rev)	CTGAAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacC4G (Fwd)	ACCACAGTGCTAGAGACTGGAACAAAAACCCTGCAGTC
12SpacC4G (Rev)	CTCTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacC4T (Fwd)	ACCACAGTGCTATAGACTGGAACAAAAACCCTGCAGTC
12SpacC4T (Rev)	CTATAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG6A (Fwd)	ACCACAGTGCTACAAACTGGAACAAAAACCCTGCAGTC
12SpacG6A (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG6T (Fwd)	ACCACAGTGCTACATACTGGAACAAAAACCCTGCAGTC
12SpacG6T (Rev)	ATGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacA7C (Fwd)	ACCACAGTGCTACAGCCTGGAACAAAAACCCTGCAGTC
12SpacA7C (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacA7G (Fwd)	ACCACAGTGCTACAGGCTGGAACAAAAACCCTGCAGTC
12SpacA7G (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacC8T (Fwd)	ACCACAGTGCTACAGATTGGAACAAAAACCCTGCAGTC

12SpacC8T (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacT9A (Fwd)	ACCACAGTGCTACAGACAGGAACAAAAACCCTGCAGTC
12SPacT9A (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacT9C (Fwd)	ACCACAGTGCTACAGACCGGAACAAAAACCCTGCAGTC
12SpacT9C (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacT9G (Fwd)	ACCACAGTGCTACAGACGGGAACAAAAACCCTGCAGTC
12SpacT9G (Rev)	TTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG10A (Fwd)	ACCACAGTGCTACAGACTAGAACAAAAACCCTGCAGTC
12SpacG10A (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG10C (Fwd)	ACCACAGTGCTACAGACTCGAACAAAAACCCTGCAGTC
12SpacG10C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG10T (Fwd)	ACCACAGTGCTACAGACTTGAACAAAAACCCTGCAGTC
12SpacG10T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG11A (Fwd)	ACCACAGTGCTACAGACTGAAACAAAAACCCTGCAGTC
12SpacG11A (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG11C (Fwd)	ACCACAGTGCTACAGACTGCAACAAAAACCCTGCAGTC
12SpacG11C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacG11T (Fwd)	ACCACAGTGCTACAGACTGTAACAAAAACCCTGCAGTC
12SpacG11T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacA12C (Fwd)	ACCACAGTGCTACAGACTGGCACAAAAACCCTGCAGTC
12SpacA12C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12SpacA12T (Fwd)	ACCACAGTGCTACAGACTGGTACAAAAACCCTGCAGTC
12SpacA12T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12NonA1G (Fwd)	ACCACAGTGCTACAGACTGGAGCAAAAACCCTGCAGTC
12NonA1G (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12NonA3C (Fwd)	ACCACAGTGCTACAGACTGGAACCAAAACCCTGCAGTC
12NonA3C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12NonA4C (Fwd)	ACCACAGTGCTACAGACTGGAACACAAACCCTGCAGTC
12NonA4C (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12NonA4T (Fwd)	ACCACAGTGCTACAGACTGGAACATAAACCCTGCAGTC
12NonA4T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12NonA5T (Fwd)	ACCACAGTGCTACAGACTGGAACAATAACCCTGCAGTC
12NonA5T (Rev)	CTGTAGCACTGTGGTCGACCTGCAGCCCAAGCG
12NonC8G (Fwd)	GCTACAGACTGGAACAAAAAGCCTGCAGTCAACCTCGA
12NonC8G (Rev)	TTTGTTCCAGTCTGTAGCACTGTGGTCGACCTGCAG
12NonC8T (Fwd)	GCTACAGACTGGAACAAAAATCCTGCAGTCAACCTCGA
12NonC8T (Rev)	TTTGTTCCAGTCTGTAGCACTGTGGTCGACCTGCAG
12NonC9T (Fwd)	GCTACAGACTGGAACAAAAACTCTGCAGTCAACCTCGA
12NonC9T (Rev)	TTTGTTCCAGTCTGTAGCACTGTGGTCGACCTGCAG

Table S2: 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.

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 S3: Forward (Fwd) and reverse (Rev) primers for designing TPM constructs with endogenous 12RSSs. Underlined regions denote the heptamer and nonamer regions.

S9 Protein Purification

S9.1 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.

S9.2 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.

References

[1] Han L, et al. (2008) *Calibration of tethered particle motion experiments.* (Springer-Verlag, New York) Vol. 150, 1 edition, pp. 123–138.

[2] Lovely GA, Brewster RC, Schatz DG, Baltimore D, Phillips R (2015) Single-molecule analysis of RAG-mediated V(D)J DNA cleavage. *PNAS* 112(14):E1715–23.

[3] Han L, et al. (2009) Concentration and length dependence of DNA looping in transcriptional regulation. *PLoS ONE* 4(5):e5621–17.

[4] Johnson S, Linden M, Phillips R (2012) Sequence dependence of transcription factor-mediated DNA looping. *Nucleic Acids Res* 40(16):7728–7738.

[5] Chure G, Lee HJ, Rasmussen A, Phillips R (2018) Connecting the dots between mechanosensitive channel abundance, osmotic shock, and survival at single-cell resolution. *Journal of Bacteriology* 200(23):e00460–18.

[6] Carpenter B, et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software* 76(1):1–32.

[7] Gerstein RM, Lieber MR (1993) Coding end sequence can markedly affect the initiation of V(D)J. *Genes Dev* 7(7B):1459–1469.

[8] Ezekiel UR, Tianhe S, Bozek G, Storb U (1997) 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(7):4191–4197.

[9] Yu K, Lieber MR (1999) Mechanistic basis for coding end sequence effects in the initiation of V(D)J recombination. *Mol Cell Biol* 19(12):8094–8102.

[10] Ru H, et al. (2015) Molecular mechanism of V(D)J recombination from synaptic RAG1-RAG2 complex structures. *Cell* 163(5):1138–1152.

[11] Kim MS, et al. (2018) Cracking the DNA code for V(D)J recombination. *Molecular Cell* 70(2):1–13.

[12] Ru H, et al. (2018) DNA melting initiates the RAG catalytic pathway. *Nature Structural & Molecular Biology* 25(8):732–742.

[13] Lutz R, Bujard H (1997) 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(6):1203–1210.

[14] Gibson DG, et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* 6(5):343–345.